Michael Freunek
Roschistrasse 1a
3007 Bern

Data Science Project

Investigation of Patent Data on Stock Performance

22 September 2019

ABSTRACT

In this project the correlation between 8 patent data parameters and stock performance of 101 companies, listed September 2019 in the Nasdaq Composite, is investigated. The 8 patent data parameters, characterizing the patent portfolio of corresponding company in the period end of year 2004 till end of year 2009, are correlated to the stock performance in the period 04.01.2010 till 28.12.2018. The result shows that 4 patent data parameters show a correlation to the stock performance with a correlation factor between 0.37 und 0.51, the other 4 patent data parameters show no or no significant correlation to stock performance.

TABLE OF CONTENTS

..

Inhalt

1. OBJECTIVE

Investors all over the world try to identify and derive parameters (company parameters, sector parameters, local and global economy parameters, interest parameters, political parameters etc.) to optimize their investment decisions. Their final goal is to outperform the stock market (or more accurate an appropriate stock market index) by stock picking, sector picking and (investment-) timing.

The major portion of the finance industry attracts costumers that their finance products yield excess returns compared to a reference index like MSCI World, S&P500, DAX30, Euro STOXX etc.

The reason for the failure of all investors and the finance industry on the long-run (disregarding statistical exceptions) is the "efficient market hypothesis" (EMH) by Eugene Fama, Robert Shiller and Lars Hansen [1].

The EMH claims, that "all (public) available information are reflected by stock prices" or in other words, "the stock prices are the best estimation for the stock values" (semi strong efficiency).1

The EMH is divided into 1: weak-form efficiency, 2: semi-strong efficiency and 3: strong-form efficiency.

The EMH dos not claim, that the market is always efficient. That's the reason, why sometimes patterns or parameters are found retrospectively, which historically leaded to outperformance. But the EMH argues, that once a market inefficiency is observed, it is arbitraged away.

Goal of this project *is less to investigate, if patent data can be used for stock market outperformance rather to tackle the question, if the stock market is efficient regarding patent data.*

Studying financial news one can observe that patent information on a superficial level is paid attention by investors. Examples are pharma companies losing patent protection for at least one of their blockbuster drugs, technology companies apply patents, leading to "speculation" of new technology developments and markets (e.g. foldable display by Apple, autonomous driving by Alphabet (Google), drone delivery by Amazon etc.).

On the other side there are good reasons, why the stock market is possibly not efficient according to patent data.

1. Until the latest time it was unclear how "to rate" patents and patent portfolios.
2. Tools to systematically investigate patent data and patent portfolios and relate them to companies are available since a short time.

[1] For this reason, ETFs (exchange traded funds) become increasingly popular

2 METHODS

2.1 Infrastructure

- Toshiba Satellite C850-1F2, Intel Core i3-2370M, CPU 2.40 GHZ
- Operation system Windows 10 Home (Version 10.0.18362)

2.2 Tools

- **PatentSight** is a German company maintaining a tool/platform to analyze and manage patent data [2]. The database covers the database DOCDB [3] of the European Patent Office. Outstanding for PatentSight are three facts
    - research-based patent parameters PAI, TR and CI [4],
    - high level of company harmonization2,
    - reporting date concept allows to analyze historical patent data without hindsight bias (back testing)3.
- **Finanzen.net**: Finanzen.net is a German company (Axel Springer SE) offering stock data [5].
- **Jupyter Notebook** (Version 6.0.0): Interactive computing product based on the platform **Anaconda** 2019.07 (Python 3.7 Version [6]). Data analysis is performed in **Python 3** and **Excel**.
- **Microsoft Excel**: Data of PatentSight are exported in Excel.

2.3 Libraries (in Python)

- **pandas**: Library and tool for dealing with data structures, data tables, manipulating data.
- **numpy**: Library and tool for calculating with data and arrays.
- **matplotlib**: Library and tool for the visualization of data in diagrams.
- **scipy**: Library and tool for mathematical calculations (scipy.stats: probability distributions, normality tests, regression etc.)

2.4 Statistics

- **Descriptive statistics:** Descriptive statistics is used in this project to describe, organize and illustrate the data (tables, diagrams). E.g. the library “matplotlib” is used to visualize the distribution of the data in histograms. With the normality test (stats.normaltest) we investigate the distribution of the data, and furthermore determine further parameters like the values and standard deviations.
- **Inferential statistics:** Inferential statistics is used in this project to analyze the mathematical “behavior” of the data. Central part is the correlation analysis of patent data versus stock performance data.

