White Noise Resemblence of Equity Indexes across Different Industries: ANOVA Comparison

Orhan Koc
orhankoc@uw.edu

Jyunghyun Noh
jyungn@uw.edu

Siew Kit Liew
siewkit@uw.edu

March 8, 2023

Abstract

This study aims to compare the randomness or irregularity of different stock indexes, and see if different markets show different levels of irregualarity at randomly chosen time windows. Monthly price data is collected from NASDAQ, SP500, DOWJIA, and DOWJTA at random time windows with a length of 5-years. The irregularity of the collected data will be quantified using the concept of entropy. The results will be compared using Analysis of Variance to see if at least on of the markets are less irregular than the others. The assessment is conducted using One-Way ANOVA, with 1 independent variable on 4 different levels, blocked with respect to time snippets. The results of the study could provide insights into the robustness and generalization of Random Walk ARIMA(0,1,0) model in stock price modeling in different industries. The findings of this study will be valuable for investors, financial analysts, and academics interested in stock price forecasting and time series modeling.

1 Introduction

Since the dawn of finance, academia and the finance industry have made numerous attempts at systematically predicting the returns on risk bearing assets. The long lasting debate of whether stock markets are predictable or not has been steered towards the former with the emergence of new models. In this paper, we wish to compare the randomness of markets of different industries to ultimately understand whether some markets are more predictable than the others.

Price forecasting is a widely studied subject usually implemented with a variety of well-tested machine learning algorithms. One of the most comprehensive ways of describing a model with forecasting capabilities is the Autoregressive integrated Moving Average, also known as ARIMA, model. The robustness of this algorithm mainly stems from the fact that it's a combination of multiple methods of modeling. One of the apparent advantages of ARIMA is it's ability to model future values solely based on past values. We will consider the simplest version of ARIMA, the random walk (0,1,0) which itself is a cumulative sum of an independent and identically distributed process, ARIMA(0,0,0). The d=1 value means we will be taking the difference of the data to convert the trending prices into stationary data. Essentially, we want to see if some markets are more random than the other. Using one-way ANOVA, we will see if there is significant difference between the entropies of different equity indexes. A time series with high approximate entropy will be more irregular and more random than a time series with low approximate entropy. [1] The equity indexes are chosen such that there is little overlap in terms of industries they represent The results of this experiment can be useful in determining which index prices resemble a random walk, in other words which industries have a return that resembles a white noise more than the others.

The data will be collected from Federal Reserve Economic Data (FRED). In order to observe difference of randomness with respect to different industries, we chose to include equity indexes of different areas of employment: Technology, Transportation, Utility etc. We include the monthly average price for SP500, NASDAQ, DOSJUA, DOWJT indexes from 2013 to 2022. We will then pick random intervals of 5 years for each observation and compare a Random Walk model to the corresponding index and record the RMSE. If all markets are equally random, than there should be no significant difference in the RMSE values for indexes, H0 is correct.

2 Theory

2.1 Random Walk

Time series analysis of stocks to predict future prices has been a focus of research since the birth of finance. Among models that use past data alone to forecast future prices, ARIMA has been the most successful alternative. Auto-Regressive Integrated Moving Average (ARIMA) is composed of two parts:

Autoregressive Regression (AR) is a special case of linear regression where the output X_t is determined by a linear combination of past values.

$$X_t = a_1 x_{t-1} + a_2 x_{t-2} + \cdots + a_p x_{t-p} + w_t \tag{1}$$

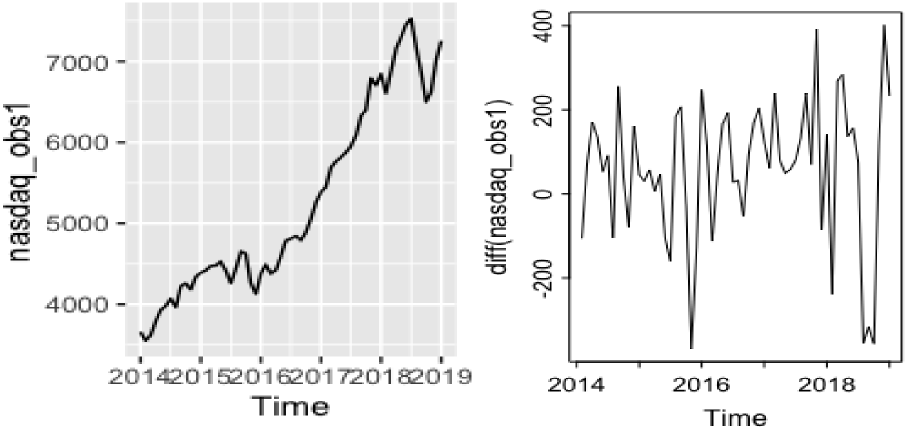
Moving Average (MA) attempts to state the mean for a period of time using a linear combination of past white noise.

$$X_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q} \tag{2}$$

Combining signal prediction of AR and noise prediction of MA together yields the relationship stated in Eq 3, and usually outperforms both AR and MA in terms of forecast accuracy.

$$X_t = a_1 x_{t-1} + a_2 x_{t-2} + \cdots + a_p x_{t-p} w_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q} \tag{3}$$

ARIMA function uses 3 parameters p,d and q: p as the lag order, d as the degree of differencing needed for stationarity and q as the order of the moving average. A special case of ARIMA is the ARIMA(0,d,0) function, which means the model has no AR or MA components, meaning the data is differentiated d times and then plotted with an independent, identically distributed process: white noise.

