A Time Series Analysis of Patent Filings and Stock Prices: A Proxy for Modeling the Relationship between Innovation and Profitability

Teddy Ganea[a*], Ky Friedman[b], and Simon Pritchard[c]

[a] *Stanford University, Department of Mathematics*
[b] *Stanford University, Doerr School of Sustainability*
[c] *Stanford University, Department of Biology*

Abstract: This study presents a novel use for patent filing data in predicting stock price as a proxy for innovation within a leading Japanese semiconductor company. After extracting feature variables out of patent filings, we fit linear models and SARIMAX models to the time series of stock price after both a simple log transform and a more complex decorrelation from a semiconductor stock market. In fitting the SARIMAX models, we employed an extensive hyperparameter grid search, testing over 14,000 configurations of order and seasonal order. We then compared each SARIMAX model to a SARIMA fit (without patent data) of the same order and seasonal order. In every model fitting, we used appropriate cross-validation techniques to compute test MSE. Our results demonstrated that decorrelation from the stock market vastly improved the performance of the models and that SARIMAX models fit on decorrelated data perform the best among the models we evaluated. Additionally, we found that for identical order and seasonal order configurations, SARIMAX with patent features as exogenous variables yielded lower test MSE in about 96% of cases than their pure SARIMA counterpart. While none of the models achieved predictive capacities great enough for market deployment, this study suggests a powerful opportunity to utilize company patent filings as a regressive variable in stock prediction models.

KEY WORDS time series, linear regression, ARIMA, SARIMA, SARIMAX, stock predictions, patents, innovation

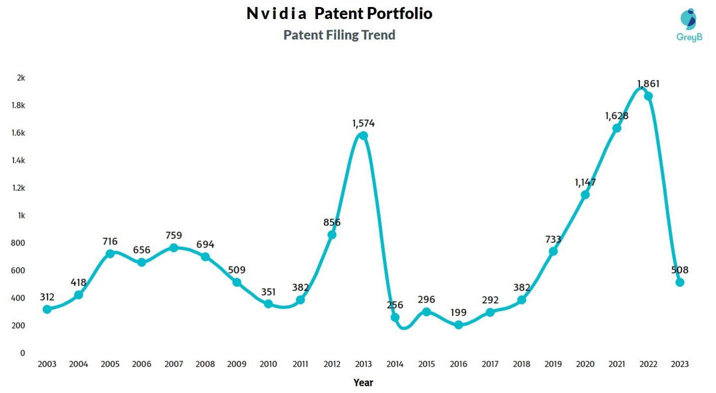
Received by Dr. Rothenhaeusler on 10th June 2024

1 Introduction

From 2022 to 2024, the value of Nvidia stock (NVDA) on the New York Stock Exchange increased by about a factor of ten, making life-changing profits for those who invested even modest retirement funds and savings. Predicting stocks remains a cornerstone of financial analysis as a result of stories like Nvidia, and the thousands that came before it. But Nvidia was an already successful and profitable company that seemingly took off along with the rest of the semiconductor industry. While some market trends and political moves were certainly insightful for long-term investors, Nvidia and similar companies have experienced a surge in innovation in large part driven by the AI revolution and the rapidly increasing demand for high-performance semiconductors which could have been utilized more explicitly in predicting this stock growth. A simple review of Nvidia's patent filings reveals a sharp increase from 199 in 2016 to 1,861 in 2022, demonstrating a meteoric rise in company innovation that was immediately preceded by a meteoric rise in company valuation.

*Correspondence to: tganea@stanford.edu

Prepared using oupau.cls *[Version: 2007/02/05 v1.00]*



Note: Patent filing year and all the patents in the portfolio are considered to plot the filing trend, Data Range Based on Publication Year (2003 - 2024).

