DEPARTMENT OF ECONOMICS AND FINANCE

SCHOOL OF BUSINESS AND ECONOMICS

UNIVERSITY OF CANTERBURY

CHRISTCHURCH, NEW ZEALAND

Predicting Chinese consumption series with Baidu

Zhongchen Song
Tom Coupé

WORKING PAPER

No. 19/2022

Department of Economics and Finance
UC Business School
University of Canterbury
Private Bag 4800, Christchurch
New Zealand

WORKING PAPER No. 19/2022

Stock Liquidity and Firm-Level Political Risk
Predicting Chinese consumption series with Baidu

Zhongchen Song[1]
Tom Coupé[1†]

December 2022

Abstract: There is a substantial literature that suggests that search behavior data from Google Trends can be used for both private and public sector decision-making. In this paper, we use search behavior data from Baidu, the internet search engine most popular in China, to analyze whether these can improve nowcasts and forecasts of the Chinese economy. Using a wide variety of estimation and variable selection procedures, we find that Baidu's search data can improve nowcast and forecast performance of the sales of automobiles and mobile phones reducing forecast errors by more than 10%, as well as reducing forecast errors of total retail sales of consumptions goods in China by more than 40%. Google Trends data, in contrast, do not improve performance.

Keywords: China, Baidu Index, Google Trends, forecasting, consumption.

JEL Classifications: C53, E21, E27

[1] Department of Economics and Finance, University of Canterbury, NEW ZEALAND

† Corresponding author: Tom Coupé. Email: tom.coupe@canterbury.ac.nz

1 Introduction

Since Google launched their keyword research and keyword search volume service Google Trends, researchers have tried to utilize this data in analyzing and predicting consumer choice and consumption economic aggregates around the world. For example Ettredge, Gerdes & Karuga, 2005; Vosen & Schmidt, 2011; Choi & Varian, 2012; Carrière-Swallow & Labbé, 2013; Woo & Owen, 2019; Yu et al., 2019. etc. These studies mostly concluded that the utilization of Google Trends can increase prediction accuracies. However, as the popularity of different search engines is highly dependent on the region and the time span, it is doubtful if these results on Google Trends can be generalized to other countries where other search engines dominate the market. Indeed, recently researchers have shifted their focus on utilizing other search engines, like Baidu, to analyze if data from other search engines can be used in a similar fashion. This strand of literature has focused on whether Baidu can help predict tourism flows (Yang et al. 2015; Li et al. 2018; Huang, Zhang & Ding, 2017; Sun et al., 2019), stock returns and stock market volatility (Shen et al., 2017; Fang et al., 2020).

We contribute to this literature by investigating whether Baidu can help to forecast macro-level consumption in China. Improvements in forecasting of macro-level consumption in China can help both public and private decision-makers. For example, better forecasts of macroeconomic consumption can help local and national governments when making decisions about fiscal policy. Similarly, private companies both in China and around the world can benefit from better forecasts of Chinese consumption when planning inventories or marketing campaigns.

We first focus on predicting total retail sales of automobile and communication appliances, published by the Chinese Statistical Bureau. When considering the purchase of a car or a mobile phone, consumers are likely to use the internet to search for information about the performance of various brands. This allows us to check whether measures of the internet search intensity for specific car and mobile phone brands contributes to predictive performance. In the US context, Choi and Varian (2012) have shown that search intensity as measured by Google Trends, increases the performance of models predicting US car sales. Similarly, Carrière-Swallow & Labbé (2013) have shown that Google Trends data improves the performance of models predicting Chilean automobile sales.

We then focus on whether Baidu can help to predict total retail sales in China. Because it is less clear which search terms can be relevant for the predicition of total sales, we compute search intensities in Baidu, for the titles of more than 1000 broad search categories that are available in Google. Such category data has been used by papers that predict aggregate consumption using Google Trends. (Vosen & Schmidt, 2011; Woo & Owen, 2019)

Data on total retail sales of Chinese automobile, communication appliances, and total retail sales are available from 2011 to 2019. We chose these two sectors because large enterprises dominate them. Therefore, a limited number of keywords can represent the entire sector. In addition, Choi & Varian (2012), the seminal study that analyzes whether using Google Trends data can improve forecasts also studied automobile sales.

They are published with a 1-month delay. This means that data for August is only available in September. In contrast, data on search intensity are available on a daily basis from Baidu. These data incorporate new information not embedded in lagged sales data. Our analysis shows that adding information from Baidu search intensities to traditional models can improve for prediction performance for sales of the two sectors. In addition, we find that predictions on total retail sales of all consumption goods in China can be improved by a substantial amount when models are augmented with Baidu Index. Further, the improved performance from Baidu data is greater than that from Google Trends or Chinese Consumer Confidence surveys.

There exists a large literature that shows Google Trends data can be used to improve forecasting accuracy. The literature that uses China's most popular search engine, Baidu Index is much less developed. The Baidu Index has huge research potential and can be used to analyze various research questions. This paper demonstrates how it can be used to monitor trends in the economy and improve the forecasting accuracy of a number of economic statistics, thus paving the way for future researchers to use the Baidu Index to predict other interesting economic development issues in the Chinese context.

This paper is structured as the following: In section 2, we provide background and literature review of this paper. In section 3, we compare the usage and function of Baidu Index and Google Trends. Section 4 details the data and model specifications used in this paper. Section 5 demonstrates the prediction results for retail sales in the automobile and communication appliances sector. Section 6 provides the prediction results of total retail sales of consumption goods. Section 7 concludes this paper.

2 Background and literature review

For many countries, private consumption is the single most important component of GDP. It, therefore, should come as no surprise that there is a long tradition of forecasting consumption. Historically, forecasters have focused on survey-based indicators like consumer sentiment or consumer confidence indices. One of the most widely cited papers in this strand of literature is Carroll, Fuhrer & Wilcox (1994). They found that lagged values of the consumer sentiment index explain about 14 percent of the variations in the growth of total real personal consumption. Other early studies using US data are Bram & Ludvigson (1998) and Howrey (2001). They also concluded that survey-based indicators help to lower the forecasting error for US private consumption. More recently, Lahiri, Monokroussos & Zhao (2015) re-examined these initial studies using higher frequency (monthly) data and more disaggregated data. They confirmed these earlier results and highlighted the efficacy of using Consumer Confidence data.

While initial studies focused on the US, other countries have also been investigated. Kwan & Cotsomitis (2007) concluded that Canada's index of Consumer Attitudes improved prediction of Canadian personal consumption. Similar results were found by Gausden & Hasan (2018) for the UK, and Juhro & Lyke (2020) for Indonesia. Dees & Brinca (2013), using European data (as well as US), found that the predictive contribution of consumer confidence was greatest when it experienced large volatility, such as during the Global Financial Crisis. There are exceptions as, for example, Cotsomitis & Kwan (2006) found that survey-based indices provided limited out of sample predictive capability for 9 European countries. However, the overall conclusion from the literature is that survey-based indicators like consumer confidence or consumer sentiment improve the forecasting accuracy of private consumption.

Given the initial question of the importance of survey-based data has largely been answered, and since data on search intensity became available through Google Trends[1],

[1] Jun et al. (2018) provide a network analysis of 657 papers that use data from Google Trends.

attention has more recently shifted to a new question: can search intensity data help to better predict economic time series?

Search engine data like Google Trends reflects people’s daily internet search behavior. As people often use search engines to research items they are interested in buying, Google Trends can track consumers’ interests. For many consumers, the internet is the main source of product information. Therefore, internet search intensity has the potential to correlate with consumers purchasing decisions. The ability of search engine data to capture people’s consumption habits thus might be useful for economic forecasts, managing stockpiles, and so on.

To illustrate, a consumer who wants to buy a product is very likely to search for information by typing a related keyword into Google. They are also very likely to conduct more research to find details about the product as they learn more. The searches consumers conduct can represent their interests in the product, and, therefore, represent their potential to buy a certain product (Bakirtas & Gulpinar Demirci, 2022). By aggregating this information, Google Trends represents the interests of a large part of the population, which in turn can be used to forecast aggregate consumption.

The first paper that suggested that internet search volume data, like Google Trends, can be useful when making economic forecasts was Ettredge, Gerdes & Karuga (2005). They showed an association between unemployment-related searches and the unemployment rate.[2]

Other studies followed. Goel et al. (2010) reported that consumer search behavior can help to predict box-office revenue for feature films and the sales of video games. Choi & Varian (2012) found that Google Trends data improved forecasts of motor vehicle and parts sales. Wu & Brynjolfsson (2015) showed that Google data can help to predict housing market sales and prices, with models incorporating Google data beating the predictions of experts from the National Association of Realtors.

Outside the US, Google trends have been found to improve forecasts of tourism flows to the Caribbean (Bangwayo-Skeete & Skeete; 2015), to Austria and Belgium (Önder; 2017), to

[2] Also, Askitas & Zimmermann (2009), D’Amuri & Marcucci (2017), Fondeur & Karame (2013), Naccarato et al. (2018), and Mihaela (2020) have investigated this link between internet search data and unemployment.

Japan (Park, Lee & Song; 2017), and to Germany (Bokelmann & Lessmann; 2019). Google Trends also has been shown to improve the forecasting accuracy of UK cinema admissions (Hand & Judge; 2012). Closer to this study, Carrière-Swallow & Labbé (2013) found Google Trends data can improve nowcasts of Chilean automobile sales.

Some studies have incorporated both survey-based measures and internet search intensity measures. Vosen & Schmidt (2011) compared the nowcasting and forecasting performance of models incorporating Google Trends and survey-based indicators and found that models incorporating internet search data outperformed the models using Survey Based Indicators. Similarly, Woo & Owen (2018) treat survey-based indicators as complementary and found evidence that Google Trends data increase the accuracy of the predictions in all sectors of consumption (durable goods, nondurable goods, and services), although the magnitude of the improvement depends on model specification.

These academic studies also have led to many practical applications. Central banks around the world have been exploring the use of Google Trends in economic models. For example, the Bank of Israel uses Google Trends to generate a monthly index that reflects the current health of the economy. This is then presented to the monetary policy committee to determine the country's interest rate. Similarly, the Bank of England uses search terms associated with the U.K.'s jobseeker allowance to predict unemployment, while the Bank of Spain uses Google Trends to predict the inflow of British tourists traveling to Spain (Morris; 2012)

Private companies have used Google Trends to conduct market research, determine inventories, produce forecasts on revenue and sales, and set up operational strategies. PwC, for example, uses Google Trends search query data to understand how popular 'Black Friday' is in South Africa (Krugel & Viljoen; 2019). Mindshare, a global media agency network uses searches for flu symptoms to create an outdoor advertisement network that raised sales by 40% (Armstrong; 2016). And news outlets like CNN and the Guardian track search volume data associated with political candidates. They use this information to better understand what their readers are interested in (Armstrong; 2016).

Compared to Google Trends, Baidu's search intensity Index has received far less attention. To date, most research using Baidu data has focused on predicting tourism flows to famous travel destinations in China. Huang, Zhang & Ding (2017) forecasted tourism flows to the Forbidden City in Beijing. Models incorporating keywords from the Baidu Index created more accurate forecasts. Similarly, Li et al. (2018) and Liu et al. (2018) use Baidu to predicting tourism flows to Beijing and Hainan Province; and Guizhou Province respectively. Yang et al. (2015) found that both Google and Baidu data improved forecasts of tourist flows to Hainan province, but that Baidu performed better, presumably due to its large market share in China. Sun et al. (2019) forecasted tourism flows to Beijing and found that the model using both Baidu and Google data performed better than models using either one by itself.

Apart from tourism flows, Baidu has been used to make forecasts about the stock market. Shen et al. (2017) reported that Baidu search data improved predictions of stock returns, and Fang et al. (2020) found that it produced more accurate forecasts of volatility in the Chinese stock market.

Baidu Index has also been used to monitor trends of the COVID-19 pandemic. Fang et al. (2021) analyzed the use of keywords “Coronavirus epidemic,” “N95 mask,” and “Wuhan epidemic,” to improve the accuracy of COVID-19 prediction models. They found that both in-sample and out-of-sample prediction accuracy significantly improved after introducing the Baidu Index.

To date, only one paper has investigated whether Baidu Index can help predict Chinese consumption. Fang et al. (2019) predicted sales of the Huawei Mate7 phone, adding 2 keywords related to the phone model to a simple AR(1) model. They found that models incorporating Baidu Index data performed better nowcasts than the AR(1) only model.

Our paper differs from previous research in that we focus on aggregate analysis of two important sectors of the Chinese consumer market, car sales and phone sales, and that we generalize our findings to the total retail sales of consumption goods in China. Further, our analysis is more extensive, employing hundreds of keywords and more advanced forecasting techniques, including the LASSO model. We also compare the benefit of using Baidu search data with that Google trends and consumer confidence data.

3 Comparison of Baidu and Google

Google Trends and Baidu Index are both data sources that reflect the search volume of users' queries. The data provided by Google Trends date back to January 2004, while the data from Baidu go back to June 2006.

FIGURES 1 and 2 show the market share of several search engines worldwide and in China, respectively, between January 2010 to July 2020. Worldwide, Google has the largest market share, by far, and its market share has been very stable during the past decade. As of July 2020, Google's market share was 92.2%. Bing and Yahoo! were 2.78% and 1.6%, respectively. Baidu stood at 0.9%.

