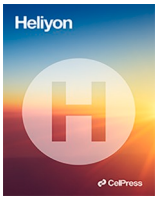
Heliyon 10 (2024) e28850

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Impact of coronavirus pandemic on stock index: A polynomial regression with time delay

Dong Bowen

Department of Applied Mathematics, Hong Kong Polytechnic University, 11 Yucai Road, Hung Hom, Hong Kong, Kowloon, China

A R T I C L E I N F O

Keywords:
Coronavirus pandemic
High-frequency data
Stock index
Polynomial regression
Time delay
Forecasting

A B S T R A C T

Motivation: Under contemporary market conditions in China, the stock index has been volatile and highly reflect trends in the coronavirus pandemic, but rare scientific research has been conducted to model the possible nonlinear relations between the two indicators. Added, on the advent that covid-related news in one time period impacts the stock market in another period, time delay can be an equally good predictor of the stock index but rarely investigated.
Objectives: To contribute to filling the gaps identified in existing research, this study models relationship between the stock market index and coronavirus pandemic by leveraging volatility in the stock market and covid data through time delay and best degree in a polynomial environment. The resultant optimal time delay and best degree model is used to derive a high-accuracy prediction of stock market index.
Novelty: In line with the possible relations, the novelty of this study is that it proposes, validates and implements polynomial regression with time delay to model nonlinear relationship between the stock index and covid.
Methods: This study utilizes high-frequency data from January 2020 to the first week of July 2022 to model the nonlinear relationship between the stock index, new covid cases and time delay under polynomial regression environment.
Findings: The empirical results show that time delay and new covid cases, when modelled in a polynomial environment with optimal degree and delay, do present better representation of the nonlinear relationship such predictors have with stock index for China. Relative to results from the polynomial regression without delay, the empirical evidence from the model with delay show that an optimal time delay of 17 weeks makes it possible to predict the stock index at high accuracy and record improvements of 16-fold or higher. The representative delay model is used to project for up to 17 weeks for future trends in the stock index.
Implication: The implication of the findings herein is that the prowess of the time delay polynomial regression is heavily dependent on instability in covid-related time trends and that researchers and decision-makers should consider modeling to cover for the unsteadiness in coronavirus cases to achieve better results.

1. Introduction

China is predicted to overtake the Unite States of America by mid-21st century in terms of national out. Due to the sharp growth recorded in China, profit oriented public and private institutions, in present-day business settings, are increasingly exposed to new

E-mail address: 21061246g@connect.polyu.hk.

https://doi.org/10.1016/j.heliyon.2024.e28850
Received 15 April 2023; Received in revised form 23 March 2024; Accepted 26 March 2024
Available online 31 March 2024

risks, growing compliance regulations [1], decreasing customer satisfaction [2], operational inefficiencies [3], fraud [4], or errors [5] that can lead to financial loss or reputational damage. In order to address these risks, there is an urgent demand for effective and sustainable decisions that utilizes innovative tools and techniques for the Chinese market. The financial and health sectors of China are among parts of the ecosystems of the nation that demand pressing solutions to issues brought about by the coronavirus pandemic [6].

The Chinese financial market and institution therein identify as discipline of operations management which handles day-to-day finance-related operations within institutions and steers them to improve in value [7]. Thus, effective decision tools and process management in businesses that uses proven methods to discover, design, model, analyze, develop, improve, and automate business processes and systems are required. Recently, under more generic terms like Business Process Management (BPM) [8], the influence of BPM is seen in the development of artificial intelligence (AI) solutions [9] specific to the banking and financial industry. Due to the surge of customer data expansion [10], competitive market retentions [11] and varying tastes and preferences of customers [12], financial institutions invest billions of dollars yearly into updating financial system procedures and capabilities through automation of processes that hitherto were performed manually by individual employees [13]. Under certain circumstances, bright expectations thwarts and business goals remain unachievable if the financial market fails to observe events, analyze, and optimize business processes through efficient tools and techniques, and refrains from receiving relevant recommendations from researchers for the enhancement of its performance indicators. It is worth mentioning that customers, who are often individuals, to consider not only performance indicators (including health records and policies implemented at the state level) in deciding to patronize financial products, particularly the stock market.

On health, recent COVID-19-related events the aftermath has had spill over effects on the stock market. The number of infections and death toll of COVID-19 pandemic has been on the rise since late December 2019 [14]. As means of controlling its spread, countries implemented purely technical measures (TMs) such as tests, research into potential vaccines, and travel restrictions [15]. Evidence from the sharp rise in infections, show that the TMs have been, for a large part, ineffective; and that the world was poorly prepared for the COVID event. Added, existing TMs are costly and time consuming [16]. Recent developments in control measures that minimize the shortfalls of traditional TMs include application-based and software-based approaches with individual-related, community-related, and government authority-run platforms. Countries including China, Singapore, Nigeria and South Africa are among the nations that have implemented various forms of the software-based COVID-19 tracking approaches [17]. Despite the prowess of existing application- and software-based approaches (EASAs), infections in countries like China are still high and the state has been implementing radical measures to combat the spread. Lessons learned from the limited impacts of TMs and EASAs are that unavailability of massive data on coronavirus-related issues [18], ineffective tracking and modeling of the variations in trends in new cases have hampered people's willingness to patronize the financial market.

Despite the inefficiencies recorded in the use EASAs in China, limited high frequency data on spread of the pandemic are tracked via combination of techniques [19]. For instance, greater proportion of the world's population have access to mobile phones and in nearly all countries, subscriber identification module (SIM) cards in mobile phones, and telephone lines are legally registered in, and tied to the personal contact details of users. For carriers of the virus, EASAs liaise regulatory bodies and instruct telecommunication companies to make available data on the historical pinging activity from towers for up to 2–3 weeks prior to the time the carrier was diagnosed positive for the COVID-19 virus [20]. Thus, a lookback effects are inculcated in EASAs and it helps ascertain massive data on carriers and their chain-like primary contacts, both for history and in real-times; which will sharply improve the precision, efficiency, and effectiveness of contact tracing.

In addition to having high frequency characteristics, there exists volatility in the stock index and new covid cases and that the expected such has high potential to exacerbate the unpredictability [21]. The inherent unpredictability implies that a sustained patronage of the stock market depends on some tenets. These tenets include investigating high frequency data for stock index and new covid cases, representative modeling of the observed non-linear trends in both variables, and high-accuracy forecasting that leverages the nonlinearity, particularly for China's stock market index and new confirmed COVID-19 cases. However, a critical review of literature show research that investigates high frequency stock index and new covid data, models of the non-linear trends observed in both variables, and proposes representative high-accuracy non-linear forecasting for China's stock market index and new confirmed COVID-19 cases, is rare. The panic emanating from COVID-19 lockdowns in China [22] driving considerable number of Chinese to limit investments in the financial markets [23] depict signs of possibly the existence of relationship between the stock market performance and new reported coronavirus cases in the country. Using the stock index as proxy for financial market and new COVID-19 cases as representative of the coronavirus pandemic, it can be established that both representations depict high frequency traits that are volatile and unpredictable. As high frequency variables, the inherent volatility necessitates that the form of relationship that possibly exist between the variables is, arguably, time-dependent and polynomial in nature. Despite the possibility of such relational links, critical review of existing research shows rare empirical evidence of studies that investigate the non-linear relationship between the stock index and new COVID-19 cases for China. These gaps motivate this research.

