Nowcasting with Big Data: is Google useful in the presence of other information?

by Xinyuan Li, London Business School

Discussant: Ajit Desai, Bank of Canada

Modelling with Big Data and Machine Learning Workshop Jointly Oragized by BoE, FRB, DAFM
26 - 27 November 2018, BoE, London

Primary Objectives

The author is trying to address the following questions:

- Is it possible to exploit new data sources, such as web-based queries (specifically, Google Trends and Google Correlates), for macroeconomic nowcasting/forecasting? (Specifically, US jobless claims and unemployment rate).

- In the presence of other variables (traditionally used predictors or regressors for forecasting), does these new datasets with their high-frequency and timely availability, improve forecast? (Specifically, US jobless claims and unemployment rate).

Primary Inferences

- Google Trends and Correlate data has potential (we need to address some issues: sampling error, seasonality, keyword research, etc.).

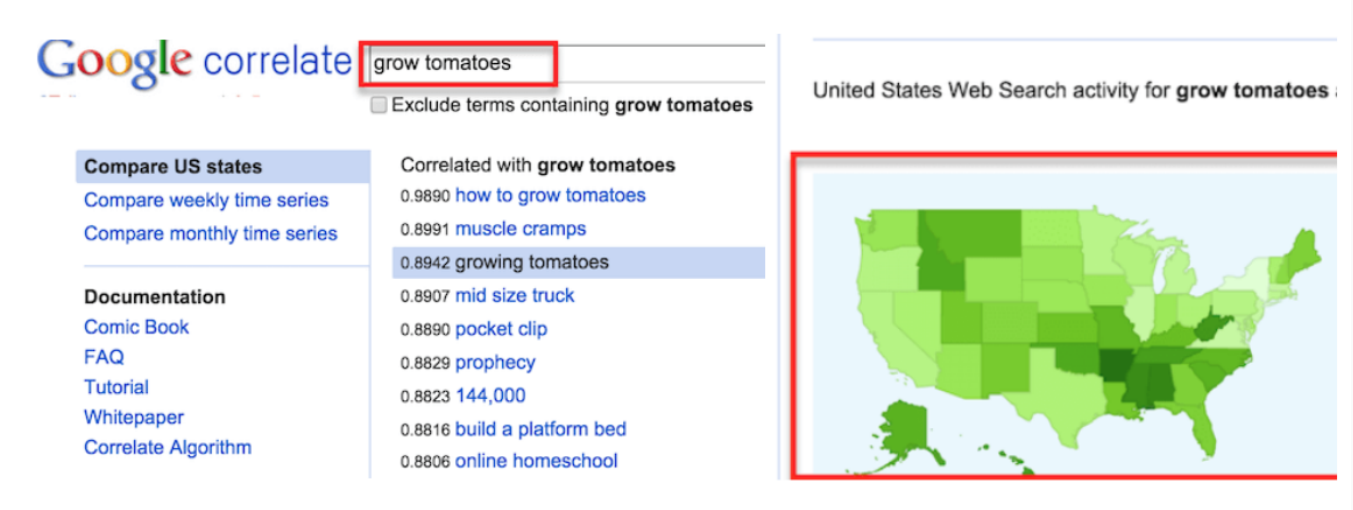
- Google Trends data hardly improves the forecast of US jobless claims or unemployment (in the presence of other variables).

- Google Correlate data significantly enhance the forecast accuracy of the initial claims and unemployment (even with the presence of other variables).

  According to Google, Correlate is a tool on Google Trends which enables users to find queries with a similar pattern to a target series.

  In other words, the pattern generates the keywords as opposed to the keywords generating the pattern as in Trends.

Google Correlate: Hidden Correlations (not Causation)



Note the terms "Mid size truck," "online homeschool," "prophecy," etc are all highly correlate with "grow tomatoes."

Those terms alone may seem not significant but one could connect them to make a meaningful story. For example, it is probably someone who drives a mid size pickup truck, lives in Oklahoma, homeschool's his/her kids, reads books like Left Behind etc.

`https://www.shivarweb.com/5374/google-correlate-for-marketing/`

Comments

- It would be interesting to use Google Trends and Correlate datasets together for forecasting US jobless claims or unemployment.

- Google Trends and Google Correlate datasets are growing every day (also, more users are using google search engine). Is it possible that these data sets could become even more useful in the future?

- There are some other search engines (bing, yahoo, duckduckgo, etc.) many users might be using or will start using. In that case, using only Google queries will bias the results?

- Seasonality in some of Google Correlate series (used) might be very different (or not present) compared to the target variables. How to deal with that?

Comments

- The technological changes happening in the Google platform (for many reasons), i.e., Google search engine is evolving every day which could influence search behavior over time and also, as search engine is becoming accessible to more users who could influence trends and correlates series (but not target), how to account the effects of these factors?

- As the number of predictors (regressors) increases (in the case where 50 correlates + other variables) could affect the performance of the model? In such cases, one could try using dimensional reduction techniques (feature abstraction, ML-based variable selection techniques).

Thank You