Journal of Big Data

METHODOLOGY

Open Access

CrossMark

A method of trend forecasting for financial and geopolitical data: inferring the effects of unknown exogenous variables

Lucas Cassiel Jacaruso[1,2*]

*Correspondence:
Jacaruso33@gmail.com;
Jacaruso@usc.edu
[2] Universität für Musik
und darstellende Kunst
Graz, Mondscheingasse 7,
8010 Graz, Steiermark, Austria
Full list of author information
is available at the end of the
article

Abstract

This paper intends to contribute to the field of trend forecasting by proposing a new forecasting approach for stock market prices and geopolitical time series data of economic, financial and geopolitical importance. Designing models which account for every possible exogenous variable of relevance to a time series in question can often be an onerous and impractical task. Instead, this paper explores a new method which uses periods of decreased significance in the variable of foremost importance as a window of opportunity to observe the possible effects other variables may be having in a general way for the purpose of trend forecasting. When the latter variables are too unquantifiable to be accounted for in a model, having the ability to nonetheless discern their overall influence can be useful for anticipating trend changes. The proposed method was used in conjunction with the existing method of exponential smoothing to generate forecasts. It was also applied alone and contrasted with the results of exponential smoothing when used separately. This paper specifically addresses the ability of the newly proposed method to forecast the upwards/downwards extrapolation of the weekly trend for 9 weeks on stock closing prices for five companies of interest (Apple Inc, Amazon.com Inc, General Electric Company, Intel Corporation, and Alcoa Corporation). It was also applied to forecasting the annual trend for 9 years of Afghan asylum seeker data. These differing areas were chosen in order to demonstrate applications in finance as well as international relations. The empirical results and 95% confidence intervals indicate a clear advantage when the newly proposed method is used both in conjunction with exponential smoothing and on its own.

Keywords: Exogenous variables, Endogenous variables, Trend extrapolation forecasting, Time series, Stock closing prices, Financial securities prices, Trading volume, Asylum seekers, Refugees

Introduction

Improving the accuracy of forecasting methods is extremely valuable in a plethora of areas including (but not limited to) financial trading, as well as anticipating geopolitical patterns which are relevant to national policies and economies. This study proposes and examines the results of a new trend forecasting method for specific financial security price and asylum seeker datasets. By now, there have been many studies on ways to integrate exogenous variables into existing forecast models in areas such as electricity prices and migrant flows [1, 2]. Others have shown the importance of forecasting migrant phenomenon, and

the shortcomings of existing methods for forecasting such trends [3]. Additionally, some have investigated the use of internet search interest for forecasting purposes. As demonstrated by Yao et al. [4], there is a negative correlation between Google search interest and crude oil prices, which even still was shown not to be useful for forecasting purposes. Another recent study came to a similar conclusion; as demonstrated by Kim et al. [5], Google search interest is not useful for predicting future abnormal returns on the Norwegian stock market. Other studies have examined the effects which exogenous variables such as news frequency have on the prices of financial securities. Lillo et al. [6] demonstrate a lag time between news sentiment and the trading activity of companies which may be useful for anticipating market movements. In their recent paper, Sun et al. [7] introduce an improved vector auto regression model and new applications for neural networks for stock forecasting. Lei [8] shows the applicability of a RS–WNN model, a type of wavelet neural network, for predicting stock trends. Finally, Ren et al. [9] show the promising results of integrating exogenous sentiment analysis into machine learning models. Virtually all studies that can be found which address similar forecasting topics, recent or not, have either attempted to find correlations between exogenous variables and the time series of interest or developed new applications of regression models/neural networks. However, there is virtually no literature which proposes actually making use of periods when the time series is under diminished influence of major known exogenous variables altogether to gain insight which can be used to make forecasts.

For the proposed method, a distinction must be made between what will be called "primary factors" and "background factors" in this paper. These terms will be used going forward. Both primary and background factors are exogenous. Primary factors can generally be described as those which influence trends in the time series values of interest on a nearly immediate timeframe. Background factors affect trends in the time series less immediately than primary factors, yet can increase the likelihood of trend changes in a way that may not be readily visible in a subjective or objective analysis. The influence of background factors can be temporarily dominated or canceled out by the effects of primary factors but may have a significant chance of ultimately influencing the trend.

The method itself identifies periods of decreased significance in values correlated with the main primary factor and observes trends that nonetheless occur in the time series in question during those said periods. Presumably, this can further isolate the effects attributable to the said background factors.

In the instance of securities trading, news is often the primary exogenous factor which causes trading volume (an endogenous variable representing the total number of buyers/sellers) to spike dramatically and forthwith influence price direction during panic buying or selling [10]. A spike in volume is likely correlated with a significant news event such as an optimistic announcement about company earnings. When trading volume is low, one can often assume a greater likelihood that price movement is not due to reaction to the latest relevant or anticipated news, but is instead due to any number of important background factors which are nonetheless difficult to account for (e.g. insider information, policy changes before they make headlines, developments in the greater business sector a company operates in, etc.).