Figure 1.1: In the years immediately preceding the sharp rise of NVDA stock, Nvidia saw a boom in patent publications. (Source: Insights by GreyB, available at: https://insights.greyb.com/nvidia-corporation-patents/)

We believe that patent filings have historically been overlooked as a powerful reflection of internal innovation and hence should be seen by financial analysts as a powerful regressor variable in their stock models. Patent data has been used previously in stock prediction models (Narin et. al., 2001) (Narin et. al., 2005) (Wu et al., 2017), but mostly just with simple linear models, typically lacking inclusion of standard time series analysis techniques. For this reason, we aimed to fill this noticeable gap in the literature and industry approach by evaluating the effect of including patent data in various models, once we bring the power of timeseries analysis to bear. Our main focus is not to identify the most powerful model which would require an even more extensive search, but instead to provide a high level analysis for the power of patent data in model development.

We tried three general classes of models: linear regressions, SARIMA, and SARIMAX. We suspected that there would be seasonal components to stock price and that SARIMA or SARIMAX would better explain than ARIMA; additionally prior work has already investigated modeling stock prices from patent data using ARIMA and cross validation techniques (Smith & Agrawal, 2015). In the linear regressions and SARIMAX, we used four feature variables from extracted from patent filings as predictors. In the case of SARIMA and SARIMAX, we explored 432 configurations of hyperparameter inputs for non-seasonal and seasonal AR components, MA components, and differencing. For every model fit, we trained the model on two separate data sets. The first used simply the log transformed stock data and the second decorrelated the stock data from the general semiconductor market before log transforming it.

All models were fit to the stock data and patent data from 2000 to 2023 of Tokyo Electron, a successful Japanese semiconductor company. We chose Tokyo Electron as our case study for two reasons: (1) they demonstrated a similar rise in stock value in the early 2020s to our leading example, and (2) we possessed extensive access to the company's patent filings.

2 Methods

2.1 Data Collection and Preprocessing

We obtained daily Tokyo Electron stock data from 2000 to 2023 from the publicly available Yahoo Finance database. Upon collection, we preprocessed the data in two methods:

- **Raw Data**: We elected to fit our model on monthly resolution to balance ensuring that the variance in patent features was reasonable between data points, while maximizing the amount of data to train our model on. To transform from daily to monthly data, we first averaged the daily closing stock price over a given month. After noticing that the monthly mean data was severely heteroscedastic, with far larger fluctuations as time (and the underlying stock price) increased, we logarithmically transformed these stock prices. This data remains highly non-stationary (ADF statistic $= -0.144$) but provided an baseline to compare with our decorrelated data.
- **Decorrelated Data**: To minimize drifts and trends in the data, we decorrelated our timeseries from the semiconductor market (specifically the PHLX Semiconductor Stock index: SOX). We chose the American SOX because, although it captures macroeconomic trends in the semiconductor market, the index is completely independent from Tokyo Electron; contrast to Japanese indices like Nikkei 225's semiconductor index, where 18% of its value comes from Tokyo Electron's share price (Nikkei Semiconductor Stock Index). After log-transforming the monthly mean stock price data for both Tokyo Electron and SOX (essentially the procedure for Raw Data), we standard-scaled each stock's timeseries by its respective mean and standard

deviation for a given time period, and then subtracted scaled SOX from scaled Tokyo Electron. It was this difference that we used as a proxy for how well Tokyo Electron was beating or underperforming the general trend of the market. While also non-stationary, this data was much more stationary (ADF statistic $= -2.19$) and serves as a simple and interpretable option for partially de-trending the data.
Worth noting is that in the decorrelation process, we chose not to convert Dollars to Yen. Keeping the respective currencies prevented us from confounding our model with changes in transnational currency strength unrelated to actual developments in the semiconductor industry.

- **Transformations for Evaluation**: For both preprocessing methods, following the model fitting, we inverted our preprocessing transformations to convert our model's predictions into units of Yen. These inverse transformations allowed us to generate test statistics (primarily MSE's) that could be easily interpreted and directly compared to actual stock price data. For **raw data**, we merely exponentiated our predictions to undo the log-transform. For **decorrelated data**, we first added back in to our predictions the previously scaled SOX data. We then rescaled this sum according to the mean and standard deviation of train Tokyo Electron data.

The raw and decorrelated data serve as a basis for our modeling, allowing us to subsequently examine how adjusting the input data to market trend impacts prediction accuracy.

