**Investigating the effects of Google search term big data on predictability of macro-economic indices**

*Shervin Sharifi Rad*

# Introduction

## Problem statement

In the real world and models with realistic assumptions, the future, almost always holds uncertainty and risk. Therefore, predicting the future is essential for mitigating these risks, planning and adaptability of organisations. For example, central banks, governments and statistical organisation gather data on a large scale in order to gain insight into the concerning governed body and to forecast the trends and movements of their economies. This then allows them plan for legislature, monetary policies and campaigns to mitigate risks of failure. []

However, the collection of data is a costly and staggering struggle which then causes the data to represent the events of the past. Hence, instead of striving for the risks of the future, through the process of nowcasting, the central banks and governments attempt on building models to predict the present.

The lag associated with the gathered data can take from weeks to months which hinders models with higher volatility and lower market memory. For example, in micro-economic, financial and logistical cases, miniature lags within the data such as that of a security pricing, could raise transactional risks and disrupt the trading process. On a larger scale such as that of central banks, system lags on issues such as employment can lead to decisions which cause volatile market conditions and send the economy in boom and bust cyclical patterns. This effect has been seen for centuries where slow government responses to market bubbles give raise to the bursting of bubbles and setting off recessions. This is due to the poor predictability of the current market conditions by the statistical organisations and the resulting slow responses to raising of the interest rates and tightening monetary policies. []

With the growth of technology, the gathering and analysis of data has been eased through artificial intelligence analyses of big data. The definition and understanding of big data remains a big question to many. Big data is essentially, large datasets that can be computationally analysed to reveal trends and patterns which can illustrate human behaviour []. Sources such as Google, MasterCard, Federal Express, UPS and many other private organisations, gather data of their services which appears to be key in understanding market movements and positions. These data range from money market figures, internet information and/or courier statistics each sharing various insights. For example, the level of market risk can be guessed from the credit default swap spreads. As the spread between the insured and the premiums raise, it can be inferred that the market is becoming riskier and insuring the same security at the same price would require higher interest payments to accommodate for the systematic risk. In addition, with the growth of e-commerce and online platforms, courier data can demonstrate the economic growth and slow-down of certain regions, employment within the area and even inflation. This can be taken from number of insights such as the frequency of deliveries, competitive pricing of the services and human capital information of the services []. However, with the changing times and greater use of internet, many transactions are made and recorded online. Hence, insight into the minds of consumers is now readily available via the searches made online.

## Approach

The focus of this paper is, therefore, on google search terms to improve the nowcasting of economic indices such as Consumer Price Index (CPI), Real Disposable Personal Income (RDPI) and to take a further step and (in-sample rolling window) forecast future trends and probabilities using a kinetic Monte Carlo simulation.

It is hypothesised that search terms within Google search engine can be used to determine the effects of consumer behaviour on these metrics which allow prediction of inflation rates, market sentiment and allows decision making on fiscal and monetary policies such as taxation and aggregate demand (fiscal policies), interest rates and money supply (monetary policies).

A question can now be raised that what search terms can be used to predict such movements? Research carried out by Ainsworth et. al. (2000) estimated the value of internet transactions to be 9% of the GDP. This poses a great opportunity for the sector to have considerable effect on the economy and thus, act as an indicator of economy’s health. Market research carried out by eMarketer (emarketer.com), has found the top 10 e-commerce websites used in the USA with their respective monthly traffic and traffic share in (July) 2018 to be summarised as the following:



As seen in table XXX the top 5 visited websites carry 94.5% of the traffic related to e-commerce. On the other hand, the top grossing e-commerce websites (2nd Quarter) in 2018 are as follows[[1]](#footnote-1):



Comparing the two tables (XXX and XXX), it can be seen that four of the 10 top grossing retailers are also in the top ten mostly visited e-commerce websites in the United States. In addition, the four companies include 88.81% of the overall online retailer searches (Amazon, eBay, Walmart, Etsy) and 60.60% of the total internet sales that has taken place in the US over 2018. Therefore, aggregating the searches for: Amazon, eBay, Walmart and Costco will provide a sufficiently reliable picture of the online retail services which can then be utilised to predict the macro-economic indices[[2]](#footnote-2).

