Models for nowcasting GDP with Google Trends:

* [Heikkinen (2019):](https://jyx.jyu.fi/bitstream/handle/123456789/66363/URN:NBN:fi:jyu-201911144867.pdf?isAllowed=y&sequence=1)
  + Note: this would work worse for places that have limited internet access
  + But majority of economic activity likely takes place where there is internet access
  + “However, studies using Google Trends data for nowcasting countries economic growth, i.e. gross domestic product (GDP), are relatively rare”
  + Lit review sources
  + “Explicitly stated, this study assumes that Google Trends data is a proxy for peoples’ interest in durable goods. For example, the more searches there are for Autos & Vehicles the consumers are signalling higher willingness to buy new cars and trucks. This increased consumption leads to increases in economic growth, as the automobile industry is a part of the GDP.”
  + ” Compared to the monthly data series, both of the official GDP statistics had a reporting lag of two months. For example, in 2019, Statistics Finland’s released fourth-quarter GDP statistics at the end of February. On the other hand, the official consumer survey data and Google Trends data are available monthly; hence, they do not have a similar time lag.”
  + “Furthermore, Google states that data do not include duplicate “searches from the same person over a short period of time”. This duplicate search term control reduces the possibility of people deliberately affecting search terms popularity. Google also specifies that Trends data include only search terms without special characters or apostrophes”
  + **“. This master’s thesis follows Götz and Knetsch (2019) paper to select appropriate initial subcategories.”**
    - I use the BoJ’s subcategory seletion
  + **“Consequently, this study applies modern dimension reduction methods, which are similar to Götz and Knetsch (2019) paper. With these dimension methods, initial 180 subcategories were compressed into 16 different broad categories”**
  + **Also used LASSO regression as a shrinkage method to include all of the predictors but penalize their inclusion**
    - “Thus, the LASSO method selects the coefficients that minimize regression residual, and meanwhile, the number of variables in the regression decreases. In other words, this method allows the researcher to use high dimensional data set in a relatively efficient manner.”
  + **“The purpose of this exercise was to simulate a real-world nowcasting situation (Stock & Watson, 2008, 1). This sort of pseudo-out-of-sample simulation is possible to estimate with two different strategies. The first strategy is “fixed” rolling window in which the sample size does not change. The second strategy is a recursive or expanding window, where the initial sample is expanding (Stock & Watson, 2008, 3–4). The first strategy demands a relatively long data series. However, because Google Trends is available since 2004, this study uses expanding window nowcasting strategy.”**
    - In this master’s thesis, the initial sample size was 20 % of the entire data, i.e. 12 periods. This initial sample size a rather short, which is because of the length of the Google Trends data, i.e. 60 periods. The forecasting exercise used both the most recent monthly data and aggregated monthly data to match the monthly data to the quarterly data. With these specifications, this thesis used following forecasting models that were estimated using ordinary least squares (OLS). Gradual description of the pseudo-out-of-sample exercise is in Appendix
  + **Run regular OLS with all categories, a LASSO that selects most important categories, the uses those categories for GDPt = Bo + B1GDPt-1 +Google Trendsit**
  + **We may want to add lags of the google trends variables too**
    - **Iterate over different kinds of models**
  + **“Similar to Finland, the inclusion of consumer confidence data also weakened Germany’s multivariate models. Therefore, univariate and multivariate results suggest that Google data is capable of capturing GDP information more effectively than consumer confidence in Germany at least”**
  + **“Unlike previous pseudo-out-of-sample exercises, this thesis’s cross-validation models are using “future” data and fixed sample size. This method allows for a longer forecasting period; hence, there are more point forecasts to evaluate and examine. More throughout depiction of this thesis’ cross-validation arrangement is in appendix 2.”**
  + **‘”**Donadelli (2015) found that Google searches have relation to policy-related uncertainty. Furthermore, policy-related Google searches are particularly popular when there are significant levels of uncertainty for economic conditions”
  + **“**Results of this master’s thesis suggest that consumer confidence data is superior for nowcasting Finland’s GDP. Finland is a relatively small country; therefore; its GDP is highly dependent on exports. Thus, Finnish Google searches could have difficulties capturing relevant information about the economy. Consequently, Google Trends data seem to work better in Germany, which is a large country with a large manufacturing sector”
* [Eichenauer et al (2021)](https://onlinelibrary.wiley.com/doi/epdf/10.1111/ecin.13049):
  + How to avoid sampling noise and construct long-run, frequency-consistent daily economic indices using Google trends
  + “As a result,daily data fail to capturelong‐run trends.Researchersthereforeface a trade‐offbetweenusinghigh‐frequencydaily seriesversustime‐consistentseries,wheresearchvolumesare comparableat differentand distantpointsin time.Businesscycleanalysisandforecastingmodels,however,typicallyrequiredata spanningmorethan a decade.”