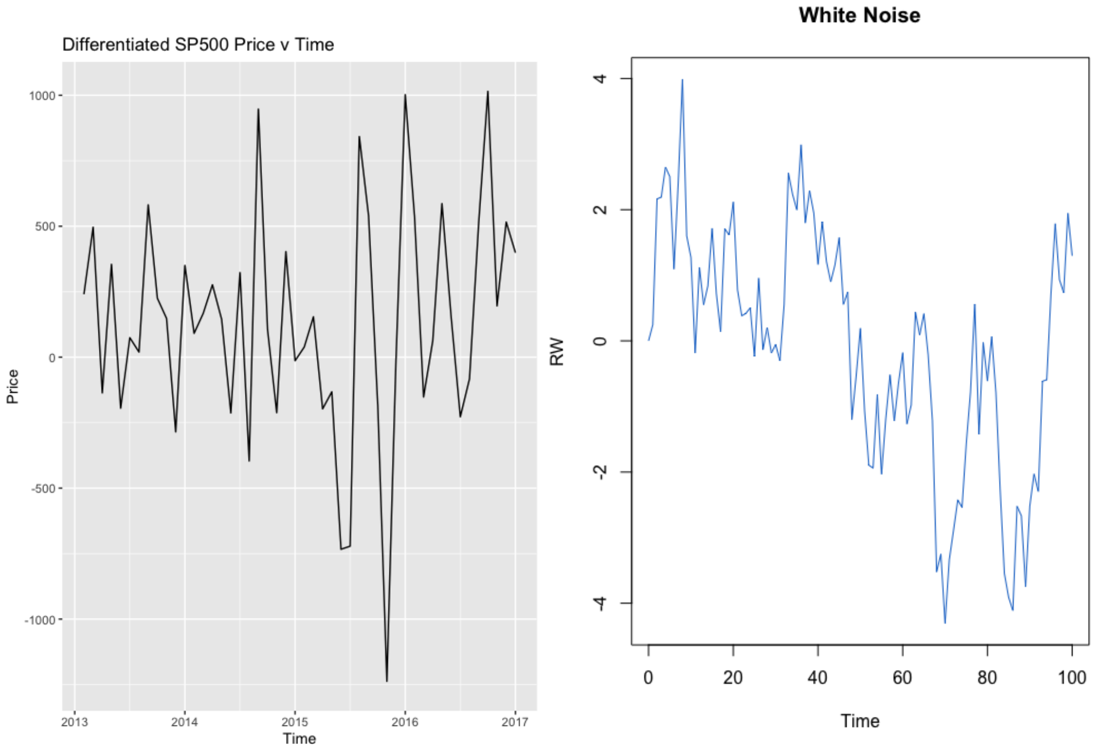


(a) Nasdaq

(b) Differenced Nasdaq

Figure 1: Original v. Stationarized

It is important to note that our model has $p = 0$ for AR component coefficient and $q = 0$ for MA component coefficient, while having a differencing order of 1. This means that we will be differencing the data once to make it stationary and calculate its approximate entropy to quantify randomness, to be able to compare it with white noise which serves as the foundation of ARIMA(p,d,q). When a time series data is stationary, it means the data has no significant change in mean or variance. Figure 1.a shows Nasdaq price vs time, and Figure 1.b shows the differentiated Nasdaq price over time, in other words a plot of periodic returns.



(a)

(b)

Figure 2: Differentiated Data v. White Noise

The resemblence in movement of white nosie vs differentiated price data can be observed in Figure 2. Many academics have also reached consensus on the success of white noise in modeling price data when seasonality and trend are stripped of the data, which is the reason why log difference is taken.[2]

2.2 One Way ANOVA

Analysis of Variance (ANOVA) is a statistical technique that is used to determine whether there is a significant difference between the means of two or more groups. The ANOVA test is used when there are multiple groups, and we want to compare the means of each group to determine whether there is a significant difference between them. The ANOVA test is based on the assumption that the data is normally distributed and that the variances of the different groups are equal.

The basic idea behind ANOVA is to compare the variation between groups to the variation within groups. If the variation between groups is much greater than the variation within groups, then there is a significant difference between the groups, and we can reject the null hypothesis.

The null hypothesis for ANOVA is that there is no significant difference between the means of the groups, and the alternative hypothesis is that there is a significant difference between the means of the groups. The test is based on the F-statistic, which is the ratio of the variation between groups to the variation within groups. If the F-statistic is large, then there is a significant difference between the means of the groups, and we can reject the null hypothesis.

There are different types of ANOVA tests depending on the number of groups being compared and the type of experimental design. We will be using a One-way ANOVA when there is only one independent variable, and it has more than two levels. The single independent variable is Equity indexes and the levels are different combination of stocks forming different equities

3 Experiment Design

Sampling Unit This study sample consisted of three different stocks as multiple levels of a single factor ANOVA design. These stocks are all featured by the huge market cap between \$10 billion and \$200 billion and their shares mostly belong to the public. In other words, the stock prices are less likely to be manipulated by nonsignificant factors and the asset is less volatile since the market is well established and therefore the data is more reliable.