China tells a different story. As shown in FIGURE 2, over the past decade, Baidu's market share in China has ranged between 50 and 80%. Currently, it stands around 70%. Baidu's biggest rival before 2013 was Google. From 2010 to 2012, Google had about 40 percent of the Chinese internet search market. During this period, Google transferred service out of mainland China due to a major hack of the company's servers and a dispute over censorship with the Chinese government. Accordingly, they redirected search queries from Google China to Google Hong Kong. However, in 2014, Google China became unavailable to mainland China users. This is clearly seen in FIGURE 2. There is a slow decrease of Google's Chinese market share in China before 2014, and almost no market share afterward. Although Google is no longer available, people from mainland China are still able to access Google by using Virtual Private Networks (VPNs). After Google's exit of the Chinese market, several other search engines began to claim a non-negligible market share but they have only been popular for a short period of time.

TABLE 1 updates and extends Vaughan & Chen's (2015) comparison of the services provided by Google Trends and Baidu Index.[3] A distinct difference between Baidu Index and Google Trends is that Google reports relative volume for a sample of Google searches. Baidu Index reports absolute volume for its whole population of searches. According to Google Trends[4], they first take a sample of the absolute search volumes. They then normalize the sample by dividing the number of searches by the total search volume for the location and

[4] Google Trends explained how the data is normalized here: https://support.google.com/trends/answer/4365533?hl=zh-Hans&ref_topic=6248052

time under consideration. The results are scaled to a range of 0 to 100, with 0 being the lowest and 100 being the highest relative search intensity value. The fact that Baidu reports absolute search volumes is important as this makes it possible to add the search volumes of various keywords, something that is not possible with Google Trends. For simplicity, in the following paragraph, we will refer to both data from Google Trends and Baidu Index as search volume data, although only Baidu Index presents absolute search volumes.

Both Google Trends and Baidu Index allow one to limit search data to (i) a specific region within a country, and (ii) a specific time period (January 2004 onwards for Google Trends; June 2006 onwards for Baidu). Further, both allow a direct comparison between up to 5 different keywords. Google only provides information on the average search volume index when there's a comparison between keywords. In contrast, Baidu Index provides both average and daily moving averages, as well as year-on-year and month-to-month growth rates of search term volumes. Both provide an extensive analysis of related searches that were conducted by people who searched for a specific keyword. It is worth noting that Baidu Index excluded searches conducted on mobile phones until December 2010.

Despite having many similarities between the services provided by both of these search engines, there are some important differences between the two. Google Trends allows its users to limit search volume to specific categories. For example, one has the option when collecting search data on Apple to limit the collection the search volumes to reflect queries for Apple the technology company, as opposed to Apple the fruit, this is very helpful when a specific keyword can represent many different objects. Meanwhile with Baidu, searches of "Apple" will produce data for both the fruit and the company.[5]

4 Data and Methodology

4.1 Data

The China Statistical Bureau publishes the series 'Total Retail Sales of Consumer Goods' each month. Total retail sales of consumer goods is the total amount of consumer goods sold directly to urban and rural residents and social groups in various sectors of the national

[5] Vaughan & Chen (2015) explain the detailed matching mechanism difference between Google Trends and Baidu Index.

economy. To facilitate data collection, the Chinese Statistics Bureau divides retail sales of consumer goods into two categories: sales from big enterprises and sales from small businesses. Sales from big enterprises are further broken down into categories.

In this study, we first focus on retail sales in the automobile and communication appliances sectors, and then generalizes our findings to the total retail sales of consumption goods in China. We choose the two sectors for the following reasons. First, they together account for a relatively large share of total retail sales in China; about 11% in 2019. Second, sales in these two sectors are dominated by big enterprises. The resulting sales figures are thus likely to reflect total industry sales. Third, sales are concentrated among a relatively small number of big brands. That means that a limited number of keywords can represent the entire sector. Fourth, an influential early study using internet search volume, Choi & Varian (2012), also studied automobile sales.

FIGURES 3 and 4 report time series data on sales in the automobile and communication appliance sectors, respectively. Both sectors show strong growth in the past decade, though recent growth has slowed somewhat. In addition, there is clearly seasonality in both sectors, with sales highest in November and December and relatively low in January and February. To control for this behavior, our analysis includes both monthly dummies and quadratic time trends.

Another feature that we want to control for is the lag between actual retail sales, and when the data are published and available. The sales data are published monthly. Data for the previous month are published in the middle of the current month. This means that if we want to nowcast sales data at the end of a given month, say August, we can only use sales data from July. Baidu Index and Google Trends make it possible to use data from August. This makes it possible to produce better predictions.

4.2 Baseline models

Our analysis produces both nowcasts and 1-month ahead forecasts of automobile and communication appliance sales in China. The difference between nowcasting and forecasting is that nowcasting aims to predict the value for August at the end of August. Forecasting aims to predict the value for August at the beginning of August, when the most recently available data are for June.

More formally, the baseline model for Nowcast and 1-month ahead Forecast are:

Nowcast:

$$C_t \ = \ \alpha C_{t-1} + \beta_1 Date + \beta_2 Date^2 + \beta_{3\ to\ 12} Month\ Dummies \qquad (1)$$

Forecast:

$$C_t \ = \ \alpha C_{t-2} + \beta_1 Date + \beta_2 Date^2 + \beta_{3-12} Month\ Dummies \qquad (2)$$

where C_t is the natural logarithm of total sales for automobile or communication appliances, in real terms[6] at time t; and $Date$ is the time trend. We also include ten monthly dummies to account for seasonality. Though automobile and communication appliance consumption are reported monthly, the values for January and February are in most years combined into one value. Accordingly, we divide the values for January and February by 2 and use this value as a separate month, treating each year as having 11 months.

We use the above baseline model to run expanding window nowcasts and forecasts. In our expanding window predictions, an additional observation is included to train the model as we move from one time period to the next. For example, when we make a nowcast for the time period *t*, we use data before *t* to train the model, but when we make a prediction for *t+1*, data for *t* is added into the training period to train the model.

We calculate RMSFE (Root Mean Square Forecasting Error) each time after we run the expanding window nowcasts and forecasts to measure the performance of the models. RMSFE is the standard deviation of the prediction errors:

$$RMSFE_i = \sqrt{\frac{1}{p}\sum_t (f-o)^2} \qquad (3)$$

Where i referes to the model used, p is the length of the evaluation period, f is the prediction, and o is the observed value. [7]

[6] The sales for automobile and communication appliances are deflated using the prices index available at the National Bureau of Statistics. The link is: https://data.stats.gov.cn/english/easyquery.htm?cn=A01. Logs of sales are also used in Choi and Varian (2012).

[7] We test the accuracy of the forecasts by calculating RMSFE, which is in line with other papers that predict consumption using internet search data. E.g. Vosen & Schmidt (2011), Carrière-Swallow & Labbé (2013), Woo & Owen (2018).

To investigate the added value of including information from Baidu, we will augment these baseline models with search query data. We next discuss how we collected the search query data.

4.3 Collection of search query data

The keywords for which we obtain the search intensity series from Baidu Index are chosen based on brands and models of automobile and communication appliances, as well as the combination of these keywords. For example, in the keywords for communication appliances, we included "Huawei" "Mate10" as well as "Huawei Mate10" as search terms (mainly Chinese languages are used as keywords). Many of these keywords don't form a valid search term when combined together either because the combination is not being searched for or because Baidu (or Google) didn't record any data for this search term[8]. In addition, some keywords that are associated with buying a new car or a new phone are also included in the search terms, for example: "Car insurance" or "Phone cases"

While collecting keywords, we used brands and models of automobiles and phones from www.autohome.com.cn and www.zol.com.cn. These 2 websites are widely used in China and they contain detailed information on brands and models of automobiles and phones that are being sold in China. Other keywords like "car insurance" or "phone cases" are mostly chosen by suggested searches and related searches. The full list of keywords is available on Dataverse[9]. Overall, we included 470 search terms for automobile sales and 727 search terms for communication appliances from Baidu Index.

One problem with some search terms for communication appliances is that they are used only during a short time period, with almost no searches done outside this peak period. This is not surprising as phone models are often popular only for a short time. FIGURE 5 shows the search volume for different models of iPhones between 2011 and 2019. Most of the search volumes for each keyword are highly concentrated around a certain time, and the search volumes before or after this specified time are small. If we run the model with each of these search terms separately this won't be very useful because each of these search terms only

[8] For Google, the warning is mostly that there's not enough data, for Baidu, the warning is "Keyword "XXX" is not included or recorded by Baidu Index", followed by an option to purchase a keyword. Baidu will then start to record it.

[9] See the following link for the full list of keywords, https://doi.org/10.7910/DVN/YT25IP, Harvard Dataverse.

provides information for prediction for a short amount of time. Even if they do have a correlation with sales and belong to the model, this correlation would've been washed away by the small search volumes around other time periods. To solve this problem, we added all the search term for each series of a product together to create a variable that has a long-lasting effect, (for example, Huawei produce several series of phones and they present a new model under this series each year, like the Nova series and the Mate series, we added up all the searches for each model under a series separately as a single variable, in the end, we have 1 variable for the Nova series and 1 variable for the Mate series. As a result, we aggregated the keywords into 86 Baidu Index variables for communication appliances.

We will estimate both the baseline models and the models augmented with Baidu search term series, using both OLS and Lasso methodologies, searching over various specifications to find the model that gives the most accurate nowcasts and forecasts.

4.4 OLS estimations

Equations 4 and 5 show the augmented nowcasting model and forecasting models which incorporate Baidu Index series.

Nowcast:

$$C_t = \alpha C_{t-1} + \beta_1 Date + \beta_2 Date^2 + \beta_{3\ to\ 12} Month\ Dummies + \beta_{13} Baidu_t \tag{4}$$

Forecast:

$$C_t = \alpha C_{t-2} + \beta_1 Date + \beta_2 Date^2 + \beta_{3-12} Month\ Dummies + \beta_{13} Baidu_{t-1} \tag{5}$$

In the models above, the lags of C_t, $Date$, $Date^2$, and $Month\ Dummies$ are the same as the baseline model in Equations 1 and 2, while $Baidu_t$ stands for the different specifications of Baidu Index we put into the models.

Note that in the nowcasting model, we are able to use Baidu data for that month, while for the forecasting model, only Baidu data from the previous month is available. In the nowcasting model, we're nowcasting the total sales data at the end of a certain month. By then, Baidu data for that specific month is already available. However, in the forecasting model, we're forecasting sales data for next month, which means only Baidu data for the previous month is available.

Because of the large quantity of Baidu search terms, OLS models do not have enough degrees of freedom to estimate equations 3 and 4 when all of the search terms are included separately. For example, for the automobile sector, in our total sample, we have 470 Baidu Index keywords but only around 100 observations to train the model, so when all 470 search terms are included separately, we would indeed have more explanatory variables than observations.

We explore several ways to reduce the number of observations.

First, because Baidu records all search volumes in exact numbers, we can simply add up all the search terms series, though this means that we are not using all of the information that Baidu Index provides.

Second, we use principal component analysis to calculate factor loadings of Baidu Index and use the first 8 principal components for both automobile search terms and communication search terms, which accounts for around 70 percent of the variation in the respective Baidu Index search terms.

Third, we follow a procedure similar to Ginsberg et al. (2009) and run a regression with each Baidu series separately and find the individual series that adds most to the baseline model during the training period. That is, we run the following OLS model:

$$\begin{aligned} C_t \;=\; & \alpha C_{t-1} + \beta_1 Date + \beta_2 Date^2 + \beta_{3\ to\ 12} Month\ Dummies \\ & + \beta_{13} Single\ Baidu\ Keyword \qquad (6) \end{aligned}$$

where each of the individual keywords is included in the model along with the baseline variables. We then select the series of the keyword that gives the highest adjusted R square in the training sample, and add this series to the baseline model to nowcast and forecast. We iterate this procedure using the expanding window, re-selecting at each stage the keyword with the highest adjusted R square for each time period. In addition, we use this method to choose the 3 keywords that, individually, gives the highest adjusted R square, and then evaluate the RMSFE of a model that includes these 3 terms together.

Besides using models that only use contemporaneous values of the Baidu index series, we further experiment with models that in addition add up to 3 lags for the nowcasting models and up to 4 lags for the forecasting models of these series.

4.5 Lasso estimations

Lasso models do not suffer from the “large number of explanatory variable problem” we described above for OLS. Lasso is an acronym for “least absolute shrinkage and selection operator” and it’s a method popular in prediction and model selection. Lasso is useful when predicting the value of the outcome variables when the number of regressors is large relative to the number of observations in the dataset (Tibshirani (1996)).

For example, a regression model with multiple regressors may take the form of:

$$Y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon \tag{7}$$

To find a solution for the above model while keeping it less complicated, Lasso set a penalty for the sum of the absolute values of the coefficients, specifically it minimizes:

$$\frac{1}{2N}\left(y - x\beta'\right)'\left(y - x\beta'\right) + \lambda \sum_{j=1}^{p} \left|\beta_j\right| \tag{8}$$

The first term of equation 8 represents the same value that OLS minimizes, which is the in-sample prediction error. The second term of equation 6 is a penalty that is controlled by the operator λ which increases as more variables are included in the model and the model gets more complex. λ also causes Lasso to omit variables because although none of the coefficients are likely to be exactly 0, the penalty operator λ drives the small coefficients to 0 as equation 6 gets minimized when the model is being estimated. The complexity of the model is set by λ, specifically, the larger the λ, the less complex the model will be, while when $\lambda = 0$, the model is the most complex.