  + “The two limitationsarise from a combinationof the followingthreefactors:(i) For privacyreasons,Googleonlyprovidesan indexof searchvolumes,ratherthan the actualnumberof searches.Scalingof the indexvarieswith thechosentime window:withineach window,the indexlies in the rangebetweenzero and 100. Only long time windowsallowcomparingmagnitudesof specificeventsover time or studyinglong‐runtrends.(ii) GSV defaultto weekly(monthly)data for time spanslongerthan 9 months(5.25 years).In combination,(i) and (ii) implythat it is not possibleto directlyextractlong‐rundaily GSV data from GoogleTrendsthat is consistentwith the long‐runtrendcapturedbythe monthlydata. (iii) For a chosentime frame,GSV data are basedon a randomsub‐sampledrawnby Google”
  + “In a secondstep, wecombinethe informationfrom monthlyand weeklyseriesinto a singledaily seriesthat is consistentwith the weeklyand monthlyseries.For this purpose,we applyChowand Lin's (1971)disaggregationroutinetwice.First,to disag-gregatethe monthlyseriesof GSV to weeklydata and, second,to disaggregatetheseweeklyresultsto daily data”
  + “herefore,althoughGoogleindexesthe searchvolumessuch that withina time windowthe highestvalueis normalizedto 100 and the lowestvalueto zero, the long‐termtrendand relativechangesare still capturedcorrectlyin the monthlyseries.This is not the case for daily and, to a lesserextent,for weeklydata”
* [Seabold and Coppola (2015):](https://openknowledge.worldbank.org/bitstream/handle/10986/22655/Nowcasting0pri0n0to0Central0America.pdf?sequence=1&isAllowed=y)
  + “These keywords were chosen ex ante with the belief that they contain relevant information that will allow us to use them as a proxy for consumer behavior and beliefs. Obtaining real-time insights into consumer behavior allows us to better predict price changes all other things equal. In some sense, the Trends data takes the place of traditional consumer-sentiment surveys. The keywords that we have chosen are listed in Table 1.”
  + “The second thing to note in figures 1 and 2 are that many of the observations for a single draw of the Google Trends data are exactly zero. These zero observations present two difficulties in particular – one conceptual and one practical. First, conceptually, these zeros suggest a lack of signal where presumably there should be some. As we collect more daily samples of the data, this problem becomes less and less, again assuming that the signal is well approximated by the mean. However, this problem does not disappear. Looking at the early parts of both series, there are still observations which are zero even at the mean”
  + “Second, as a practical problem, some of the Google Trends data contain strong seasonal components. Studies such as Carri`ere-Swallow and Labb´e [2013] alleviate the effects of seasonality in the trends data by using year-over-year percent changes for them as well as the series to forecast. However, if the base year is zero, we would lose this entire year of data.”
  + “After handling the sampling dimension, we apply transformations to smooth the data for each keyword and attempt to better identify the signal from the noise, given the nature of the search data. Here, we take several different approaches. First, we apply a simple exponential smoothing model with additive errors to the data.”
  + “The second takeaway is the importance of order identification in ARIMA modeling. This is perhaps not a surprise for any forecaster, but the successful results here using automatic techniques are encouraging. If a forecaster were to focus on fewer series and apply the Box-Jenkins methodology rather than relying on automatic model selection procedures it might be possible to outperform the benchmark models further.”
* [Combes and Bortoli (2016):](http://www.carmaconf.org/carma2016/wp-content/uploads/pdfs/4226.pdf)
  + “Despite the automatic selection of many Google Trends, it appears that forecasts’ accuracy is not significantly improved with adding more variables”
  + “Applied treatments are not very well documented but series are supposedly corrected accordingly to a trend resulting from an increase in popularity of the search engine itself. They are normalized too so that their maximum always equals 100, which means that they might be revised between one extraction at a certain date and another one later on and that direct comparison between two distinct series is not possible.”
