Future predictions using Google Trends Data

Tariq Kalim, Anas El Kassimi, Khalil Merzouk
CS-401, Applied Data Analysis, EPFL, Switzerland

***Abstract*—Google Trends helps users visualise trends in people's search behaviour within Google Search, News, Images, Shopping and Youtube. It can improve forecasts of the current level of activity for different economic time series. Predicting the present is useful as it may help identify turning points in these economic time series before publication. This study investigates Google trends present predictive power veracity claimed by the paper *Predicting the Present with Google Trends*, its future predictive power on economic time series and on the VIX, also called the fear index, a relevant financial index.**

I. Introduction

The famous economist Yogi Berra once said : "It's tough to make predictions, especially about the future." This research paper aims to validate the present predictive power of google trends and demonstrate its future predicting power. First, we develop rigorously a baseline model on "Motor Vehicles and Parts Dealers" and add trends data to confirm **present** "predictiveness". Then using the same time series we investigate this time **future** "predictiveness" on "Motor Vehicles and Parts Dealers". Finally we try to predict the VIX index with and without google trends to show how impactful it can be.

II. Related work

In 2012, Hyunyoung Choi and Hal Varian published the paper [1]. They found that simple seasonal AR models that include relevant Google Trends variables outperform models that exclude these predictors. But the paper did not give any details on model selection that could let us think that the model chosen enhanced the results and falsely demonstrating that trends data had added value.
Since no particular attention was given to time series issues, as an extension, we explore its structure and apply rigorous model selection. Furthermore we pursue the extension by predicting the future as described earlier.

III. Datasets Description

On this project, we work with two datasets. the "merged_autos" and the Yahoo Data for VIX. the first one contains the monthly sales of motor vehicles and parts dealers (more details on [1]). It goes from 01/01/2004 to 01/07/2011. We first transform the data by computing the log sales. the second one contains the daily VIX index close price from 2019 to 2020 extracted using Yahoo API.
The daily VIX contains a lot of noise due to trading. Daily google trends data used for VIX prediction is also noisy (because keywords depend on various factors(workday or weekend)). Hence we resample both data weekly, last value of the week for the VIX as we are interested in predicting the changes and the mean of the week for trends data. We then merge the trends with the VIX data into a single dataframe.

IV. Motor Vehicles and Parts Dealers

A. Methods

The sales time series is not stationary as we can see in Fig.1 with the associated Augmented Dickey-Fuller test p-value, a test for the null hypothesis that a unit root is present in a time series i.e the time series is not stationary.

We fail to reject the null hypothesis, hence we should make the time series stationary in order to build an AR overall model. Consequently, we use the following transformation broadly used in the financial industry: log(S(t)/S(t-1))
The time series obtained looks stationary from 2004 to 2008, from 2008 to 2014 and not on the whole period. This is confirmed by Augmented Dickey-Fuller test p-values on Fig.2.

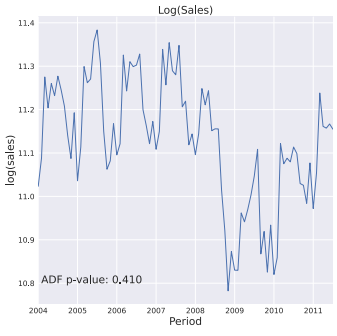

Fig. 1: Log(Sales)

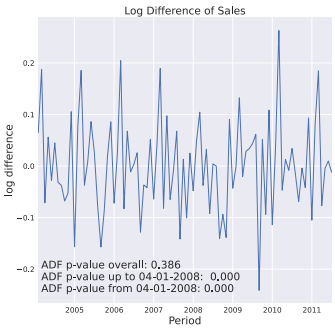

Fig. 2: Difference of Log(sales)

Having a time series that is stationary on both periods separately and since we are building an AR model, we need the partial auto-correlation function for both periods Fig3.

	coef	std err	z	P>—z—	[0.025	0.975]
intercept	-0.0005	0.006	-0.086	0.931	-0.011	0.010
sales.L6	-0.1731	0.076	-2.289	0.022	-0.321	-0.025
sales.L12	0.6927	0.073	9.477	0.000	0.549	0.836
suvs	0.5545	0.139	3.999	0.000	0.283	0.826
insurance	-0.2083	0.092	-2.262	0.024	-0.389	-0.028

TABLE I: Exogenous AR model summary

	Baseline	trends	Improvement
Paper's overall model	6.34%	5.66%	10.5%
Paper's recession model	8.86%	6.96%	21.5%
Our overall model	4.75%	4.30%	9.47%
Our recession model	3.93%	3.63%	7.6%

