## The potential of the Google Trends data

### What is Google Trends?

Google Trends is a tool provided by Google. Google Trends reports an index of search activity. Google Trends supports users with two options, one is time-series and the other is geographical location. In both two options, there are three steps to measure the index.

In time-series option. In the first step, we choose a query that relates to the question we concern about and a period at a particular area. In the second step, Google measures the proportion of the number search of queries including the question in total queries in the chosen period. If the period is less than three months, the result is daily. If the period is larger than three months, the result is weekly. In the third step, Google ranks levels of proportion by day or week, the index is calculated by dividing daily/weekly proportion by the highest day/week proportion and time to 100. It means, according to Google definition[[1]](#footnote-1): “A value of 100 is the peak popularity for the term. A value of 50 means that the term is half as popular. A score of 0 means there was not enough data for this term.”

In the geographical location option, three steps are similar. Firstly, we choose a query about the question we concern about and geography at a particular time. In the second step, Google measures the proportion of the number search of queries including the question in total queries in chosen geography. In the third step, Google ranks levels of proportion by location, the index is calculated by dividing location proportion by the highest location proportion and time to 100. And also, according to Google's definition, we have: “100 is the location with the most popularity as a fraction of total searches in that location, a value of 50 indicates a location which is half as popular. A value of 0 indicates a location where there was not enough data for this term.”

**The query can be specific keyword, topic, or categories. Generally, topic and categories have been more favored than specific keyword recently. The reason is specific keyword may lead to ambiguity problem (Woloszko, 2020). For example,** the "Apple" keyword is related to not only the fruit but also the company. Or other instance is point out by Baker & Fradkin (2013), they need "jobs" keyword to measure job search, however, they realized that the keyword includes the term of personal name, "Steve Jobs". Some particular keywords have multiple meanings, thereby it leads to the risk that the researcher probably chooses the wrong term, and it is dangerous for the research (Stephens-Davidowitz & Varian, 2014). **Meanwhile,** the topics are created by combining multiple requests made on Google stand on their purpose and meaning and taking into account where users click (Woloszko, 2020). It allows to single out searches related to the term directed by the researcher and removes the keywords that belong to other terms. For example, if we choose the topic "Apple (company)", it combines searches for keywords such as "iPhone", "iPad", "apple watch", and "MacBook"; and it understandably excludes fruit term.

Google classifies search queries into about 1200 categories[[2]](#footnote-2). They are constructed as a 5-level hierarchical classification and aggregate together all searches related to each of them. The categories are assigned to individual searches by a probabilistic algorithm (Varian & Choi, 2011). For example, the queries would be allocated to the category **Vehicle Tires** which is a subcategory of **Auto Parts** which is a subcategory of **Automotive**. Another reason helping topics and categories outperformance specific keywords is they are harmonized across language (Woloszko, 2020). For instance, in term of a specific keyword, if we want to investigate the keyword "Birthday", we have to input "Sinh nhật" for Vietnamese, or "お誕生日**" for Japanese; meanwhile, the topic "Birthday" is similar in term of both Vietnamese and Japanese, and other languages. The categories, which are included in the Google Trends dataset, perform the same as the topic. With the number of topics and categories, the Google Trends dataset is believed that tends to cover a large area of economic fields (Stephens-Davidowitz & Varian, 2014). Furthermore, the data has been collected since 2004, and from all countries using Google, it has the ability to provides a large panel dataset for research activity (Woloszko, 2020). Therefore, there is an advancement in the implementation of the Google Trends index in research.**

**However, there are some concerns warned by literature when applying the Google Trends index. The first is careful when interpreting the trends in search behavior in the case of long-term (**Stephens-Davidowitz & Varian, 2014)**. For instance, U.S. searches related to the keyword "science" are downward since 2004. It does not mean that the interest in science declines. The reason is the Google searchers' composition has changed through time. In 2004, colleges and universities where searches on science were popular used the internet heavily. At the present, the composition of internet users has become broader. Chapter 3 will present how to resolve this problem. The second is watchful in conclusions based on the relative value of two searches at the national level (**Stephens-Davidowitz & Varian, 2014). For instance, in the U.S., the number of searches related to "Jewish" is 3.2 times more than "Mormon". It is not able to conclude that the Jewish population is 3.2 times larger than the Mormon population. It may be the Jewish people uses the internet in higher proportions than the Mormon people. Generally, the Google Trends data is more valuable for relative comparisons.

**So far, this section has already introduced some basic information about Google Trends. It presented what Google Trends is, what the Google Trends index stands for, how the data is collected and calculated. Furthermore, it shows the reason why recent research prefers topics and categories to a specific keyword, thereby pointing out the advantage of the Google Trends data application. On the other hand, potential pitfalls are warned. Next, several empirical research applying the Google Trends will be provided to show the potential of this dataset. These papers were written before 2020, my purpose is to point out the large area fields that Google Trends can cover. The application of this dataset for the Covid-19 pandemic will be highlighted in section 2.2.**

### The application of Google Trends data

One of the pioneers' efforts to manipulate the Google Trends data is the work of Choi & Varian (2011)[[3]](#footnote-3). The authors claim that the queries from Google search are often correlated with multiple economic factors and may be worth short-term economic prediction. They constructed a model to use Google search series for predicting some topics related to the economy: Motor vehicles and parts, Initial claims for unemployment benefits, Travel, and Consumer confidence. The predictors of the model of Choi & Varian (2011) are the Search Volume Indexes (SVIs) of Google search categories related to the concern topics ("Trucks & SUVs" and "Automotive Insurance" for Motor vehicles and parts; "Local/Jobs" and "Society/Social Services/Welfare & Unemployment" for Initial claims for the unemployment benefits; subcategory "Hong Kong" under the category "Vacation Destinations" for Travel; and "Crime & Justice", "Trucks & SUVs", and "Hybrid & Alternative Vehicles" for Consumer confidence), while the dependent variables are the indexes reported by official sources (such as U.S. Census Bureau for Motor vehicles and parts, U.S. Department of Labor for Initial claims for unemployment benefits, The Hong Kong Tourism Board for Travel, and Roy Morgan Consumer Confidence Index for Australia for Consumer confidence). The results showed that the queries added to help the model fitter. The authors found out that the SVIs are not useful to forecast the variations but can be valuable to detect turning points. They do not believe that Google Trends is valuable for *predicting the future*, but it is really helpful for *predicting the present*.