## Background and Literature review

To date, numerous researches has been carried out to evaluate the relevance of big data to economic context. Choi and Varian 2009 utilised Google Search Insights (GSI) data to predict metrics such as unemployment claims, automobile demand and vacation destinations. Similar to Real Disposable Personal Income, automobile demand can prove an economy’s growth through purchase of long-term assets. However, this highly depends on the economic conditions such as interest rates which can cause tendency to invest, save or even make borrowing more difficult. In addition, car sales may not be a true reflection of growth in urban areas as more residents may choose public transport due to ease of transport rather than considering it a luxury. All in all, the paper proved that the in-sample season Auto-Regressive forecasts of the indices which included Google search data outperformed models that did not include the data by around 5 to 20%. []

A paper published by the European Central Bank (Nymand-Andersen et. al. 2018), used an aggregate of car brands and dealerships to predict number of car sales in the euro area. Using a Vector autoregressive (VAR) and Autoregressive Distributed Lag (ARDL) models, the paper predicted the volume of car sales in both short and long term and proved a bi-directional relationship between the indicator and search volumes. The paper also proved that the Google Trends data, concatenated with: euro area household disposable income, euro area industrial confidence indicator, euro area harmonised inflation rate for cars and euro area household savings reduces forecasting errors up 131% in comparison to a baseline model. []

Guzman (2011) used Google data as predictor of inflation. The paper was able to create an index based on Google search terms which predicted inflation up to 12 months. This was compared to survey based indicators which proved to be highly biased and costly and lagged in gathering of data. Therefore, the real time availability, cost and efficiency of the search term models proved to be a huge success in utilising big data in forecasting macro-economic indices.

Preis et. al. (2010) examine the use of search data for measuring consumer sentiment.

Schmidt and Vosen (2009) and Lindberg (2011) use retail sales and consumption metrics through the use of internet datasets. The study by Schmidt and Vosen lacked category classifications and the search queries appear to have skewed the final results. Despite this, the inclusion of search terms appeared to improve in-sample forecasting of the indices. Another limitation with this paper was the time series limitation as at the time, the data collected by Google, did not amass to a reliable statistical significance. The paper however, suggested that seasonal adjustment to the Google data would increase the accuracy of the year-on-year (YoY) growth perspectives. []

# Data reliability and assumptions

The market under investigation was chosen as the US mainly due to the availability and accessibility of data. In addition, the US is deemed to be one of the largest markets with implications on almost all globalised economies which signifies its importance. On the other hand, to this, the sources which provide the data such as Federal Reserve Economics Data (FRED), has stringent rule book (System of National Accounts, SNA) on the collection and processing of the data which maintains the consistency of data collection and processing. Under the SNA, the major restrictions within the data collection were listed as []:

* Illegal activities: These cash flows will affect the subject under study; however, the data will not be available since they are conducted out of public sight.
* Family member care and loans, do-it-yourself projects and housework cash flows. By nature, these activities are self-contained and if included would undermine the long standing analytical purposes of economic data.

In addition, since the Google Trends data is not adjusted (YoY/seasonally/monthly), the macro-economic indicators have also been selected without any adjustments to remove bias and errors.

## Real Disposable Personal Income (RDPI)

Personal income is defined as the income that persons receive in return for their provision of labour, land and capital plus transfer receipts, less contributions for government social insurance. Upon this definition, real disposable income would be the personal income plus benefits income, less direct taxes and government charges when inflation adjusted.

This metric helps in predicting the consumption trends since with an increase in disposable income, household which allows for monetary policy planning and interest rate changes to control inflation in the near future.

## Consumer Price Index (CPI)

Consumer Price Index is the calculated change in price of a pre-determined basket of goods. The basket of goods comprises of hundreds of household essentials such as eggs, bread milk, etc, chosen and updated each year by the central bank. The change in price of the basket of goods is an indication of inflation. The aim of most governments with steady growth of the economy in mind is to maintain the level of inflation around 2% []. The metric also allows for the government to decide on legislature such as minimum wages and savings rate to ensure the economy is growing as a whole which makes CPI one of the most important indices.

Figure XXX shows the CPI , meaning that the CPI in 1982 was at 100 and any gains after that is represented in percentages. For example, CPI in January 2004 is 185.2% meaning there was an 85.2% increase in the price of the basket of goods since 1982.

The discrepancies in the basket of goods can be neglected since the variation would be reflected in all economic calculations. This is because CPI is used as a parameter for inflation rate which is reflected in the systematic risk of the market. Therefore, whilst calculating the Weighted Average Cost of Capital (WACC) in any financial and micro-economic context, the effects will be reflected and normalised.

## Google Trends

Google trends gathers data on any searched term on the Google search engine and provides granulated information categorised by region, time, category and data type (such as YouTube videos, images, news search and Google shopping). There are 30 categories on the top-level and 250 at the second level which uses an Artificial Intelligence natural language processing engine. The key words regarding “E-Commerce services” can be found under either “Shopping”, “Business and Industrial – Business Services” and “Computers and Electronics – Enterprise Technology”. It was decided that choosing the shopping subcategory, would include a wider retail domain instead. This also removes the possibility of excluding relevant data due to mis-categorisation and granularization of the data but also would focus the searches on retailers. For example, “Amazon” can be found under “travel” and other irrelevant categorised and would therefore skew the correlations further in the study.