  + Not a lot of explanation of methodologies but good baseline take
  + See also: <file:///C:/Users/16094/Downloads/mars2015E_d2.pdf>
* Choi and [Varian:](https://www.frbsf.org/economic-research/wp-content/uploads/sites/4/Varian-part_1.pdf)
  + Model with economic index comprising lagged variables, google trends, and other exogenous variables
  + Claim a BSM and a Kalman regression seems to work well
  + Helped predict unemployment rate for men
* [Stackoverflow](https://stackoverflow.com/questions/57709398/using-simple-models-on-google-trends-data-to-predict-something-doesnt-work-as-e):
  + Code to do baseline ML prediction using google trends
  + Also see for a possibly better SA process: <https://facebook.github.io/prophet/docs/seasonality,_holiday_effects,_and_regressors.html#seasonalities-that-depend-on-other-factors>
  + <https://www.kaggle.com/code/teyang/covid-19-google-trends-auto-arima-forecasting>
  + <https://silentsingerz.github.io/resume/images/Time%20Series%20Analysis.html>
* [Boone et al.:](https://static1.squarespace.com/static/5b9e942a8f5130f854dbef81/t/5be9a6208985835f898e3a0f/1542039073203/incorporating-google-trends-data-into-sales-forecasting.pdf)
  + AR model
* [Austin et al. (2021):](C://Users/16094/Downloads/wpiea2021295-print-pdf.pdf)
  + Explanation of Google Trends access, coverage, TaC, Conceptual alignment
  + “To illustrate how a “Google Trend” is calculated consider the following example. Assume there are 10,000 searches in week 1 in a region and that 1,000 are related to restaurants. The level of interest in restaurants is therefore 1,000/10,000=.1. Assume that each week we measure the level of interest in restaurants (e.g., week 2=.08, week 3=.09) as illustrated in Table 6. The weekly level of interest in restaurants is indexed to the week with the highest level of interest (week 4 in our example). Using search activity as a proxy for demand for restaurant services the trend would be interpreted as an indication that demand for restaurant services was increasing in the first four weeks, stable over the next three weeks and declining in the final weeks. This provides valuable information about turning points in activity.”
  + “s.It therefore seems appropriate that if we want to aggregate Google Trends to monitor current economic trends, we should aggregate them using the ISIC Rev 4. Classifying Google Trends according to this classification will facilitate the use of this information to improve the frequency and timeliness of economic indicators.”
    - Probably a step too far
  + “As noted earlier the category “Consumer Electronics” is comprised of topics such as “Sony,” “Fortnite,” “PlayStation,” “Xbox,” “Apple,” “Canon” etc”
  + “The benefit of the Google Trends data is that users have access to a long and high frequency time series. These data are particularly useful in helping understand turning points and are intended to be combined with and benchmarked to official measures to improve their timeliness and frequency. Therefore, the emphasis of the series will generally be on the current period. While the emphasis is on the current period a long time series is required to establish relationships and models with existing official measures of economic activity. Since there are many factors that can influence search intensity a 5-year moving intervals is used and the weekly trends are smoothed using a five-week moving average”
  + **“Finally, often the series exhibit lag effects and therefore for certain series – such as travel type series where vacation interest precedes the trip some consideration should be given to lagging the series. This needs to be done on a case-by-case basis. ”**
  + Even with the above methodology, there still may be significant sampling noise for smaller countries for certain categories. Taking multiple samples on a weekly basis and taking the average of SVIs as the current SVI may reduce this variance as suggested in Woloszko (2020)
  + As a result, rescaling SVI\_a over SVI\_b using the ratio between the two series based on only one observation is probably not a good idea. A possible solution would be to multiply SVI\_b by the mean of the ratio of SVI\_a/SVI\_b taken over all common observations.
    - Applicable for when I link 2019 to 2020
* [Nakazawa (2022):](https://www.boj.or.jp/en/research/wps_rev/wps_2022/data/wp22e09.pdf)
  + “ In Japan, it takes about one and a half months for the first preliminary estimates of GDP to be released. The attempt to forecast the current GDP in real time using the most up-to-date information is called "GDP nowcasting," and it has been attracting increasing attention worldwide in recent years”
  + “, the Federal Reserve Bank of Atlanta and the Federal Reserve Bank of New York operate nowcasting models called "GDP Now" and "Nowcasting Report," respectively. They update their forecasts and post them on their websites”
  + “However, the performance of the existing nowcasting models, which only use monthly and quarterly indicators, has been deteriorating since the spread of COVID-19 in 20202 . This is partly due to the repeated introduction and then lifting of public health measures, which has increased the amplitude of economic fluctuations. The existing models which use traditional monthly and quarterly economic data with a (albeit relatively short) lag before publication, means the most up to date situation is not fully incorporated into the forecast values.”
  + “They find that the constructing model outperforms models using only traditional monthly and quarterly economic data in the forecasting accuracy. Woloszko (2020) also constructs a GDP nowcasting model for 46 countries using daily available Internet search volume data.”