As an effort to contribute to filling the gaps identified, this research entails two innovations. First, the study proposes non-linear modeling of observed trends under polynomial regression environment with time delay for the stock index and new covid cases in China. The polynomial model with time delay is estimated separately for the stock index and new recorded COVID-19 cases, and jointly for the two variables. Second, the optimal time delays observed from the polynomial regression for the two variables [24] are used as input data to model, propose, test, validate and implement a high-accuracy polynomial model for forecasting the stock index up to 17 weeks ahead. In the midst of recent panic in China, the findings of this study would contribute to minimize the sizable number of Chinese who are deciding to decrease their participation of financial assets.

The remainder of the study is organized into four sections namely materials and methods, results and discussions, limitations and future research, and conclusion.

2. Materials and methods

2.1. Materials

This study uses stock index from the Shanghai Stock Exchange and newly and locally confirmed corona-virus cases (i.e., excluding imported ones) in China as materials for empirical analysis. Although the New York Stock Exchange, London Stock Exchange, and several other exchange markets exist, this study focuses on China, the covid-19 pandemic, and extreme variability in Chinese stocks. As such, the Shanghai Stock Exchange and newly and locally confirmed corona-virus cases are considered. The data on the stock index were extracted from Tong Hua shun [25], a widely used mobile application in China to access information and data on the stock market. On the new covid cases, the data were extracted from the official website of National Health Commission of People's Republic of China [26] for it is reliable according to state data-reporting standards. The data on stock index covers weekly timeline for a five-day week (i.e., Monday to Friday), whereas that of newly confirmed cases contains the total number of days in a whole week (i.e., Monday to Sunday). The five-day working week data for the stock index is used because the market operates from Monday to Friday excluding weekends and public holidays. Thus, no data exists for Saturday and Sunday on the stock index. Unlike the stock market, corona-virus cases have no definitive timeline and that new cases can be reported in any day of the week; hence, data on the stock index is extracted to cover Monday to Sunday. It must be noted that the data spans 2020 week 1–2022 week 27 (i.e., 131 weeks in total). The 131 weeks produces 655 data points for the stock index and 917 for new covid cases. The original data for the 131 weeks has been deposited in the supplementary materials under the file name Original Data. From the Original Data, it can be seen that there are a total of 51 and 19 entries of missing data on the stock index and new covid cases, respectively. The missing values are due to three reasons namely festivities (i.e., Spring Festival, Tomb-Sweeping Day, Dragon Boat Day, Mid-Autumn Day and Chinese National Day), public holidays (i.e., Labour Day, New Year) and absence of official records.

2.2. Data processing

As mentioned, the stock index values cover weekly timeline and the Grubbs test [27] was used to group them. The Grubbs test was adopted because according to the Grubbs Critical Value Table, the technology is applicable to data set with 3–100 entries; thus, the 3 to 5-day data fulfills the assumption. The mathematical expressions for utilizing the Grubbs test are presented in Eqs. (1) through (5):

$$x_i^{ave} = \frac{1}{n}\sum_{j=1}^{n} x_i^j \tag{1}$$

$$\triangle x_i = \left| x_i^j - x_i^{ave} \right| \tag{2}$$

$$x_i^{\max} = \max\{\triangle x_i\} \tag{3}$$

$$s_i = \sqrt{\frac{\left(x_i^j - x_i^{ave}\right)^2}{n-1}} \tag{4}$$

$$G_i = \frac{x_i^{\max}}{s_i} \tag{5}$$

where x_i^j is the stock index of the j^{th} entry of the i^{th} week, max is the maximum number, *ave* is the average, s_i is the standard deviation and G_i is the Grubbs statistic.

Python was used for the Grubbs test and the algorithms from Jupyter notebook is deposited in supplementary materials. It is worth mentioning that due to the 3 or higher entries requirement, the Grubbs test was applied to the weeks with at least 3 days of data. The test was used to both group and clean the data. The data cleaning procedure was necessary to cater for outliers. An entry is considered as outlier if from the Grubbs test, $G_i > G_{0.95}(n)$ where $G_{0.95}(n)$ is a critical value. For all values that were found to be outliers, such entries were deleted.

After the data was cleaned (i.e., grouped and checked for outliers), the remaining set was checked for missing entries. To get better use of each weekly data set, filling the missing data is, arguably, a proper way to address the issue. Relative to the stock market, the newly confirmed covid numbers is random and independent; thus, it is convincing to only fill the missing data for the stock index. Multiple imputation was used as a randomized approach to fill the missing data.

Among several approaches of multiple imputation, miceforest is chosen and adopts multiple imputation by chained Eqs. (MICE) [28]. The MICE involve a few steps. First, the original data are copied and saved into four different data files. The four files are named as *datafile1*, *datafile2*, *datafile3* and *datafile4*. Second, using *datafile1*, the technology arranges the data after the Grubbs test into five columns with each column representing a day (i.e., Monday to Friday). The MICE approach checks for missing values per column and randomly selects values already present in the column to fill sections with missing entries. This procedure is repeated for all 5 columns to produce data that contains no missing values from Monday to Friday. Third, holding four columns in *datafile1* constant, the random values in the fifth column imputed in the previous step are intentionally deleted. Fourth, using random forest applied to the four columns held constant in *datafile1*, the missing values (intentionally deleted) are predicted. The predicted values are considered as

imputed data. Assuming that Monday is the column whose entries were intentionally deleted and that Tuesday-Friday were held constant, the random forest technique is used to predict and impute the predictions as entries to filling the missing values in the Monday column. Fifth, once the Monday entries in *datafile1* are completed, the 'delete one apply random forest to the other four' technique is adopted to predict and fill in the missing entries for Tuesday, Wednesday, Thursday and Friday. Sixth, the procedure is repeated for three iterations using *datafile1*. The six procedures described above are replicated to the other three data files (i.e., *datafile2*, *datafile3* and *datafile4*). The results from the third iteration of each data file are saved. Correlation analysis is adopted in choosing the final multiple imputation data with no missing entries. Here, the imputed data that produces a correlation matric closest to that of the original data is considered the best fit.

After the multiple imputation process, it was observed that all the 131 weeks of data on stock index had no missing entries. The 5-day data per week for stock index was then aggregated into one value per week using the pandas rolling function [29]. Leveraging the model proposed by Theophilus et al. [29] and as depicted in Eqs. (6) and (7), the mathematical expressions used in the rolling function methods:

$$x_i^{ave} = \frac{1}{5}\sum_{j=1}^{5} x_i^{j} \tag{6}$$

$$ST_{ad}^{i} = \frac{1}{4}\left(x_{i-3}^{ave} + x_{i-2}^{ave} + x_{i-1}^{ave} + x_i^{ave}\right) \tag{7}$$

where x_i^j is the stock index of the j^{th} entry of the i^{th} week after multiple imputation, ST is stock index, and ad is adjust data. It can be observed that for the average of the stock index each week after multiple imputation (i.e., x_i^{ave}), i must be positive integer. In these cases, for the adjust data of stock index, it starts from week 4–131 (i.e., 128 weeks in total).

The non-missing data on new covid cases was also formatted into 128 weeks, i.e., from week 4–131. Unlike the data for stock index, that for new covid cases was unpredictable and it is better to use the sum of each week to aggregate into one value per week. These multiple imputations, rolling function and aggregation procedures produce 128 inputs of processed data with no missing entries both for the stock index and new covid cases.