The same concept can also apply to forecasting trend changes in migrant phenomenon. The primary factor is assumed to be violence or instability which directly and physically displaces populations. Background variables which are much more challenging to accurately quantify may include population sentiment/morale, local knowledge of developing situations, or economic changes before they are officially recorded. A rise in emigration even while there was diminished activity in the reports of direct conflict may expose the effects of those background factors and signal a proclivity for emigration numbers to increase on a longer timeframe.

There are certain limitations to this method. First, it only focuses on forecasting the upwards/downwards extrapolation of the trend, rather than a specific prediction interval. Also, it does not reveal what the background factors are, but may only shed light on the effect they may be having on the time series of interest.

As was shown by Makridakis et al. [11] the most successful forecasting results from the 2018 M4 Forecasting Competition were combinations of forecasting methods. In that light, this paper examines the accuracy of the proposed method used alongside simple (or single) exponential smoothing [12]. The hypothesized result is as follows: using the newly proposed method, both in conjunction with exponential smoothing and on its own, will yield a higher accuracy rate than exponential smoothing used by itself when applied on the same sample size to generate a forecast.

Methods and results

The proposed method of understanding the effects of unknown exogenous variables

The relationship between primary and background exogenous variables and their corresponding endogenous variables can broadly be mapped out as follows:
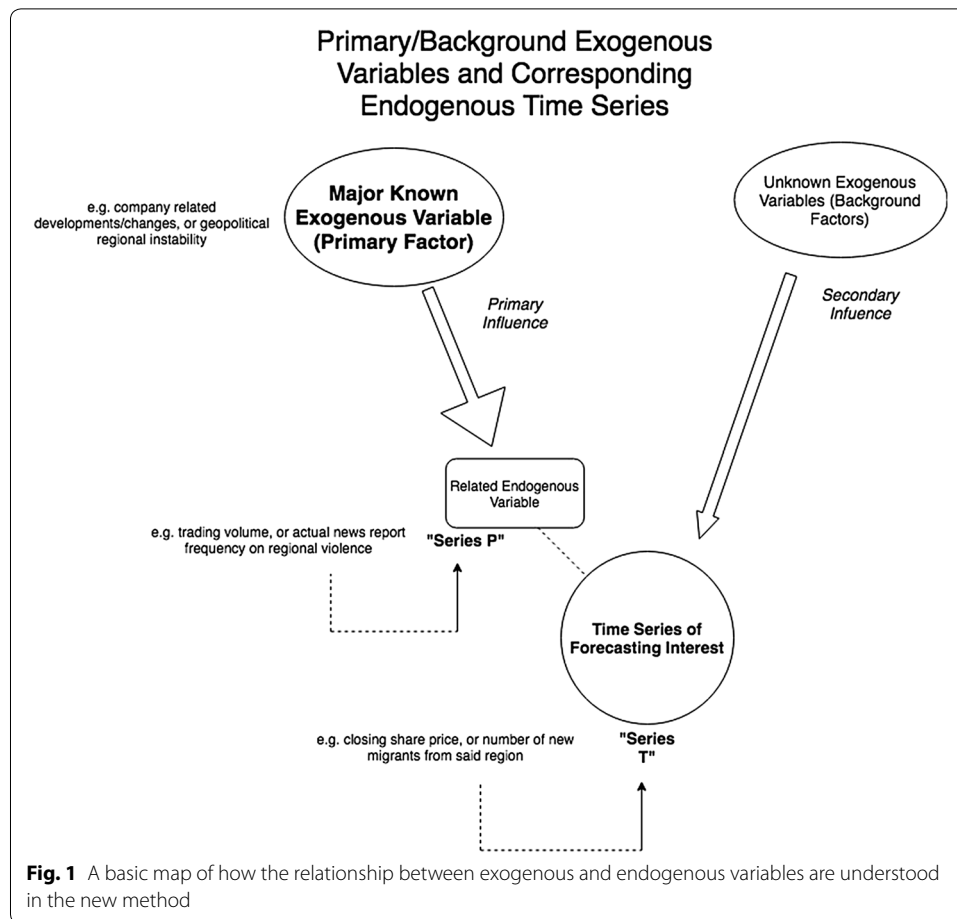
Series P is the one correlated with the endogenous variable directly representative of the primary factor. For this study, series P was the trading volume for the stocks of interest or the frequency of conflict reports in Afghanistan. Series T is the time series to forecast, i.e. the actual stock price or the number of asylum seekers of Afghan origin worldwide. The temporal resolution and timeframe in both series P and T were strictly identical in every forecast, an essential prerequisite for the application new method. Figure 1 outlines the premise of the newly proposed method: when the primary factor (represented by series P) is less significant, the positive or negative change in the time series of interest (series T) may be more attributable to background factors. Subsequently, the forecast is based on the effect background factors are presumably having. Each forecast is generated by identifying the minimum value in series P (as a percentage the said minimum value represents of the mean of series P) and observing the change that occurs in series T at that same point in time (as a percent change from the previous value in series T).

