Stock Market Global Nexus

Augusto Chang, Ernesto Mealla, Alfredo Celedón
Georgia Institute of Technology
This manuscript was compiled on December 8, 2023

We aim to explore the intricate interplay of global stock markets by analyzing the Istanbul Stock Exchange dataset in the UCI Machine Learning Repository comprised with data from eight major international stock market indexes: S&P, DAX, FTSE, NIKKEI, BOVESPA, MSCE_EU, and MSCI_EM from June 5, 2009, to February 22, 2011. We will employ multiple statistical methods, including multiple linear regression, time series analysis, and non-seasonal Box-Jenkins models, to understand the relationships between these indexes and their influence on one another. The study seeks to uncover patterns, dependencies, and trends within the global financial ecosystem during the specified timeframe.

Description of Dataset: For this project, the dataset consists of daily returns (from closing to closing) for each of the mentioned indices over a period of 18 months from June 5, 2009, to February 22, 2011. Data is collected from imkb.gov.tr and finance.yahoo.com according to data source owners. The features employed will be the eight indices in the data set, the date for which the return values are recorded and season of the year (an additional categorical feature derived from date classification)

Data Headers: Each data sample (row) shows the returns for the 8 stock market indices for a given day.

All the data for the indices has a floating-point data type and is displayed to over 6 decimal places of accuracy.

- Istanbul Stock Exchange National 100 Index (XU100)
- Standard & Poor's 500 Return Index (S&P)
- Stock Market Return Index of Germany (DAX)
- Stock Market Return Index of the UK (FTSE)
- Stock Market Return Index of Japan (NIKKEI)
- Stock Market Return Index of Brazil (BOVESPA)
- MSCI European Index (MSCE_EU)
- MSCI Emerging Markets Index (MSCI_EM)
- Date: (datetime dd/mm/yyyy) shows the day for which the sample data is being collected. This is a unique identifier for each row
- Season: (string: winter, spring, summer, fall) shows the season corresponding to the date on the northern hemisphere

Data Pre-Processsing: To preposition the date into meteorological seasons we used the northern hemisphere cut offs (given that the feature is going to be used to analyze the USA's S&P. Cut offs:

- Spring: March 1 to May 31
- Summer: June 1 to August 31
- Fall: September 1 to November 30
- Winter: December 1 to February 28

To perform the MLR, we aggregated the day-to-day variances for each of the indices cumulatively, this helped us portray the total difference over time and therefore, make trends more visible.

Exploratory Analysis:

The summary statistics table presents the mean, standard deviation, and range for each index. Notably, some indices exhibit a positive mean, suggesting an upward trend over time, while others show a downward trend. This observation provides an initial insight into the directional movements of the indices.

Statistic	ISE	SP	DAX	FTSE
Mean	0.00039	0.00053	0.00078	0.00020
Standard Deviation	0.02112	0.01409	0.01456	0.01266
Max	0.10062	0.06837	0.05895	0.05032
Min	-0.08472	-0.05426	-0.05233	-0.05482
	NIKKEI	BOVESPA	EU	EM
Mean	-0.00007	-0.00013	0.00016	0.00032
Standard Deviation	0.01485	0.01575	0.01299	0.01050
Max	0.06123	0.06379	0.06704	0.04780
Min	-0.05045	-0.05385	-0.04882	-0.03856

Figure 1: Summary Statistics Table

Overview of Methods: This project employs three main statistical methods to analyze the data:

- Multiple Linear Regression (MLR)
- Time Series Regression (TSR)
- Non-Seasonal Box-Jenkins

Multiple Linear Regression Analysis

In this analysis, the S&P index serves as the dependent (response) variable, while the 7 other stock market indices (XU, DAX, FTSE, NIKKEI, BOVESPA, MSCE_EU, MSCI_EM) act as independent (explanatory) variables. The goal is to explore the

relationship between the S&P index and its dependence on various global and regional indices. To enhance the precision of coefficient estimates and improve generalization, we utilized data spanning the entire time-frame from Jan 5, 2009, to Feb 22, 2011.