Analytical techniques were applied to the processed data to estimate the functional relationship that exist among the stock index, new covid cases and the possible impacts of time delay. Although several techniques such as ARDL, quantile regression, multiple linear regression, etc., exist, this study selects polynomial regression with degree and time delay due because none of such other modeling techniques cover the objective of incorporating delay and degree. Without accurately incorporating degree and delay, leveraging other methods increases the chances that model would mistake the objective and weak at minimizing the impact of hidden variables; both of which existing research advices against [see[1]]. For comparison purposes, this study estimates the functional relationship both for time delay and without time delay. Given that ST is stock index, ad is adjust data, del is time delay, brd is best regression degree, i is the week number, NC is new covid cases without time delay, q denotes the set of integers from 0 to brd, a to e are constants and $NC^{'}$ is the new covid cases with time delay, the empirical relationships were estimated via polynomial environment as [see Tsai et al. [30]; Ostertagová [31]] in the series of equations from Eqs. (8) through to (15):

$$ST_{ad} = \sum_{q=0}^{brd_{ST,i}} a_q i^q \tag{8}$$

$$NC = \sum_{q=0}^{brd_{NC,i}} b_q i^q \tag{9}$$

$$ST_{ad} = \sum_{q=0}^{brd_{ST,i}} c_q (i + del_{ST})^q \tag{10}$$

$$NC = \sum_{q=0}^{brd_{NC,i}} d_q (i + del_{NC})^q \tag{11}$$

$$del_{NC,ST} = del_{NC} - del_{ST} \tag{12}$$

$$NC_i^{'} = NC_{i-del_{NC,ST}} \tag{13}$$

$$ST_{ad} = \sum_{q=0}^{brd_{ST,NC^{'}}} e_q NC_i^{'q} \tag{14}$$

[1] Riley, Patrick. "Three pitfalls to avoid in machine learning." Nature 572.7767 (2019): 27–29. https://doi.org/10.1038/d41586-019-02307-y.

for estimating regression with time delay; and

$$ST_{ad} = \sum_{q=0}^{brd_{ST,NC}} f_q NC_i^q \tag{15}$$

for without time delay.

It is worth adding that deriving an optimal time delay (i.e., *del*) is integral to discovering the polynomial regression with time delay that best fits the data. In deriving the degree, the mean square errors (MSEs) derived from plotting the true versus fitted polynomial regressions for the covid cases and stock index were saved. Two groups of saved MSEs were plotted against regression degrees and for each group the degree that corresponds to the minimum MSE was saved and regarded as best regression degree (i.e., *brd*).

Under the two best regression degrees for the covid cases and stock index (i.e., $brd_{NC,i}$ and $brd_{ST,i}$), MSEs derived from plotting the true versus fitted polynomial regressions with different integer time delay were saved. The two groups of saved MSEs were plotted against different integer time delay and for each group, the time delay that corresponds to the minimum MSE was saved and regarded as optimal time delay (i.e., *del*). By this approach, two optimal time delays (i.e., del_{ST} and del_{NC}) were identified. The optimal time delay for the final predictive model (i.e., $del_{NC,ST}$) was estimated as the difference between the delays from covid cases and stock index.

2.3. Test of model precision

To derive a better representation of the high frequency and volatile data, the two polynomial regressions (i.e., with and without time delay) were compared. The comparisons were made by applying the two polynomial models to three different sets of data derived from the sample. The data (i.e., stock index and new covid cases) used in the three tests with and without delay are displayed in Table 1.

The first set contains the data of the two groups (i.e., stock index and new covid cases) from week 4–60, divided into model and validation sets. For the model set, depending on the model selected (i.e., either with or without delay), the data for modeling spans week 4 -$(60 - del_{NC,ST})$ for polynomial regression without time delay and $(4 + del_{NC,ST})$-$(60 - del_{NC,ST})$ for with time delay. The empirical models derived from the model sets are used to predict two groups of validation data; i.e., from week $(61 - del_{NC,ST})$ to 60. The two predictions from with and without delay are compared with the true values and the differences are used to estimate the MSEs for the two polynomial regressions. The corresponding MSEs are compared and the polynomial regression that records the least MSE is regarded as a better model. Similar approaches are also applied in Tests 2 and 3. Upon establishing that polynomial regression with time delay fits the high frequency volatile data best, the resulting empirical model is implemented to project the stock index for up to seventeen weeks.

3. Results and discussions

3.1. Descriptive statistics and data processing

The empirical results on descriptive metrics of both new coronavirus cases and the stock index are presented in Table 2 and the rolling function graph of the processed data is depicted in Fig. 1. On new covid cases, the study reports that a total of 371 days (i.e., 7 times 53) were recorded and that a deduction of extra days due to overlap of calendar days produces 366 for 2020. Out of the 371 entries, 19 (i.e., ~5%) of the values are missing and such phenomenon are ascribed to unavailability of data due to public holidays. Similar observations are made for 2021 and 2022 with ~1.89% and ~3.7% of missing data, respectively. From Table 2, on day-on-day basis, the minimum number of new covid cases in China was reported to be zero (0) for 2020 and 2021 and one (1) for 2022. Thus, there were days for which no new cases of coronavirus infections were reported. The trend is significantly different from maximum parameters. For instance, basing on the weeks with data entries, the maximum parameter for 2020 is 15,152 (i.e., Monday, Wednesday, February 12th) which is considerably different from 182 (i.e., Monday, December 27th) for 2021 and 5646 (i.e., Thursday, April 28th) in January–July 2022. The variations between the maximum values year-by-year supports the indication that total number new covid cases depicts a form of unpredictability. The dynamism in the variations is also depicted by the standard deviation which stood at 1046.06 in 2020, 36.00 in 2021 and 985.84 in 2022. The positive skewness for all three years (i.e., 9.39 in 2020, 1.96 in 2021 and 2.00 in 2022) indicate that greater proportion of the daily new covid cases are skewed to the right of the mean.

Relative to the new covid cases and as reported in Table 2, there are considerable differences on the descriptive metrics of stock

Table 1
Testing and validation of polynomial regression with and without time delay.

Test	Model	Overall Data	Model data	Validation data
Test 1	With delay	4–60	$(4 + del_{NC,ST})$- $(60 - del_{NC,ST})$	$(61 - del_{NC,ST})$-60
	No delay	4–60	4-$(60 - del_{NC,ST})$	$(61 - del_{NC,ST})$-60
Test 2	With delay	44–90	$(44 + del_{NC,ST})$-$(90 - del_{NC,ST})$	$(91 - del_{NC,ST})$-90
	No delay	44–90	44-$(90 - del_{NC,ST})$	$(91 - del_{NC,ST})$-90
Test 3	With delay	74–120	$(74 + del_{NC,ST})$-$(120 - del_{NC,ST})$	$(121 - del_{NC,ST})$-120
	No delay	74–120	74-$(120 - del_{NC,ST})$	$(121 - del_{NC,ST})$-120

Table 2
Descriptive statistics on the variables.