In practice, the newly proposed method is defined as follows:

Let PX = min [P1...PN − 1, PN]

$$A = \left(\frac{(PX)}{\mu[P1 \ldots PN-1, PN]} + \left| \frac{(TX)-(T(X-1))}{(T(X-1))} \right| \right) \times 100$$

where A is the significance of the forecast indicator.

Fig. 1 A basic map of how the relationship between exogenous and endogenous variables are understood in the new method

A greater change in series T while the value of series P is at a typical minimum will result in a higher *A* value. The change in series T is expressed as absolute value, since any change is weighted positively in the final resulting *A* value. However, whether the change in series T was positive or negative must be separately considered in order to determine if the forecast is for a rising or declining trend:

When $(((TX)-(T(X-1))) \div (T(X-1))) \times 100 < 0$, the forecast is for a declining trend.

When $(((TX)-(T(X-1))) \div (T(X-1))) \times 100 \geq 0$, the forecast is for a rising trend.

Simple exponential smoothing

Simple exponential smoothing will be used in the testing process outlined later, and is defined as follows:

$$St = \alpha yt - 1 + (1 - \alpha)St - 1, \quad t > 0$$

where α is the smoothing constant (or damping factor), t is the time period, and $0 > \alpha < 1$.

For this study, the damping factor α was always 0.9 whenever exponential smoothing was applied.

Data

The data used was 19 weeks of daily closing prices and trading volume on the 5 stocks of interest listed on the NYSE [13] or NASDAQ [14]. These stocks of interest were AA (Alcoa Corp.), AAPL, (Apple, Inc.), AMZN (Amazon.com, Inc.), GE (General Electric Company), and INTC (Intel Corp.). All market data was sourced from Commodity Systems, Inc. [15], and accessed via Yahoo Finance [16]. For the asylum seeker forecasts, monthly data was used for conflict reports and asylum seekers worldwide originating in Afghanistan from 1999 to 2015. Asylum seeker figures were obtained from the United Nations High Commissioner for Refugees (UNHCR) [17]. The Global Database of Events, Language and Tone (GDELT) event database [18] was used to obtain the internet-wide frequency of conflict reports in Afghanistan and queried with standard SQL via Google BigQuery [19]. Asylum seeker figures were separately reported by country of origin and country of residence in the UNHCR dataset and had to be combined to obtain monthly worldwide totals.

Timeframes for forecasting the upwards/downwards extrapolation of the trend

For each stock market forecast, series P always consisted of 5 data points representing the daily trading volume for each day over the course of 1 business week. Series "T" always consisted of the daily closing prices for the same days of that week.

For asylum seeker data, series P always consisted of the total monthly news reports of conflict in Afghanistan for 1 year, and series T represented monthly asylum seekers of Afghan origin worldwide for that same year.

For stocks, the forecasting horizon was for 1 week ahead (Friday through Friday). For asylum seeker data, the forecasting horizon was for 1 year ahead (December through December). More specifically, the forecast only anticipated if the following discrete Friday or December value would be greater than or less than the previous Friday or December value (the last value in the time series used to actually generate the forecast). It did not, for example, anticipate if the mean value of the next week or year would rise or decline. For stocks, the training period was from 4/2/2018 to 6/22/2018, and the testing period was from the week of 7/9/2018 to 9/10/2018. If a week contained a holiday on which the market was closed, that week was omitted from the training or testing period. Therefore, there were always 9 weeks of testing for stocks. For asylum seeker data, the training period ranged from January 1999 through December 2005, and the testing period ranged from 2006 to 2014. The training period was used to establish the mean and minimum accurate A value for each dataset, essential for the testing process detailed later.

All forecast training and testing was done with the constraint of using 1 week of stock data to anticipate the trend 1 week ahead, and 1 year of refugee data to anticipate the trend 1 year ahead.

Working examples

For additional clarity regarding the newly proposed method, two examples of the newly proposed method are provided below.

For the data in the Fig. 2, the trading volume (series P) is [29017700, 28408600, 22431600, 22889300, 25124300]. Series T represents the daily closing prices [168.804779,

Fig. 2 Apple stock daily closing prices and trading volume for the week of April 9th, 2018

171.981339, 171.177277, 172.864822, 173.4505]. *PX* is the minimum of series P, which is 22431600. At that same point in time (4/11/18), the value in series T (or *TX)* is 171.177277. Therefore, the newly proposed method as outlined previously is as follows.

$$ A = \left(\frac{(PX)}{\mu[P1\ldots PN-1, PN]} + \left| \frac{(TX)-(T(X-1))}{(T(X-1))} \right| \right) \times 100 $$