  + The Benchmark model is a forecast combination model, which is obtained by simply averaging the following three forecasts: (i) estimates from the Bridge model; (ii) estimates from the CMIDAS (Combined Mixed-Data Sampling) model; and (iii) economists' GDP forecasts (in JCER ESP forecast survey)
  + The first condition is that the categories are considered to be directly related to shortterm economic fluctuations. Many of the 1,132 categories listed above are not considered to satisfy this condition, such as "astronomy" and "obesity." Therefore, out of the 1,132 categories, we select 252 categories that are considered a priori to be related to short-term economic fluctuations, such as categories related to the demand for specific goods and services (e.g., "wine" and "weddings") and categories that are considered to reflect broad business sentiment (e.g., "bankruptcy" and "retail trade"). As the second condition, we exclude search volume data whose values significantly depend on when the data is downloaded. According to Medeiros and Pires (2021), the data values of Google Trends is based on sampled data. Thus, if the number of searches for a given word or category is significantly scarce, the data can change significantly depending on the time point at which the data is downloaded. Since such a problem affects the reproducibility of the results, we eliminate these "volatile" series. Specifically, we take multiple samples of the same search count data at different times, and we use categories for which the standard deviation between the samples was less than a certain level14. The number of categories that satisfy both the first and second conditions is 217.
  + In the Google Trends data, there are some discontinuities due to changes such as that of definition. To deal with this, we adjust data with reference to Woloszko (2020)15. Seasonal adjustments are conducted to the adjusted series. We call them search volume index for each category in this paper. To eliminate their long-term trend, we use the deviation from the trend calculated by the Hodrick-Prescott filter16 .
  + Thus, the number of samples is less than the number of explanatory variables, making the OLS estimation impossible. In this paper, we avoid this problem by using the Elastic Net estimation, a machine learning technique, to estimate coefficients and select variables appropriately at the same time.
  + Selected "Best models" improve the forecasting accuracy mainly for the 2-month prior forecast compared to the Benchmark model and enable us to nowcast GDP earlier and more accurately.
* [Buell et al. (2021):](C://Users/16094/Downloads/wpiea2021124-print-pdf.pdf)
  + The Machine Learning (ML) approach implemented in this paper addresses common shortfalls of existing nowcasting approaches in the following ways:11 1) Implementation of five different ML regression models with cross-frequency skip-sampling method and the inclusion of comprehensive cross-validation checks 2) Integration of mixed and higher frequency data (daily and biweekly, in addition to monthly and quarterly), including alternative data such as Google Trend search terms 3) Ability to handle a wide array of variables (compared to OLS or other parametric modelling approaches), while avoiding overfitting through regularization and parameter tuning 4) Greater flexibility to generate model predictions without the need to wait for all variables to be updated at month or quarter end (e.g., models will update nowcast predictions for quarter-end GDP given any incremental amount of data as it becomes available)
  + **Their methods are pretty advanced**
  + Recommendations:
    - First, the variables should be transformed to growth rates so that each time series is relatively stationary • Second, high–frequency data should be aggregated into weekly or monthly intervals to reduce the runtime of a single filtration.24 • Third, we recommend the inclusion of additional variables to improve the factor’s explanatory power.
  + this paper uses three different evaluation metrics: a rolling score, a recession score, and a cross validation score
  + For the three countries assessed, the linear models (Elastic Net, SVR) perform better when predicting recessions, but the tree-based models (Random Forest, XGBoost) perform better when predicting non-recessionary periods. This result is in line with the well-known bias variance tradeoff between the expressiveness of a model and over-fitting (see Athey 2018).26 The tree-based ML models are more flexible than their linear model counterparts and can capture more complex relationships. Since most periods in the training data are non-recessionary, tree-based models in general have better performance in predicting non-recessionary periods. However, tree-based models display overfitting due to higher variance and therefore underperform in predicting the less frequent recession events. Linear ML models, on the other hand, exhibit lower variance and lower bias because of their simpler functional form, and thus out-perform in predicting recessions
  + **To aid in the interpretation of the models presented in this report, as well as in the prioritization of data collection by the fund, rankings of variable importance (i.e., feature importance) for the top two ML models for each country (determined by the cross validation RMSE score) are included.**
  + the nowcast methods presented in this paper can be compared to a benchmark Bayesian vector autoregression (B-VAR) with shrinkage; doing so will further demonstrate how a combination of different methods can improve the prediction performance for different countries
  + Further investigation into appropriate estimation of values for missing variables (e.g., simple mean or mode, k nearest neighbor) may improve prediction performance for the three-pass regression method. Additional refinements to the Kalman filtration to develop a daily or weekly high frequency latent factor that is correlated with GDP change would improve the latent factor’s usability and help build a leading economic indicator for quarterly GDP.