Variable/metric	*Original data*			*Post-multiple imputation*		
	Year			Year		
	2020	2021	2022	2020	2021	2022
New covid cases						
Count	371	371	189			
Missing	19	7	7			
Minimum	0.00	0.00	1.00			
Maximum	15152.00	182.00	5646.00			
Average	240.88	23.57	632.02			
Standard deviation	1046.06	36.00	985.84			
Skewness	9.39	1.96	2.00			
Stock index						
Count	265	265	135	655 (131 weeks 5 days/week)		
Missing	22	22	17	0		
Minimum	2703.33	3385.54	2957.68	2703.33		
Maximum	3474.92	3661.09	3651.89	3731.69		
Average	3147.05	3558.16	3318.93	3344.93		
Standard deviation	234.58	73.96	171.58	250.25		
Skewness	−0.20	−0.05	−0.02	−0.77		
Adjusted data (post multiple imputation and aggregated)						
				2020–2022		
New covid cases						
Count				128		
Minimum				0		
Maximum				31,433		
Average				1627.52		
Standard deviation				4846.33		
Skewness				3.88		
Stock index						
Count				128		
Minimum				2795.69		
Maximum				3650.16		
Average				3347.06		
Standard deviation				243.60		
Skewness				−0.80		

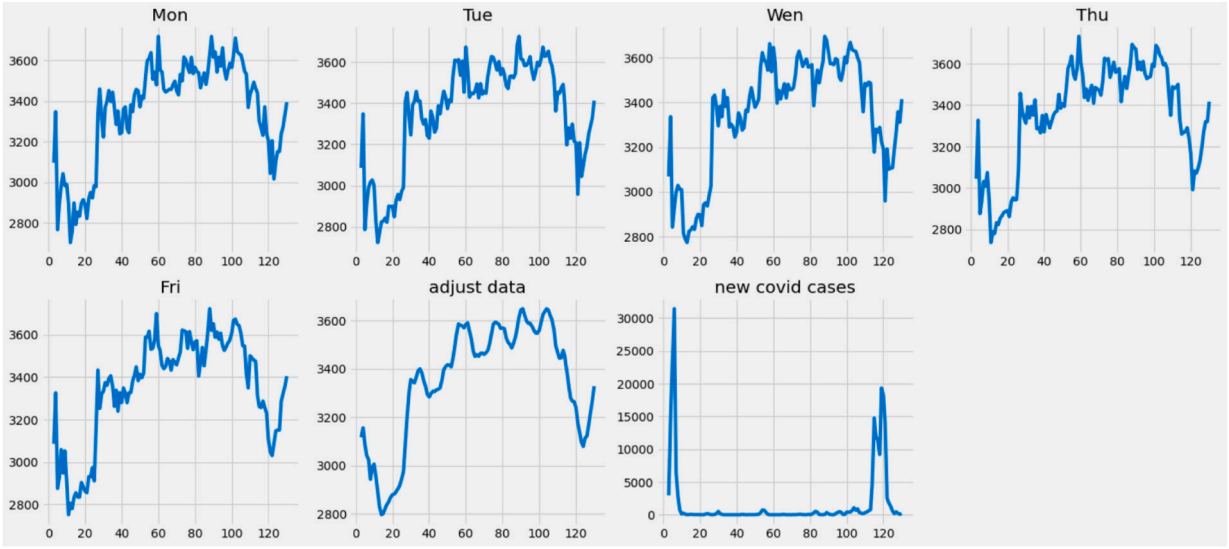


Fig. 1. Rolling function of the adjusted data.

index. For instance, the study reports a total of 265 days (i.e., 5 days times 53 weeks) of data for 2020 which is a little over 100 less than the 371 reported for new covid cases. The difference can be traced to variations among the total number of days used in the estimation. Each week, there are 7 days to report new covid cases but because the stock market operates on a 5-day work week (i.e., Monday to

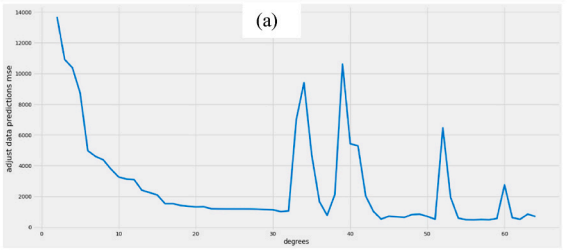
Friday), the stock index can only be recorded five times. One similarity among the stock index and new covid cases is missing values. Unlike covid that is not affected by holidays, the stock market does. Holidays such as Spring Festival, Tomb-sweeping Day, Labour Day, Dragon Boat Day, Chinese National Day, New Year and weekends did affect the number of data entries for the stock index. Another similarity is overlapping of days (i.e., days in a particular week that extends into other weeks or calendar years) as such were recorded both for new covid cases and stock index.

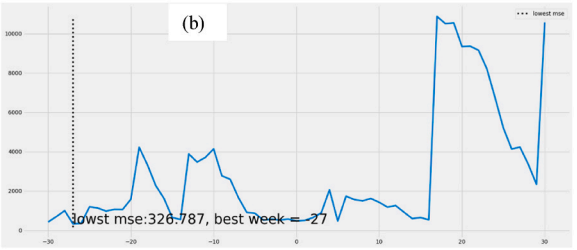
Setting aside the similarities, the study finds that a total of 265, 265 and 135 entries of data were recorded on stock index for 2020, 2021 and 2022, respectively. Out of the entries, 22 (i.e., ~8.3%), 22 (i.e., ~8.3%) and 17 (i.e., ~12.59%) of the values were found to be missing for 2020, 2021 and 2022, respectively. As stated earlier, the missing data are attributed to unavailability of data on weekends and public holidays. One dissimilarity between the descriptive statistics on new covid cases and stock index is non-zero minimum. Using the day-on-day Monday-Friday non-missing entries, the minimum stock index recorded in China was 2703.33 for 2020, 3385.54 for 2021 and 2957.68 for 2022, which presents a stacking difference for the 0-0-1 recorded for new covid cases. Here, the results show that there was no day for which the stock index was zero (0) in China. The maximum of the stock index stood at 3474.92 for 2020, 3731.69 for 2021 and 3651.89 for 2022 with corresponding averages of 3147.05, 3558.16 and 3318.93 for the three years. The standard deviation of the stock index is estimated to be 234.58 for 2020, 73.69 for 2021 and 171.58 for 2022. Though the averages and standard deviations of the stock index are meaningfully different from that of new covid cases, the variations within the means and dispersions year-by-year indicates the existence of unpredictability in the stock market. Compared with the positive skewness recorded for new covid cases, all the three years (i.e., 2020, 2021 and 2022) were negative for the stock index. This translates that for 2020, 2021 and 2022, greater proportion of the stock index are skewed to the left of the average.

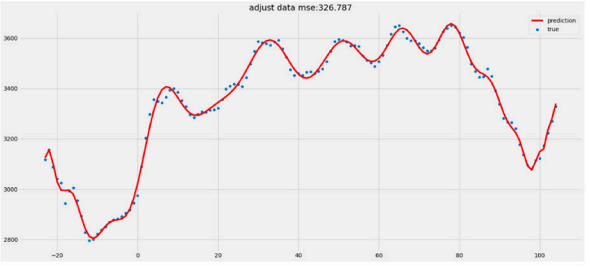
As depicted in Table 2, the stock index data had missing values whose fraction stayed below 15% (i.e., 8.30% for 2020 and 2021 and 12.59% for 2022). Multiple imputation technologies were utilized to fill in the missing stock index data and the processed data entailed 131 weeks with 655 day-on-day entries (i.e., Monday-Friday for 131 weeks) of non-missing values. Results from the post-imputation show a minimum of 2703.33, maximum of 3731.69, average of 3344.93, standard deviation of 250.25, and skewness of −0.77.