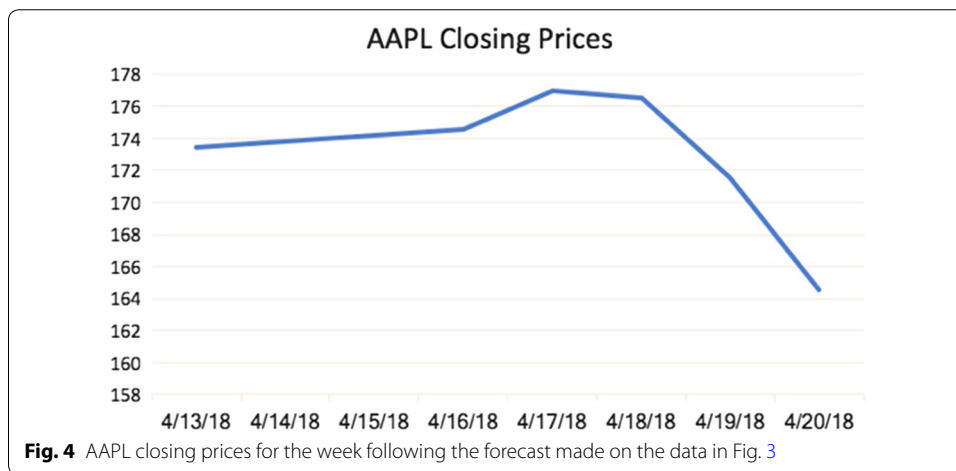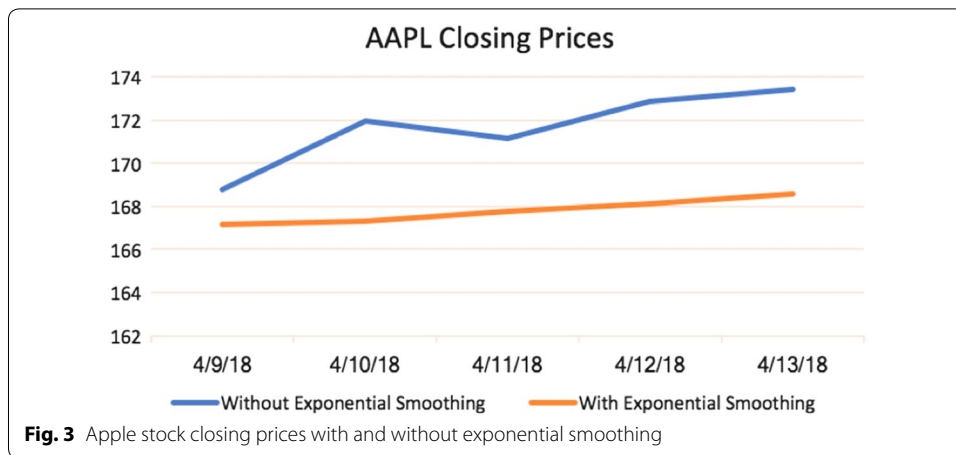$$ A = ((0.877115) + |((171.177277) - (171.981339)) \div ((171.981339))|) * 100 $$

$$ A = 88.179 $$

As $(((TX) - (T(X-1))) \div (T(X-1))) \times 100$ or $((171.177277) - (171.981339))/((171.981339)) * 100 < 0$, a forecast for a declining trend on a 1 week timeframe was produced. On the same week, applying exponential smoothing on the closing prices indicated the following rising trend (Fig. 3).

In this example, exponential smoothing indicated a rising trend whereas the new method indicated a declining trend on the price data for 1 week ahead.

For the subsequent week, AAPL stock prices closed as follows (Fig. 4).

In this case, the new method accurately forecast the declining trend.

Fig. 3 Apple stock closing prices with and without exponential smoothing

Fig. 4 AAPL closing prices for the week following the forecast made on the data in Fig. 3

The same forecasting method was applied to data on asylum seekers of Afghan origin the same way (Fig. 5).