[2] Company harmonization means to respect all subsidiaries for correct patent ownership allocation.
[3] With some discussible restrictions (see chapter 5.1).

3 Data

This project is based on patent data by PatentSight [2] and finance data by finanzen.net [5]. We will finally correlate patent data versus finance data for companies listed in the Nasdaq Composite. The companies are manually selected from the Nasdaq Composite list [7] fulfilling the following conditions: In 2019 more than 20 active patent families, patent data available since (at least) 2004, quoted on stock exchange (at least) since 01.01.2010.

1. **Patent data by PatentSight** [2] for the period 31.12.2004 till 31.12.2009, exported to the Excel file "Module_1_stock_patent_stock_source_data_freunek":
    - Owner (= company name, data type string),
    - Reporting Date (= cutoff date, Reporting dates = 31.12.2004, 31.12.2005, 31.12.2006, 31.12.2007, 31.12.2008 and 31.12.2009),
    - Portfolio Size (PS) [3] (= number of active patent families, data type integer),
    - Patent Asset Index (PAI) [3] (accounts for CI and PS, data type float),
    - Technology Relevance (TR) [3] (data type float), and
    - Competitive Impact (CI) [3] (mean value of corresponding patent portfolio, accounts for TR and market coverage, data type float).

2. Based on the data exported by PatentSight, in this project **4 additional parameters** are calculated (in the Excel file):
    - Mean value of the TR in the period 31.12.2004 - 31.12.2009,
    - Relative gradient of the TR in the period 31.12.2004 - 31.12.2009,
    - Relative gradient of the PAI in the period 31.12.2004 - 31.12.2009,
    - Patent Performance Index (PPI) for 31.12.2009 (calculated from the mean value of the TR in the period 31.12.2004 - 31.12.2009, relative gradient TR in the period 31.12.2004 - 31.12.2009 and the relative gradient PAI in the period 31.12.2004 - 31.12.2009 according to the method of the z-score)[4]

3. **Stock data** by Finanzen.net [5]. Stock data are manually written into the Excel file "Module_1_stock_patent_stock_source_data_freunek":
    - Stock prices for 04.01.2010 (first exchange price 2010)
    - Stock prices for 28.12.2018 (last exchange price 2018)
    - Stock return (= (Stock price for 28.12.2018 / Stock price for 04.01.2010 - 1) · 100)

The data are summarized in the Excel file "Module_1_stock_patent_stock_correlation_freunek", which will be the basis for the Python code, where we finally correlate the following data:

[4] Here, no reference could be cited. This will be published soon by the author of this project.

Stock return vs. PS (31.12.2009), PAI (31.12.2009), TR (31.12.2009), mean value of the TR in the period 31.12.2004 - 31.12.2009, relative gradient of the TR in the period 31.12.2004 - 31.12.2009, relative gradient of the PAI in the period 31.12.2004 - 31.12.2009 and PPI for 31.12.2009.

A snippet of the patent data is given in fig. 1. The frequency distributions as histogram plots of the patent and stock data can be seen in fig. 2.

	Mean value TR 5 years	Gradient PAI	Gradient TR	relative Gradient PAI	relative Gradient TR	Portfolio Size	Patent Asset Index	Technology Relevance
count	101.000000	101.000000	101.000000	101.000000	101.000000	101.000000	101.000000	101.000000
mean	2.455089	218.710791	-0.112896	0.056995	-0.047600	1762.663366	4777.389740	2.244605
std	1.117561	961.750828	0.218913	0.107597	0.073575	4096.961218	10796.924778	1.106276
min	0.803500	-1787.805520	-0.578575	-0.521174	-0.212440	9.000000	49.054294	0.899822
25%	1.706920	2.564171	-0.226581	0.004005	-0.098688	116.000000	343.428262	1.522357
50%	2.236996	24.664789	-0.121493	0.054380	-0.054296	362.000000	1082.149942	2.003379
75%	2.915675	109.191510	-0.007593	0.108362	-0.004691	950.000000	2515.498048	2.482704
max	7.561052	6642.229489	1.225187	0.394842	0.162245	24378.000000	63111.674705	8.916876