3.2. *Polynomial regression with and without delay*

Results of the polynomial regression with time delay for stock index and new covid cases are depicted in Figs. 2 and 3, respectively. As polynomial functions, the best degree and time delay were estimated to derive the optimal models for each of the two variables. On the stock index, the results show that the best degree of 56 (see Fig. 2(a)). The degree is traced to the least mean square error (MSE) and the results is clearly depicted in Fig. 2(a) with MSE of 472.64. The time delay for stock index is also marked by the lowest MSE on the week-MSE graph depicted in Fig. 2(b). The results show that delay of 27 weeks (i.e., −27) is the best time delay for the polynomial stock index function. As shown in Fig. 2(b), the dotted vertical line marks the lowest MSE of 326.79 at delay 27; thus, a time delay of 27 weeks best fits the nonlinear polynomial stock index model. Merging the best degree (of 56) and time delay (of −27) produces the







(c) (d)

Fig. 2. Polynomial function for time and stock index. (a) Showcasing that the best degree for stock index is 56. (b) The least mean square error that corresponds to the best degree in (a) to help detect the best delay. (c) Optimal polynomial regression function derived from merging the best degree in (a) and time delay in (b) (d) Predicted values (i.e., red line). of the optimal degree and time delay polynomial regression function versus the true data (i.e., blue dots).

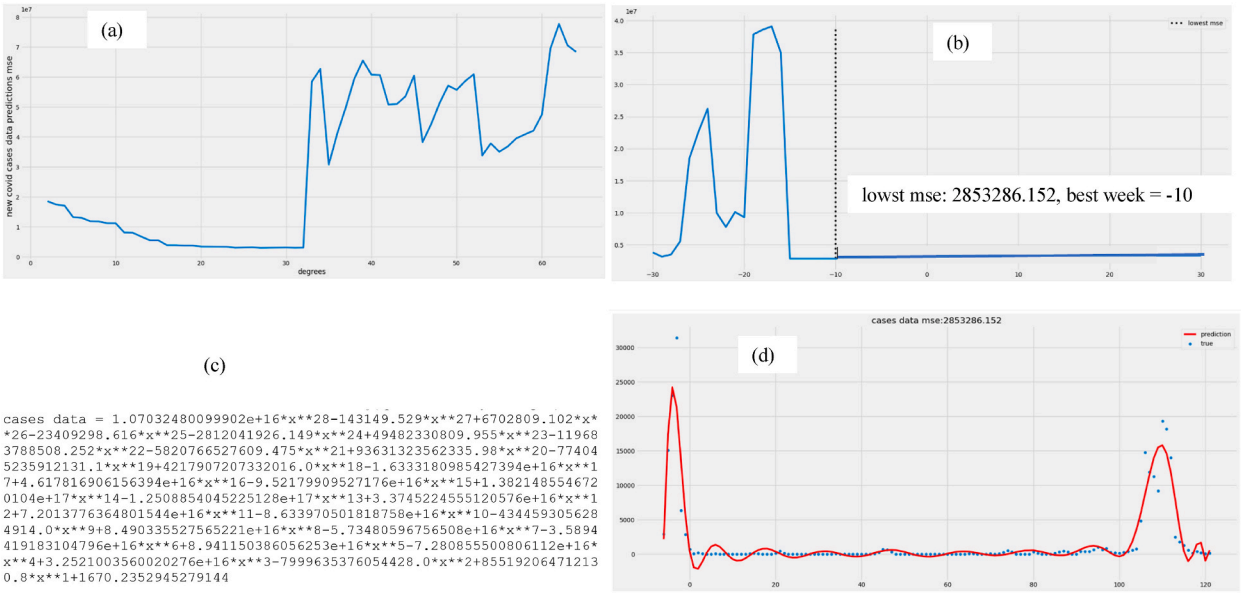


Fig. 3. Polynomial function for time and new covid cases. (a) The best degree of the new covid cases [is 27]. (b) The best time delay for the covid variable. (c) Optimal polynomial regression function from merging the best degree in (a) and time delay in (b). (d) Original versus predicted covid data.

optimal polynomial regression function displayed in Fig. 2(c). The empirical model in Fig. 2(c) is applied to the original data to test the representativeness and the results corresponding to the necessary computations are shown in Fig. 2(d). From 2(d), the predicted values (i.e., red line) of the optimal degree and time delay polynomial regression function closely mimic the true data (i.e., blue dots). The closeness of the predicted to true values translates into a low MSE of 326.79, which connotes that the polynomial regression function that encapsulates optimal degree and time delay provides a good fit to the high-frequency and volatile stock index in Chinese stock market.

Compared with the stock index, the results on new covid cases show lesser best degree and delay yet larger MSEs for both parameters. The best degree of the new covid cases is 27 (see Fig. 3(a)) compared with 56 for the stock index. Just as in the cases of the stock index, the best degree is traced to the least MSE and the results is clearly depicted in Fig. 3(a) with MSE of 300789.51. The time delay for the covid variable is also marked by the lowest MSE on the week-MSE graph depicted in Fig. 3(b). The results show that delay of 10 weeks (i.e., −10) is the best time delay for the polynomial covid function. As shown in Fig. 3(b), the dotted vertical line marks the lowest MSE of 2853286.15 at delay 10; thus, a time delay of −10 weeks best fits the nonlinear polynomial model for new coronavirus cases in China. Merging the best degree (of 27) and time delay (of −10) produces the optimal polynomial regression function displayed in Fig. 3(c). To be in line with the stock index cases, the empirical model in Fig. 3(c) is applied to the original covid data to test the representativeness and the empirical results corresponding to the required calculations are shown in Fig. 3(d). From 3(d), the predicted values (i.e., red line) of the optimal degree and time delay polynomial regression function closely mimic the original new covid data (i.e., blue dots). The nearness of the predicted new covid cases to true values, though with quite large MSE (of 2853286.15), translates that the polynomial regression function that captures optimal degree and time delay provides a good fit to the high-frequency and extremely volatile trends in coronavirus cases in China. The finding of high MSE is ascribed to the extreme volatility and unpredictability in new reported covid cases.

Despite the relatively high MSEs, both polynomial models are reflective of the data and are, thus, merged into one model with optimal time delay of 17 {i.e., stock index (−10) minus new covid cases (−27); −10 – (−27) = 17}. The accuracy of the optimal stock index *f* (optimal time delay of new covid) model is tested for prediction purposes.

3.3. Test of the accuracy of polynomial regression for forecasting

In testing the accuracy of the merged model, the 128 weeks (i.e., Week 4–131) of data were split into three sets to test the forecast precision of the polynomial regression function. Results on the functions with and without delay are displayed in Fig. 4 and Table 3.