We use Stata’s built-in “Lasso” command, which allows various λ selection methods (CV selection, Adaptive Lasso and plugin formula), grid settings for CV selection and adaptive Lasso, fold of the selection, etc. We tried multiple settings and model specifications and compared the performance of multiple Lasso models. In this study, we will only present the model using CV selection to decide the penalty operator λ because CV selection selects the λ that gives the lowest root mean square forecasting errors.

To have a fair comparison of the OLS and Lasso methods, we first run the same specifications as we ran for OLS. We first estimate the model that includes the sum of the

Baidu Index series, then estimate a model that includes the first 3 principal components, and finally, run the model which includes all Baidu series.[10]

As Lasso provides a way for model selection when a big amount of regressors is included in the model, we also experiment with models that add both lagged terms of the Baidu series and interaction terms into the model. The interactions we include interact with the baseline variable and Baidu Index search terms.[11]

4.6 Predictions using Google Trends.

Although we are primarily interested in what happens when adding Baidu information to the baseline models, we also evaluate the predictive performance of adding search intensities from Google Trends.

In the literature that tries to predict consumption related aggregates with internet search volume data, Google Trends is definitely the most used data source. Although Google is no longer widely used in China since Google quit the Chinese market in 2014, for completeness we also check whether Google Trends can serve as a good forecasting tool in China's automobile and communication appliances sales.

We collected keywords for Google Trends in a similar fashion as we did with Baidu Index, though ended up with less Google Trends series, because more keywords are recorded and valid in Baidu Index, partly due to the fact that many of our keywords are in Chinese.

At the same time, we included extra series based on Google Trends' 'categories'. For example, if you select the category "automobile and cars", Google Trends will aggregate the data for all searches that fit this category. In addition to the keywords, we added all the categories and sub-categories under "automobile and cars", "internet and telecommunications" as well as other categories that could be associated with automobile and communication appliances consumption like "Shopping", "Travel", "Games" etc.

[10] As a robustness check, we also run the Lasso models with the top 1 and top 3 variables to see how the results of these models compare to the other models. The results of these models don't change the results presented so far.

[11] Because up to 3 lags of Baidu Terms will be included in the model, to keep our data comparable, in all of the models in the empirical analysis we exclude the first three time periods.

Overall, we included 190 variables for automobile sales and 327 variables for communication appliances from Google Trends. As we mentioned earlier, search queries associated with communication appliances are highly concentrated around specific periods. While for Baidu, we therefore aggregated some series by simply summing, this is not possible for Google Trends series as Google Trends reports relative search volumes rather than absolute volume. We analyze the predictive performance of the Google Trend series using the same nowcast and forecast models (equations 4 and 5) we used to analyze the predictive performance of the Baidu Index.

There is one exception, however: since we cannot add the Google Trends series, we cannot run the regression with the sum of all the Google Trends data as we did for the Baidu Index.

4.7 Predictions using the Consumer Confidence Index.

As explained above in the literature review, the existing literature often compares internet search data with survey-based indicators like consumer confidence index. (Carroll, Fuhrer & Wilcox (1994), Bram & Ludvigson (1998), Howrey (2001), etc.)

In our final analysis, we analyze the predictive performance of adding the consumer confidence index to the baseline model, to compare the predictive power of consumer confidence with Baidu Index and Google Trends. Note that unlike search intensity data, data for the CCI are available with a delay and hence enter as a lagged variable in the models:

Nowcast:

$$C_t = \alpha C_{t-1} + \beta_1 Date + \beta_2 Date^2 + \beta_{3\ to\ 12} Month\ Dummies + \beta_{13} CCI_{\mathrm{t-1}} \tag{9}$$

Forecast:

$$C_t = \alpha C_{t-2} + \beta_1 Date + \beta_2 Date^2 + \beta_{3-12} Month\ Dummies + \beta_{13} CCI_{\mathrm{t-2}} \tag{10}$$

This is illustrated in the above equations: in the nowcasting model, we can only use the CCI of the previous month, while in the forecasting model we need to lag CCI twice.

5 Empirical results: Sectoral retail sales

5.1 Nowcasting results

Table 2 shows the RMSFE of nowcasts of retail sales using different regression methodologies and specifications.

We start our analysis using the “long sample” to run the models, using data from 2011 to 2019. Each model is thus trained initially using 7 years of data (January 2011 to December 2017), further adding one more observation into the training period each time a prediction is made. The data between January 2018 to December 2019 is used as the testing sample, to cross validate the accuracy of the forecasts.

The top panel shows the results for the OLS models. The bottom panel shows the results for the Lasso models. The nowcast results of automobile sales are shown in the left panel of the table, while the results for communication appliances are shown in the right panel. Besides the absolute RMSFE, for the models that include information from Baidu, I also show the reduction in RMSFE relative to the RMSFE of the baseline model (OLS or Lasso). Positive numbers show the percentage improvement in predictive performance, negative numbers mean that adding Baidu information decreased predictive performance.

If we look at the OLS model that adds, to the baseline model, the sum of the Baidu Indices of the search terms. The left panel of Table 2 shows that including the sum of Baidu Index into the baseline model can improve nowcasting performance for automobile sales: including the contemporaneous Baidu sum improves forecasting accuracy by 3.45%. However, the same panel shows that adding Baidu information is not a guarantee to get a more accurate forecast: if lags of the Baidu sum are added, in addition to the contemporaneous values, the RMSFEs become worse than the RMSFE of the baseline model.

In the case of communication appliances (the right panel of Table 2), I find that including the Baidu sum always improves the nowcasts, and that including 3 lags of the sum of Baidu Index into the model improves the accuracy the most, reducing the RMSFE by 7.54% compared to the baseline model.

Using the sum of the Baidu series is unlikely to exploit all available information, so next, I try alternative ways of adding the information from Baidu. Including the first 3 principal components.

When using, instead of the sum of the search terms, the PCA factors of Baidu Index, I find that nowcasting accuracy of automobile sales can be further improved, when using up to 1 lag of PCA factors of Baidu Index, this improves predictive accuracy by 11.25%. Similarly, for the automobile sales, I also found that adding Baidu PCA factors contribute to predictive performance. However, the decrease in RMSFE for the PCA models are smaller for the communication appliances sector.

Next, rather than aggregating the Baidu series, I analyze what happens if I add, to the baseline model, the individual keyword series that gives the highest adjusted R square during the training period. Table 2 shows that when only keyword with the highest adjusted R square is included for each training period, I do not see a reduction in RMSFE for the automobile sector. However, if I add the 3 series, that individually gives the highest adjusted R square in the training period, jointly into one model, I get a reduction in the nowcast errors of 10.53% as compared to the baseline model.

For sales of communication appliances, including the Baidu Index keyword series with the highest adjusted R square reduces nowcasting errors by about 12.86%, while including the 3 best individual series jointly, nowcasting errors are reduced by about 13.84%.[12] Note that both these improvements are smaller improvements than the improvement I obtained when using the PCA method.

Next, I turn to the Lasso models at the bottom half of Table 2. In theory, Lasso models should be able to do better, because, unlike OLS, Lasso models do not force us to select ex-ante which individual series to include. Instead, Lasso models use the data to select the best models.

Similar to the results from the OLS models, I again find that adding Baidu information to the baseline model can improve forecasting accuracy. If the sum of the Baidu indices is added, I improve forecasting accuracy by about 9% for the automobile sales and about 7% for the communication appliances sales.

[12] Note that the RMSFE of the nowcasts is sometimes the same when an extra lag of Baidu Index is included. This happens when the extra lag of the Baidu Index series does not improve forecasting accuracy over the best model with one less lag.

The best Lasso model for automobile sales is the model that adds PCA factors of the Baidu series. Specifically, the Lasso model for automobile sales that adds 1 lag of Baidu series PCA to the baseline model performed best, leading to a reduction of the RMSFE by 20.36%.

Adding the Baidu series jointly into the model also improves the model RMSFE by about 10% to 20%. Adding interactions between the Baidu series and the base model variables reduces the RMSFE further to about a 20% improvement compared to the base model.

Note further that adding all Baidu series jointly to the base model does not work for communication appliances as it has worse predictive accuracy than the base model. For communication appliances, the best Lasso model is the model that includes the sum of the Baidu series, rather than all series individually. Hence, comprehensive models are not always the better models.

In fact, Table 2 shows that, for both communication appliances and automobile sales, the overall best model is not the more complex Lasso model. The model with the lowest RMSFE is in both cases, an OLS model. They are the PCA factor model for automobile sales (a RMSFE of 0.0495, compared to the best baseline (the OLS baseline model) of 0.0558, an 11.25% improvement), and the top 3 OLS model for communication appliances (a RMSFE of 0.0963, compared to 0.1118 for the best baseline model (the OLS baseline model), a 13.84% improvement).[13]

One possible reason for the relatively poor performance of the Lasso model is that the Lasso model can have difficulties handling highly correlated variables (Hastie, Tibshirani & Wainwright (2015)). Theoretically, when one has a large enough sample size, highly correlated explanatory variables will not cause problems. However, in our sample, there are at most 10 years of monthly data, so the sample size is relatively small.[14]

Summarizing our findings so far, the evidence suggests that nowcasting of Chinese consumption series can be improved by including search intensity information from Baidu,

[13] While, for automobile sales, the Lasso model with all individual Baidu series shows the highest improvement over the OLS baseline model, the Lasso baseline model has a higher RMSFE than the baseline OLS model, allowing the best OLS model to be the overall best model even for automobile sales.

[14] Hastie, Tibshirani & Wainwright (2015) suggest using elastic nets rather than Lasso when variables are highly correlated. We also experimented with elastic net models but none of the elastic net models with Baidu information outperformed the Lasso baseline model. The results of the elastic net predictions are not listed here but are available upon request.

but also that there is no guarantee that adding such information will always improve predictive performance. In fact, we find that the best models are relatively simple models with some Baidu information rather than models with lots of Baidu series.

5.2 Forecasting results

Table 3 shows the RMSFE and the reduction in RMSFE of the various forecasting models. Similar to Table 2, the top panel shows the forecasting results for OLS models while the bottom panel shows the results for Lasso models.

Overall, we observe the following pattern: OLS models with the sum and PCA factors of Baidu Index don't help much when forecasting sales in the automobile sector, and help somewhat when forecasting sales in communication appliances. But when we include the top 1 and top 3 most useful Baidu variables in the baseline model, there is a bigger reduction in forecasting errors. For the automobile series, the best OLS model incorporates the top 1 Baidu variables, reducing the RMSFE by 14.70% compared to the baseline model. For the communication appliances sales, the model with the best individual Baidu series, reduces RMSFE by 12.34% compared to the baseline model without Baidu information.

As for the Lasso models, Lasso models do better than OLS models for automobile sales but do worse for communication appliances sales. The best model incorporates 3 additional lags of the Baidu Index series, reducing the RMSFE by 19.42% compared to the baseline Lasso model for the automobile sales.

The results presented so far are based on data starting in 2011. However, the market share of Baidu was substantially lower in the early years of the sample because of the competition of Google, so Baidu search volumes tend to be low compared to more recent years. This structural change can affect the forecasting ability of the forecasting models. To check this, we will run the following analysis using the "short sample", focus on the period since 2015, after Google quit China.

5.3 Limiting our sample period to 2015–2019

Tables 4 and 5 show the RMSFE for expanding window nowcasts and forecasts of OLS and Lasso models when using data from 2015 to 2019. In this analysis, the model is thus trained initially using 3 years of data (January 2015 to December 2017), further adding one more observation into the training period each time a prediction is made (expanding window).

Table 4 shows that Baidu Index series contain extra information that can help nowcast sales in both the automobile sector and, to a lesser extent, the communication appliances sector. While including Baidu information does not always improve forecasting accuracy over the baseline model, the models with the lowest RMSFE indeed again include Baidu series.

When using the shorter time period, the best model for automobile sales is the Lasso model that includes 1 lag of the individual Baidu series PCA factors. This model has a RMSFE of 0.0454, an improvement of about 44% over the baseline Lasso model and an improvement of about 24% over the OLS baseline model. The best model for communication appliances is the OLS model that includes 3 lags of the sum of Baidu index but in this case, adding Baidu information only improves forecast accuracy by about 1.5% compared to the OLS baseline model.

Table 5 shows the prediction results for the shorter sample for both sectors and presents evidence that including the Baidu series in predictive models for Chinese consumption series can improve predictive accuracy. For the automobile sector, the model with the lowest RMSFE is the OLS model incorporating the top 3 Baidu series, reducing forecasting errors by 24.39% relative to the best baseline model. For the communication appliances, the best OLS and the best Lasso models give similar improvements in accuracy over the best baseline model, improving forecasts about 3%.

5.4 Empirical results: Google Trends

So far, we have focused on whether Baidu search data can help to improve predictions. In this section, we use search intensities from Google Trends for the period 2015 to 2019. While Google was no longer available in mainland China after 2015, it could still be accessed using VPNs, so some data are available.

Tables 6 and 7 show the nowcasting and forecasting results for the baseline and augmented models. The results of the baseline models are exactly the same as in Tables 4 and 5. The RSMFSs and reductions in prediction errors for the OLS and Lasso models are listed, where once again, up to 3 lags for the nowcasting models and 4 lags for the forecasting models are included. The left and right panels correspond to the prediction results for the automobile and communication appliances sectors, respectively. Using PCA factors of Google Trends in Lasso model is able to reduce RMSFE by a little. Other than that, there is hardly any reduction in RMSFE from the Google augmented models. Adding Google Trends information thus does not improve predictive accuracy by much, especially when compared to the Baidu Index augmented models.