Multiple Linear Regression Model I :

$Y = 0.0034 - 0.334x1 + 0.3781x2 + 0.6944x3 + 0.2264x4 + 0.0157x5 - 0.1951x6 + 0.0462x7$

Our analysis using python produced the following results, which are also shown in Fig 8 in appendix:
The y-intercept of our model was 0.0034, and its significance was assessed through a p-value of 0.239, indicating that it was not significantly different from zero. Moving on to the coefficients associated with each independent variable, these values elucidate the impact of each predictor on the dependent variable. For a detailed breakdown of these coefficients, please refer to the first column of Fig 8: MLR I Model Output in the appendix. The associated p-values are located in the fourth column. As we aimed for a 95% confidence interval, we established a significance threshold of 0.05, implying that any coefficient with a p-value exceeding this threshold was deemed statistically insignificant. In our assessment, ISE, DAX, FTSE, NIKKEI, and EU demonstrated significance within the model, while other variables did not meet our criteria. This aligns with our F-statistic, which yielded a p-value of 0, suggesting the presence of at least one significant variable in the model. Furthermore, the Adjusted R-squared value, at 99.1%, underscores a strong fit of regression of our model, signifying its ability to explain 99.1% of the variance in the dependent variable.

Multiple Linear Regression Model II:

$Y = 0.0098 - 0.011x1 + 0.3810x2 + 0.3810x3 + 0.6948x4 + 0.2402x5 - 0.1814x6$

For this model, we incorporated the variables found to be significant in our Multiple Linear Regression Model I. These include the ISE, DAX, FTSE, NIKKEI, and EU. The results of this analysis are presented in Fig 9: MLR II model output in the appendix.

The y-intercept for our model held a value of 0.0098 which was found to be significantly distinct from zero, as indicated by its p-value of 0. The coefficients of the independent variables can be seen in the first column of Figure 9 in the appendix. Associated p-values for each coefficient are located in the fourth column, where a 95% confidence interval was applied. Based on our criteria we concluded that the ISE coefficient was not significant enough while DAX, FTSE, NIKKEI, and EU were all statistically significant. Again, this observation aligns with our F-statistic, yielding a p-value of 0 meaning that at least one variable is significant. Then the Adj. R-squared of the model suggests that the model explains 99.1% of the variance in the dependent variable.

Despite the reduction in the number of independent variables to reduce overfitting, both MLR models exhibit the same high explanatory power, as evidenced by the 0.991 Adjusted R-squared values.

Diagnostic Checking MLR II: To assess the random error assumptions in the model, we conducted a Q-Q plot and an Anderson-Darling test. The resulting high p-value of 0.456 led us to fail to reject the null hypothesis, indicating no significant departure from normality and confirming that residuals are normally distributed. This finding is further supported by the histogram of residuals in Figure 3, which illustrates an approximate mean of 0, satisfying the mean-zero test.

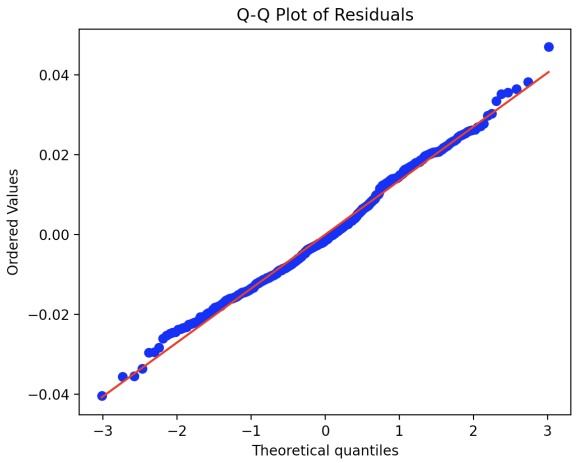

Figure 2: Quantile-Quantile plot of Residuals

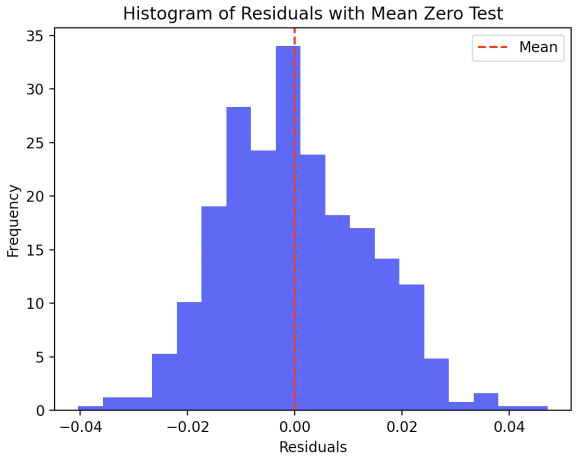

Figure 3: Histogram of Residuals

Time Series Regression Analysis

Our focus for the time series regression model is centered on the S&P index, our goal is to estimate the coefficients of the independent variables and model their relationship with the S&P over time. By leveraging time-ordered data, we aim to fit and evaluate the model, enabling us to forecast and predict future values accurately. This analysis will provide insights into the trends and patterns of the S&P index over time.