Figure 1: Table snippet of used patent data

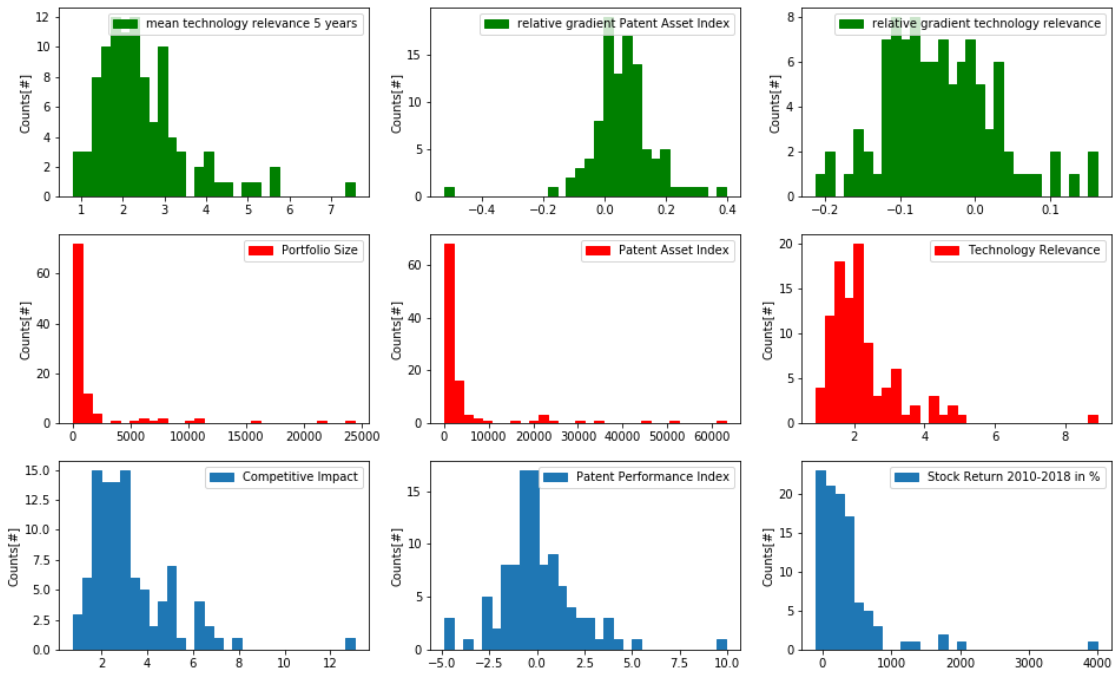

Figure 2: Histograms of patent data (period 2004 – 2009) and stock performance (period 2010 – 2018) for selected companies of Nasdaq Composite

4 METADATA

The metadata required to reproduce the analysis are the description of the patent parameters and stock data. These metadata are given as heading of the corresponding parameter-column in the Excel file "Module_1_stock_patent_stock_source_data_freunek". All further calculations required are given in the Excel file as well. The Excel file is stored on GitHub repository.

5 DATA QUALITY

In this project we do not deal with measured data. For this reason, we have no quality issues with measurement procedures, measurement uncertainties and noise. In this sense the used data are "exact". Nevertheless, there are some effects, which could affect the quality of the data and finally the quality of the results:

5.1 Patent data

1. It is not clear, if every company used in the analysis is perfectly harmonized, i.e. that possibly some patent families are not correctly accounted to a company
2. In the patent system there are time periods, where it is not clear, if a patent family is active or inactive. PatentSight tends to count such patent families as active.
3. The (current) status of a company (in particular mergers, acquisitions) is calculated back in time as former status.

5.2 Stock data

1. Exchange prices for stocks can differ.
2. Spreads (bid-price and ask-price)
3. Dividends are not respected.

For **patent data** we expect point 3 as major data uncertainty. It is unclear about the influence on the correlation results. Unfortunately, we have no possibility to improve the quality at this point.

For **stock data** we expect point 3 as major data uncertainty, too. Respecting that typical dividends (in the technology- and pharma-sector) amount to 1-3 % p.a., the uncertainty should be within 20 % return (in total). In view of stock returns in the range of 100 – 4000 %, the data uncertainty in stock data should hardly affect the correlation results. The quality of the stock data can easily be improved.

A further problem that could occur are database errors.

6 DATA FLOW

The following diagram shows the data flow diagram for the data used in this project.