In this example, series P is the monthly reports of conflict in Afghanistan for 2007 [2578, 2970, 4157, 5431, 7079, 7126, 5696, 4286, 5078, 4523, 4676, 3587]. Series T is the number of Asylum seekers worldwide of Afghan origin for that same year [875, 631, 785, 754, 780, 717, 641, 693, 834, 870, 876, 825].
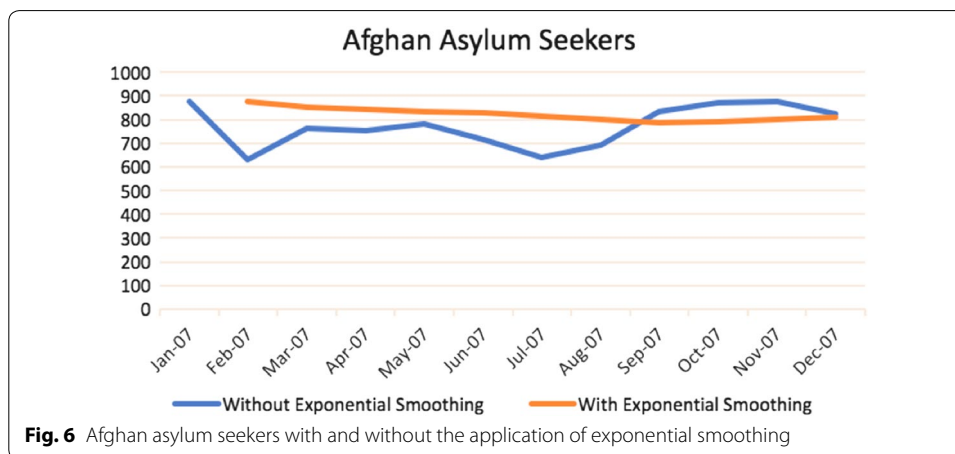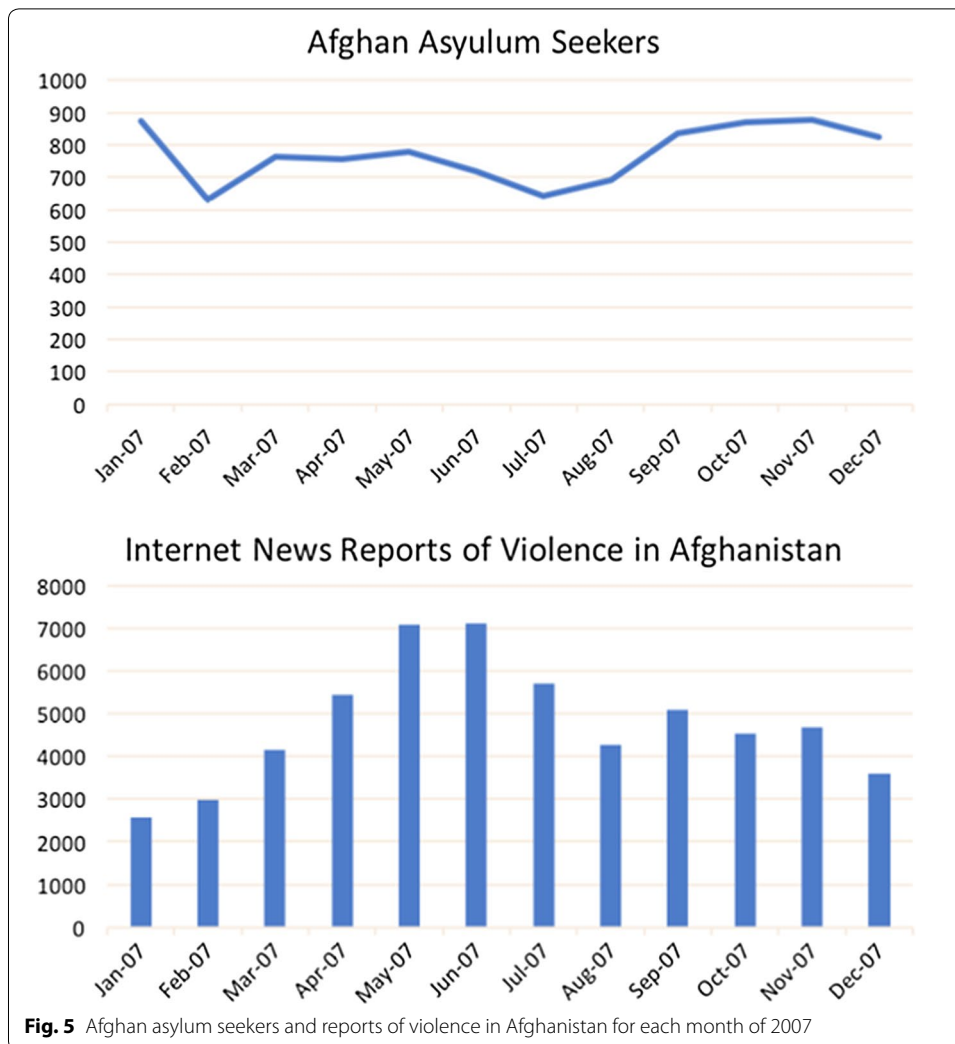
In this example, the lowest value in series P was in January of 2007, when the value in series T was 875. The previous value in series T for December of 2006 was 814, which must be known to find $(T(X-1)$. When the lowest point in series P corresponds with the first value in series T, the last value for the previous year (or week in the case of stocks) was always used as $(T(X-1)$.
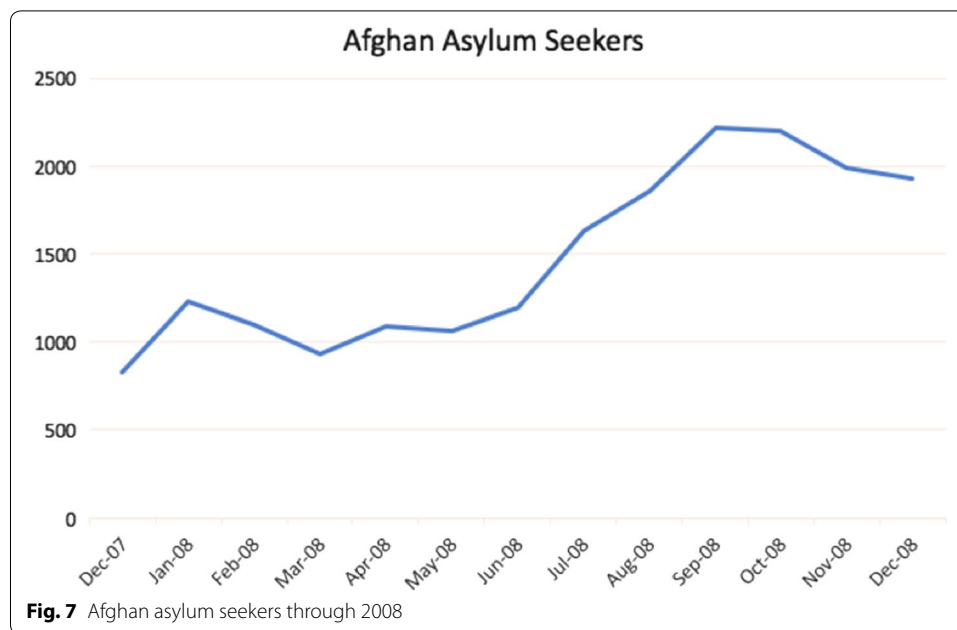
After applying the formula, $A = 61.5901.$ $(((TX) - (T(X-1))) \div (T(X-1))) \times 100 > 0$ in this instance, therefore a forecast for a rising trend on a 1 year timeframe was produced.