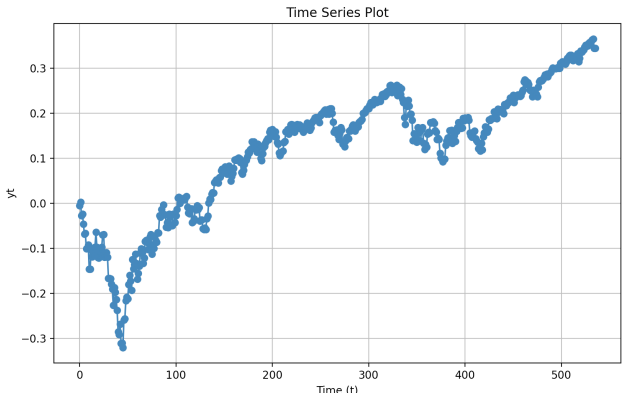

Figure 5: Time Series Plot

By plotting the Time Series, we can observe an initial downward trend. However, after that, the overall data tends to increase over time. A year consists of 260 data points (only weekdays); however, the data does not display clear evident seasonality. Upon inspection, it appears that the data exhibits constant variation and so no transformation is needed.

To gain a clearer understanding of the trends, we constructed a time series regression model incorporating both trend and seasonal variations. In this model, we introduced binary variables to represent the four seasons, with Winter serving as the base quarter.

Binary variables:

- Qtr1: Winter
- Qtr2: Spring
- Qtr3: Summer
- Qtr4: Fall

Y = - 0.0770 + 0.0008t - 0.0282q2 - 0.0434q3 - 0.0021q4

Our analysis summary results are shown in Figure 10: TSR model output in the appendix.

Analyzing the results of the time series regression model, we observe that the coefficient for parameter t is 0.0008 with a p-value of 0, indicating a small but significant upward trend over time. All independent variables, except for q4, exhibit statistical significance based on their p-values. The negative coefficients for q2 and q3 suggest that the index tends to decrease in Spring and Summer compared to its base value in Winter. Our Adjusted R-squared value is 80.1%, indicating that the model explains 80.1% of the variance. The F-statistic, with a p-value close to 0, demonstrates that at least one variable in the model is statistically significant.

These findings enhance our understanding of the trends and seasonal patterns influencing the S&P index in our time series regression model. However, we posit that an alternative model, such as the non-seasonal Box-Jenkins model, may potentially provide a better fit for our data.

Non-Seasonal Box-Jenkins Model Analysis

Our goal with the Box-Jenkins Model is to uncover the inherent patterns within the S&P index, treated as a univariate time series. To achieve this, we utilize a combination of ARIMA (AutoRegressive Integrated Moving Average) components, employing a mixed AR-MA model. This approach aids in identifying autocorrelation functions and facilitates the forecasting of future S&P return values. The following steps outline the process used for this analysis.

1. Making S&P data stationary: Fortunately, our raw data was already in first-difference form. By plotting this first difference over time as seen in Fig 6 above we observe a consistent mean and variance, coupled with the results of a Dickey-Fuller test yielding a p-value essentially of 0 (rejecting the null hypothesis), we can confidently assert that the data is indeed stationary.

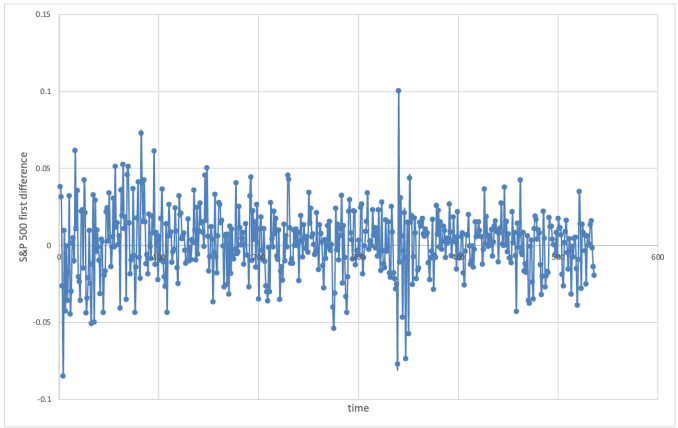

Figure 6: First difference (raw) over Time plot

2. Tentative Identification model: Then, we proceeded to plot the autocorrelation and partial autocorrelation functions, as illustrated in Fig. 7. The autocorrelation plot revealed a spike at lag 1, followed by an abrupt cutoff, indicative of a potential AutoRegressive (AR) parameter of 1. Additionally, the partial autocorrelation plot exhibited a damped exponential decay, suggesting a gradual decline with a cut-off point around lag 4. Therefore out tentative model is an ARIMA(1,0,4).