Company data like

- company name,
- subsidiaries,
- merger, acquisitions

and patent data like

- technical data like patent classes,
- bibliographic data like owner, inventor, title, abstract,
- data like priority date, application date, publications date

are collected, allocated and harmonized by PatentSight. There, the data (patent data) can be exported to an Excel file.

Finanzen.net collects data from stocks and exchanges. These data are combined with the patent data in the Excel file and finally evaluated in the Jupyter Desktop.

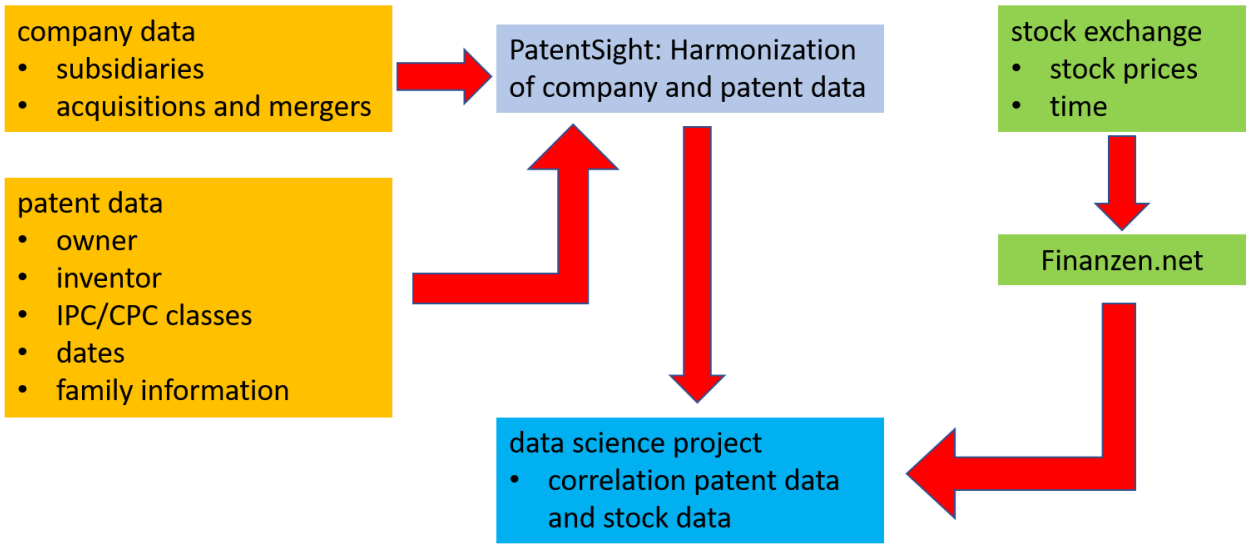

Figure 3: data flow diagram

7 DATA MODELS

7.1 Conceptual

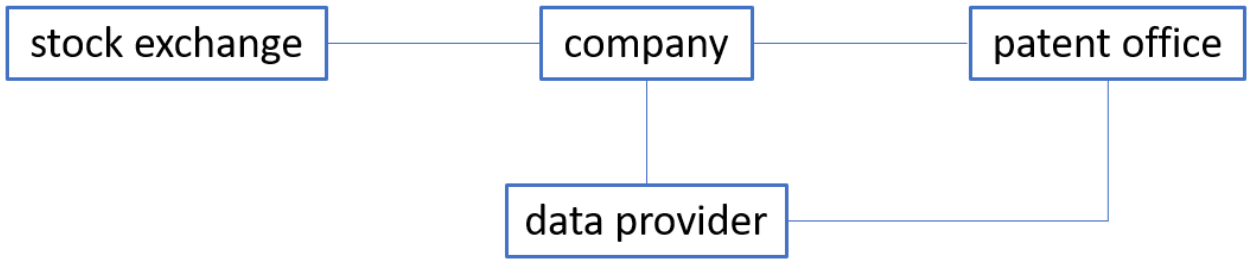

Figure 4:Conceptual data model

7.2 Logical

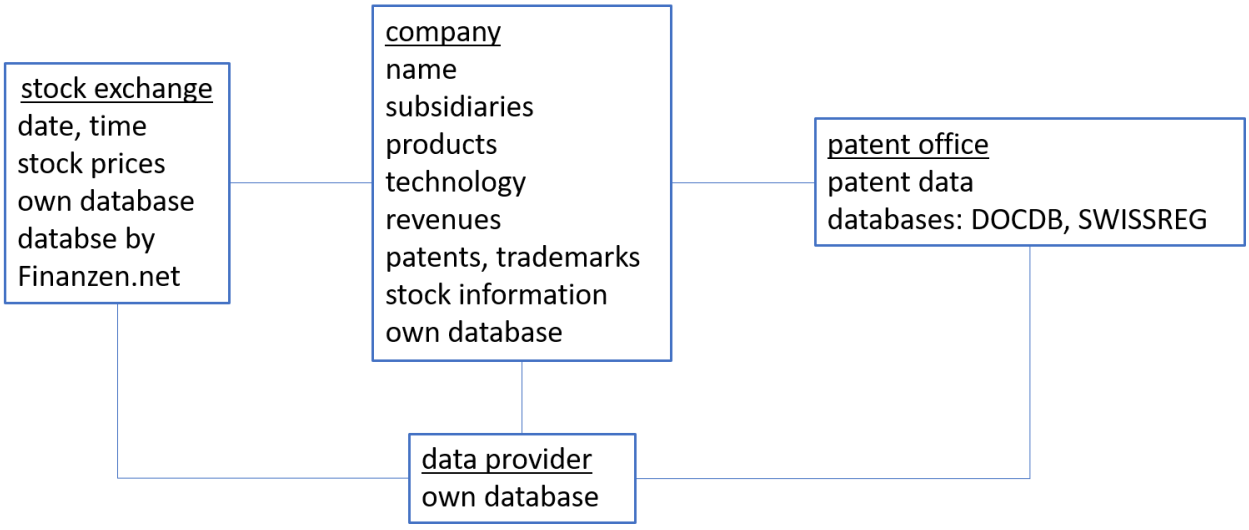

Figure 5: Logical data model

7.3 Physical (infrastructure needs)

The physical infrastructure needs are quite low. The performance of the computer (see chapter "Infrastructure") is sufficient, the demand of storage amounts to less than 2 MB.

8 RISKS

- There is a risk of data loss, the loss of the documents, figures or the loss of the Python code. This project, the data, the code, the documents and the figures are saved locally on the Toshiba laptop hard disk and simultaneously on the Microsoft cloud **OneDrive** (hard disk and cloud are permanently synchronized online). For this reason, the risk of data loss is rather small.
- Possible mistakes by performing this research project are due to the researcher, e.g. confusing about columns and lines in the tables, wrong allocation of companies in PatentSight and Finanzen.net, Python code errors, or mistakes by adopting some data manually into the Excel files.
- A retrospective useless documentation of the project, making it finally unreproducible.
- Further risks see also chapters “Conclusions” and “Data Quality”.

The impact of the (realized) risks could very between very small to making the whole project totally useless. Whereas the impact of “some” data errors may affect the results a trifle, the loss of data or a useless documentation prevents a reproduction of the results.