We obtained Tokyo Electron patent filing data from VALUENEX, providing us with the full Japanese and English patent filings from 1986 to 2023. We utilized the patent data after performing a feature extraction on three versions of the English patent filings: the abstract, the full text, and the full text with the abstract upweighted. In trying each, we asked a critical question above how future models may rely on deep information about the patents encoded in the full text versus more surface level knowledge that could be scraped easily from the abstract.

2.2 Feature Extraction and Time Lag

As a textual dataset, a single patent cannot be used explicitly in a regressive model. Instead, we needed to perform a feature extraction from the patents in order to quantify the patents in some fashion. We binned our patent data on a monthly basis based on publication date, maintaining an equal time interval with our stock data. Based on previous work by one of our authors for VALUENEX, we utilized four features meant to represent distinct characteristics of a company's patent filings for each time interval. All features are based on an algorithm that contextually represents patents in a 2-dimensional coordinate space. The features are:

- **Novelty**: Measured by analyzing the change in the center of mass of patents over time to proxy shifts in the company's innovation focus.
- **Volume**: Simply the count of new patent filings in a given time interval.
- **Breadth**: Measured by averaging the distance of patents from the center of mass as a proxy for how spread out the company's innovation focus is at a given interval of time.
- **Depth**: Used a density estimation to measure the concentration of patents in the text space.

These features provide a monthly time series representing the company's general patent filings. However, filing a patent at a given time is both a signal of the company's current internal performance and their potential future performance (when they begin utilizing the patent). For this reason, we introduced a time lag component in which we offset the alignment of the stock price with the patent data on yearly intervals (0-10 years), attempting to identify if there is a delayed impact of innovation on stock prices, and if so, the length of this delay.

3 Model Building

As mentioned, we fit three classes of models:

- **Linear Regression**: Simple linear models were applied using all four patent features as predictors; this approach served primarily as a baseline for comparison with the more complex time series models.
- **SARIMA**: These models served as a control for the performance of the time series models since they do not include the patent data in their prediction, instead relying only on the classic seasonal and non-seasonal autoregressive and moving average models to fit and predict time series fluctuations.
- **SARIMAX**: Our most complex models, these incorporated the patent features as exogenous variables in capturing the autoregressive and moving average components of the stock data.

All three classes of models were fit on both the Raw and Decorrelated Data to test the impact of eliminating market trends from our Tokyo Electron data.

Since the linear models and SARIMAX models include the patent data, each model was trained on all three feature extraction types (abstract, full text, full text with weighted abstract). Then, each one of these models was trained using time lags of 0 years through 10 years for the patent data. So, for a simple linear model or a given SARIMA (p, d, q, P, D, Q, S) configuration, there are a total of thirty-three models. Since our SARIMA models don't use patent features as a regressor there was only one possible model for each SARIMA configuration.

We evaluate the models using the mean squared error (MSE) in squared Yen, obtained through the timeseries cross-validation algorithm shown in class. Specifically, for all valid values of n, we fit our models on the first n years in the data and then predicted on the months in year $n + 1$. For numerical stability, we began our timeseries cross-validation run on the first 5 years (20% of the data) as a warm-up period. After obtaining these predictions, we postprocessed them into units of Yen as described in 2.1, and then computed the MSE across months.

This protocol meant that every tested model was fitted 18 times, one for each year between 2005 and 2023. In total, we tested over 28,512 models with over 513,216 fits, requiring our team to run our code on 9 computers for about 20 hours.

When decorrelating Tokyo Electron from the semiconductor industry, we avoided data leakage by recalculating the fitting data for each time window since our scaling parameters were entirely dependent on which period was being trained on.

4 Results

4.1 Impact of Model Class and Decorrelation on Model Accuracy

Our most critical analysis deals with comparison of the three model classes (linear, SARIMA, and SARIMAX) and the two time series data sets (raw and decorrelated). Our expectation was that SARIMAX would perform better than SARIMA and linear models since it included both the patent data as well as the time series flexibility. Furthermore, we expected the decorrelated data to serve as a more powerful training set when compared to the raw data.