Meanwhile, exponential smoothing indicated the following declining trend (Fig. 6).

Actual asylum application for the subsequent year (2008) was as follows (Fig. 7).

In this case, the new method also accurately forecast the rising trend.

Fig. 5 Afghan asylum seekers and reports of violence in Afghanistan for each month of 2007

Fig. 6 Afghan asylum seekers with and without the application of exponential smoothing

Fig. 7 Afghan asylum seekers through 2008

The testing process

Accuracy was measured simply as the percentage of forecasts that were true on the given timeframe out of the total number of forecasts made for the testing period of each time series. All forecasts were classified as either true or false after examination of the subsequent week (for stock prices) or year (for asylum seeker figures). Afterwards, statistical confidence intervals were established for all testing on stock data. The purpose of the testing was to establish the accuracy of the various approaches for using the newly proposed method outlined below, both in conjunction and in contrast with exponential smoothing.

The forecast testing process for each set of data individually was to.

1. Use the training period to establish the mean and minimum A value for the newly proposed method which accurately forecast a rising or declining trend for series T.
2. Apply the newly proposed method on the testing period; whenever A was below the mean accurate value established in the above process, exponential smoothing was deferred to instead. Accuracy was measured for this joint approach. In the results section, this approach will be referred to as the "combined method".
3. Use the minimum accurate A value from the training period as the threshold (instead of the mean accurate A value). If A was below that threshold during the testing period, the forecast was negated. By this is meant that the initial forecast was anticipated to be false if A was below the aforementioned threshold, and the opposite trend was projected instead. Accuracy was measured for this approach. In the results section, this will be referred to as "low A threshold".

4. Establish the accuracy of the newly proposed method alone on the testing period without an *A* threshold. In the results section, this will be referred to as the "new method alone".
5. Establish the accuracy of exponential smoothing used alone on the testing period as a baseline.

Results

Presented below are the results of the tests as per the process previously described.

All results are presented in the order of most accurate to least accurate. The 95% confidence interval for the stock price forecasting results is also displayed. The same percentage values occur because all testing periods consisted of nine forecast tests, which were classified as either true or false.

Afghan asylum seeker trend forecasting

Below are the test results for Afghan asylum seeker figures.

Mean accurate *A* value from training period: 78.7525.

Method	Percent accuracy (%)
New method alone	77.78
Low A threshold	44.44
Exponential smoothing	44.44
Combined method	33.33

Apple stock

Below are the test results for Apple stock, listed as AAPL on NASDAQ.

Mean accurate *A* value: 77.2926.

Method	Percent accuracy (%)
Low A threshold	88.89
New method alone	77.78
Combined method	77.78
Exponential smoothing	33.33

Amazon stock

Below are the test results for Amazon stock, listed as AMZN on NASDAQ.

Mean accurate A value: 74.1332.

Method	Percent accuracy (%)
Low A threshold	88.89
New method alone	88.89
Combined method	88.89
Exponential smoothing	66.67

General Electric stock

Below are the test results for General Electric, listed as GE on the NYSE.

Mean accurate A value: 76.1532.

Method	Percent accuracy (%)
Combined method	88.89
Low A threshold	66.67
New method alone	66.67
Exponential smoothing	56.56

Alcoa Corp. stock

Below are the test results for Alcoa Corporation stock, listed as AA on NASDAQ.