5.5 Empirical results: Consumer Confidence Index (long sample)

Survey-based indicators like the Consumer Confidence Index (CCI) and the Consumer Sediment Index are often linked with forecasting sales and consumption before other datasets like Baidu Index and Google Scholar are incorporated. Therefore, in the following section CCI in China is used to predict sectoral retail sales.

Tables 8 and 9 show prediction accuracy and decrease in forecasting errors of CCI augmented models relative to the baseline models, using data between 2011-2019. The results provided limited evidence that CCI improves the nowcasting accuracy of automobile and communication appliance sales. CCI also doesn’t seem to improve forecasts of communication appliances sales, as the OLS baseline model has the lowest RMSFE for all these models. When forecasting automobile sales, results are somewhat improved by including CCI information. The best model, the Lasso model that includes the CCI, improves accuracy by about 12.5% over the baseline Lasso model, and about 6% over the baseline OLS model. If we compare the added value of CCI to the added value of the Baidu series, however, the added value of the CCI is smaller, as the Baidu series is able to reduce a bigger percentage of the forecasting errors in several model specifications.

5.6 Empirical results: CCI and Baidu Index (long sample)

In our previous results, we showed that models with Baidu Index or CCI information can sometimes produce more accurate results than the baseline models. In the following section, we explore if by adding both Baidu Index and CCI information into the models, prediction accuracy can be improved, and how do these improvements compare to our previous models. Prediction results using the long sample are presented in Table 10 and Table 11.

To see if there's indeed any added value when both Baidu Index and CCI are included, firstly we compare the results from Table 2 (nowcasting with Baidu but without CCI) and Table 10 (nowcasting with Baidu and CCI), it's evident that in most cases results in Table 2 are better than that in Table 10, this indicates that in many cases models incorporating only Baidu Index does a better job compared to the models using both CCI and Baidu Index. In terms of the forecasting results, however, if we compare the forecasting results of Table 3 (Baidu, no CCI) with Table 11 (Baidu and CCI), for the automobile sector the best performing forecasting model with the smallest RMSFE is the model that includes 3 additional lags of both CCI and Baidu Index, which improves accuracy by 22.85% over the baseline Lasso model, and about 17.2% over the baseline OLS model. This is not surprising because both the results in Table 3 and Table 9 suggest that when 3 additional lags of CCI or 3 additional lags of Baidu Index increases forecasting accuracy by a lot. However, the same cannot be said for the communication appliances sector, as none of the models performed better than the baseline model when both CCI and Baidu Index are included.

5.7 Comparison

Table 12 shows the models that yield the smallest RMSFE in their specification, and the respective reduction in RMSFEs compared to the baseline models. The results suggest that in almost all specifications, models which incorporate the Baidu Index produce smaller prediction errors compared to the baseline models. In addition, the Baidu Index is able to reduce nowcasting and forecasting errors by around 11 to 44 percentage points for the automobile sector, and around 10 percentage points for the communication appliances sector.

In addition, our results show that in most cases, CCI is not very useful in predictions in both sectors. However, this is subject to the specification of the model used. In a few scenarios, there is a bigger improvement in prediction accuracy when both Baidu Index and CCI are incorporated.

6 Empirical results: Total Retail Sales

6.1 Collection of data and methodology

In the previous analysis, we focused on predicting only two sectors of the retail sales of consumer goods in China. The reasons why only these two sectors are chosen are detailed in section 4.

Although the previous results already indicated that keyword series related to brands, models, and related searches are able to reduce prediction errors in retail sales of the two sectors, it is interesting to explore whether our findings can be generalized to the total retail sales of all consumer goods. Government agencies and policy makers are indeed often more interested in such aggregated measures, tracking the entire consumer demand, which in turn provides an overall indicator for economic health and domestic consumption.

The difficulty associated with using Baidu Index to predict total retail sales in China lies in the ambiguity of the related keywords. For our previous models, we were able to assemble a list of keywords by using the brands and models that are associated with both sectors. However, it is less clear how to come up with a comprehensive list of keywords that could correlate with total retail sales.

This, however, doesn’t seem to be an issue when using Google Trends, because unlike Baidu Index, Google Trends provides the unique function of limiting keywords into a specific category. Using the example in section 3 of this paper, one has the option when collecting search data on Apple to limit the collection the search volumes to reflect queries for Apple, the technology company, as opposed to Apple the fruit. This function is not only very helpful when a specific keyword can represent many different objects, but it also comes in handy as a measure of the popularity of aggregate searches conducted under this category. To be precise, when one selects a category without imputing a specific keyword, Google Trends will show an aggregate measure for the search volume of all the related keywords under this category for the chosen time span. Google Trends has 1132 categories and sub-categories as of Feb, 2021, which provides a comprehensive categorical system that measures all the searchers conducted by its users. This category data has been used by past papers to predict aggregate consumption using Google Trends. (Vosen & Schmidt, 2011; Woo & Owen, 2019)

This unique function in Google Trends, although unavailable in Baidu Index, provides us with the opportunity to construct a list of keywords based on the titles of these comprehensive categories. Specifically, we used the names of the categories and sub-categories as keywords to collect search volume series from Baidu Index, and used these series to nowcast and forecast total retail sales of consumer goods in China. We started with a list of 1132 keywords, corresponding to the 1132 categories and sub-categories from Google Trends, and started gathering data by imputing these keywords in Baidu Index, many of these keywords don't form a valid search on Baidu, when this happens, Baidu usually issues the following warning: "Keyword "XXX" is not included or recorded by Baidu Index", followed by an option to purchase a keyword. Baidu will then start to record it after the keyword is purchased. As a result, only 982 of the keywords yield useable search volume series. In the following section, we attempt to nowcast and forecast total retail sales using these 982 keywords series from Baidu Index[15].

Figure 6 reports time series data on total retail sales in China. Once again, we included monthly dummies and quadratic time trends to account for the seasonality and time trends in the data.

The model specifications are similar to the ones previously used in this paper. Specifically, the following methods are used:

First, by using the sum of the Baidu Index series.

Secondly, by adopting principal component analysis to transform all the Baidu Index series that has a correlation coefficient above 0.9, into factor loadings.

Thirdly, by running a regression with each Baidu series separately and find the series that individually adds most to the baseline model during the training period.

The models are calibrated using both OLS and Lasso, allowing up to 3 lags for nowcasting and 4 lags for forecasting to explore any information embedded in the lagged series of Baidu Index. The specification of the models is consistent with the previous methodologies.

[15] See the following link for the full list of Google categories, https://doi.org/10.7910/DVN/YT25IP, Harvard Dataverse.

6.2 Nowcasting results:

Similar to our previous tables, Table 13 shows the RMSFE of nowcasts of total retail sales using different regression methodologies and specifications, with the top panel shows the results for the OLS models, and the bottom panel shows the results for the Lasso models. Both absolute RMSFE and the reduction in RMSFE relative to the RMSFE of the baseline model (OLS or Lasso) are listed.

As the table suggests, in most cases adding Baidu Index into the baseline model can improve nowcasting performance. The best OLS model is the one that adds, to the baseline model, 3 series that individually increase the adjusted R square the most. Specifically, when 3 additional lags of Baidu Index are included, this model is able to reduce 24.51% of the nowcasting errors.

In the case of Lasso models, we find that including the Baidu Index series always improves the nowcast, and that including 3 lags of the individual Baidu Index series into the model improves the accuracy the most, reducing the RMSFE by 42.28% compared to the baseline Lasso model. This model is also the model with the overall lowest RMSFE, which is 0.0203.

6.3 Forecasting results:

Table 14 shows the forecasting results for total retail sales. This table is structured similarly to our previous results.

In terms of the OLS models, we found that in all cases, the incorporation of Baidu Index series decreases forecasting errors, and this decrease seems to be more prominent as more lags are included in the models. Specifically, if the sum of Baidu Index series is added to the model, the forecasting accuracy is improved by around 3% to 10%. If principal components are added to the models, forecasting accuracies are improved by between 12% to 36%. While if 3 series that individually increased the in sample adjusted R square the most are added to the model, this reduces around 46% to 52% of the forecasting errors. The model with the smallest RMSFE is the model that adds, to the baseline model, 3 series that individually increases the most adjusted R square in the training period, at most this specification (with 2 additional lags) is able to reduce 52.38% of the forecasting errors.

Similarly, the results of the Lasso models also indicated that in all cases Baidu Index series decrease RMSFE. The Lasso model with individual factors is able to reduce 34.15% of the forecasting errors.

In summary, the evidence suggests that when Baidu Index series are included in the models, in most cases, we see a prominent improvement in prediction accuracy. To be precise, in some specifications, the inclusion of Baidu Index is able to reduce 42% of the nowcasting errors and 52% of the forecasting errors compared to the baseline models.

7 Conclusion

This paper contributes to the literature that analyzes whether 'big data', in this case, search intensity series from search engines, can improve economic forecasts. We start this analysis by presenting a literature review on how the use of internet search engine data has developed in the past decade, followed by a comparison between Google Trends and Baidu Index. We then detailed our forecasting methodology and results.

In contrast to previous research that has focused on Google Trends, this study provides a comprehensive analysis of the potential of the Baidu Index, the leading search engine in China, to improve forecasts of Chinese sectoral and aggregate consumption data.

Our results indicate that search intensity from Baidu Index contains information about the futures sales of both the automobile sector and the communication appliances sector, as well as about future overall retail sales. Incorporating search intensity data from Baidu can substantially improve the nowcasts and forecasts of the sales, by more than 10% for the auomobile and phone sales and by more than 40% for total sales. In addition, our results indicated that simple models that incorporated Baidu Index typically perform better than complicated models with a lot of Baidu Index series, and that OLS models mostly do a better job in nowcasting and forecasting than LASSO models. When comparing the added value of Google Trends and CCI to Baidu Index, we show that Baidu Index augmented models performed better than Google Trends models or CCI models.

In addition, we show that for the Chinese consumption series analyzed here, Baidu information improves predictive accuracy more than either Google Trends information or information from the consumer confidence index.

Our results also show that internet search engine data is useful in forecasting consumption aggregates in the Chiese context. As existing literature main focused on using Google Trends to improve forecasting accuracy, they mainly focused on developed economies where Google dominated the search engine market. (Ettredge, Gerdes & Karuga, 2005; Vosen & Schmidt, 2011; Choi & Varian, 2012; Woo & Owen, 2019; Yu et al., 2019) However, as the popularity of different search engines is highly dependent on the region and the time span, it is doubtful if the results of the existing literature on Google Trends can be generalized to developing economies like China, where the most popular search engine is Baidu. We contribute to the existing literature by using both Baidu and Google in the Chinese context. Our results indicate that the forecasting models' accuracy is significantly improved when Baidu Index is incorporated into the models.

These results suggest that both private companies and government organizations in China could benefit from analyzing whether their operational decisions can be improved by adding information from Baidu to their forecasting models. For example, forecasting retail sales more accurately can help private companies to optimize their resources and formulate inventories accordingly to meet the changes in demand. More accurate total retail sales of consumption goods also means policy makers can foresee the trends in future consumption and adjust policy decisions.

This paper is the first to look at using Internet search engine data like Baidu Index to prediction aggregate consumption series in China. Future work in this area may find it useful to test if Baidu Index can be used to forecast other aggregate economic series in a developing country like China. In addition, other machine learning mechanisms can also be used to test if the results can be further improved upon.

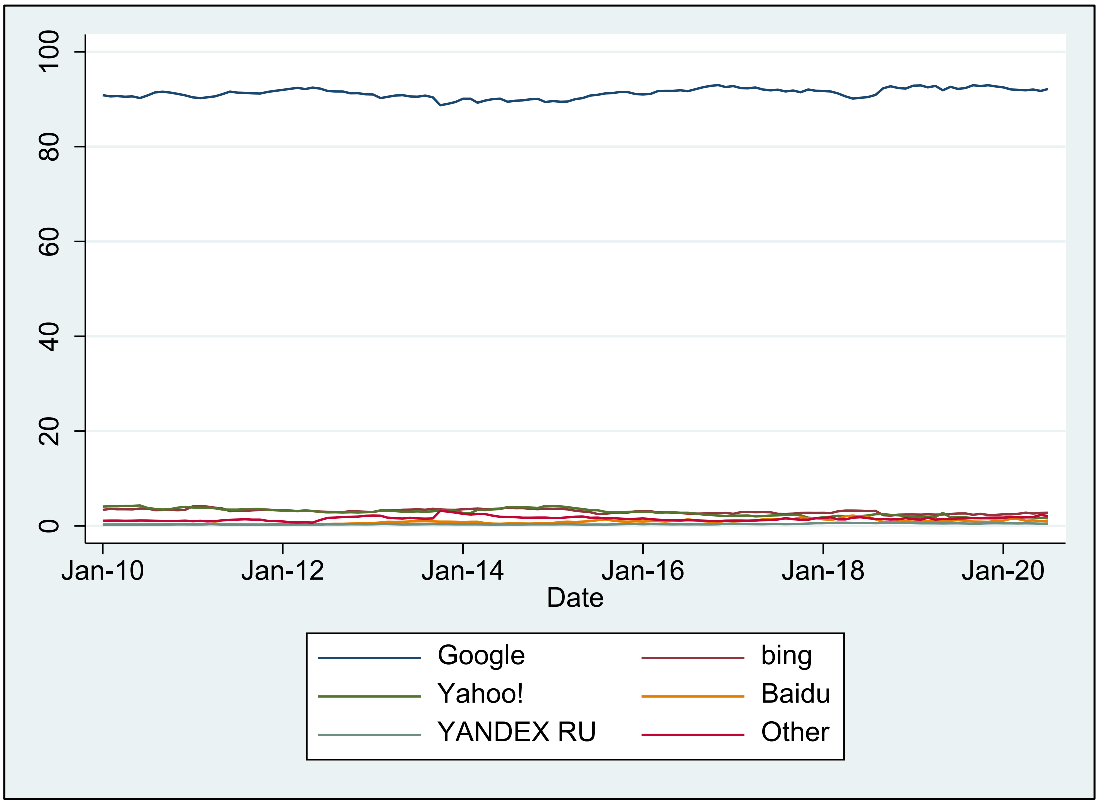
TABLE 1
Comparison between Google Trends and Baidu Index