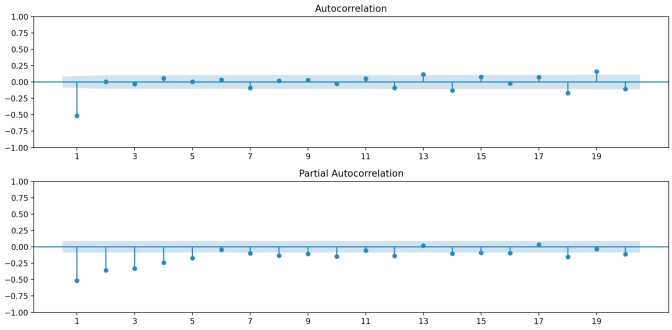

Figure 7: AC and PAC function plots

3. Estimation of Parameters: ARIMA (1,0,4)

y = - 04568ar.L1 - 0.4914ma.L1 - 0.4136ma.L2 + 0.0151ma.L3 + 0.0931ma.L4

For this analysis, we utilized Python again to fit a mixed ARIMA model with a tentative order of (1,0,4) shown in figure 11 in the appendix - featuring 1 AutoRegressive (AR) term, no differencing (given the series' already stationary nature), and 4 Moving Average (MA) terms. Upon close examination, as seen in column 5 of fig 11, our analysis revealed that none of the coefficients achieved statistical significance, with all p-values exceeding 0.05, except for the maL4 coefficient.

Diagnostic Checking: The Ljung-Box test statistic, compared to a chi-square distribution with corresponding significance and degrees of freedom (0.05 and 1), fell way below the critical value. As a result, we failed to reject the null hypothesis, suggesting an absence of significant autocorrelation at the tested lags.

Collectively, these findings indicate that the current (1,0,4) mixed ARIMA model might not be the optimal fit for the S&P time series and therefore, attempting to perform predictions would not be appropriate.

Concluding Remarks:

Throughout this project, our primary objective has been a thorough analysis and comparison of eight globally significant stocks spanning from January 5, 2009, to February 22, 2011. To uncover the intricate correlations among these stocks across various seasons, we applied three distinct methods: Multiple Linear Regression, Time Series Regression, and Non-Seasonal Box-Jenkins.

Our findings indicate a robust correlation between the US stock market and key international markets, specifically those of the UK (FTSE), Germany (DAX), Japan (NIKKEI), and the EU. Notably, shifts in these markets, whether positive or negative, are reflected in the US stock market or vice versa. We can conclude correlation not causality. The Time Series Regression unveiled intriguing seasonality in the S&P 500 over a year, suggesting a slight increase during winter compared to Spring or Summer. Finally, In our exploration of the Non-Seasonal Box-Jenkins Model, however, we found that it might not be the best fit for our data. Therefore we believe that further analysis should be done in order to find a model that fits the data.

For transparency and to encourage replication, the data, code used and results in this report are available for testing in the following repository: GitHub Repository.

Limitations and Future Work:

We have identified 3 main limitations in our project:
Limited Data Size: The dataset spans just over 2 years, which may not be sufficient to clearly reveal seasonality patterns. A more extended timeframe could provide a more comprehensive understanding of temporal trends.
Outdated Data: Our data extends back to 2011, making it over 10 years old. If we want to forecast any future values, obtaining more recent and up-to-date data is crucial for reflecting the current status of the stock market.
Complexity of Stock Markets: Stock markets inherently operate as complex and chaotic systems, posing challenges for accurate modeling. Our dataset, collected just after the 2008 stock market crisis during a period of recession, further heightened the complexity. This presented a challenge given our basic knowledge in statistical modeling. Moving forward, we aim to continue learning and explore more sophisticated models to better capture the intricacies of stock market behavior.

Acknowledgments

We want to thank professor Shihao Yang for his teachings and feedback during the class and project.

Oguz, A. (2013). *Istanbul Stock Exchange*. UCI Machine Learning Repository.

https://archive.ics.uci.edu/dataset/247/istanbul+stock+exchange

Bikos , K. (n.d.). *Defining Seasons*. Seasons: Meteorological and Astronomical. https://www.timeanddate.com/calendar/aboutseasons.html

Appendix:

Here are the model summary results using statsmodel.api library in python for reference.