Mean accurate A value: 68.4996.

Method	Percent accuracy (%)
New method alone	77.78
Low A threshold	55.56
Combined method	33.33
Exponential smoothing	22.22

Intel stock

Below are the test results for Intel stock, listed as INTC on NASDAQ.

Mean accurate A value: 79.0145.

Method	Percent accuracy (%)
New method alone	77.78
Low A threshold	44.44
Combined method	33.33
Exponential smoothing	33.33

Confidence intervals for all methods on stock market data

Below is the 95% CI for the accuracy of each method on the stocks of interest, based on the test results presented above.

95% CI		
Method	**Upper bound (%)**	**Lower bound (%)**
New method alone	87.53	68.03
Low A threshold	94.57	44.2
Combined method	100	28.7
Exponential smoothing	65.32	19.52

Discussion

For tests on stock market data, the 95% confidence interval suggests the accuracy of the newly proposed method is the most reliable for investing applications, and is entirely within the profitable range. However, results of the combined method and low A threshold approaches yield confidence intervals which are not reliable for investing applications, as the lower bound is below 50% in both cases. The confidence intervals for the accuracy of exponential smoothing show the lowest range in contrast with all other approaches tested within the parameters of this study, confirming the initial hypothesis.

Based on the test results on stock market data, the first question that comes to mind is why the results are more promising for Apple and Amazon stock. All three approaches which make use of the new method (new method alone, combined method, and low A threshold) perform better on AAPL and AMZN than they do on other stocks. The mean accuracy of these three approaches is 81.4833% for AAPL stock, and 88.89% for AMZN stock. For all other stocks, the mean accuracy of the same three approaches is 74.0767% or less. Regarding this point, the work of Yao et al. [4] and is invoked again, as well as that of Kim et al. [5]. Both of these studies suggest that Google search interest is not useful for forecasting trends in prices or returns for various tradable securities including crude oil and stocks. The test results of this paper suggest an altogether different kind of relationship with Google search interest, however. Results tentatively suggest a positive
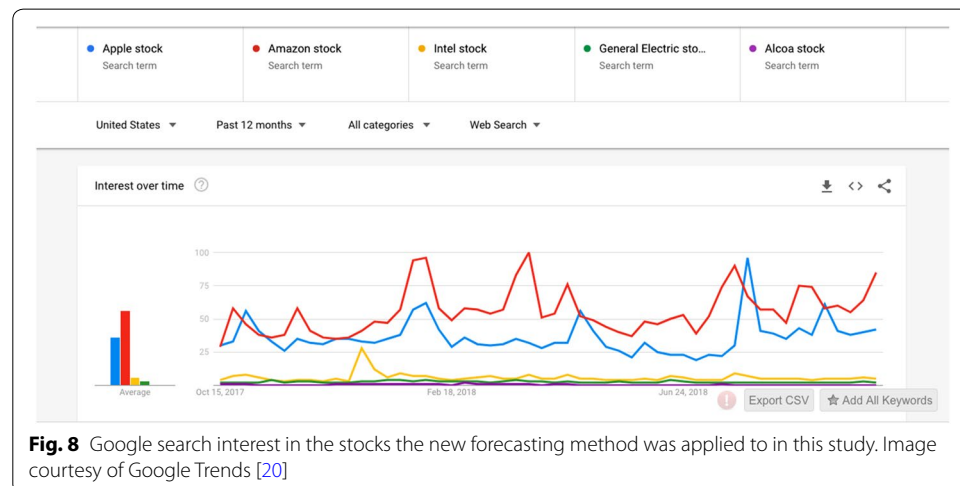

Fig. 8 Google search interest in the stocks the new forecasting method was applied to in this study. Image courtesy of Google Trends [20]

correlation between Google Index search interest and the reliability of the newly proposed method of trend forecasting. Google Index search interest on each of the 5 stocks analyzed in this paper is presented below.

As can readily be seen in Fig. 8, "Amazon stock" is ranked highest for interest as a search term on Google, followed by "Apple stock". The most logical reason for this potential correlation is that increased Google Index search interest in a stock indicates more trading interest from the general masses of less-informed traders. This can make a stock more susceptible to being bought and sold on behalf of a large number of non-professional traders likely to react to news, whereas more informed institutions and traders operate on sophisticated strategy, in-depth research, and even inside industry knowledge on the same stock. This may create a more pronounced distinction between reactionary news-based trading and strategic informed trading, which may accentuate the difference between the so-called primary and background factors which influence the price. Therefore, these stocks may better suit the new method presented in this paper. The other stock test results, however, did not correspond with the Google search interest in the same order. Additional investigation will be needed on this potential correlation in the future. For Afghan asylum seeker trend forecasts, the newly proposed method alone was more accurate than exponential smoothing alone by 33.34%.