Features	***Google Trends***	***Baidu Index***
• Limit to a specific country	Yes	No, only China
• Limit to a specific region within that country	Yes	Yes
• Limit to a specific time period	Yes, earliest Jan.2004	Yes, earliest June. 2006 for PC and Dec.2010 for phones.
• Limit to a specific category	Yes	No
• Maximum number of terms that can be compared	5	5
• Search volume reported	Relative volume	Absolute volume
• Average search volume	Reports average across the sample period	Reports both average and daily moving average across the sample period.
• Report total search volume of several terms	Yes	Yes
• Method of matching	Partial matching	Complete matching
• Related searches	Yes	Yes

Features	***Google Trends***	***Baidu Index***
• Related searches	Yes	Yes
• Demography of the people	Only shows the region where the searches are from	Region, age, gender, and information on what sectors are people interested in when they search for a certain keyword.
• Separate searches from different user platforms (PC or phones)	No	Yes
• Show the news headlines related to the search terms	No	Yes, but only when there's a spike in the search volume.
• Limit to a specific search option (News, Pictures, etc.)	Yes	No
• Measure of popularity amongst internet users and news outlets	No	Yes

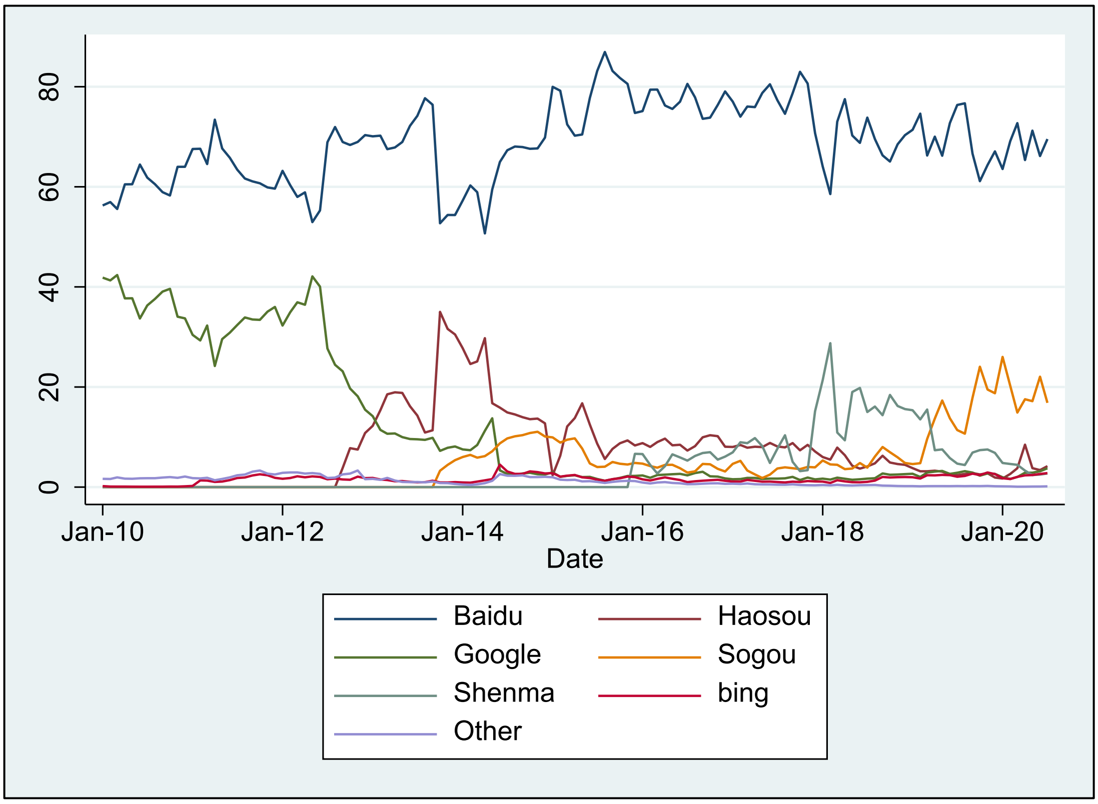
Source: updated from Vaughn and Chen (2015), Google Trends, Baidu Index.

FIGURE 1
Search engine market share worldwide



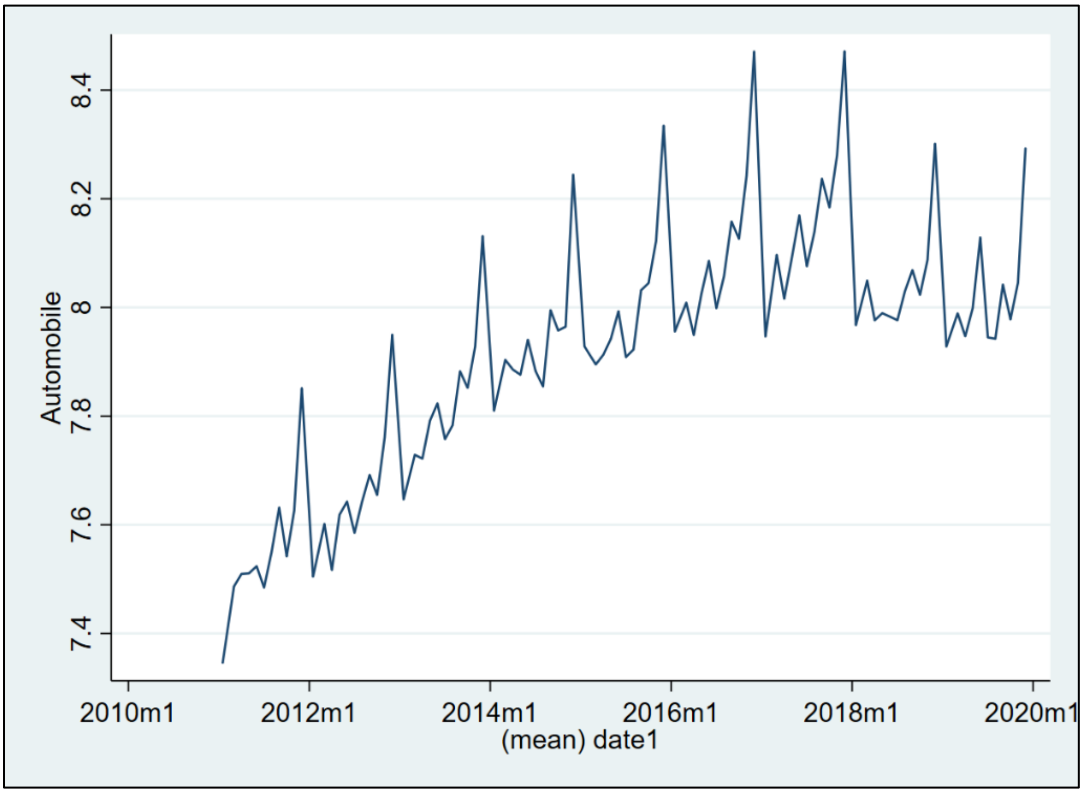
(Source: Statcounter GlobalStats 2020)

FIGURE 2
Search engine market share in China



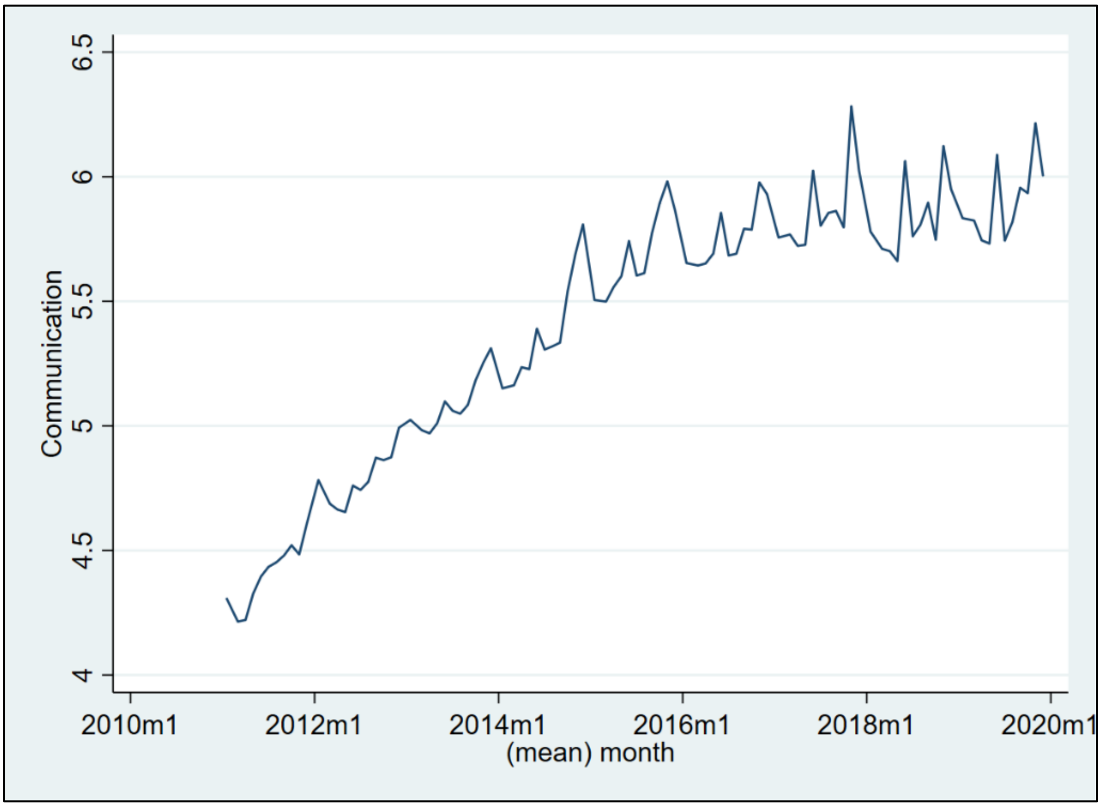
(Source: Statcounter GlobalStats 2020)

FIGURE 3
Natural logarithm of automobile sales in China



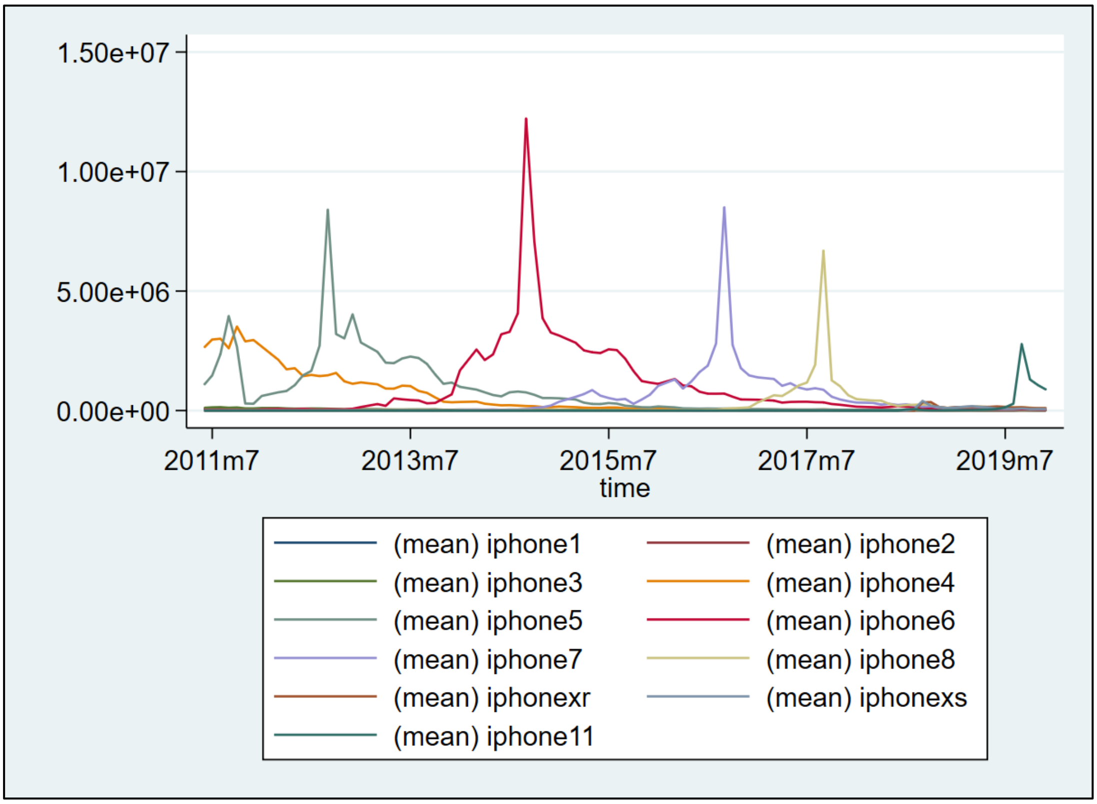
(Source: National Bureau of Statistics of China)