9 PRELIMINARY STUDIES

Goal of this study is to evaluate whether the stock market is efficient with reference to patent data. For this purpose we correlate the patent data *mean TR 5 years (period 2004 - 2009), relative gradient PAI (period 2004 - 2009), relative gradient TR (period 2004 - 2009), PS (year 2009), PAI (year 2009), TR (year 2009), CI (year 2009) and PPI (year 2009)* vs. *stock returns (period 2010 – 2018)* for selected companies (currently) listed in the Nasdaq Composite (selection criteria see chapter “Data”).

The correlation coefficients are given in table 1.

Whereas no (or only a very weak) correlation seems to be given for the 4 patent parameters *relative gradient PAI (period 2004 - 2009), relative gradient TR (period 2004 - 2009), PS (year 2009), PAI (year 2009)*, the correlation coefficient for the patent parameters *mean TR 5 years (period 2004 - 2009), TR (year 2009), CI (year 2009) and PPI (year 2009)* varies between 0.37 and 0.51.

It is little wonder that there seems to be no correlation for the patent parameters *PS (year 2009), PAI (year 2009)*, since they simply hold for “large” companies. And known from the stock market is the fact, that “large” companies do not tend to outperform smaller companies in general (disregarding some market phases).

Table 1: Correlation coefficients of patent data vs. stock performance

correlated parameters	correlation coefficient (np.corrcoef)

mean TR 5 years / stock return	0.51
relative gradient PAI / stock return	0.13
relative gradient TR / stock return	0.15
PS / stock return	-0.07
PAI / stock return	-0.06
TR 2009 / stock return	0.51
CI 2009 / stock return	0.41
PPI 2009 / stock return	0.37

In figures 4-7 we can see the scatter plots as well as the regressions lines for the parameter pairings *mean TR 5 years / stock return* (fig. 4), *TR 2009 / stock return* (fig. 5), *PPI 2009 / stock return* (fig. 6) and *CI 2009 / stock return* (fig. 7). The parameters of the regression lines are given in table 2.