Conclusion

In this paper, the results of a newly proposed trend forecasting method which identifies the effects of unknown exogenous variables are examined when used in conjunction, and in comparison, with exponential smoothing. The empirical results confirm the original hypothesis, which anticipated combining the newly proposed method with exponential smoothing would be more successful than using exponential smoothing alone to forecast the upwards/downwards extrapolation of the trend. Surprisingly, however, a more useful confidence interval is indicated when the new method is used alone without an A value reliability threshold. The confidence intervals suggest that the approach presented in this paper can be an additional tool for growing wealth more effectively, and anticipating geopolitical events more acutely. The possible link between Google search interest and the reliability of the newly proposed method will be further explored on a wider range of financial securities. Furthermore, investigating integration of the newly proposed method into machine learning models will also be a worthwhile future endeavor.

Abbreviations

Stock ticker symbols
AAPL: Apple, Inc; AMZN: Amazon.com, Inc; GE: General Electric Company; INTC: Intel Corporation; AA: Alcoa Corporation.

Databases
UNHCR: United Nations High Commissioner for Refugees; GDELT: The Global Database of Events, Language and Tone.

Statistic
CI: confidence interval.

Authors' contributions
LJ developed the new method presented in this paper, obtained and prepared the data, tested results, and authored the manuscript. The author read and approved the final manuscript.

Author details
[1] University of Southern California, Los Angeles, CA, USA. [2] Universität für Musik und darstellende Kunst Graz, Mondscheingasse 7, 8010 Graz, Steiermark, Austria.

Publisher's Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References
1. Gianfreda A, Grossi L. Forecasting Italian electricity zonal prices with exogenous variables. Energy Econ. 2012;34(6):2228–39. https://doi.org/10.1016/j.eneco.2012.06.024.
2. Fertig M, Kahanec M. Projections of potential flows to the enlarging EU from Ukraine, Croatia and other eastern neighbors. IZA J Migr. 2015. https://doi.org/10.1186/s40176-015-0029-8.
3. Wilson T. Can international migration forecasting be improved? The case of Australia. Migr Lett. 2017;14(2):285–99.
4. Yao T, Zhang Y-J. Forecasting crude oil prices with the Google Index. Energy Procedia. 2017;105:3772–6. https://doi.org/10.1016/j.egypro.2017.03.880.
5. Kim N, Lučivjanská K, Peter M. Google searches and stock market activity: evidence from Norway. Finance Res Lett. 2018. https://doi.org/10.1016/j.frl.2018.05.003.
6. Lillo F, et al. How news affects the trading behaviour of different categories of investors in a financial market. Quant Finance. 2014;15(2):213–29. https://doi.org/10.1080/14697688.2014.931593.
7. Sun H, Xu J. Improved approaches for financial market forecasting based on stationary time series analysis. In: 2018 IEEE 3rd international conference on Big Data analysis (ICBDA), 2018. https://doi.org/10.1109/ICBDA.2018.8367703.
8. Lei L. Wavelet neural network prediction method of stock price trend based on rough set attribute reduction. Appl Soft Comput. 2018;62:923–32. https://doi.org/10.1016/j.asoc.2017.09.029.
9. Wu R, Liu T. Forecasting stock market movement direction using sentiment analysis and support vector machine. IEEE Syst J. 2018. https://doi.org/10.1109/JSYST.2018.2794462.
10. Trading Volume: What It Reveals about the Market. Rediff. www.rediff.com/money/special/trading-volume-what-it-reveals-about-the-market/20090703.htm. 2009. Accessed 15 Sept 2018.
11. Makridakis S, et al. The M4 competition: results, findings, conclusion and way forward. Int J Forecast. 2018;34(4):802–8. https://doi.org/10.1016/j.ijforecast.2018.06.001.
12. 6.4.3.1. Single exponential smoothing. 1.3.3.30. Weibull Plot. www.itl.nist.gov/div898/handbook/pmc/section4/pmc431.htm. Accessed 17 Sept 2018.
13. The New York Stock Exchange. The New York Stock Exchange | NYSE. www.nyse.com/index. Accessed 19 Sept 2018.
14. NASDAQ's Homepage for Retail Investors. NASDAQ.com. www.nasdaq.com/. Accessed 20 Sept 2018.
15. Commodity Systems Inc. ReDim Statement. www.csidata.com/. Accessed 20 Sept 2018.
16. Yahoo Finance—Business Finance, Stock Market, Quotes, News. Yahoo! Finance, Yahoo!. finance.yahoo.com/. Accessed 21 Sept 2018.
17. United Nations. The UN Refugee Agency. UNHCR. www.unhcr.org/. Accessed 22 Sept 2018.
18. The GDELT Project. GDELT. www.gdeltproject.org/. Accessed 22 Sept 2018.
19. BigQuery—Analytics Data Warehouse | BigQuery | Google Cloud." Google, Google. cloud.google.com/bigquery/. Accessed 23 Sept 2018.
20. Google Trends. Google. trends.google.com/. Accessed 17 Oct 2018.