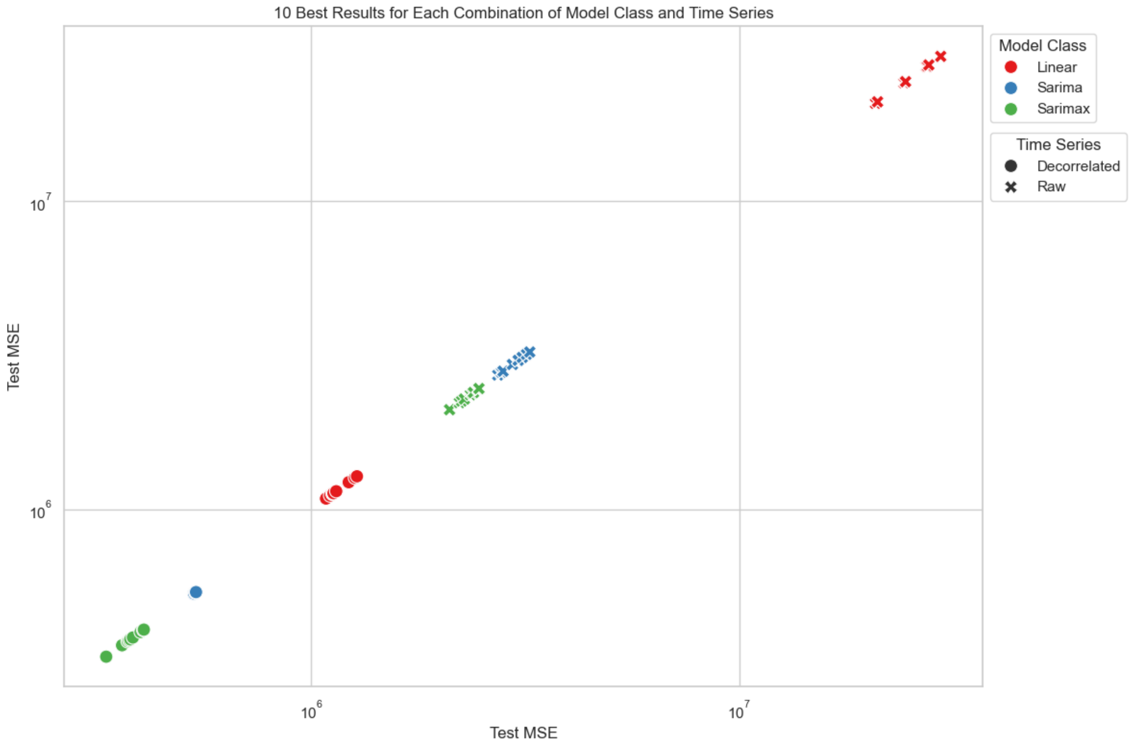


Figure 4.1.1: Results of the ten best models from each class and time series data set configuration

We examined the ten best achieved models within each model class and time series data set pairing, as measured by test MSE. We chose to focus on just these optimal models because of the high variability in performance across model classes depending on how many models were trained, and of high variability in model performance in general. For example, since we fit 33 SARIMAX models for every SARIMA model (see section 3), the test MSEs of SARIMAX models had enormously higher variance, with some models performing so poorly that they achieve test MSEs 10^{50} times worse than the best achieved. These outlying failures do not represent the true potential for the models, so we chose only to focus on the best models.

Our results show that, for a given time series data set, the linear models performed worse than the SARIMA models which performed worse than the SARIMAX models. Secondly, our results show that, for a given model class, the decorrelated data consistently yielded lower test MSE for the best models. Importantly, decorrelating Tokyo Electron from the broader semiconductor market significantly improved model performance.

In a direct comparison of SARIMA and SARIMAX models, we compared the best achieved test MSE for a given order and seasonal order configuration between the two model types. We tested a total of 432 hyperparameter configurations of order and seasonal order and found that SARIMAX was able to achieve a lower test MSE in 414 instances, outperforming SARIMA in 95.8% of the configurations. SARIMAX also performed better relatively to SARIMA on decorrelated compared to raw data.