From Table 3, the data used for test 1 span weeks 4–60. Depending on the model selected (i.e., either with or without delay), the data for modeling span 4–43 weeks for polynomial regression without time delay and 21–43 for with delay. The empirical models that correspond to the modelled data are depicted in Supporting Material to Fig. 4 [Optimal function for test 1] for which the MSE for the model with delay (20044.32) is observed to be lesser than that without delay (127382.20). Results from implementing the optimal function for test 1 to predicting weeks 44–60 also depict smaller MSE for the polynomial stock model with delay (i.e., 2202.99) relative to that without delay (i.e., 40717.6). The comparison of the predicted values for with and without delay are plotted against the true

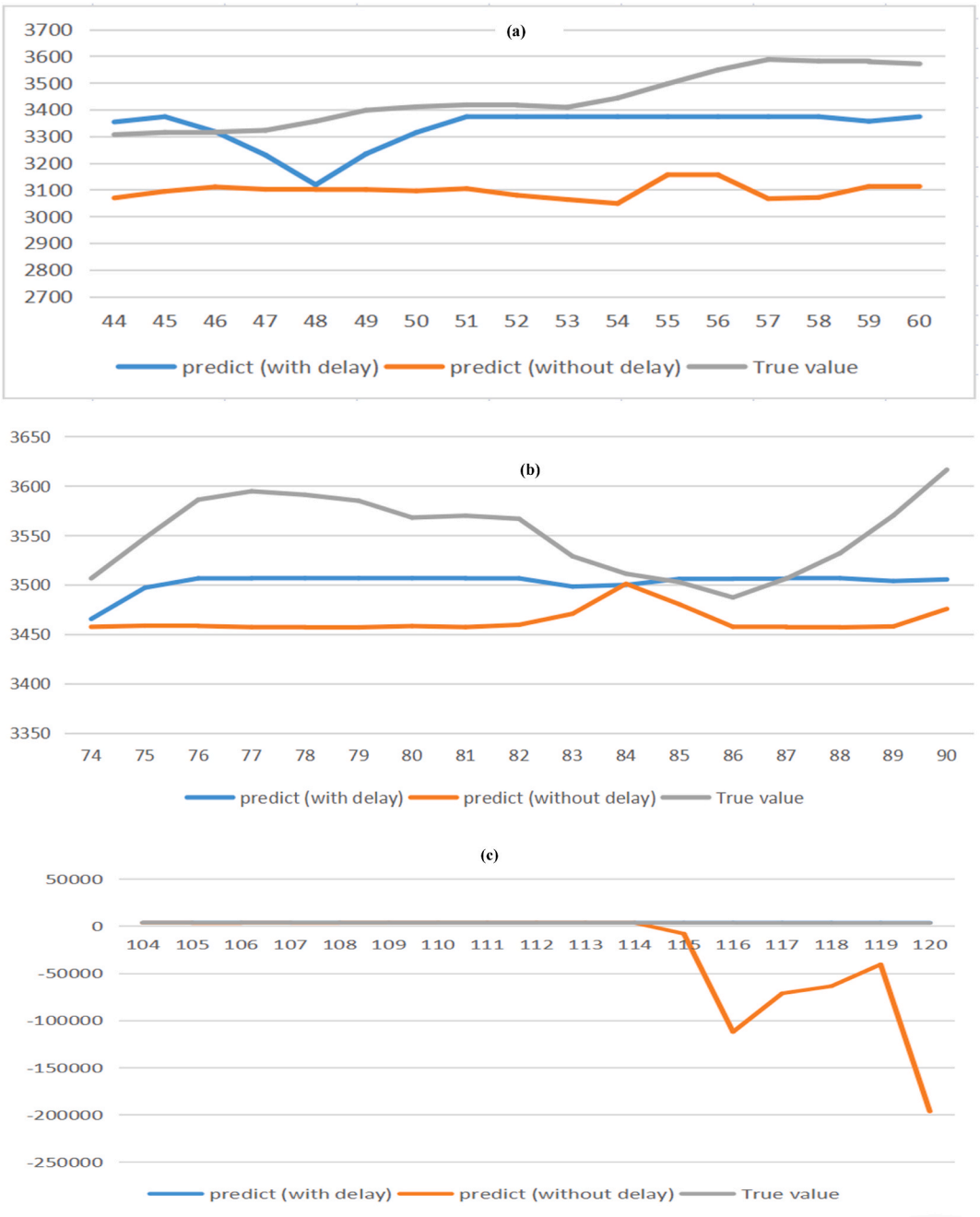


Fig. 4. Polynomial stock function with and without delay. (a) Test 1: line graph of predictions for test 1; (b) Test 2: line graph of predictions for test 2; (c) Test 3: line graph of predictions for test 3. Note: For (a), (b) and (c), all values on the y-axis represent stock index and that on y-axis for weeks.

Table 3
Testing, validation and comparison of polynomial regression for stock index with and without time delay.

Test	Model	Overall Data	Model data	Best degree	Model MSE	Validation data	Validation MSE	Conclusion
Test 1	With delay	4–60	21–43	9	20044.32	44–60	2202.992	better
	No delay	4–60	4–43	8	127382.2	44–60	40717.6	
Test 2	With delay	44–90	61–73	2	3631.329	74–90	2251.322	better
	No delay	44–90	44–73	2	9504.876	74–90	7059.19	
Test 3	With delay	74–120	91–103	3	27656.06	104–120	908.1108	better
	No delay	74–120	74–103	2	3.85E+09	104–120	1626.452	

stock index values and are displayed in Fig. 4(a). The evidence from the model MSEs, validation MSEs and the graph for test 1 all indicate that the polynomial stock function that accounts for time delay provides better representation to the true values of the stock index.

The prowess that the polynomial stock function with time delay has over that without delay is observed in other two tests. In test 2, that uses data spanning weeks 44–73 for modeling and 74–90 for validation, the empirical models that correspond to the modelled data are depicted in Supporting Material to Fig. 4 [Optimal function for test 2] for which the MSE for with delay (3631.329) is observed to be lesser than that without delay (9504.876). Results from implementing the optimal function for test 2 to predicting weeks 74–90 also depict smaller MSE for the polynomial stock model with delay (i.e., 2251.322) relative to that without delay (i.e., 7059.19). The comparison of the predicted values for with and without delay for test 2 are plotted in Fig. 4(b). Similarly, in test 3 {see Table 3 and Fig. 4(c)}, the MSEs both from modeling and validation sets are found to be significantly less than from without delay regression. Using MSEs from the model set, the polynomial regression with time delay presents improvement of between ~3-fold (i.e., test 2) and more that 1000-fold (i.e., test 3). On the validation dataset, the regression with delay equally presents improvement relative to without delay by between ~2-fold (i.e., test 3) and ~18-fold (i.e., test 1).

Thus, from the testing and validation, the empirical results show that the polynomial regression with delay better represents the stock index data and thus, is implemented for making future projections.

3.4. Implementation of polynomial regression with time delay for forecasting

To reiterate and buttress the finding that the polynomial regression with delay represents the stock index data better than without delay, the forecasting model was estimated both for with and without delay. The empirical models are presented in the appendix (see Appendix I). For MSEs of 15303.08 for with delay and 49865.845 without delay, the former presents improvements of ~3-fold. The representative with delay function is used to make projections of up to 17 weeks into the future of the stock index. The results are reported in Table 4.