  + Finally, future work can include additional visualizations of nowcasts to aid policymaking, such as computation of density nowcasts, generation of predictive distributions (e.g., alternative future sample paths) for GDP growth point estimates, and further decomposition of how different data categories contribute to the nowcast outcome
* [Richardson et al. (2018):](https://www.bis.org/ifc/publ/ifcb50_15.pdf)
  + The majority of the ML algorithms also outperform the other two commonly used statistical benchmarks, namely the factor model and the small Bayesian VAR model.
* [Wolosko (2020):](https://www.oecd-ilibrary.org/docserver/6b9c7518-en.pdf?expires=1669247413&id=id&accname=guest&checksum=739078B261A0B4BFF5C413D1D8716F6A)
  + High-frequency and big data are often subject to limitations as the original purpose of their collection is usually not scientific analysis. These caveats call for specific attention and statistical preprocessing. Google Trends variables are transformed to year-over-year growth rate in order to remove seasonality. Breaks occurring in January 2011 and January 2016 caused by changes in the data collection process are addressed by smoothing the year-on-year growth rates.
  + **Finally, as the Google Search user base has increased dramatically since 2004, the relative search intensities of most search categories decrease over time. The methodology used to filter out the long-term bias as well as the detailed preprocessing treatments are described in detail in Annex A.**
  + The algorithm captures non-linearities that are likely to be key when there are extreme movements in GDP, but which are difficult to estimate with more conventional econometric approaches. Cross-country differences related to Google Search’s market penetration or institutional settings are flexibly captured as the neural network allows for all possible interactions between Google Trends variables and country dummies.
  + This paper uses a neural panel model, which exploits a large sample of observations from 46 countries while capturing cross-country heterogeneity. Neural networks are able to handle heterogeneity in the data as long as country dummies are included. A neural network whose architecture incudes an intermediate layer with enough neurons (in our case, 100) can flexibly model each possible interaction between Google Trends variables and country dummies. Each neuron takes as input signals from Google Trends variables and country dummies, and returns a non-linear function of the weighted sum of these inputs. As a result, the model can capture country-specific elasticities
  + : Shapley values are the contributions of a variable to the GDP growth estimate predicted by the model. Variables are ranked by importance, and for each variable. Each point correspond to an observation (that is a given month \* a given country) and its colour depends on the value of the variable
  + **Explanation of all of the things that need to be addressed in pre-processing**
  + **Gives code for running model**
* [Bantis et al. (2022):](C://Users/16094/Downloads/SSRN-id3860194.pdf)
  + “Using more disaggregated Google Trends data than its “categories” is not beneficial”
  + “owever, a critical question we address is whether these findings for advanced economies, such as Germany and the Euro area, are replicated for emerging economies. There are opposing reasons suggesting that Google search data might be more or less valuable for less developed economies. In such economies, traditional data sources may be of a lower quality, or information may not be available or may be fragmented. This suggests Google Trends data may fill a void and be more valuable. Against this, the Google Trends data may themselves be less useful if lower rates of internet usage make the information less representative.”
  + “The results of our forecasting exercises provide a number of insights. First, as most of the literature suggests, dynamic factor models successfully incorporate new information as it becomes available, with forecast errors tending to decrease as we move from forecasting to nowcasting and backcasting”
  + “Second, for the U.S., estimates of factor models outperform a simple autoregressive benchmark at all horizons, while for Brazil, they primarily outperform the benchmark at nowcasting and backcasting. Third, factor models that utilize both economic indicators and Google Trends categories outperform by far the benchmark in both countries, establishing the value of “Big Data” in the form of Google Trends data for now(fore)casting GDP growth. Fourth, benefits from performing variable selection before the computation of common factors tend to arise mainly at one-quarter-ahead forecast horizons (h = 1), and their performance decreases as we incorporate more data”
  + **Pretty advanced**
* [Medeiros and Pires (2021):](https://www.econstor.eu/bitstream/10419/249731/1/td683.pdf)
  + Talks about how google trends data depends on when you pull it
  + “First of all, it is important to mention that very popular terms (e.g. COVID-19 in 2020) do not vary so much among different samples. However, in many situations our terms of interest are not these very popular ones. So how to overcome the possible problem shown in the plots above? The answer is very simple. By gathering many different samples and averaging across every term, one can get a more reliable time series of that term”
  + **LASSO models that are simple**
* <https://www.tandfonline.com/doi/full/10.1080/15427560.2021.1913160?needAccess=true>