Equipment As the intention of our study is to test the validity of ARIMA model, we utilized 5 years of dependable data for input from Kaggle. Since we are dealing with time series data and model adequacy, we used `library(tseries)` to process time series data and `library(ggplot)` to visualize White Noise. In order to calculate the approximate entropy of time series data, we used `library(TSEntropies)`. The experiment will use Analysis of Variance, where observations (time intervals) will be used as blocks to mitigate the differences of approximate entropy due to times of volatility. Time intervals will be of same length, chosen at random.

Stock — Observations				
Equity Index	t_1	t_2	t_3	t_4
NASDAQ	$y_{1,1}$	$y_{1,2}$	$y_{1,3}$	$y_{1,4}$
SP500	$y_{2,1}$	$y_{2,2}$	$y_{2,3}$	$y_{2,4}$
DOWJU	$y_{3,1}$	$y_{3,2}$	$y_{3,3}$	$y_{3,4}$
DOWJT	$y_{3,1}$	$y_{3,2}$	$y_{3,3}$	$y_{4,4}$

Table 1: Observations Table

Hypothesis 0: *Observed index stocks are of same irregularity, one market is not significantly more random than the others.* $\mu_1 = \mu_2 = \mu_3 \ldots \mu_n$

Hypothesis 1: *Observed index stock prices are not of same irregularity, at least one market is significantly more random than the others.* $\mu_a \neq \mu_b$ *where* $\mu_a, \mu_b \in \{\mu_1, \mu_2, \mu_3, \ldots, \mu_n\}$

Dependent Variable of this experiment is the approximate entropy of the log differentiated price for each market.

Independent Variable of this experiment are stock prices, with 4 different levels included to represent different industries.

4 Data Processing

4.1 Preprocessing

Granularity 10 years monthly price data fetched from Fred for S&P500 companies and loaded into R workstation. To address known performance issues of ARIMA, daily data was converted into monthly price average data with a frequency of 12 from FRED's Graphical User Interface.

Stationary Data ARIMA(0,0,0) works best on data with no trend and constant variance. Most financial instruments, including stock prices have a trend with non-constant variance. In order to prepare the data for ARIMA timeseries forecast, we will compute the difference of logs for each data set and conduct a Dick & Fuller's (DF) test. The results of the function `adf.test(data)` from `library(tseries)` on the differenced timeseries data `diff(ts_obs1)`, will yield a p-value where

Hypothesis 0: $p > 0.05$, *Time series data is not stationary.*

Hypothesis 1: $p < 0.05$, *Time series data is stationary*

meaning time series with a DF result of $p < 0.05$ are workable with ARIMA. For time series that failed DF test the first differentiation will be differenced again.

Date Date format should be converted to MM-DD-YY in order to create a timeseries object R can work with.

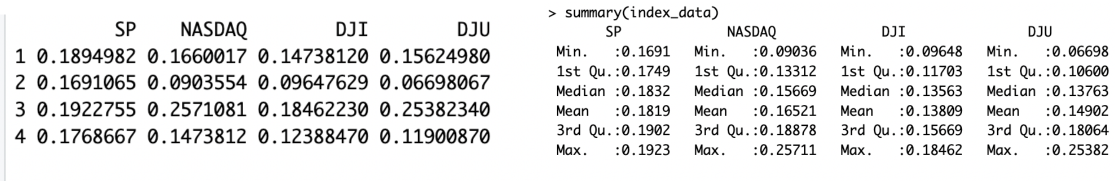
4.2 Data Collection

In order to test the generalization of white noise modeling, we will be testing index prices of securities from different industries. We will be looking at SP500, NASDAQ, DJIA, DJTA

A quantitative difference regarding securities of different industries is the change in volatility and trend strength. This apparent difference in volatility among indexes was not chosen by the experimenters, but rather a natural result due to the change in risk appetites of the investors of corresponding industries. Trend and seasonality is expected to be mitigated using logarithm and differentiation.

5 Results

The recorded entropy for each index at each time window can be seen in Figure 3.a and the statistical summary for the entropy table can be seen in Figure 3.b. We can see that SP had a larger mean entropy than the rest of the markets, followed by Nasdaq.



```
         SP    NASDAQ        DJI        DJU
1 0.1894982 0.1660017 0.14738120 0.15624980
2 0.1691065 0.0903554 0.09647629 0.06698067
3 0.1922755 0.2571081 0.18462230 0.25382340
4 0.1768667 0.1473812 0.12388470 0.11900870
```

(a) Entropy table

```
> summary(index_data)
       SP            NASDAQ             DJI               DJU         
 Min.   :0.1691   Min.   :0.09036   Min.   :0.09648   Min.   :0.06698  
 1st Qu.:0.1749   1st Qu.:0.13312   1st Qu.:0.11703   1st Qu.:0.10600  
 Median :0.1832   Median :0.15669   Median :0.13563   Median :0.13763  
 Mean   :0.1819   Mean   :0.16521   Mean   :0.13809   Mean   :0.14902  
 3rd Qu.:0.1902   3rd Qu.:0.18878   3rd Qu.:0.15669   3rd Qu.:0.18064  
 Max.   :0.1923   Max.   :0.25711   Max.   :0.18462   Max.   :0.25382  
```

(b) Table summary

Figure 3

The table acquired by adjoining the entropy values of different indexes as columns are then converted into LONG form to conduct ANOVA using `melt` function from `reshape2` package. The resulting ANOVA results can be seen in Table 2:

	Df	Sum Sq	Mean Sq	F Value	Pr(>F)
Markets	3	0.00440	0.00147	0.469	0.710
Residuals	12	0.0376	0.00313		
Total	15	0.042			