FIGURE 4
Natural logarithm of communication appliances sales in China



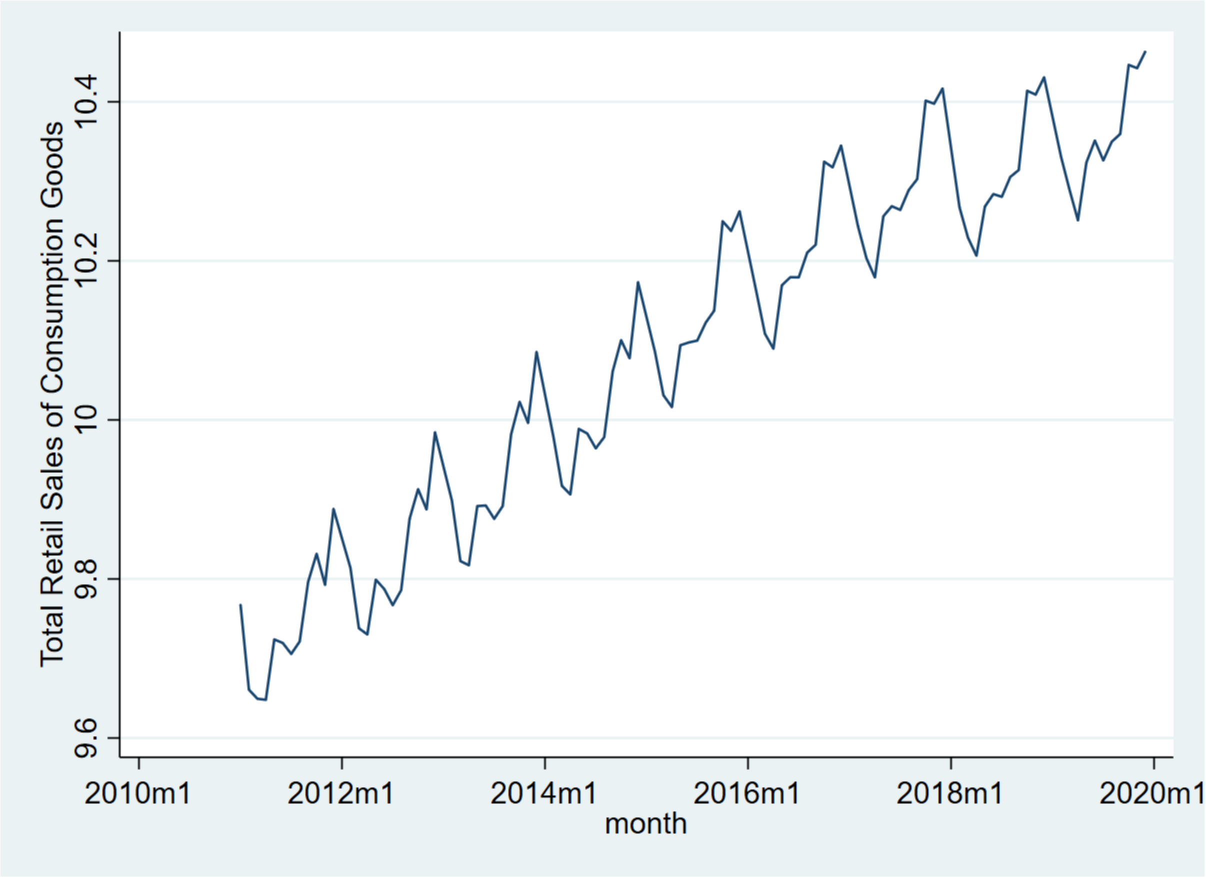
(<u>Source</u>: National Bureau of Statistics of China)

FIGURE 5
Search Volumes for iPhone Related Keywords In China



(<u>Source</u>: Baidu Index)

FIGURE 6
Natural logarithm of total retail sales in China



(<u>Source</u>: National Bureau of Statistics of China)

TABLE 2 Nowcasting with Information from Baidu (long sample)

		Automobile		***Communication***	
	Lags	***RMSFE***	***Reduction***	***RMSFE***	***Reduction***
A) OLS					
Baseline		0.0558		0.1118	
Sum	L0	0.0538	**0.0345**	0.1077	0.0364
	L1	0.0570	-0.0224	0.1055	0.0558
	L2	0.0598	-0.0719	0.1036	0.0731
	L3	0.0622	-0.1147	0.1034	**0.0754**
PCA	L0	**0.0495**	**0.1125**	0.1103	0.0132
	L1	0.0504	0.0957	0.1104	0.0128
	L2	0.0531	0.0481	0.1110	0.0070
	L3	0.0579	-0.0387	0.1100	0.0161
Best series	L0	0.0642	-0.1505	0.0974	**0.1286**
	L1	0.0634	-0.1367	0.0974	0.1286
	L2	0.0637	-0.1420	0.1065	0.0475
	L3	0.0623	-0.1176	0.1065	0.0475
Top 3 series	L0	0.0500	0.1033	0.0983	0.1205
	L1	0.0499	**0.1053**	**0.0963**	**0.1384**
	L2	0.0539	0.0340	0.1043	0.0674
	L3	0.0538	0.0352	0.1043	0.0674
B) LASSO					
Baseline		0.0649		0.1177	
Sum	L0	0.0592	**0.0885**	0.1090	**0.0736**
	L1	0.0608	0.0636	0.1188	-0.0093
	L2	0.0627	0.0351	0.1159	0.0151
	L3	0.0657	-0.0120	0.1166	0.0093
PCA	L0	0.05174	0.2032	0.1175	0.0016
	L1	0.05171	**0.2036**	0.1195	-0.0157
	L2	0.0540	0.1684	0.1185	-0.0071
	L3	0.0609	0.0617	0.1117	**0.0507**
Individual Factors	L0	0.0580	0.1067	0.1459	-0.2397
	L1	0.0568	0.1250	0.1596	-0.3564
	L2	0.0599	0.0775	0.1713	-0.4560
	L3	0.0533	**0.1794**	0.1811	-0.5391
Interactions	L0	0.0522	**0.1965**	0.1138	**0.0325**
	L1	0.0567	0.1268	0.1342	-0.1405
	L2	0.0523	0.1950	0.1216	-0.0339

	L3	0.0535	0.1760	0.1509	-0.2826

NOTE: L stands for the number of lagged Baidu series. L1 means both lags 0 and 1 are included. Forecasting models use one more lag than nowcasting models. Sum includes the sum of Baidu series as an additional variable to the baseline model. PCA adds the first 8 principal component. Best series adds the Baidu series that gives the highest adjusted R square in the training period. Top 3 series adds the 3 series that individually gives the highest adjusted R square in the training period. Individual factors adds all series separately, while interactions in addition interacts these separate series with the baseline variables.

TABLE 3 Forecasting with Information from Baidu (long sample)

		Automobile		***Communication***	
	Lags	***RMSFE***	***Reduction***	***RMSFE***	***Reduction***
A) OLS					
Baseline		0.0665		0.1102	
Sum	L0	0.0671	-0.0086	0.1015	0.0789
	L1	0.0697	-0.0472	0.0987	0.1044
	L2	0.0728	-0.0937	0.0973	**0.1176**
	L3	0.0778	-0.1688	0.0979	0.1121
PCA	L0	0.0666	-0.0008	0.1162	-0.0543
	L1	0.0661	0.0067	0.1169	-0.0606
	L2	0.0648	0.0259	0.1168	-0.0593
	L3	0.0736	-0.1059	0.1142	-0.0364
Best series	L0	0.0713	-0.0722	**0.0966**	**0.1234**
	L1	0.0665	0.0009	0.1094	0.0070
	L2	0.0625	0.0603	0.1094	0.0070
	L3	**0.0567**	**0.1470**	0.1094	0.0070
Top 3 series	L0	0.0649	0.0249	0.0983	**0.1085**
	L1	0.0634	0.0466	0.1036	0.0600
	L2	0.0571	**0.1412**	0.1036	0.0600
	L3	0.0601	0.0966	0.1036	0.0600
B) LASSO					
Baseline		0.0714		0.1110	
Sum	L0	0.0685	0.0407	0.1169	-0.0539
	L1	0.0623	**0.1276**	0.1167	-0.0514
	L2	0.0652	0.0861	0.1190	-0.0724
	L3	0.0687	0.0374	0.1195	-0.0767
PCA	L0	0.0677	0.0510	0.1232	-0.1103
	L1	0.0748	-0.0488	0.1160	-0.0453
	L2	0.0714	-0.0001	0.1229	-0.1080
	L3	0.0708	0.0078	0.1170	-0.0545
Individual Factors	L0	0.0866	-0.2134	0.1568	-0.4133
	L1	0.0753	-0.0545	0.1529	-0.3777
	L2	0.0604	0.1537	0.1781	-0.6050
	L3	0.0575	**0.1942**	0.1651	-0.4881
Interactions	L0	0.0730	-0.0224	0.1620	-0.4597
	L1	0.0783	-0.0968	0.1686	-0.5194
	L2	0.0601	**0.1577**	0.1993	-0.7958
	L3	0.0610	0.1449	0.1927	-0.7362

NOTE: L stands for the number of lagged Baidu series. L1 means both lags 0 and 1 are included. Forecasting models use one more lag than nowcasting models. Sum includes the sum of Baidu series as an additional variable to the baseline model. PCA adds the first 8 principal component. Best series adds the Baidu series that gives the highest adjusted R square in the training period. Top 3 series adds the 3 series that individually gives the highest adjusted R square in the training period. Individual factors adds all series separately, while interactions in addition interacts these separate series with the baseline variables.

TABLE 4 Nowcasting with Information from Baidu (short sample)

		Automobile		***Communication***	
	Lags	***RMSFE***	***Reduction***	***RMSFE***	***Reduction***
A) OLS					
Baseline		0.0601		0.0777	
Sum	L0	0.0556	**0.0750**	0.0815	-0.0494
	L1	0.0581	0.0329	0.0792	-0.0193
	L2	0.0585	0.0257	0.0782	-0.0074
	L3	0.0613	-0.0196	**0.0765**	**0.0142**
PCA	L0	0.0461	**0.2333**	0.0816	-0.0512
	L1	0.0464	0.2284	0.0845	-0.0876
	L2	0.0462	0.2305	0.0994	-0.2801
	L3	0.0467	0.2222	0.0994	-0.2801
Best series	L0	0.0621	-0.0328	0.1644	-1.1174
	L1	0.0620	-0.0315	0.1673	-1.1550
	L2	0.0580	**0.0345**	0.1767	-1.2758
	L3	0.0580	0.0345	0.1041	-0.3408
Top 3 series	L0	0.0631	-0.0495	0.1245	-0.6037
	L1	0.0640	-0.0657	0.1300	-0.6742
	L2	0.0611	-0.0165	0.1367	-0.7602
	L3	0.0638	-0.0610	0.0971	-0.2506
B) LASSO					
Baseline		0.0815		0.0845	
Sum	L0	0.0659	**0.1919**	0.0847	-0.0026
	L1	0.0676	0.1705	0.0815	0.0357
	L2	0.0713	0.1249	0.0786	0.0702
	L3	0.0729	0.1056	0.0769	**0.0903**
PCA	L0	0.0479	0.4124	0.0874	-0.0348
	L1	**0.0454**	**0.4430**	0.0829	0.0189
	L2	0.0472	0.4214	0.0930	-0.1008
	L3	0.0469	0.4245	0.0861	-0.0184
Individual Factors	L0	0.0659	0.1919	0.2264	-1.6794
	L1	0.0627	0.2307	0.1061	-0.2558
	L2	0.0526	**0.3548**	0.1171	-0.3861
	L3	0.0552	0.3234	0.1199	-0.4188
Interactions	L0	0.0718	0.1198	0.2165	-1.5622
	L1	0.0945	-0.1587	0.1183	-0.3998
	L2	0.0761	**0.0667**	0.1267	-0.4992
	L3	0.1068	-0.3107	0.1494	-0.7677

NOTE: L stands for the number of lagged Baidu series. L1 means both lags 0 and 1 are included. Forecasting models use one more lag than nowcasting models. Sum includes the sum of Baidu series as an additional variable to the baseline model. PCA adds the first 8 principal component. Best series adds the Baidu series that gives the highest adjusted R square in the training period. Top 3 series adds the 3 series that individually gives the highest adjusted R square in the training period. Individual factors adds all series separately, while interactions in addition interacts these separate series with the baseline variables.