4.2 Influence of Time Lag on Predictive Performance

We had been interested to see the effect of shifting the stock filing time series relative to the stock data on model performance. Based on previous exploratory and external work (Vitt & Xiong, 2015), there is evidence that patent filings possess a ”time lag” before they render financial impact for a company.

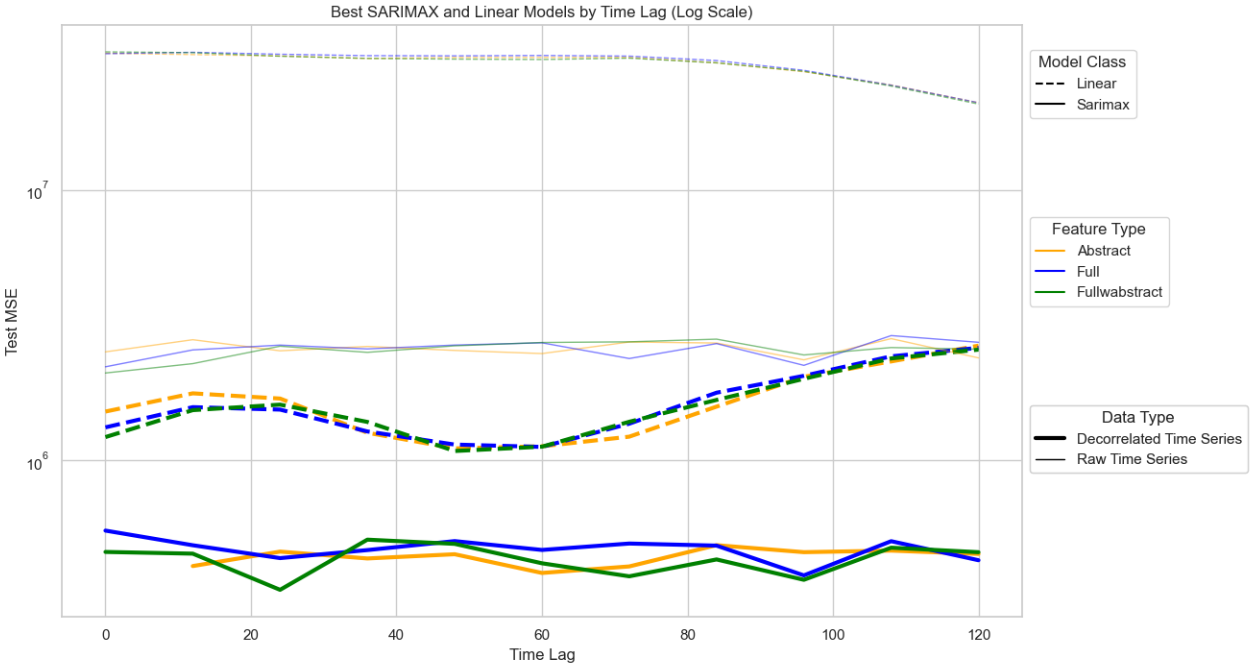


Figure 4.1.2: Comparison of the best achieved test MSE for a given model class, feature type, time series type, and time lag (in months)

Recall that, since SARIMA models did not incorporate patent data, they do not have a time lag component. When we compare the results of the linear and SARIMAX models, we again confirm the importance of decorrelation on model performance. We notice a difference in the general shape and behavior of the relationship between time lag and test MSE depending on model class. For the SARIMAX models, best achieved test MSE does not appear to have a clear trend with time lag. There are more smooth and identifiable trends between test MSE and time lag for the linear models, but these trends differ depending on which time series data set was used.

4.3 Impact of Feature Type on Model Performance

It was important that we examine how the patent text data used to extract the feature data impacts the model performance. We had two conflicting narratives: (1) the shorter text data (such as abstract) would give greater weight to the key words that distinguish patents from one another and hence perform better when capturing innovation features such as breadth and depth, and (2) the longer text data (such as full text) would give the feature data a more extensive view of the patents themselves and improve the models.