In addition, the data type was selected as web search, since Google shopping is a relatively new launch by Google and is still under developed; hence most searches would be to redirect to the retail websites. The time frame was chosen to include the greatest overlap between the macro-economic index and the Google trends data. This was from the beginning of trends data collection in January 2004 since all macro-economic indices date back to the 20th century.

The data is then presented as a percentage, normalised to the greatest number searched hits within the chosen time frame within the geographical area chosen as seen in figure XXX.

A major limitation with Google Trends data is the lack of monthly web and mobile app traffic. For example, if the website is saved in browser cookies and redirects the search to the website before landing on google, the search term will not appear in in the Google trends data. In addition, in recent years, many e-commerce companies have developed mobile based platforms which is also excluded from the internet search terms. Therefore, in this study, the effects of the e-commerce on the economy is only limited to the highly searched companies on google where in some cases may not result in a transaction. However, this highly depends on the number of these anomalies and the limitation can be ignored if cross-correlation between the economic indicators can be found with the google search terms. Also, it is possible that other websites include links to products on the retail website which would also go undetected since they are not being searched directly through Google. This also goes for product searches where the search results would provide links to the e-commerce website. Out of all anomalies, this is assumed to be the greatest, judging by human behaviour.

This is why, the companies have been selected based on traffic volume and not by sales, since the availability of the data and accessibility to private website traffic is infeasible; in addition to the post processing and data cleansing required.

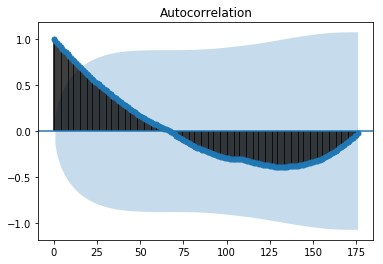
# Preliminary work

## Data fitting

Why moving average and not AR or ARIMA?

Justification of rolling window frame?

## Autocorrelation



The confidence interval represented by a cone has a default value of 95%.

A difficulty posed for future steps in the research is the size and number of data points available for the index. As mentioned previously, the monthly data can only be taken from the year XXX (even though data is available from XXX) to comply with the Google Trends dataset. However, having 177 datapoints to start with means that after XXX shifts in the lag boundary, the overlapping index and Google Trends data would be diminished to only XXX. This would seriously diminish the reliability of the assessment.

## Non-linear dependence

Financial markets have proven that asset prices move independently during bull markets, whereas in bear markets, all asset prices lose value together.

Evaluating cross-correlation on non-linear multivariate datasets can have irregularities in the sense that the forecast may work in bull markets. For example, imagine a random variable , uniformly distributed around zero, and a variable . This clearly shows ’s dependence on ; however, the correlation coefficient between them will be zero because it only detects linear dependencies between the two. This effect has rarely been spoken in previous papers written on the matter.

Although, looking at macro-economic indicators will mitigate this effect since indices such as CPI and RDPI are always upwards trending due to inflationary pressures and growth rates. However, in unstable and stagnant economies such as those in recession or deflation, forecasts may prove to misjudge the trends. In these cases, the linear correlation over-estimate dependencies in non-crisis periods and under-estimate correlations in crisis periods.

In a financial context, research by Ang et. al. (2001) and Patton (2002) showed that non-linear dependences demand a higher risk premium due to high correlation with bad market conditions.

# References

1. Bureau of Economic Analysis U.S. Department of Commerce, 2017, National Income and Product Accounts handbook, Chapter-2, p.10-p.19
2. U.S. Bureau of Economic Analysis, Real Disposable Personal Income [DSPIC96], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/DSPIC96, November 30, 2018.
3. Ainsworth, I., 2000. E-commerce: the wave of the future. Canadian Investment Review, 13(1), p.12.
4. Guzman, G., 2011. Internet search behavior as an economic forecasting tool: the case of inflation expectations. Journal of economic and social measurement, 36(3), pp.119–167.
5. <https://www.statista.com/statistics/324582/b2c-e-commerce-as-percentage-of-gdp-usa/>
6. <https://www.businessinsider.com/biggest-online-shopping-sites-list-2018-4?r=US&IR=T>
7. <https://www.emarketer.com/Chart/Top-10-US-Companies-Ranked-by-Retail-Ecommerce-Sales-Share-2018-of-US-retail-ecommerce-sales/220521>

1. The top 10 grossing retailers table excludes private companies, includes all internet transactions regardless of method of payment or fulfilment (standing order payments). [↑](#footnote-ref-1)
2. Note that the websites domain (.com) has been removed when selecting the search term as it does not represent the true searches made by users. [↑](#footnote-ref-2)