Table 2: Regression line parameters slope and intersection

parameters	slope	intercept
mean TR 5 years / stock return	238	-261
TR 2009 / stock return	241	-216
CI 2009 / stock return	122	-67
PPI 2009 / stock return	95	325

With the eye the data look rather uncorrelated. A problem, effecting an increasing correlation coefficient, could the data points with very high stock return (> 1000 %).

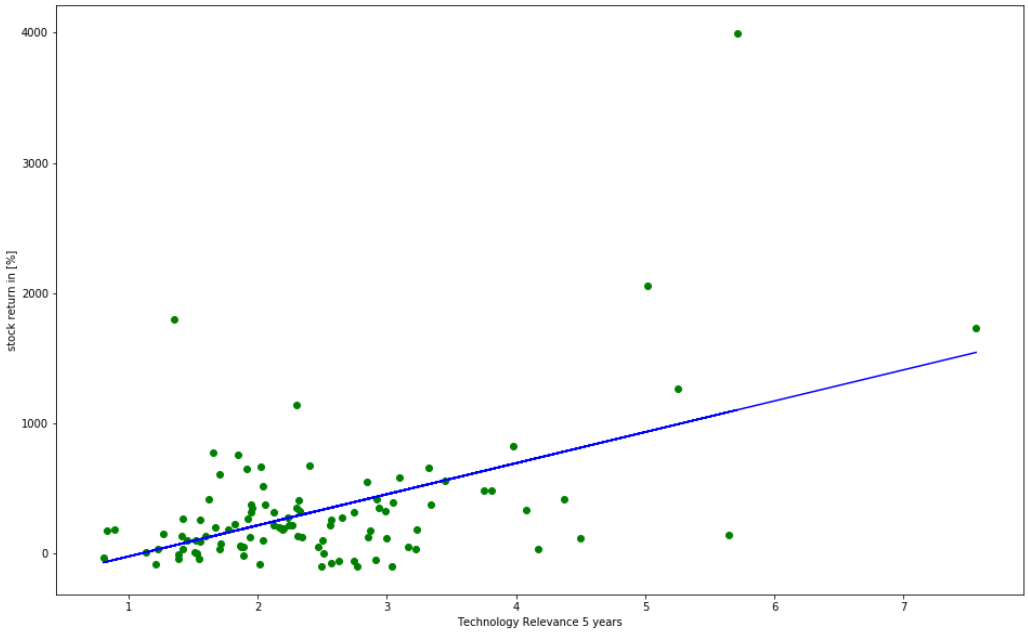

Figure 6: Scatterplot and regression line of mean TR for 5 years and stock return