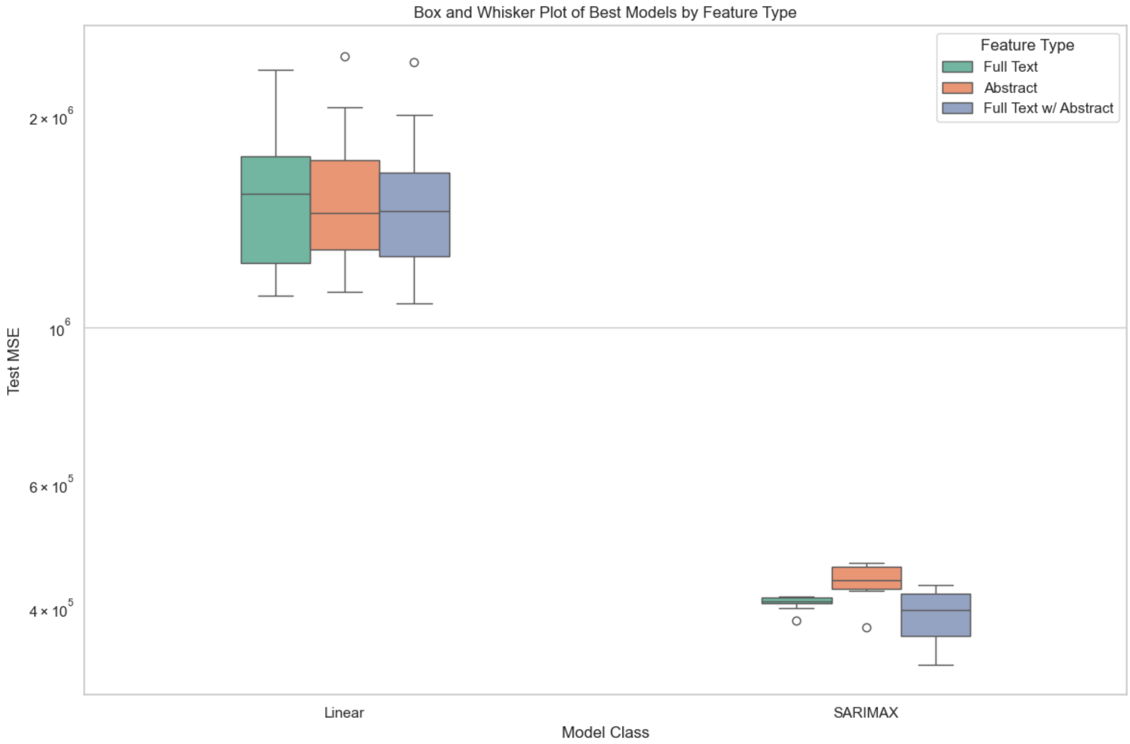


Figure 4.1.3: Comparison of the ten best linear models and SARIMAX models for a given feature type

Our results do not suggest a significant difference between models trained on the three different feature types (abstract, full text, and full text with abstract). The model class has a greater role in determining test MSE than the feature type, which often cannot be used to discern the best models of a given model class from one another.

5 Discussion

5.1 Relevance of Patent Data

Our findings underscore the opportunity to integrate patent filing time series into existing and novel stock prediction algorithms for improved prediction accuracy. We showed that the inclusion of patent features as exogenous variables to a SARIMA model improves our final predictive capacity consistently across hundreds of hyperparameter configurations. This result is meaningful because of our time series cross-validation regimen; additional features in linear regression (used in the X in SARIMAX) will almost always reduce *test* performance unless they are truly predictive. This suggests that patent data can provide unique insight into a company's performance that has real impacts on financial performance.

However, our results suggest that the type of patent filing text provided to the model is not particularly important. Since a model of a given class can perform nearly equally well regardless of the feature type provided, we believe that models can generally be trained on any of the three feature types to similar results.

We also note that not all patent data is the same. In this case study, we looked at a technology company in which patents may more directly reflect innovation and future product sales than other industries. Similarly, our models make no attempt to discern the value of given patents; one patent filed may have a far greater impact on stock price, a relationship we do not attempt to identify here.