TABLE 5 Forecasting with Information from Baidu (short sample)

		Automobile		*Communication*	
	Lags	*RMSFE*	*Reduction*	*RMSFE*	*Reduction*
A) OLS					
Baseline		0.0707		0.0785	
Sum	L0	0.0638	**0.0971**	0.0783	0.0027
	L1	0.0659	0.0676	**0.0759**	**0.0325**
	L2	0.0673	0.0475	0.0762	0.0286
	L3	0.0685	0.0314	0.0780	0.0058
PCA	L0	0.0692	0.0209	0.0803	-0.0232
	L1	0.0660	0.0661	0.0815	-0.0388
	L2	0.0667	0.0562	0.0848	-0.0805
	L3	0.0667	0.0562	0.0925	-0.1788
Best series	L0	0.0599	**0.1523**	0.0901	-0.1473
	L1	0.0615	0.1295	0.0867	-0.1049
	L2	0.0611	0.1366	0.0957	-0.2194
	L3	0.0616	0.1294	0.0957	-0.2194
Top 3 series	L0	0.0671	0.0507	0.0932	-0.1874
	L1	0.0583	0.1758	0.0886	-0.1291
	L2	**0.0535**	**0.2439**	0.0858	-0.0935
	L3	0.0584	0.1739	0.0858	-0.0935
B) LASSO					
Baseline		0.0789		0.0823	
Sum	L0	0.0716	0.0929	0.0805	0.0220
	L1	0.0719	0.0898	**0.0759**	**0.0775**
	L2	0.0711	**0.0997**	0.0781	0.0512
	L3	0.0717	0.0920	0.0785	0.0466
PCA	L0	0.0698	0.1154	0.0838	-0.0180
	L1	0.0676	0.1436	0.0845	-0.0269
	L2	0.0668	**0.1533**	0.0907	-0.1025
	L3	0.0734	0.0698	0.0904	-0.0991
Individual Factors	L0	0.0781	0.0113	0.1092	-0.3275
	L1	0.0803	-0.0166	0.1184	-0.4393
	L2	0.1020	-0.2916	0.1164	-0.4151
	L3	0.0905	-0.1468	0.1150	-0.3971
Interactions	L0	0.1228	-0.5551	0.1382	-0.6800
	L1	0.0876	-0.1093	0.1149	-0.3958
	L2	0.0829	-0.0499	0.1018	-0.2371
	L3	0.0858	-0.0868	0.1124	-0.3661

NOTE: L stands for the number of lagged Baidu series. L1 means both lags 0 and 1 are included. Forecasting models use one more lag than nowcasting models. Sum includes the sum of Baidu series as an additional variable to the baseline model. PCA adds the first 8 principal component. Best series adds the Baidu series that gives the highest adjusted R square in the training period. Top 3 series adds the 3 series that individually gives the highest adjusted R square in the training period. Individual factors adds all series separately, while interactions in addition interacts these separate series with the baseline variables.

TABLE 6 Nowcasting with Information from Google (short sample)

		Automobile		***Communication***	
	Lags	***RMSFE***	***Reduction***	***RMSFE***	***Reduction***
		A) OLS			
Baseline		0.0601		0.0777	
PCA	L0	0.0681	-0.1327	0.0897	-0.1554
	L1	0.0663	-0.1026	0.0845	-0.0885
	L2	0.0664	-0.1053	0.0840	-0.0821
	L3	0.0632	-0.0512	0.0787	-0.0134
Best series	L0	0.0618	-0.0281	0.0768	0.0106
	L1	0.0667	-0.1097	0.0843	-0.0851
	L2	0.0612	-0.0181	0.0868	-0.1181
	L3	0.0623	-0.0369	0.0883	-0.1369
Top 3 series	L0	0.0615	-0.0239	0.1761	-1.2679
	L1	0.0667	-0.1098	0.0877	-0.1288
	L2	0.0707	-0.1772	0.0931	-0.1987
	L3	0.0729	-0.2135	0.0881	-0.1348
		B) LASSO			
Baseline		0.0815		0.0845	
PCA	L0	0.0727	0.1081	0.0850	-0.0062
	L1	0.0702	**0.1390**	0.0910	-0.0764
	L2	0.0837	-0.0266	0.0827	**0.0214**
	L3	0.0744	0.0869	0.0841	0.0046
Individual Factors	L0	0.1048	-0.2856	0.1234	-0.4601
	L1	0.1094	-0.3420	0.1439	-0.7029
	L2	0.1109	-0.3608	0.1496	-0.7703
	L3	0.1090	-0.3371	0.1625	-0.9233
Interactions	L0	0.1070	-0.3129	0.1247	-0.4754
	L1	0.1094	-0.3418	0.1363	-0.6133
	L2	0.1097	-0.3456	0.1409	-0.6679
	L3	0.1039	-0.2749	0.1324	-0.5673

NOTE: L stands for the number of lagged Google series. L1 means both lags 0 and 1 are included. Forecasting models use one more lag than nowcasting models. PCA adds the first 8 principal component, Best series adds the Google series that gives the highest adjusted R square in the training period. Top 3 series adds the 3 series that individually gives the highest adjusted R square in the training period. Individual factors add all series separately, while interactions in addition interacts these separate series with the baseline variables.

TABLE 7 Forecasting with Information from Google (short sample)

		Automobile		Communication	
	Lags	*RMSFE*	*Reduction*	*RMSFE*	*Reduction*
A) OLS					
Baseline		0.0707		0.0785	
PCA	L0	0.0745	-0.0542	0.0866	-0.1039
	L1	0.0740	-0.0464	0.0776	0.0108
	L2	0.0775	-0.0957	0.0797	-0.0158
	L3	0.0747	-0.0570	0.0924	-0.1777
Best series	L0	0.0800	-0.1312	0.0883	-0.1246
	L1	0.0861	-0.2171	0.0948	-0.2079
	L2	0.0852	-0.2050	0.0946	-0.2051
	L3	0.0852	-0.2050	0.0874	-0.1138
Top 3 series	L0	0.0761	-0.0756	0.0896	-0.1413
	L1	0.0799	-0.1296	0.0893	-0.1380
	L2	0.0853	-0.2060	0.0893	-0.1372
	L3	0.0892	-0.2622	0.0990	-0.2615
B) LASSO					
Baseline		0.0789		0.0823	
PCA	L0	0.0911	-0.1540	0.0767	0.0676
	L1	0.0849	-0.0758	0.0764	**0.0713**
	L2	0.0888	-0.1252	0.0853	-0.0362
	L3	0.0841	-0.0652	0.0852	-0.0349
Individual Factors	L0	0.0898	-0.1378	0.1241	-0.5087
	L1	0.0946	-0.1987	0.0818	**0.0060**
	L2	0.0987	-0.2507	0.1693	-1.0579
	L3	0.1016	-0.2869	0.1550	-0.8831
Interactions	L0	0.0942	-0.1930	0.1324	-0.6092
	L1	0.0864	-0.0947	0.1538	-0.8696
	L2	0.1036	-0.3120	0.1454	-0.7674
	L3	0.1047	-0.3258	0.1263	-0.5352

NOTE: L stands for the number of lagged Google series. L1 means both lags 0 and 1 are included. Forecasting models use one more lag than nowcasting models. PCA adds the first 8 principal component, Best series adds the Google series that gives the highest adjusted R square in the training period. Top 3 series adds the 3 series that individually gives the highest adjusted R square in the training period. Individual factors add all series separately, while interactions in addition interacts these separate series with the baseline variables.

TABLE 8
Nowcasting with CCI Information (long sample)

		Automobile		*Communication*	
	Lags	*RMSFE*	*Reduction*	*RMSFE*	*Reduction*
		A) OLS			
Baseline		0.0558		0.1118	
CCI	L1	0.0575	-0.0306	0.1162	-0.0395
	L2	0.0565	-0.0135	0.1161	-0.0382
	L3	0.0574	-0.0293	0.1167	-0.0438
	L4	0.0576	-0.0327	0.1205	-0.0780
		B) LASSO			
Baseline		0.0649		0.1177	
CCI	L1	0.0654	-0.0077	0.1306	-0.1100
	L2	0.0636	**0.0212**	0.1152	**0.0212**
	L3	0.0637	0.0183	0.1192	-0.0128
	L4	0.0637	0.0198	0.1388	-0.1798

NOTE: L stands for the number of lagged CCI series. L2 means both lags 1 and 2 are included. Forecasting models goes back more lag than nowcasting models.

TABLE 9
Forecasting with CCI Information (long sample)

		Automobile		*Communication*	
	Lags	*RMSFE*	*Reduction*	*RMSFE*	*Reduction*
A) OLS					
Baseline		0.0665		0.1102	
CCI	L2	0.0659	0.0092	0.1141	-0.0353
	L3	0.0660	0.0087	0.1173	-0.0642
	L4	0.0665	0.0004	0.1226	-0.1122
	L5	0.0645	0.0303	0.1253	-0.1365
B) LASSO					
Baseline		0.0714		0.1110	
CCI	L2	0.0705	0.0122	0.1133	-0.0208
	L3	0.0645	0.0969	0.1149	-0.0358
	L4	0.0643	0.0994	0.1453	-0.3098
	L5	**0.0624**	**0.1250**	0.1450	-0.3063

NOTE: L stands for the number of lagged CCI series. L3 means both lags 2 and 3 are included. Forecasting models go back more lags than nowcasting models.

TABLE 10 Nowcasting with CCI and Baidu (long sample)

		Automobile		***Communication***	
	Lags	***RMSFE***	***Reduction***	***RMSFE***	***Reduction***
		A) OLS			
Baseline		0.0558		0.1118	
Sum	L0	0.0544	0.0238	0.1126	-0.0077
	L1	0.0569	-0.0201	0.1107	0.0101
	L2	0.0616	-0.1041	0.1096	0.0192
	L3	0.0660	-0.1829	0.1109	0.0076
PCA	L0	0.0547	0.0193	0.1038	0.0715
	L1	0.0535	0.0398	0.1059	0.0524
	L2	0.0558	-0.0014	0.1086	0.0283
	L3	0.0585	-0.0491	0.1099	0.0171
Best series	L0	0.0654	-0.1726	0.1019	**0.0888**
	L1	0.0634	-0.1366	0.1027	0.0812
	L2	0.0649	-0.1633	0.1086	0.0284
	L3	0.0666	-0.1949	0.1099	0.0166
Top 3 series	L0	0.0526	0.0564	0.1035	0.0743
	L1	**0.0502**	**0.1001**	0.1042	0.0679
	L2	0.0566	-0.0151	0.1087	0.0272
	L3	0.0595	-0.0664	0.1072	0.0406
		B) LASSO			
Baseline		0.0649		0.1177	
Sum	L0	0.0596	**0.0827**	0.1261	-0.0716
	L1	0.0602	0.0723	0.1311	-0.1141
	L2	0.0628	0.0333	0.1260	-0.0711
	L3	0.0683	-0.0521	0.1357	-0.1534
PCA	L0	0.0543	0.1632	0.1219	-0.0362
	L1	0.0531	0.1830	0.1250	-0.0625
	L2	0.0556	0.1436	0.1261	-0.0721
	L3	0.0573	0.1171	0.1167	0.0081
Individual Factors	L0	0.0591	0.0903	0.1456	-0.2373
	L1	0.0575	0.1143	0.1587	-0.3488
	L2	0.0599	0.0775	0.1713	-0.4560
	L3	**0.0505**	**0.2225**	0.1827	-0.5526
Interactions	L0	0.0585	0.0995	0.1138	0.0325
	L1	0.0647	0.0037	0.1342	-0.1405
	L2	0.0647	0.0044	0.1216	-0.0339
	L3	0.0528	**0.1874**	0.1515	-0.2876

NOTE: L stands for the number of lagged Baidu series. L1 means both lags 0 and 1 are included. Forecasting models use one more lag than nowcasting models. Sum includes the sum of Baidu series as an additional variable to the baseline model. PCA adds the first 8 principal component. Best series adds the Baidu series that gives the highest adjusted R square in the training period. Top 3 series adds the 3 series that individually gives the highest adjusted R square in the training period. Individual factors adds all series separately, while interactions in addition interacts these separate series with the baseline variables.

TABLE 11 Forecasting with CCI and Baidu (long sample)

		Automobile		*Communication*	
	Lags	*RMSFE*	*Reduction*	*RMSFE*	*Reduction*
A) OLS					
Baseline		0.0665		0.1102	
Sum	L0	0.0671	-0.0092	0.1068	**0.0312**
	L1	0.0710	-0.0670	0.1069	0.0299
	L2	0.0754	-0.1330	0.1077	0.0227
	L3	0.0780	-0.1717	0.1114	-0.0110
PCA	L0	0.0683	-0.0269	0.1189	-0.0788
	L1	0.0661	0.0068	0.1211	-0.0988
	L2	0.0663	0.0036	0.1210	-0.0978
	L3	0.0700	-0.0517	0.1181	-0.0712
Best series	L0	0.0631	0.0519	0.9073	0.0927
	L1	0.0623	0.0634	1.0013	-0.0013
	L2	0.0601	0.0962	1.0450	-0.0450
	L3	**0.0583**	**0.1236**	1.0294	-0.0294
Top 3 series	L0	0.0703	-0.0562	**0.9438**	**0.0562**
	L1	0.0616	0.0740	0.9744	0.0256
	L2	0.0635	0.0453	0.9899	0.0101
	L3	0.0592	**0.1104**	1.0162	-0.0162
B) LASSO					
Baseline		0.0714		0.1110	
Sum	L0	0.0683	0.0428	0.1200	-0.0810
	L1	0.0632	**0.1142**	0.1170	-0.0539
	L2	0.0662	0.0724	0.1229	-0.1077
	L3	0.0679	0.0481	0.1301	-0.1726
PCA	L0	0.0670	0.0615	0.1260	-0.1355
	L1	0.0657	0.0793	0.1284	-0.1572
	L2	0.0670	0.0606	0.1263	-0.1381
	L3	0.0664	0.0694	0.1199	-0.0807
Individual Factors	L0	0.0866	-0.2134	0.1568	-0.4133
	L1	0.0753	-0.0545	0.1529	-0.3777
	L2	0.0604	0.1537	0.1780	-0.6046
	L3	**0.0551**	**0.2285**	0.1517	-0.3670
Interactions	L0	0.0730	-0.0224	0.1620	-0.4597
	L1	0.0783	-0.0968	0.1686	-0.5194
	L2	0.0601	**0.1577**	0.1993	-0.7958
	L3	0.0638	0.1056	0.1862	-0.6784

NOTE: L stands for the number of lagged Baidu series. L1 means both lags 0 and 1 are included. Forecasting models use one more lag than nowcasting models. Sum includes the sum of Baidu series as an additional variable to the baseline model. PCA adds the first 8 principal component. Best series adds the Baidu series that gives the highest adjusted R square in the training period. Top 3 series adds the 3 series that individually gives the highest adjusted R square in the training period. Individual factors adds all series separately, while interactions in addition interacts these separate series with the baseline variables.