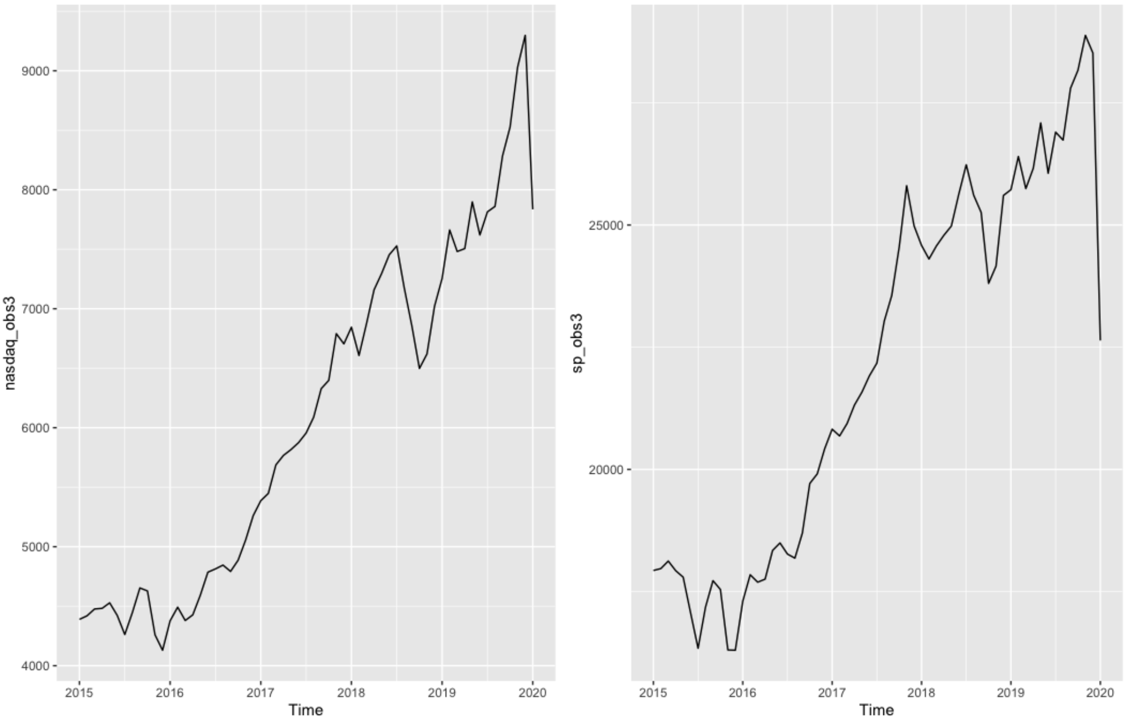
Table 2: ANOVA Table

It's important to note in Figure 3.a, the range of values pertaining to row 3 have a significantly larger mean of entropy. Row 3 data refers to price data that was collected between 2015 - 2020 for 4 different indexes. The prices collected during this time window were largely affected by the selling pressure due to COVID virus. The following Figure shows the original prices plotted against time:

The sharp drop due to COVID sell-off can be better seen when the time series is differentiated once. The experimenters did not remove the aforementioned unhealthy results to keep the sampling process random. See Figure 5 for the log differentiated, `diff(log(sp_obs1))` price plots.

We analyzed 4 different levels, which resulted in 3 degrees of freedom for markets and 12 degrees of freedom for residuals given we had 4 observations. From the F Table in Appendix A, we observed the critical Value for F with the following parameters, Df1 = 3, Df2 = 12, $\alpha = 0.05$ which yields:

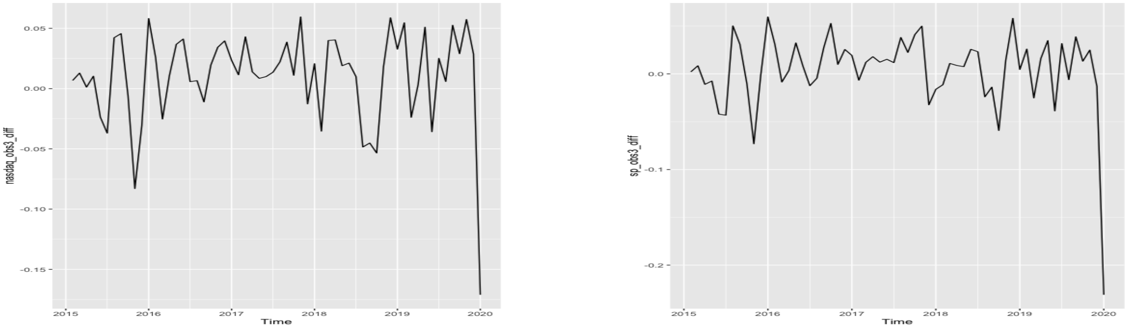
$$F_c = F_{0.05,3,12} = 3.49$$



(a) NASDAQ

(b) SP500

Figure 4

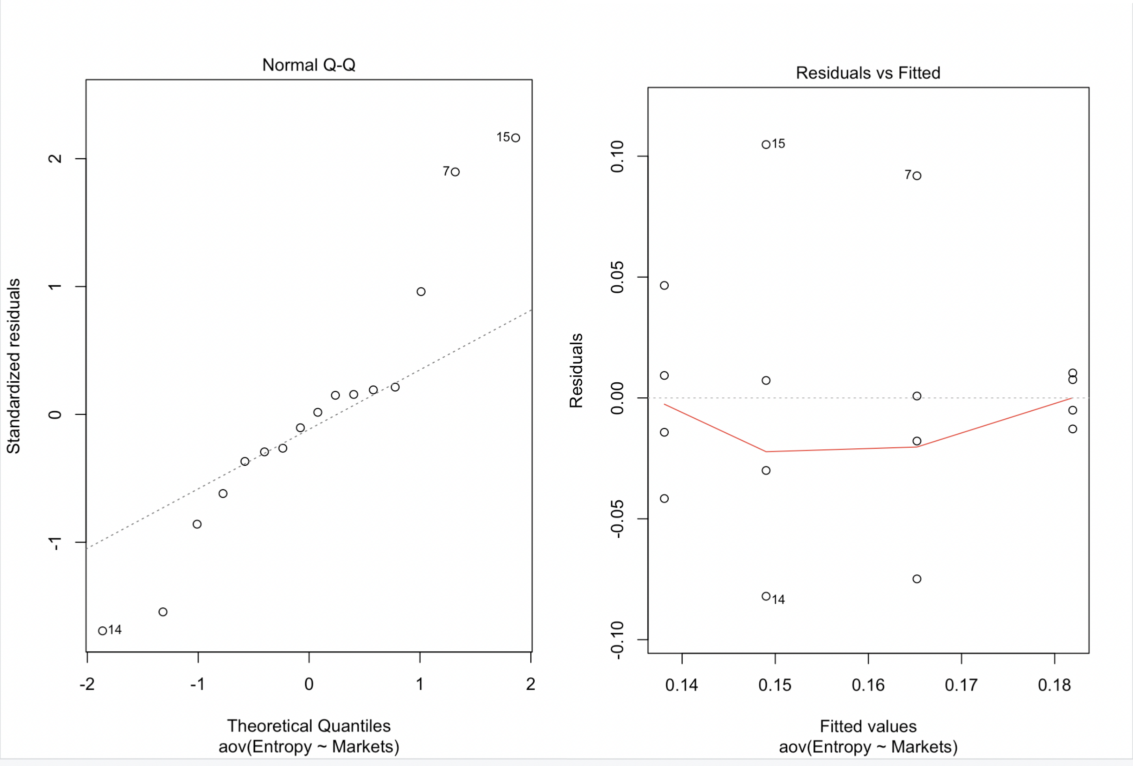


(a) Differentiated NASDAQ

(b) Differentiated SP500

Figure 5

since $F < F_c$, we conclude there are no significant difference between markets in terms of entropy.



(a) Q-Q Plot

(b) Residual v Fitted

Figure 6: Adequacy Checking

QQ Plot The points follow the line closely, we can assume the data is normally distributed.

Residual v Fitted The points appear to be random, with no apparent trend. There is a slight curve seen in the middle of the plot, but the evidence isn't enought to conclude non-normal distribution for this experiment.

6 Discussion

6.1 Evaluation

We have recorded the entropy values for NASDAQ, SP500, DOWJIA and DOWJTA markets observed for 5 years, replicated for four times. We have stationarized the log scaled data to calculate and compare the approximate entropy for each stock, of each time window. We have conducted Analysis of Variance for this data to compare the variance between different markets to the variance within; in order to make a conclusion about whether entropy changes from market to market.

6.2 Limitations & Future Work

As is the case for most experiments, we expect to see more accurate results with a replication on a higher scale.

Observations Due to time constraints of course schedule, we were only able to conduct the experiment for 4 different time windows. In order to observe the appearance of Random Walk, we need to collect more observations to make sure the hypothesized independent and identically distributed behavior of markets converge. Even though one way anova has a minimum required numer of observations of 3, estimating the randomness inherently requires a large number of observations conducted on a large numer of IV levels.

Levels Also the number of levels of independent variable, equity indexes, were limited to four - which in return increases the room for error. If we had included more levels of our independent variable,

Overlap As much as we tried to pick indexes so that they have small overlap in terms of stocks included, the prominent indexes such as SP500 include a wide range of popular assets, which are also chosen by other indexes. This small but direct correlation between levels of independent variable hinders the accuracy of the experiment.

7 Conclusion

In this paper, we have conducted ANOVA on the observations of entropy calculations for 4 different equity indexes; repeated the observation for randomly selected time windows out of the available data for four different times. We aimed to find out if the entropy for each equity index is the same across different industries. The ticker with higher entropy would resemble white noise returns or random walk price movement.

According to our ANOVA results, we couldn't find signficant entropy difference between the aforementioned markets, with respect to 5-yr time windows chosen. The findings are parallel to the theoru of modeling financial markets with Random Walk or modeling stock market returns with white noise. We conclude that there is no significant evidence to prove the set of markets: NASDAQ, SP500, DOWJIA, DOWJTA have an outlier in terms of approximate entropy or irregularity.