5.2 Importance of Data Processing

Adjusting stock price data for general market trends was critical for improving model performance. This result was expected since the raw data was heavily non-stationary, especially when compared to our decorrelated data. Intuitively, decorrelating allowed our model to focus on the effect of company-specific innovation on stock price performance, by accounting for the complex slew of macroeconomic and industry-wide factors encapsulated in our general semiconductor index. However, our decorrelation method is far from industry standard, suggesting that a greater focus on de-trending - and more generally, on preprocessing - the data could yield sizable gains in model performance.

We also examined the role of time lag between patent filing and financial impact. Our results cannot conclusively determine a precise temporal link between patent filings and stock impact, especially because the impact of time lag varies with the model type and specific configuration. For example, when trained on decorrelated data, linear models appear to demonstrate a four to five year time lag between patent filing and financial impact, whereas the best SARIMAX models do not appear to demonstrate any precise relationship between test MSE and time lag. Other work, external to this study, noticed a consistent improvement in model performance around the eight year time lag, suggesting a long-term delay in financial benefit. This behavior is noticeable in our study, but is not as pronounced in other work. Contrary to our expectations, we did not see a shorter term improvement in model performance. In other unpublished work, we have noticed an improvement in model performance using a three year time lag, a result not observed in this study. Time lag may prove to be important, but its effect cannot easily be disentangled from other hyperparameters.

5.3 Effectiveness of SARIMAX Models

Our SARIMAX models outperformed both the linear models and SARIMA models when trained on the same data. This supports our hypothesis that the inclusion of time series analysis and patent data when predicting stocks can improve prediction accuracy. It should come as no shock that the linear models failed to keep up with SARIMA and SARIMAX models, but it is interesting that a SARIMAX model fit to raw data performs worse than a linear model fit to decorrelated data, indicating that decorrelation may have a more significant impact than model choice, further highlighting the need for strong decorrelation techniques.

6 Conclusion

Our study demonstrates the potential future role of patent filing time series as a component of successful stock price prediction models. We suggest that patent filings provide insight into a company's technological innovation which in turn plays a driving role in financial performance. The superior performance of SARIMAX models with decorrelated stock data further reveals that time series analysis techniques, detrending, and the inclusion of patent data can all improve performance over simple and even complex models. We hope to continue this research by expanding our model comparisons and improving our decorrelation techniques before integrating common predictive variables into our models.

Acknowledgment

This work would not have been possible without access to exclusive patent publication data provided by VALUENEX. Additional thanks are due for the STATS 207 teaching team who provided valuable comments on our initial research project proposal.

References

[1] Narin, F., Breitzman, A., & Thomas, P. (2005). Using patent citation indicators to manage a stock portfolio. In H. F. Moed, W. Glänzel, & U. Schmoch (Eds.), Handbook of Quantitative Science and Technology Research: The Use of Publication and Patent Statistics in Studies of S&T Systems (pp. 553–568). Springer Netherlands. https://doi.org/10.1007/1-4020-2755-9_26

[2] Narin, F., Thomas, P., & Brietzman, T. (2001). Using patent indicators to predict stock portfolio performance. https://www.researchgate.net/publication/307583065_Using_patent_indicators_to_predict_stock_portfolio_performance

[3] Smith, M., & Agrawal, R. (2015). A Comparison of Time Series Model Forecasting Methods on Patent Groups. https://ceur-ws.org/Vol-1353/paper_13.pdf

[4] Vitt, C. A., & Xiong, H. (2015). The impact of patent activities on stock dynamics in the high-tech sector. 2015 IEEE International Conference on Data Mining, 399–408. https://doi.org/10.1109/ICDM.2015.95

[5] Wu, S.-Q., Tsao, C.-C., Chang, P.-C., Fan, C.-Y., Chen, M.-H., & Zhang, X. (2017). A study of patent analysis for stock price prediction. 2017 4th International Conference on Information Science and Control Engineering (ICISCE), 115–119. https://doi.org/10.1109/ICISCE.2017.34