From Tables 4 and it is estimated that the stock index would hit 3324.64 in week 148. All else equal, week 148 corresponds to late October to early November 2022. Thus, from the last week in October to the first week in November 2022, the stock index for that week in China is expected to reach 3324.64. Although the accuracy of the underlying polynomial regression with delay model is high, results from the projections of the first ten weeks (out the 17 predicted ones) are considerably different from expectations. The extreme positive and negation stock index projections for the first 10 weeks are ascribed to the optimal 17-week delay effect. Due to the 17-week optimal time delay, the prediction in weeks 132–141 leverage the highly unstable new covid cases recorded from 2022 week 11–2022 week 20. Such finding is an indication that the prowess of the time delay polynomial regression is heavily dependent on instability in covid-related time trends and that researchers should consider modeling to cover for the unsteadiness in coronavirus cases to achieve better results.

It is important to note that the projections that are presented in this study are not compared with other studies that have been conducted for the reasons that are listed below. To begin, neither the Shanghai Stock Exchange nor the National Health Commission of the People's Republic of China nor Tong Hua shun (i.e., the primary sources of data) publishes estimates of stock indexes that take into account time delay(s) and best degree(s). Secondly, there is research that suggests that the stock index on the Shanghai Stock Exchange makes use of stochastic processes, bottom-up processes, top-down processes, and other similar processes. These processes work with distinct conditions, assumptions, and rules in comparison to polynomial regressions for time delay. Thus, comparing stock index predictions that are modelled based on optimal time delay and best degree in a polynomial environment with others that do not account delays and degrees turns into committing mistakes that Riley [32] suggest avoiding.

Table 4
Stock index projections.

Week	Stock index (projection)
132	−27301241.74
133	1.72279E+11
134	1.84627E+11
135	1.26371E+11
136	18,961,503,617
137	−1.46448E+13
138	−7.64451E+12
139	3.16464E+11
140	−30255.93
141	−14324.36
142	2564.51
143	3374.64
144	3330.45
145	3404.7
146	3314.76
147	3447.89
148	3324.64

4. Limitations and future research

Despite the innovations in the study, there are few limitations that future studies may explore. First, the study uses the basic linear association based on correlation between the variables as premise in modeling the polynomial relationship of the stock index model with new covid cases, optimal delay and degree as predictors. Thus, though the stock index prediction model utilizes the independent variables, no causal relationship was conducted. Future research could explore the causal relationships, then utilize the causation model for making predictions about the stock index.

Second, the stock predictions as conducted in this paper produce projections on weekly basis. The study fails to partition the weekly forecasts into day-on-day values which makes it difficult to figure out the indexes for each day on the week. Future researchers leverage the limitation to propose strategies and decomposition models to determine the stock index per workday per week.

5. Conclusion

Against the backdrop that under contemporary market conditions in China, the stock index in the country has been volatile and highly reflective of trends in the coronavirus pandemic, but rare evidence of scientific research to fill in such gap, and that covid-related news in one time period impacts the stock market in another period, this study contributes to filling such gaps.

Noting that under contemporary market conditions in China, the stock index has been volatile and highly reflect trends in the coronavirus pandemic, t*he motivation of this study is to contribute to rare* scientific research conducted to model the possible nonlinear relations between the two indicators. Another motivation is to help fill the gap that existing research has failed that time delay can be an equally good predictor of the stock index.

In line with the possible relations, the novelty of this study is that it proposes, validates and implements polynomial regression with time delay to model nonlinear relationship between the stock index and covid. This research proposes non-linear modeling of observed trends under polynomial regression environment with time delay for the stock index and new covid cases in China. The polynomial model with time delay is estimated separately for the stock index and new recorded covid cases, and jointly for the two variables. Added, the optimal time delays observed from the polynomial regression for the two variables are used as input data to model, propose, test, validate and implement a high-accuracy polynomial model for forecasting the stock index up to 17 weeks ahead. The study utilizes high-frequency data from January 2020 to the first week of July 2022 to model the nonlinear relationship between the stock index, new covid cases and time delay under polynomial regression environment. The empirical results show that time delay and new covid cases, when modelled in a polynomial environment with optimal degree and delay, do present better representation of the nonlinear relationship such predictors have with stock index for China. Relative to results from the polynomial regression without delay, the empirical evidence from the model with delay show that an optimal time delay of 17 weeks makes it possible to predict the stock index at high accuracy and record improvements of 16-fold or higher. The representative delay model is used to project for up to 17 weeks for future trends in the stock index. The are two advantages of the proposed new model. First, proposed model can easily be replicated to other countries. The characteristics and procedures in the selection of optimal time delay and best degree are closely tied to nature of the data. The data-driven trait of the proposed model presents an advantage in terms of replicability to other countries. Second, the proposed model also presents dynamism in nonlinear modeling. Evidence from the three out-of-sample tests conducted in the study show varied optimal degrees as well as time delays. The variability in optimal degrees and time delays allow for dynamism in utilizing the model for varied purposes. The implication of the findings herein is that researchers and decision-makers should consider modeling to cover for the unsteadiness in coronavirus cases to harness the prowess of utilizing time delay polynomial regression for critical decisions.

This research paper thoroughly examines a notable void in the current body of knowledge, specifically delving into the correlation between the stock index volatility in China and the patterns observed during the coronavirus pandemic. With extensive research in mind, this study aims to delve into the intricate connections between the stock index and the effects of the coronavirus pandemic. Additionally, it seeks to address the gap in current research by showcasing the effectiveness of time delay as a predictor of the stock index. This study is significant due to its direct connection to the current market conditions in China. The stock index in this region has shown considerable fluctuations, which closely align with the patterns observed during the coronavirus pandemic. This research stands out due to its innovative approach in utilizing polynomial regression with time delay to effectively model the complex correlation between the stock index and COVID-19 cases. This approach brings a fresh perspective to the intricate dynamics between these two indicators.

The contributions to the literature are highly significant. The study presents a new approach and successfully demonstrates its validity and implementation. With a deep understanding of the subject matter, the author introduces polynomial regression with time delay as a modeling technique. This innovative approach sheds new light on the complex relationship between the stock index and pandemic-related factors. Based on extensive research and analysis of high-frequency data from January 2020 to July 2022, the results clearly demonstrate the enhanced accuracy achieved by implementing the suggested model. Notably, the utilization of an optimal time delay of 17 weeks further enhances the model's performance. With this discovery, one can confidently predict stock index trends up to 17 weeks in advance. This valuable tool equips decision-makers with the necessary information to navigate market volatility and uncertainties caused by the pandemic.

One of the major strengths of the proposed model is its ability to be replicated in different countries. With its data-driven approach, the model can easily adapt to various datasets, making it applicable in markets beyond China. In addition, the model's versatility is enhanced by the variability in optimal degrees and time delays, allowing it to be used for different purposes and in various contexts. Nevertheless, as with any research study, it is important to take into account certain limitations. The efficacy of the model relies

heavily on the specific data characteristics and the distinct circumstances surrounding the pandemic and market conditions. In addition, the study's timeframe, which extends from January 2020 to July 2022, might not encompass long-term trends or shifts in market dynamics. To further enhance the study, it would be valuable to extend the research period and perform cross-validation using datasets from various countries. This would allow for a more comprehensive evaluation of the model's applicability.

Overall, this research provides a significant contribution to the scholarly and practical comprehension of the connection between the stock index and the coronavirus pandemic in China. With the proposal and validation of a new polynomial regression model with time delay, this study addresses important gaps in scientific research. Additionally, it provides a reliable and adaptable tool for predicting stock index trends in the midst of uncertainties caused by the pandemic.

Data availability statement

The data used in this study have been attached as supplementary materials.