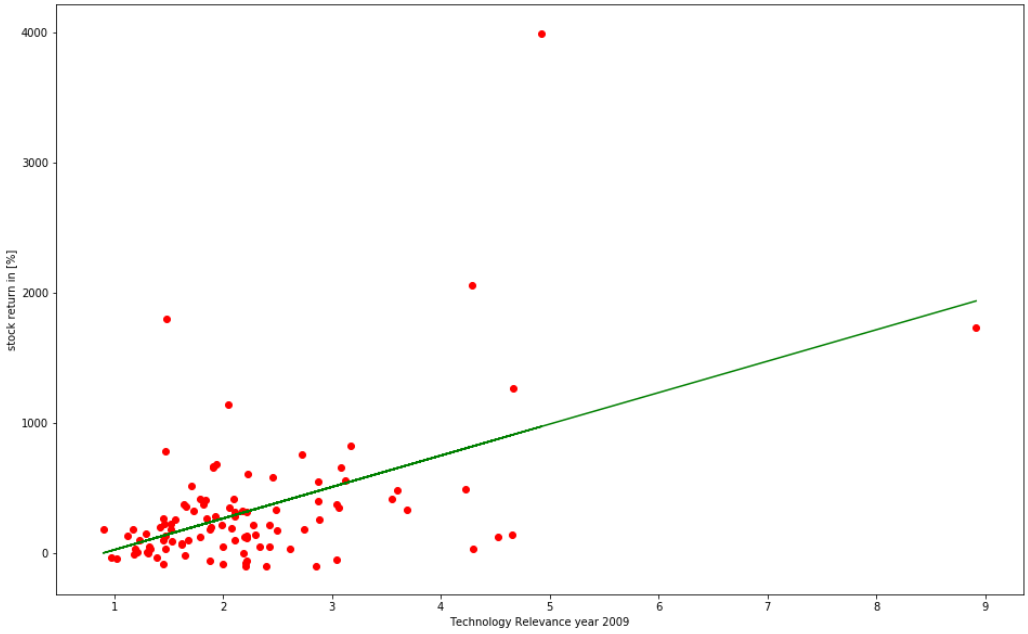

Figure 7: Scatterplot and regression line of TR in year 2009 and stock return

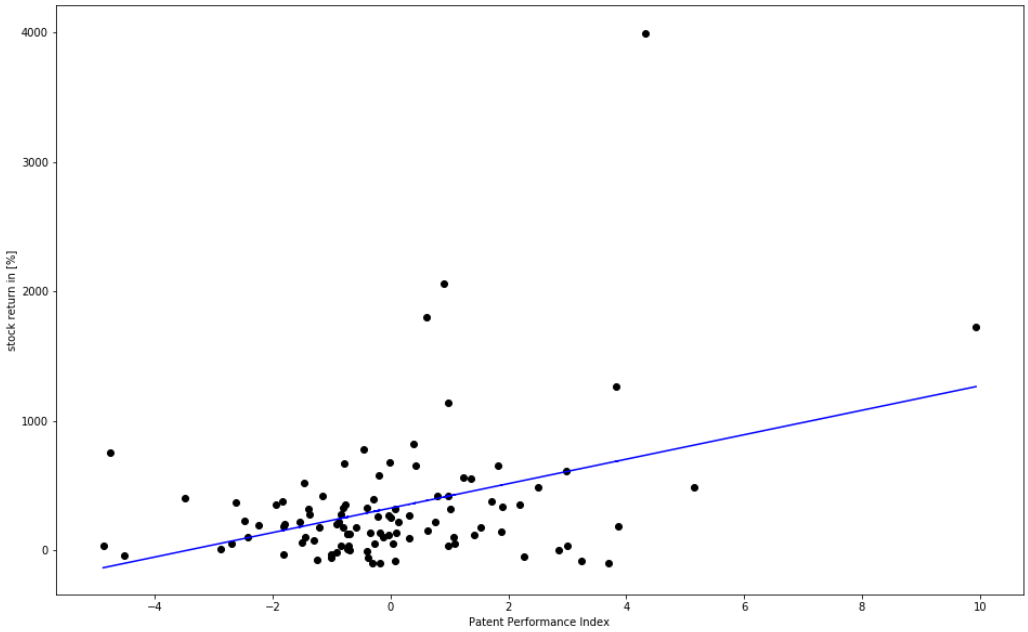

Figure 8: Scatterplot and regression line of PPI in year 2009 and stock return

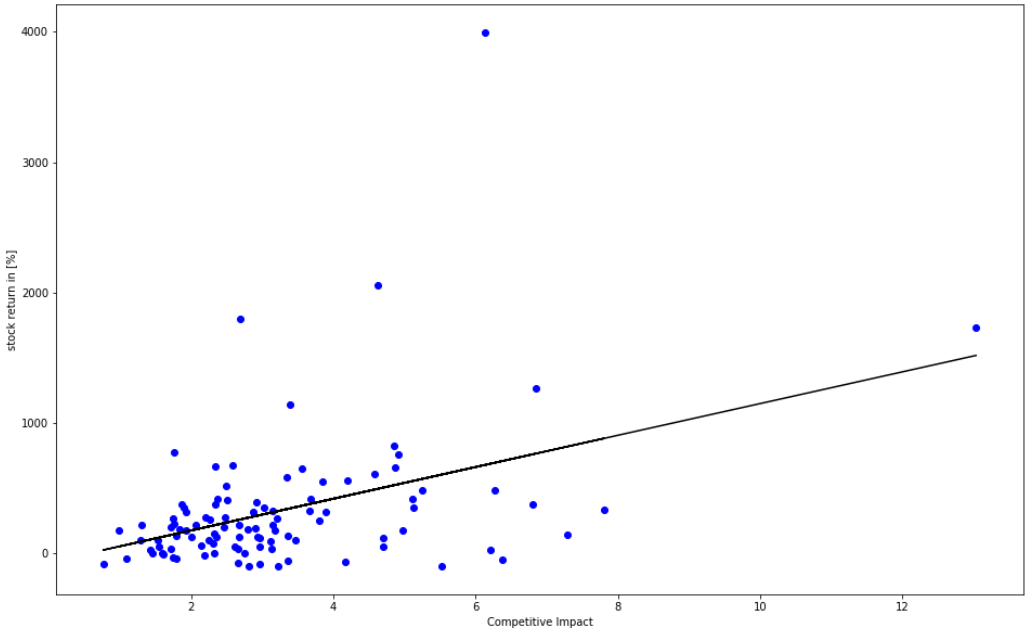

Figure 9: Scatterplot and regression line of CI in year 2009 and stock return

10 CONCLUSIONS

The results of this patent-stock research project disclose a rather mixed picture.

The patent and stock data, except for the relative gradient of the TR, do not follow a normal distribution (normality test see Jupyter Notebook “stock_project.ipynb” and fig. 2). A partial explanation for this could be the fact, that the number of 101 companies is to small bearing in mind that the active patent portfolio size varies in the range 9 - 24,000 in the year 2009.

The correlation analysis of the patent parameters “relative gradient PAI, relative gradient TR, PS and PAI” do not show a correlation to stock data. The patent parameters “PPI, CI, mean TR 5years and TR” show a small correlation with coefficients varying between 0.37 and 0.51, eventually indicating a stock market inefficiency, at least (somewhere) in the period 2005-2018.

This should be motivating to continue this research project, respecting the following weak points of the present analysis:

1. As stock return only the stock price was considered and not the total return (including stock price and dividends).
2. The companies are not grouped according to patent portfolio size and industrial/technical sector. That means in this analysis we compare healthcare companies with tech-companies like Apple, Microsoft, Amazon etc.
3. Statistical outliers, especially in the stock return data (return > 1000 %) could have a greater impact on the correlation parameters.
4. The analysis was performed for a very specific time period instead of a rolling period.
5. Due to the fact we take the current constellation of the Nasdaq Composite, we possibly have a bias of retrospective “successful” companies.

These problems should be tackled by future work.

11 REFERENCES

1. https://en.wikipedia.org/wiki/Efficient-market_hypothesis (status 2019/09/21)

2. https://www.patentsight.com/de/ (status 2019/09/21)