One of the very first papers using Google Trends is Suhoy (2009). She tried to replicate the work of the previous paper for several factors of Israeli economic for the purpose that evaluates the value of applying query indices as economic predictors. Based on the concept of growth cycles, the author believed that the popularity dynamics (represented by particular category of searches on Google) coincide with economic dynamics, therefore SVIs have ability to predict the economic growth. The number of topics concerned is larger than Choi & Varian (2011): Industrial production, Retail trade, Revenue of trade and services, Consumer imports, Exports of services, and The employment rate. The categories applied are not only outnumber ("Automotive", "Business", "Home & Garden", "Beauty & Personal Care", "Food & Drink", "Travel", "Real Estate", "Shopping", "Industries", "Computers & Electronics", and "Finance & Insurance) but also more details because of number of subcategories (31 queries are used). **Suhoy's work proved that query categories are valuable for growth-cycle prediction, especially these subcategories: "Recruiting and Staffing", "home appliances", "Travel", "Real estate", "Food and Drink", and "Beauty and Personal care".**

**In a similar manner used as the input variables of the prediction models, the Google Trends index also is approached as the outcome. Aiming to discover how accurate search indicators represent the peak of the business, Chen et al. (2014) applied several keywords including "Foreclosure", "Layoff", and "Recession", and they investigated how the queries reacted to the fluctuation of the economy during the 2007-2008 recession in the U.S. By using the Markov-switching model, they found out that SVIs are valuable for nowcasting real-time economic activity. They believed that Google search queries are worth pointing out the 2007-2008 peak turning point of the business cycle with both timeliness and accuracy.**

**Baker & Fradkin (2017) used Google Job Search Index (GJSI) as an explained variable to investigate the impact of unemployment insurance policy on labor markets. They highlighted multiple advantages of Google Trends data: (1) this data is free, (2) it is able to measure the impacts in real-time, (3) it is less suffered by sampling bias. The author demonstrated that GJSI has a strong link to job search from many surveys. This article concluded GJSI does react to the potential benefit duration of unemployment insurance, but the impact is small. As mentioning the data from surveys, Vosen & Schmidt (2011) had a considerable effort to compare survey-based indicators and Google Trends. They forecasted the private consumption (represented by consumer growth) based on predictors collected from two surveys, which are the University of Michigan's Consumer Sentiment Index (MCSI) and the Conference Board's Consumer Confidence Index (CCI). And then they compare the result to the prediction outcome of the model applying 56 weekly categories related to consumption from Google Trends (the authors compute monthly averages to fit with consumption data structure). Their founding confirmed that the prediction of Google factors outperforms the outcome of survey-based models. Google factors have the potential to offer significant benefits to forecasting private consumption.**

**In the attempt to predict the financial market volatility, Balakrishnan & Dixit (2013) applied 850 keywords that are strongly related to finance query the New York Times Article Search API to predict the value of the Future of the S&P500 Index traded at the Chicago Mercantile Exchange (the data is collected from Financial Time Series).** The authors concluded that Google Trends data for financially relevant keywords can predict well the market volatility. Especially, they showed that using a very small subset (5 keywords) that had the most related provided the best results.

**The last decade saw the broad area of economic fields of researchers applying the Google Trends data, such as Mccalum & Bury (2013) and Ficetola (2013) in environmental economics, or Martinez et al. (2021) in agricultural economics. Considering the search behavior has a strong link to the interests about the environment of the public, Macclum & Bury (2013) tried to investigate the trend of Google search queries related to the environment in the period from 2004-2009. They used data of 19 series that relate to environment term to study. And the authors found out that the strong negative slopes of search indicators during the period. They concluded that the interest of the public in environmental topics declines. Ficetola (2013) points out the mistake of Macclum & Bury (2013). The author highlighted the first concern which is said at the beginning of this section. Additionally, Ficetola (2013) included topics empathizing computer science, other disciplines (astrophysics), and leisure. They found out that all topics' search value decreased during the period. They concluded that the interest in the environment was not truly decreasing, the core reason is the increase of search for terms unrelated to, and advised that to be careful about the complexity of analyzing the internet search tools' data.**

**Martinez et al. (2021) tried to evaluate how helpful Google Trends data is as a predictor of grain prices. The explanatory is the queries related to three topics: Maize, Corn, and Soybean. While the outcome variables are the agricultural futures contracts which had the underlying assets were Soybean and Corn, in the period from January 2004 to August 2019. The result provides some pieces of evidence that Google Trends is valuable to predict grain prices.**

1. https://support.google.com/trends/answer/4365533?hl=en [↑](#footnote-ref-1)
2. https://github.com/pat310/google-trends-api/wiki/Google-Trends-Categories [↑](#footnote-ref-2)
3. It is the updated version of Choi & Varian (2008). Both articles have the title is "Predicting the Present with Google Trends". Since Choi & Varian (2008), multiple works about Google Trends had been released before the updated version, but I still introduce the Choi & Varian (2011) first because it and the previous version provided the basic approach for following papers of same field. [↑](#footnote-ref-3)