CRediT authorship contribution statement

Dong Bowen: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e28850.

References

[1] Chamzallari M., Chantziaras A., Grose C. The impact of COVID-19 on firm stock price volatility. In: Sklias, P., Polychronidou, P., Karasavvoglou, A., Pistikou, V., Apostolopoulos, N. (eds) Business Development and Economic Governance in Southeastern Europe. Springer Proceedings in Business and Economics. Springer, Cham.
[2] Y. Sun, M. Wu, X. Zeng, Z. Peng, The impact of COVID-19 on the Chinese stock market: sentimental or substantial? Finance Res. Lett. 38 (2021) 101838.
[3] X. Chen, Z. Wang, X. Li, Z. Liu, K. Li, The impact of covid-19 on the securities market: evidence from Chinese stock and bond markets, Procedia Comput. Sci. 187 (2021) 294–299.
[4] A.A.M. Gamal, A.A. Al-Qadasi, M.A.M. Noor, N. Rambeli, K. Viswanathan, The impact of COVID-19 on the Malaysian stock market: evidence from an autoregressive distributed lag bound testing approach, J. Asian Finance, Econ. Business 8 (7) (2021) 1–9.
[5] O. Guedhami, A. Knill, W.L. Megginson, L.W. Senbet, The dark side of globalization: evidence from the impact of COVID-19 on multinational companies, J. Int. Bus. Stud. 53 (2022) 1603–1640.
[6] N. Li, Y. Zhu, The impact of COVID-19 on stock market in China, Front. Econ. China 16 (4) (2021) 42–58.
[7] A. Ullah, X. Zhao, A. Amin, A.A. Syed, A. Riaz, Impact of COVID-19 and economic policy uncertainty on China's stock market returns: evidence from quantile-on-quantile and causality-in-quantiles approaches, Environ. Sci. Pollut. Control Ser. (2022).
[8] T. Kumeka, P. Ajayi, O. Adeniyi, Is stock market in Sub-Saharan Africa resilient to health shocks? J. Financ. Econ. Pol. 14 (4) (2022) 562–598.
[9] K. Young, S.-A. Yeoh, M. Putman, S. Sattui, R. Conway, E. Graef, A. Kilian, M. Konig, J. Sparks, M. Ugarte-Gil, L. Upton, F. Berenbaum, S. Bhana, W. Costello, J. Hausmann, P. Machado, P. Robinson, E. Sirotich, P. Sufka, J. Yazdany, J. Liew, R. Grainger, Z. Wallace, A. Jayatilleke, The Global Rheumatology Alliance. The impact of COVID-19 on rheumatology training—results from the COVID-19 Global Rheumatology Alliance trainee survey, Rheumatol. Adv. Pract. 6 (1) (2022) rkac001.
[10] M. Maguire, R. Bree, S.H. Moore, Munro, The impact of COVID-19 on Irish higher education: special issue Part 1: all Ireland, J. High Educ. 12 (3) (2020).
[11] T.T. Nguyen, The impact of COVID 19 on the banking and financial market in Vietnam: difficulties and proposed solutions, Article No.: 2, in: AADNIC-ABMECR '20: Proceedings of the 2nd Africa-Asia Dialogue Network (AADN) International Conference on Advances in Business Management and Electronic Commerce Research, 2020, pp. 1–4.
[12] M.K. Shoji, M.J. Venincasa, J. Sridhar, The impact of the COVID-19 pandemic on ophthalmology resident perceptions of clinical experience, surgical training, and personal life, J. Acad. Ophthalmol. 13 (2) (2021) e288–e297.
[13] G. Espinosa-Paredes, E. Rodriguez, J. Alvarez-Ramirez, A singular value decomposition entropy approach to assess the impact of Covid-19 on the informational efficiency of the WTI crude oil market, Chaos, Solit. Fractals 160 (3) (2022) 112238.
[14] C. Martz, R. Powell, B. Wee, The impact of COVID-19 lockdowns on youth relationships with nature: a socio-spatial perspective, Child. Youth Environ. 32 (1) (2022).
[15] T.H.L. Tu, The impact of COVID-19 on individual industry sectors: evidence from Vietnam stock exchange, J. Asian Finance, Econ. Business 8 (7) (2021) 91–101.
[16] F.J. Nusantara, The impact of COVID-19 pandemic information on sectoral stock performance during lockdown and new-normal: an evidence from Indonesia stock exchange, J. Int. Conf. Proc. 4 (1) (2021).
[17] Z. Luo, An analysis of COVID-19's impact on Japanese stock market returns using daily growth in cases and death (2), in: Proceedings of the 2022 International Conference on Business and Policy Studies, 2022, pp. 165–169.
[18] Q. Wang, L. Liu, J.L. Zhao, et al., Pandemic or panic? A firm-level study on the psychological and industrial impacts of COVID-19 on the Chinese stock market, Financ. Innov. 8 (36) (2022).
[19] M.J. Coffey, S. Kerns, S. Sanghani, L. Wachtel, The impact of COVID-19 on brain stimulation therapy, Psychiatr. Clin. 45 (1) (2022) 123–131.
[20] A. Arafa, N. Alber, The impact of coronavirus pandemic on stock market return: the cases of the MENA region, Int. J. Econ. Finance 12 (12) (2020).
[21] A. Nepp, O. Okhrin, J. Egorova, Z. Dzhuraeva, A. Zykov, What threatens stock markets more - the coronavirus or the hype around it? Int. Rev. Econ. Finance 78 (2022) 519–539.

[22] S. Conway, A. Kirresh, A. Stevenson, M. Ahmad, The impact of COVID-19 on cardiology training, Br. J. Cardiol. 28 (1) (2021).
[23] S. Dhar, I. Bose, Emotions in Twitter communication and stock prices of firms: the impact of Covid-19 pandemic, Decision 47 (2020) 385–399.
[24] V.M. Matushok, S.A. Balashova, The estimation of the financial crisis impact on the Russian stock market, Radiobiologiia (2011) 39–49.
[25] Tong Huashun Finance,https://www.10jqka.com.cn/.
[26] Official website of the National Health Commission, PRC, http://www.nhc.gov.cn/.
[27] A.S. Rahman, K. Haddad, A. Rahman, Identification of outliers in flood frequency analysis: comparison of original and multiple Grubbs-Beck test, Int. J. Environ. Ecol. Eng. 8 (12) (2014) 840–848.
[28] P. Royston, I.R. White, Multiple imputation by chained equations: implementation in stata, J. Stat. Software 45 (4) (2011) 1–20.
[29] E.H. Theophilus, C.R.E. Coggins, P. Chen, E. Schmidt, M.F. Borgerding, Magnitudes of biomarker reductions in response to controlled reductions in cigarettes smoked per day: a one-week clinical confinement study, Regul. Toxicol. Pharmacol.: Regul. Toxicol. Pharmacol. 71 (2) (2015) 225–234.
[30] C.Y. Tsai, J. Kim, F. Jin, M. Jun, M. Cheong, F.J. Yammarino, Polynomial regression analysis and response surface methodology in leadership research, Leader. Q. 33 (1) (2022) 101592.
[31] E. Ostertagová, Modelling using polynomial regression, Procedia Eng. 48 (2012) 500–506.
[32] Pa Riley, Three pitfalls to avoid in machine learning, Nature 572 (7767) (2019) 27–29.