TABLE II: Paper's and Our's MAE comparison

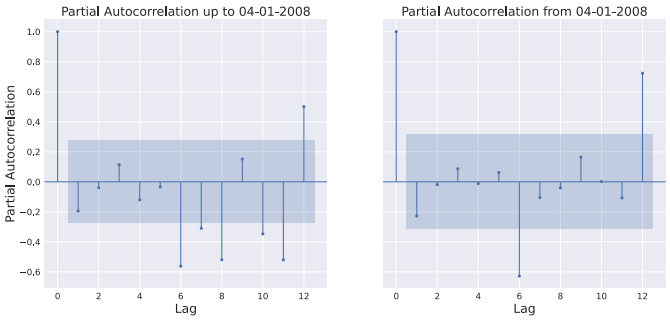

Fig. 3: Log(Sales) Partial Auto-correlation

Both periods show significance for lags 6 and 12 (The first period has more significant lags : 7, 8, 10 and 11). The choice of lag 12 was a good choice in the paper[1]. In order to be consistent with the rolling model of the paper, we chose the lags that are significant in both periods: 6 and 12.

We fit first the AR model with lags 6 and 12 (baseline model) and secondly by adding trends data ("suvs" and "insurance") for the whole period as in-sample. We then use a rolling leave-one-out prediction as in the paper[1], for the baseline model and trends model. Furthermore, the Granger-causality test for detecting if there is instantaneous causality between sales, suvs and insurance, will enable us to provide a robust validation of our findings.

Finally, we construct a future predicting model on the autos using the same baseline AR model, but this time we will use trends data from previous month to predict the current month.

B. Results and Findings

1) Predicting the present: All predictors chosen show statistical significance in-sample (small p-values) Table I. Let us now dive in the rolling out-sample results summarized by the MAE Table II. By fitting the rolling model, the MAE of the baseline model is 4.75% and the MAE of the trends model is 4.30%.

The new baseline model improved the out-sample predictions, In the paper the MAE was 6.34% and with our rigorously chosen model we obtain 4.75% i.e a 25% improvement. We can also clearly see that Trends data improves our predictions

	sales_x	suvs_x	insurance_x
sales_y	1.0	0.0000	0.0
suvs_y	0.0	1.0000	0.0
insurance_y	0.0	0.0007	1.0

TABLE III: Granger-Causality Test Matrix

MAE of the baseline model	4.75%
MAE of the trends model	4.22%
MAE improvement	11.20%

TABLE IV: Out-sample Future Predictions MAE

by 9.47% even with a better AR model. For the recession period, the MAE of the baseline model is 3.93% while the MAE of the trends model is 3.63%. We found that the improvement is only 7.6% compared to the new baseline model by adding trends data. This shows that Trends have present predictive power and that it is not redundant with the information contained in the sales time series. Let us validate our results with a statistical test: Granger-causality test Table III.

The matrix is read this way: does $suvs_x$ causes instantaneous movement of $sales_y$. We can see that we have p-values of almost 0 for "suvs" and "insurance" on "sales". This test enables us to reject the null hypothesis that there is no instantaneous causality of trends on sales. Finally by building rigorously a better model and using Granger-causality test we can now affirm that Google trends data has present predictive power claimed by the paper.

2) Predicting the future: Using the same procedure as before but considering as exogenous variables for the baseline model(AR with lags 6 and 12) "insurance" lag 1 and 2 to predict sales of next month. We did not consider the "suvs" lagged as they are not statistically significant. The MAE of the baseline model is 4.75% while the MAE of the trends model is 4.22%, the MAE improvemes by 11.20% as we can see in Table IV.

We managed to perform better than a simple AR model by building a model that includes google trends data. Consequently, we can see that google trends can be useful to predict future economic indexes.

Next we are going investigate if it can help predict future values of financial market: The VIX (Ticker on Yahoo:^ VIX).

V. Predicting the future: VIX

A. Methods

The first challenge is to determine which keywords for trends data to use to predict the VIX. The VIX is known as the fear index pricing current and expected US market volatility. The VIX reacts to various events: covid-crisis, expected recession, financial crisis... We thought of the following as they are relevant and can be source of market volatility from 2019 to 2020: recession, war, stock market,

	coef	std err	z	P>—z—	[0.025	0.975]
const	0.2255	0.675	0.334	0.738	-1.098	1.549
ar.L1	-1.2116	0.219	-5.530	0.000	-1.641	-0.782
ar.L2	-0.4739	0.217	-2.186	0.029	-0.899	-0.049
ma.L1	1.3990	0.184	7.590	0.000	1.038	1.760
ma.L2	0.7802	0.174	4.488	0.000	0.439	1.121
sigma2	18.4302	1.902	9.691	0.000	14.703	22.158

TABLE V: VIX ARMA model summary

conflict, terrorism, mortgage, bankrupcy, debt, elections, economy, protest, shooting, food bank, loan, virus, crisis.