TABLE 12 Model comparison

OLS				
	Automobile		***Communication***	
	Best Model	***Reduction***	***Best Model***	***Reduction***
Nowcast (Long Sample)	Baidu PCA	0.1125	Baidu Top 3 Series	0.1384
Forecast (Long Sample)	Baidu Top 1 Series	0.1470	Baidu Top 1 Series	0.1234
Nowcast (Short Sample)	Baidu PCA	0.2333	Baidu Sum	0.0142
Forecast (Short Sample)	Baidu Top 3 Series	0.2439	Baidu Sum	0.0325
Lasso				
	Automobile		***Communication***	
	Best Model	***Reduction***	***Best Model***	***Reduction***
Nowcast (Long Sample)	Baidu Individual Factors with CCI	0.2225	Baidu Sum	0.0736
Forecast (Long Sample)	Baidu Individual Factors with CCI	0.2285	Baseline	0
Nowcast (Short Sample)	Baidu PCA	0.4430	Baidu Sum	0.0903
Forecast (Short Sample)	Baidu PCA	0.1533	Baidu Sum	0.0775

Note: Sum includes the sum of Baidu series as an additional variable to the baseline model. PCA adds the first 8 principal component. Best series adds the Baidu series that gives the highest adjusted R square in the training period. Top 3 series adds the 3 series that individually gives the highest adjusted R square in the training period. Individual factors adds all series separately.

Table 13 Nowcasting with Information from Baidu (long sample)

Total Retail Sales of Consumer Goods	***Lags***	***RMSFE***	***Reduction***
OLS			
Baseline		0.0349	
Sum	L0	0.0324	0.0709
	L1	0.0352	-0.0070
	L2	0.0353	-0.0113
	L3	0.0352	-0.0085
PCA	L0	0.0296	0.1533
	L1	0.0319	0.0871
	L2	0.0308	0.1172
	L3	0.0268	**0.2314**
Best series	L0	0.0317	0.0919
	L1	0.0334	0.0446
	L2	0.0302	0.1358
	L3	0.0307	0.1214
Top 3 series	L0	0.0286	0.1803
	L1	0.0309	0.1156
	L2	0.0482	-0.3821
	L3	0.0264	**0.2451**
LASSO			
Baseline		0.0352	
Sum	L0	0.0320	0.0920
	L1	0.0330	0.0619
	L2	0.0332	0.0556
	L3	0.0326	0.0751
PCA	L0	0.0263	0.2534
	L1	0.0292	0.1698
	L2	0.0289	0.1793
	L3	0.0248	0.2968
Individual Factors	L0	0.0260	0.2623
	L1	0.0259	0.2653
	L2	0.0219	0.3770
	L3	**0.0203**	**0.4228**

NOTE: L stands for the number of lagged Baidu series. L1 means both lags 0 and 1 are included. Forecasting models use one more lag than nowcasting models. Sum includes the sum of Baidu series as an additional variable to the baseline model. PCA adds the first 8 principal component. Best series adds the Baidu series that gives the highest adjusted R square in the training period. Top 3 series adds the 3 series that individually gives the highest adjusted R square in the training period. Individual factors adds all series separately, while interactions in addition interacts these separate series with the baseline variables.

Table 1 Forecasting with Information from Baidu (long sample)

<u>*Total Retail Sales of Consumer Goods*</u>	*Lags*	*RMSFE*	*Reduction*
OLS			
Baseline		0.0466	
Sum	L0	0.0420	0.0994
	L1	0.0444	0.0483
	L2	0.0451	0.0314
	L3	0.0448	0.0381
PCA	L0	0.0369	0.2083
	L1	0.0410	0.1202
	L2	0.0341	0.2685
	L3	0.0298	0.3602
Best series	L0	0.0351	0.2469
	L1	0.0335	0.2813
	L2	0.0325	0.3017
	L3	0.0297	0.3637
Top 3 series	L0	0.0251	0.4609
	L1	0.0241	0.4825
	L2	**0.0222**	**0.5238**
	L3	0.0225	0.5162
LASSO			
Baseline		0.0435	
Sum	L0	0.0396	0.0884
	L1	0.0409	0.0588
	L2	0.0407	0.0638
	L3	0.0382	0.1221
PCA	L0	0.0352	0.1912
	L1	0.0411	0.0536
	L2	0.0345	0.2062
	L3	0.0310	0.2876
Individual Factors	L0	0.0421	0.0304
	L1	0.0297	0.3174
	L2	0.0296	0.3194
	L3	0.0286	**0.3415**

NOTE: L stands for the number of lagged Baidu series. L1 means both lags 0 and 1 are included. Forecasting models use one more lag than nowcasting models. Sum includes the sum of Baidu series as an additional variable to the baseline model. PCA adds the first 8 principal component. Best series adds the Baidu series that gives the highest adjusted R square in the training period. Top 3 series adds the 3 series that individually gives the highest adjusted R square in the training period. Individual factors adds all series separately, while interactions in addition interacts these separate series with the baseline variables.

References

Askitas, N., & Zimmermann, K. (2009). *Google Econometrics and Unemployment Forecasting* (No. 4201). Institute of Labor Economics (IZA).

Armstrong, P. (2016). Why your business should be using Google Trends. Retrieved 27 December 2020 from https://www.forbes.com/sites/paularmstrongtech/2016/04/01/why-your-business-should-be-using-google-trends/?sh=129725d98268

Bangwayo-Skeete, P. F., & Skeete, R. W. (2015). Can google data improve the forecasting performance of tourist arrivals? mixed-data sampling approach. *Tourism Management (1982), 46*, 454-464. doi:10.1016/j.tourman.2014.07.014

Bakirtas, H., & Gulpinar Demirci, V. (2022). Can Google Trends data provide information on consumer's perception regarding hotel brands?. *Information Technology & Tourism*, *24*(1), 57-83.

Bokelmann, B., & Lessmann, S. (2019). Spurious patterns in google trends data - an analysis of the effects on tourism demand forecasting in germany. *Tourism Management (1982), 75*, 1-12. doi:10.1016/j.tourman.2019.04.015

Bram, J., & Ludvigson, S. C. (1998). Does consumer confidence forecast household expenditure? A sentiment index horse race. *Economic Policy Review*, *4*(2).

Carrière-Swallow, Y., & Labbé, F. (2013). Nowcasting with Google Trends in an emerging market. *Journal of Forecasting*, *32*(4), 289-298.

Carroll, C. D., Fuhrer, J. C., & Wilcox, D. W. (1994). Does consumer sentiment forecast household spending? If so, why?. *The American Economic Review*, *84*(5), 1397-1408.

Choi, H., & Varian, H. (2012). Predicting the present with Google Trends. *Economic record*, *88*, 2-9.

Cotsomitis, J. A., & Kwan, A. C. (2006). Can consumer confidence forecast household spending? Evidence from the European commission business and consumer surveys. *Southern Economic Journal*, 597-610.

D'Amuri, F., & Marcucci, J. (2017). The predictive power of Google searches in forecasting US unemployment. *International Journal of Forecasting*, *33*(4), 801-816.

Dees, S., & Brinca, P. S. (2013). Consumer confidence as a predictor of consumption spending: Evidence for the United States and the Euro area. *International Economics*, *134*, 1-14.

Ettredge, M., Gerdes, J., & Karuga, G. (2005). Using web-based search data to predict macroeconomic statistics. *Communications of the ACM*, *48*(11), 87-92.

Fang, J., Gozgor, G., Lau, C. K. M., & Lu, Z. (2020). The impact of Baidu Index sentiment on the volatility of China's stock markets. *Finance Research Letters*, *32*, 101099.

Fang, J., Wu, W., Lu, Z., & Cho, E. (2019). Using Baidu index to nowcast mobile phone sales in China. *The Singapore Economic Review*, *64*(01), 83-96.

Fang, J., Zhang, X., Tong, Y., Xia, Y., Liu, H., & Wu, K. (2021). Baidu index and COVID-19 epidemic forecast: evidence from China. *Frontiers in public health*, *9*, 685141.

Fondeur, Y., & Karamé, F. (2013). Can Google data help predict French youth unemployment?. *Economic Modelling*, *30*, 117-125.

Gausden, R., & Hasan, M. S. (2018). An assessment of the contribution of consumer confidence towards household spending decisions using UK data. *Applied Economics*, *50*(12), 1395-1411.

Ginsberg, J., Mohebbi, M., Patel, R. et al. (2009). Detecting influenza epidemics using search engine query data. Nature 457, 1012–1014.

Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M., & Watts, D. J. (2010). Predicting consumer behavior with Web search. *Proceedings of the National academy of sciences*, *107*(41), 17486-17490.

Guzman, G. (2011). Internet search behavior as an economic forecasting tool: The case of inflation expectations. *Journal of economic and social measurement*, *36*(3), 119-167.

Hand, C., & Judge, G. (2012). Searching for the picture: Forecasting UK cinema admissions using google trends data. *Applied Economics Letters, 19*(11), 1051-1055. doi:10.1080/13504851.2011.613744

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). Statistical learning with sparsity: the lasso and generalizations. *CRC press*.

Howrey, E. P. (2001). The predictive power of the index of consumer sentiment. *Brookings papers on economic activity*, *2001*(1), 175-207.

Huang, X., Zhang, L., & Ding, Y. (2017). The Baidu Index: Uses in predicting tourism flows–A case study of the Forbidden City. *Tourism Management*, *58*, 301-306.

Juhro, S. M., & Iyke, B. N. (2020). Consumer confidence and consumption expenditure in Indonesia. *Economic Modelling*, *89*, 367-377.

Jun, S., Yoo, H. S., & Choi, S. (2018). Ten years of research change using google trends: From the perspective of big data utilizations and applications. *Technological Forecasting & Social Change, 130*, 69-87. doi:10.1016/j.techfore.2017.11.009

Krugel, L. & Viljoen, C. (2019). Post-Black Friday analysis. Retrieved 27 December 2020, from https://www.pwc.co.za/en/press-room/post-black-friday-analysis.html

Kwan, A. C., & Cotsomitis, J. A. (2006). The usefulness of consumer confidence in forecasting household spending in Canada: A national and regional analysis. *Economic Inquiry*, *44*(1), 185-197.

Lahiri, K., Monokroussos, G., & Zhao, Y. (2016). Forecasting consumption: The role of consumer confidence in real time with many predictors. *Journal of Applied Econometrics*, *31*(7), 1254-1275.

Li, S., Chen, T., Wang, L., & Ming, C. (2018). Effective tourist volume forecasting supported by PCA and improved BPNN using Baidu index. *Tourism Management*, *68*, 116-126.

Liu, Y., Tseng, F., & Tseng, Y. (2018). Big data analytics for forecasting tourism destination arrivals with the applied vector autoregression model. *Technological Forecasting & Social Change, 130*, 123-134. doi:10.1016/j.techfore.2018.01.018

Mihaela, S. (2020). Improving unemployment rate forecasts at regional level in romania using google trends. *Technological Forecasting & Social Change, 155*, 120026. doi:10.1016/j.techfore.2020.120026

Morris, D., 2012. *Google Searches Central Banks' Latest Economic Tool*. [online] The Financial Post. Available at: <https://financialpost.com/news/economy/google-searches-central-banks-latest-economic-tool.> [Accessed 27 December 2020].

Naccarato, A., Falorsi, S., Loriga, S., & Pierini, A. (2018). Combining official and Google Trends data to forecast the Italian youth unemployment rate. *Technological Forecasting and Social Change*, *130*, 114-122.

Önder, I. (2017). Forecasting tourism demand with Google trends: Accuracy comparison of countries versus cities. *International Journal of Tourism Research*, *19*(6), 648-660.

Park, S., Lee, J., & Song, W. (2017). Short-term forecasting of Japanese tourist inflow to South Korea using Google trends data. *Journal of Travel & Tourism Marketing*, *34*(3), 357-368.

Shen, D., Zhang, Y., Xiong, X., & Zhang, W. (2017). Baidu index and predictability of Chinese stock returns. *Financial Innovation*, *3*(1), 4.

Sun, S., Wei, Y., Tsui, K. L., & Wang, S. (2019). Forecasting tourist arrivals with machine learning and internet search index. *Tourism Management, 70*, 1-10.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 5*8(1), 267-288.

Vaughan, L., & Chen, Y. (2015). Data mining from web search queries: A comparison of google trends and baidu index. *Journal of the Association for Information Science and Technology*, *66*(1), 13-22.

Vosen, S., & Schmidt, T. (2011). Forecasting private consumption: survey-based indicators vs. Google trends. *Journal of forecasting*, *30*(6), 565-578.

Woo, J., & Owen, A. L. (2019). Forecasting private consumption with Google Trends data. *Journal of Forecasting*, *38*(2), 81-91.

Wu, L., & Brynjolfsson, E. (2015). The future of prediction: How Google searches foreshadow housing prices and sales. In *Economic analysis of the digital economy* (pp. 89-118). University of Chicago Press.

Yang, X., Pan, B., Evans, J. A., & Lv, B. (2015). Forecasting Chinese tourist volume with search engine data. *Tourism Management*, *46*, 386-397.

Yu, L., Zhao, Y., Tang, L., & Yang, Z. (2019). Online big data-driven oil consumption forecasting with Google trends. *International Journal of Forecasting*, *35*(1), 213-223.