3. https://www.epo.org/searching-for-patents/data/bulk-data-sets/docdb.html#tab-1 (status 2019/09/21)

4. H. Ernst, N. Omland, The Patent Asset Index – A new approach to benchmark patent portfolios, World Patent Information 33, p. 34-42, 2011

5. https://www.finanzen.net/ (status 2019/09/21)

6. https://www.anaconda.com/distribution/ (status 2019/09/21)

7. https://ch.marketscreener.com/NASDAQ-COMP-4944/einzelwerte/

12 Appendix

12.1 Swiss National Science Foundation data management plan

Investigation of Patent Data on Stock Performance

Description

This project aims to investigate the stock market efficiency with reference to patent data.

Institution

This project takes place st the advanced training “Data Science” at the university of Bern by Sigve Haug 2019-2020.

Data Collection

The data needed/produced in this work can be divided into three parts:

1. Patent data by the data provider PatentSight
2. Data, calculated based on the PatentSight data
3. Stock return data by Finanzen.net

How will the data be collected or created?

The patent data by PatentSight will be exported to an Excel file. The stock data will be manually written into the same Excel file. The calculated data are as well calculated in the Excel file. The final result data will be exported to a second Excel file and used for the Python code.

Documentation and Metadata
What documentation and metadata will accompany the data?

The documentation is this report. It will be saved in the same GitHub repository. The metadata are given in the corresponding Excel files.

How will you manage copyright and Intellectual Property Rights (IPR) issues?
IPR rights could arise with the patent data of PatentSight. They are not publicity available; a license is required. Also, with a license the data have to be kept secure. A consultation of Carsten Guderian and Marco Richter from PatentSight on Tuesday 10.09.2019 yields the permission to use the data for this project.

Storage and Backup
How will the data be stored and backed up during the research?
During the project, the data, documents and files stored simultaneously locally on hard disk and the Microsoft Cloud OneDrive. Finally, the data will also we stored on the GitHub repository.