References

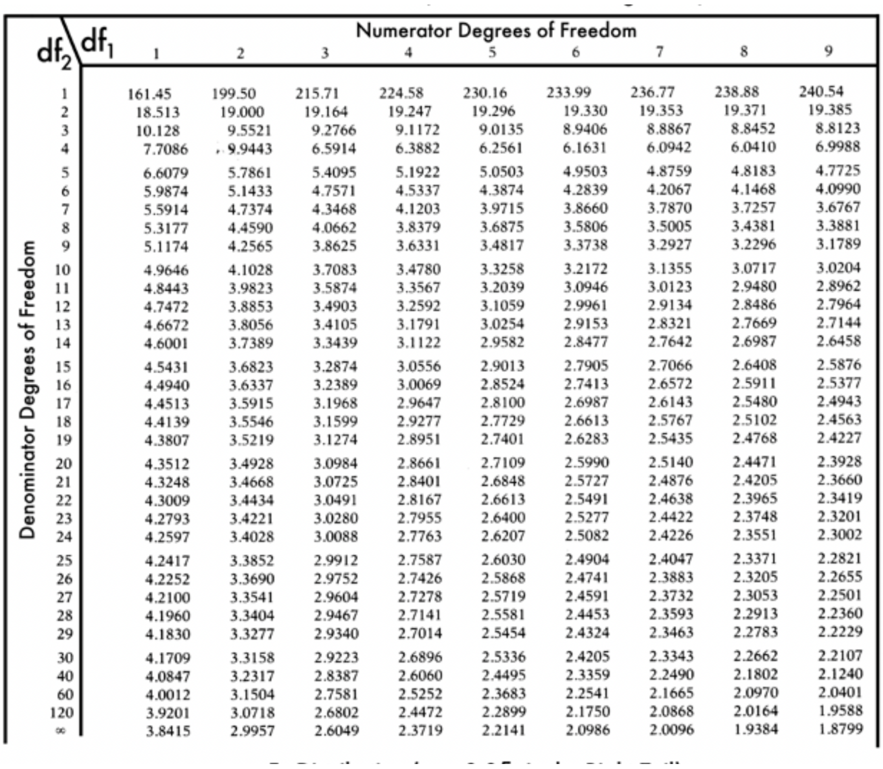
[1] Pincus S, Kalman RE. Irregularity, volatility, risk, and financial market time series. Proc Natl Acad Sci U S A. 2004 Sep 21;101(38):13709-14. doi: 10.1073/pnas.0405168101. Epub 2004 Sep

9. PMID: 15358860; PMCID: PMC518821.

[2] Pincus, S. M. Approximate entropy as a measure of system complexity. Proceedings of the National Academy of Sciences 88, 22972301 (1991).

Appendices

A F Table



Denominator Degrees of Freedom		Numerator Degrees of Freedom								
	df2\df1	1	2	3	4	5	6	7	8	9
	1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54
	2	18.513	19.000	19.164	19.247	19.296	19.330	19.353	19.371	19.385
	3	10.128	9.5521	9.2766	9.1172	9.0135	8.9406	8.8867	8.8452	8.8123
	4	7.7086	9.9443	6.5914	6.3882	6.2561	6.1631	6.0942	6.0410	6.9988
	5	6.6079	5.7861	5.4095	5.1922	5.0503	4.9503	4.8759	4.8183	4.7725
	6	5.9874	5.1433	4.7571	4.5337	4.3874	4.2839	4.2067	4.1468	4.0990
	7	5.5914	4.7374	4.3468	4.1203	3.9715	3.8660	3.7870	3.7257	3.6767
	8	5.3177	4.4590	4.0662	3.8379	3.6875	3.5806	3.5005	3.4381	3.3881
	9	5.1174	4.2565	3.8625	3.6331	3.4817	3.3738	3.2927	3.2296	3.1789
	10	4.9646	4.1028	3.7083	3.4780	3.3258	3.2172	3.1355	3.0717	3.0204
	11	4.8443	3.9823	3.5874	3.3567	3.2039	3.0946	3.0123	2.9480	2.8962
	12	4.7472	3.8853	3.4903	3.2592	3.1059	2.9961	2.9134	2.8486	2.7964
	13	4.6672	3.8056	3.4105	3.1791	3.0254	2.9153	2.8321	2.7669	2.7144
	14	4.6001	3.7389	3.3439	3.1122	2.9582	2.8477	2.7642	2.6987	2.6458
	15	4.5431	3.6823	3.2874	3.0556	2.9013	2.7905	2.7066	2.6408	2.5876
	16	4.4940	3.6337	3.2389	3.0069	2.8524	2.7413	2.6572	2.5911	2.5377
	17	4.4513	3.5915	3.1968	2.9647	2.8100	2.6987	2.6143	2.5480	2.4943
	18	4.4139	3.5546	3.1599	2.9277	2.7729	2.6613	2.5767	2.5102	2.4563
	19	4.3807	3.5219	3.1274	2.8951	2.7401	2.6283	2.5435	2.4768	2.4227
	20	4.3512	3.4928	3.0984	2.8661	2.7109	2.5990	2.5140	2.4471	2.3928
	21	4.3248	3.4668	3.0725	2.8401	2.6848	2.5727	2.4876	2.4205	2.3660
	22	4.3009	3.4434	3.0491	2.8167	2.6613	2.5491	2.4638	2.3965	2.3419
	23	4.2793	3.4221	3.0280	2.7955	2.6400	2.5277	2.4422	2.3748	2.3201
	24	4.2597	3.4028	3.0088	2.7763	2.6207	2.5082	2.4226	2.3551	2.3002
	25	4.2417	3.3852	2.9912	2.7587	2.6030	2.4904	2.4047	2.3371	2.2821
	26	4.2252	3.3690	2.9752	2.7426	2.5868	2.4741	2.3883	2.3205	2.2655
	27	4.2100	3.3541	2.9604	2.7278	2.5719	2.4591	2.3732	2.3053	2.2501
	28	4.1960	3.3404	2.9467	2.7141	2.5581	2.4453	2.3593	2.2913	2.2360
	29	4.1830	3.3277	2.9340	2.7014	2.5454	2.4324	2.3463	2.2783	2.2229
	30	4.1709	3.3158	2.9223	2.6896	2.5336	2.4205	2.3343	2.2662	2.2107
	40	4.0847	3.2317	2.8387	2.6060	2.4495	2.3359	2.2490	2.1802	2.1240
	60	4.0012	3.1504	2.7581	2.5252	2.3683	2.2541	2.1665	2.0970	2.0401
	120	3.9201	3.0718	2.6802	2.4472	2.2899	2.1750	2.0868	2.0164	1.9588
	∞	3.8415	2.9957	2.6049	2.3719	2.2141	2.0986	2.0096	1.9384	1.8799