To reduce the number of trends keywords,we decided to select the trends that are correlated with the VIX to remove any insignificant trend. Hence, we run a linear regression and check for statistical significance i.e pvalue ≤ 0.05 with exogenous being a keyword and endogenous being the VIX.

To apply time series theory, we check if our data is stationary. Hence, we run the Augmented Dickey-Fuller test. We then make the VIX stationary by log difference and trends data by difference as we cannot compute the log since some values are 0. We run again the ADF test to validate stationarity of the transformed time series.
Following the same framework in the previous section, we build a baseline model and models containing trends considering all the data as in-sample and rolling our model for out-sample predictions.

1) Autoregressive-moving-average "ARMA" using only VIX data to have a baseline to compare when adding trends data.

2) Vector-autoregressive "VAR" model using trends data. To determine its order we use the AIC criterion.

3) Vector-autoregressive-moving-average, "VARMA" model with the same order of the baseline model. It enables us to compare them since it must contain the baseline model.

Finally we are going to compare the results to determine if trends data has any future predictive power of weekly VIX.

B. Results and Findings

The selection of keywords through statistical significance of correlation results to the following keywords: recession, stock market, mortgage, food bank, loan, virus, crisis.
As a baseline model The ARMA(2,2) model shows statistical significance for all the regressors Table V. We chose this model after testing different parameters and it is the only one that shows significance for all the regressors by maximizing the numbers. The in-sample predictions Fig.4 of the ARMA(2,2) model struggles to predict big jumps as expected since VIX reacts brutally and considering only its past values give poor in-sample predictions.

MAE of the baseline model	4.01
MAE of the trends model VAR	8.03
MAE of the trends model VARMA	5.58
MAE improvement VAR	-100.08%
MAE improvement VARMA	-39.12%

TABLE VI: Out-Sample VIX Predictions MAE

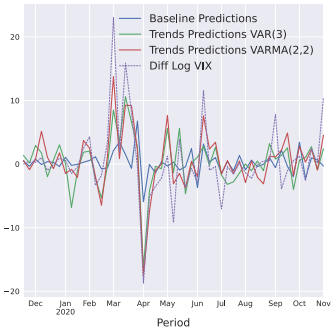

Fig. 4: Insample VIX Predictions

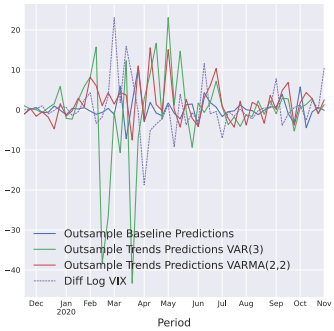

Fig. 5: OutSample VIX Predictions

Using the AIC criterion to build the VAR model with trends data we build a VAR(3). As explained Above we build the VARMA(2,2) model containing trends data. For in-sample predictions, both models react better than the baseline model to big jumps as we can see on Fig.4.

We are interested in predicting the VIX around the Covid crisis, as a rolling window we took 40 for the three models so that first prediction date is 2019-11-17. The results will show if it is possible to predict the VIX movement around this turbulent period. We can see that MAE(not in %) Table VI is high for all models and adding google trends results to lower performance of out-sample predictions. The plot 5 shows that the out-sample prediction is very bad!
Our hypothesis is that the market contains much more information and reacts way before people start looking for keywords. This shows how hard it is to predict financial markets. Several theories around market efficiency focus on the fact that the market data contains all available information. We are also aware that this data is available live in a high frequency and we imagine if there is any opportunity, big institutions will exploit it and make it disappear especially when using data as accessible as google trends to make predictions.

VI. Conclusion

The study confirms the claim of the Google paper: present predictive power of google trends on the "monthly sales of motor vehicles and parts dealers" during normal periods and recessions. Furthermore, we showed that trends data has future predictive power on the same economic indicator. These results are interesting as we believe trends data can predict other economic indicators before publication ! Moreover, we showed that predicting the weekly VIX is challenging using trends data as no improvement was found using VAR and VARMA models.

REFERENCES

[1] H. V. HYUNYOUNG CHOI, "Predicting the present with google trends," *THE ECONOMIC RECORD*, vol. 88, pp. 2–9, jun 2012.