OLS Regression Results

Dep. Variable:	cumulative_sp	R-squared:	0.991
Model:	OLS	Adj. R-squared:	0.991
Method:	Least Squares	F-statistic:	8532.
Date:	Mon, 27 Nov 2023	Prob (F-statistic):	0.00
Time:	20:56:33	Log-Likelihood:	1550.6
No. Observations:	536	AIC:	-3085.
Df Residuals:	528	BIC:	-3051.
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0034	0.003	1.178	0.239	-0.002	0.009
cumulative_ise	-0.0334	0.010	-3.443	0.001	-0.052	-0.014
cumulative_dax	0.3781	0.026	14.631	0.000	0.327	0.429
cumulative_ftse	0.6944	0.039	17.617	0.000	0.617	0.772
cumulative_nikkei	0.2264	0.015	15.356	0.000	0.197	0.255
cumulative_bovespa	0.0157	0.024	0.650	0.516	-0.032	0.063
cumulative_eu	-0.1951	0.052	-3.742	0.000	-0.298	-0.093
cumulative_em	0.0462	0.034	1.354	0.176	-0.021	0.113

Omnibus:	5.969	Durbin-Watson:	0.690
Prob(Omnibus):	0.051	Jarque-Bera (JB):	5.828
Skew:	0.250	Prob(JB):	0.0542
Kurtosis:	3.101	Cond. No.	144.

Figure 8: MLR I model output

OLS Regression Results

Dep. Variable:	cumulative_sp	R-squared:	0.991
Model:	OLS	Adj. R-squared:	0.991
Method:	Least Squares	F-statistic:	1.179e+04
Date:	Mon, 27 Nov 2023	Prob (F-statistic):	0.00
Time:	21:13:18	Log-Likelihood:	1546.3
No. Observations:	536	AIC:	-3081.
Df Residuals:	530	BIC:	-3055.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0098	0.002	5.178	0.000	0.006	0.014
cumulative_ise	-0.0111	0.006	-1.838	0.067	-0.023	0.001
cumulative_dax	0.3810	0.017	22.138	0.000	0.347	0.415
cumulative_ftse	0.6948	0.040	17.558	0.000	0.617	0.773
cumulative_nikkei	0.2402	0.013	18.063	0.000	0.214	0.266
cumulative_eu	-0.1814	0.049	-3.716	0.000	-0.277	-0.085

Omnibus:	5.029	Durbin-Watson:	0.701
Prob(Omnibus):	0.081	Jarque-Bera (JB):	5.130
Skew:	0.232	Prob(JB):	0.0769
Kurtosis:	2.882	Cond. No.	125.

Figure 9: MLR II model output

OLS Regression Results

Dep. Variable:	yt	R-squared:	0.801
Model:	OLS	Adj. R-squared:	0.800
Method:	Least Squares	F-statistic:	534.9
Date:	Mon, 27 Nov 2023	Prob (F-statistic):	1.20e-184
Time:	18:49:20	Log-Likelihood:	713.96
No. Observations:	536	AIC:	-1418.
Df Residuals:	531	BIC:	-1397.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0770	0.007	-10.410	0.000	-0.092	-0.062
t	0.0008	1.87e-05	42.763	0.000	0.001	0.001
Qtr2	-0.0282	0.008	-3.612	0.000	-0.044	-0.013
Qtr3	-0.0434	0.008	-5.745	0.000	-0.058	-0.029
Qtr4	-0.0021	0.008	-0.266	0.790	-0.017	0.013

Omnibus:	46.990	Durbin-Watson:	0.050
Prob(Omnibus):	0.000	Jarque-Bera (JB):	60.485
Skew:	-0.698	Prob(JB):	7.34e-14
Kurtosis:	3.870	Cond. No.	1.32e+03

Figure 10: TSR model output

SARIMAX Results

Dep. Variable:	SP	No. Observations:	536
Model:	ARIMA(1, 0, 4)	Log Likelihood	1507.903
Date:	Mon, 04 Dec 2023	AIC	-3001.806
Time:	13:08:38	BIC	-2971.817
Sample:	0	HQIC	-2990.074
	- 536		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
const	8.698e-06	9.1e-05	0.096	0.924	-0.000	0.000
ar.L1	-0.4568	0.358	-1.276	0.202	-1.158	0.245
ma.L1	-0.4914	0.350	-1.406	0.160	-1.176	0.194
ma.L2	-0.4136	0.336	-1.231	0.218	-1.072	0.245
ma.L3	0.0151	0.041	0.369	0.712	-0.065	0.095
ma.L4	0.0931	0.035	2.677	0.007	0.025	0.161
sigma2	0.0002	8.61e-06	24.410	0.000	0.000	0.000

Ljung-Box (L1) (Q):	1.25	Jarque-Bera (JB):	197.70
Prob(Q):	0.26	Prob(JB):	0.00
Heteroskedasticity (H):	0.29	Skew:	0.26
Prob(H) (two-sided):	0.00	Kurtosis:	5.93

Figure 11: Non-Seasonal Box Jenkins model output