(a) F Table

B R Script

```
library(forecast)
library(ggplot2)
library(tseries)
library(TSEntropies)
library(reshape2)
setwd("/Users/orhankoc/Documents/ARIMA_RCBD/scipts")
```

```
RW <- arima.sim(model= list(order = c(0, 0, 0)), n=60)
plot.ts(RW,main="White␣Noise", col=4)
autoplot(diff(sp_obs2)) + labs( y="Price", title="Differentiated␣SP500␣Price␣v␣Time")

################################################################################
#                                    ARIMA
#
################################################################################
# independent variable 1: SP500, monthly data
#
#______________________________________________________________________________#
# https://fred.stlouisfed.org/series/SP500#0
SP = read.csv("SP500.csv")
SP_ts <- ts(SP$DJIA, start=c(2013, 1), end=c(2022, 3), frequency=12)
# minimum sample is n=50 recommended for ARIMA
sp_obs1 <- window(SP_ts, c(2014), c(2019, 1))
sp_obs2 <- window(SP_ts, c(2012), c(2017, 1))
sp_obs3 <- window(SP_ts, c(2015), c(2020, 1))
sp_obs4 <- window(SP_ts, c(2013), c(2018, 1))

# Step 2: Check for stationarity
autoplot(sp_obs1)
autoplot(sp_obs2)
autoplot(sp_obs3)
autoplot(sp_obs4)

sp_obs1_diff <- diff(log(sp_obs1))
sp_obs2_diff <- diff(log(sp_obs2))
sp_obs3_diff <- diff(log(sp_obs3))
sp_obs4_diff <- diff(log(sp_obs4))

autoplot(sp_obs1_diff)
autoplot(sp_obs2_diff)
autoplot(sp_obs3_diff)
autoplot(sp_obs4_diff)

# check if difference of log data is stationary for SP500 using Dick Fulley test
#H0: The time series is non-stationary.
```

```
#HA: The time series is stationary.
adf.test(sp_obs1_diff) # is stationary
adf.test(sp_obs2_diff) # is NOT stationary
adf.test(sp_obs3_diff) # is NOT stationary
adf.test(sp_obs4_diff) # is stationary

# APPROXIMATE ENTROPY:
ApEn(RW, r = 0.1*sd(RW))
ApEn(sp_obs1_diff, r = 0.1*sd(sp_obs1_diff)) # 0.1894982
ApEn(sp_obs2_diff, r = 0.1*sd(sp_obs2_diff)) # 0.1691065
ApEn(sp_obs3_diff, r = 0.1*sd(sp_obs3_diff)) # 0.1922755
ApEn(sp_obs4_diff, r = 0.1*sd(sp_obs4_diff)) # 0.1768667

################################################################################
# independent variable 2: NASDAQ, monthly data
#
#______________________________________________________________________________#
# https://fred.stlouisfed.org/series/NASDAQ100#0
NASDAQ = read.csv("NASDAQ100.csv")
nasdaq_ts <- ts(NASDAQ$NASDAQ100, start=c(2013,1), end=c(2022, 3), frequency=12)

nasdaq_obs1 <- window(nasdaq_ts, c(2014), c(2019, 1))
nasdaq_obs2 <- window(nasdaq_ts, c(2012), c(2017, 1))
nasdaq_obs3 <- window(nasdaq_ts, c(2015), c(2020, 1))
nasdaq_obs4 <- window(nasdaq_ts, c(2013), c(2018, 1))

# Step 2: Check for stationarity
autoplot(nasdaq_obs1)
autoplot(nasdaq_obs2)
autoplot(nasdaq_obs3)
autoplot(nasdaq_obs4)

nasdaq_obs1_diff <- diff(log(nasdaq_obs1))
nasdaq_obs2_diff <- diff(log(nasdaq_obs2))
nasdaq_obs3_diff <- diff(log(nasdaq_obs3))
nasdaq_obs4_diff <- diff(log(nasdaq_obs4))

autoplot(nasdaq_obs1_diff)
```

```r
autoplot(nasdaq_obs2_diff)
autoplot(nasdaq_obs3_diff)
autoplot(nasdaq_obs4_diff)

# check if difference of log data is stationary for SP500 using Dick Fulley test
#H0: The time series is non-stationary.
#HA: The time series is stationary.
adf.test(nasdaq_obs1_diff) # is stationary
adf.test(nasdaq_obs2_diff) # is NOT stationary
adf.test(nasdaq_obs3_diff) # is stationary
adf.test(nasdaq_obs4_diff) # is stationary

ApEn(nasdaq_obs1_diff, r = 0.1*sd(nasdaq_obs1_diff)) # 0.1660017
ApEn(nasdaq_obs2_diff, r = 0.1*sd(nasdaq_obs2_diff)) # 0.0903554
ApEn(nasdaq_obs3_diff, r = 0.1*sd(nasdaq_obs3_diff)) # 0.2571081
ApEn(nasdaq_obs4_diff, r = 0.1*sd(nasdaq_obs4_diff)) # 0.1473812

################################################################################
# independent variable 3: DOWJ Transportation average, monthly data
#
#______________________________________________________________________________#
# https://fred.stlouisfed.org/series/DJTA#0
DJT = read.csv("DJTA.csv")

DJT_ts <- ts(DJT$DJTA, start=c(2013,1), end=c(2022, 3), frequency=12)
djt_obs1 <- window(DJT_ts, c(2014), c(2019, 1))
djt_obs2 <- window(DJT_ts, c(2012), c(2017, 1))
djt_obs3 <- window(DJT_ts, c(2015), c(2020, 1))
djt_obs4 <- window(DJT_ts, c(2013), c(2018, 1))

# Step 2: Check for stationarity
autoplot(djt_obs1)
autoplot(djt_obs2)
autoplot(djt_obs3)
autoplot(djt_obs4)

djt_obs1_diff <- diff(log(djt_obs1))
djt_obs2_diff <- diff(log(djt_obs2))
```

```r
djt_obs3_diff <- diff(log(djt_obs3))
djt_obs4_diff <- diff(log(djt_obs4))

autoplot(djt_obs1_diff)
autoplot(djt_obs2_diff)
autoplot(djt_obs3_diff)
autoplot(djt_obs4_diff)

# check if difference of log data is stationary for SP500 using Dick Fulley test
#H0: The time series is non-stationary.
#HA: The time series is stationary.
adf.test(djt_obs1_diff) # is stationary
adf.test(djt_obs2_diff) # is NOT stationary
adf.test(djt_obs3_diff) # is stationary
adf.test(djt_obs4_diff) # is NOT stationary

ApEn(djt_obs1_diff, r = 0.1*sd(djt_obs1_diff)) # 0.1473812
ApEn(djt_obs2_diff, r = 0.1*sd(djt_obs2_diff)) # 0.09647629
ApEn(djt_obs3_diff, r = 0.1*sd(djt_obs3_diff)) # 0.1846223
ApEn(djt_obs4_diff, r = 0.1*sd(djt_obs4_diff)) # 0.1238847

################################################################################
# independent variable 4: DOWJ Utility average, monthly data
#
#______________________________________________________________________________#
# https://fred.stlouisfed.org/series/DJUA#0
DJU= read.csv("DJUA.csv")

DJU_ts <- ts(DJU$DJUA, start=c(2013,1), end=c(2022, 3), frequency=12)
dju_obs1 <- window(DJU_ts, c(2014), c(2019, 1))
dju_obs2 <- window(DJU_ts, c(2012), c(2017, 1))
dju_obs3 <- window(DJU_ts, c(2015), c(2020, 1))
dju_obs4 <- window(DJU_ts, c(2013), c(2018, 1))

# Step 2: Check for stationarity
autoplot(dju_obs1)
autoplot(dju_obs2)
autoplot(dju_obs3)
autoplot(dju_obs4)

dju_obs1_diff <- diff(log(dju_obs1))
dju_obs2_diff <- diff(log(dju_obs2))
dju_obs3_diff <- diff(log(dju_obs3))
dju_obs4_diff <- diff(log(dju_obs4))

autoplot(dju_obs1_diff)
autoplot(dju_obs2_diff)
autoplot(dju_obs3_diff)
autoplot(dju_obs4_diff)

# check if difference of log data is stationary for SP500 using Dick Fulley test
#H0: The time series is non-stationary.
#HA: The time series is stationary.
adf.test(dju_obs1_diff) # p<0.05, Reject H0 at d=1, is stationary
adf.test(dju_obs2_diff) # p>0.05, Reject H0 at d=1, NOT stationary
adf.test(dju_obs3_diff) # p<0.05, Reject H0 at d=1, NOT stationary
adf.test(dju_obs4_diff) # p<0.05, Reject H0 at d=1, is stationary

ApEn(dju_obs1_diff, r = 0.1*sd(dju_obs1_diff)) # 0.1562498
ApEn(dju_obs2_diff, r = 0.1*sd(dju_obs2_diff)) # 0.06698067
ApEn(dju_obs3_diff, r = 0.1*sd(dju_obs3_diff)) # 0.2538234
ApEn(dju_obs4_diff, r = 0.1*sd(dju_obs4_diff)) # 0.1190087
################################################################################
#                               ONE WAY ANOVA
#
################################################################################

# ENTROPY ANOVA
SP = c(0.1894982, 0.1691065, 0.1922755, 0.1768667)
NASDAQ = c(0.1660017, 0.0903554, 0.2571081, 0.1473812)
DJI = c(0.1473812, 0.09647629, 0.1846223, 0.1238847)
DJU = c(0.1562498, 0.06698067, 0.2538234, 0.1190087)

index_data = data.frame(SP, NASDAQ, DJI, DJU)
index_data
```

```
colnames(index_data)
index_data$id <- c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16)
index_data$id

index_data$Markets

summary(index_data)

index_data_long <- melt(index_data,
                    variable.name = "Markets",
                    value.name = "Entropy")

model = aov(Entropy ~ Markets, index_data_long)
summary(model)
plot(model)
```