DOI: 10.15439/2019F239
ISSN 2300-5963 ACSIS, Vol. 18

Predicting Automotive Sales using Pre-Purchase Online Search Data

Philipp Wachter
University of Hohenheim
Schwerzstr. 35, 70599 Stuttgart, Germany
Email: philipp.wachter@uni-hohenheim.de

Tobias Widmer
University of Hohenheim
Schwerzstr. 35, 70599 Stuttgart, Germany
Email: tobias.widmer@uni-hohenheim.de

Achim Klein
University of Hohenheim
Schwerzstr. 35, 70599 Stuttgart, Germany
Email: achim.klein@uni-hohenheim.de

Abstract—**Sales forecasting is an essential element for implementing sustainable business strategies in the automotive industry. Accurate sales forecasts enhance the competitive edge of car manufacturers in the effort to optimize their production planning processes. We propose a forecasting technique that combines keyword-specific customer online search data with economic variables to predict monthly car sales. To isolate online search data related to pre-purchase information search, we follow a backward induction approach and identify those keywords that are frequently applied by search engine users. In a set of experiments using real-world sales data and Google Trends, we find that our keyword-specific forecasting technique reduces the out-of-sample error by 5% as compared to existing techniques without systematic keyword selection. We also find that our regression models outperform the benchmark model by an out-of-sample prediction accuracy of up to 27%.**

I. Introduction

IMPROVING the accuracy of sales forecasts is an important business challenge for optimizing production planning. As a decisive component of planning processes, sales forecasts form the basis of managerial decision-making. The automotive industry is characterized by a complex and uncertain business environment forcing car manufacturers to constantly improve their supply chain efficiency to stay competitive [1]. Hence, sales forecasts have become an integral component of supply chain processes. Because automotive manufacturers have implemented built-to-forecast vehicle production systems [2], accurate predictions are indispensable to ensure efficient production processes, optimize inventory levels, and improve the overall market performance [3]. Moreover, increasing product individualization places ever-higher demands on business information systems [4] and material requirements planning [5]. Inaccurate predictions can lead to inventory shortages, overstocking or unsatisfied customer demands [6].

Forecasting the future demand for durable consumer goods such as cars is challenging for three reasons. First, reliable forecasting models must integrate accurate representations of the customer buying behavior. Potential customers typically engage in online searches to determine what car to buy. Searching for pre-purchase information is regarded as an integral element of the consumer's buying behavior [7]. Extensive online research applies in particular to the purchasing process of cars. About 50% of the customers spend more than ten hours to identify the best matching vehicle for their requirements [8]. Ernst and Young report that customers devote more time for online research per-purchase of a car than for any other product [9]. Customers use different keywords and combinations of keywords to determine their choice. However, the extent to which these keyword-specific search results affect the sales performance of car manufacturers is still not known. Hence, understanding the online search behavior of customers is critical to improve forecast models. Second, fluctuating macroeconomic factors have a significant impact on automobile sales [10]. If lagged effects of economic factors are not considered in forecasting models, the forecast accuracy is further impaired. Third, in addition to the seasonal demand pattern for cars, external factors such as marketing campaigns further complicate the forecasting process.

Prior research on sales forecasting has focused on rather simple techniques that incorporate historic sales data and/or socioeconomic variables but pay little attention to information reflecting customer search behavior [10]–[12]. Subsequent approaches use customer online search data to predict car sales. Choi and Varian (2009) study a model that incorporates Google search data [13]. Their findings provide econometric evidence that using Google data can enhance the prediction accuracy of car sales. As a consequence, Google search data have become an important element of sales forecasting in this field of research [14]–[16]. Although these approaches predict car sales based on customer search data, they do not systematically select the most relevant keywords used by customers, which might lead to sales of new cars.

Against this backdrop, we propose a novel forecasting technique that combines keyword-specific customer search behavior from Google Trends with a set of economic variables for sales prediction in the automotive industry using a regression approach. To identify the most relevant keywords that customers use in Google prior to purchasing a new car, we use a backward induction approach. By using Google Ads, we identify the most relevant keywords that customers used

1This work has been partially supported by the Federal Ministry of Economic Affairs and Energy under grant ZF4541001ED8.

in Google search in the context of buying a new car. We include keywords related to new car purchases and exclude keywords associated to post-purchase and other queries unrelated with pre-purchase searches. Then, we obtain the Google Trends monthly time series of the most relevant keywords for new car sales. To validate our proposal, we use a unique dataset of car sales of a large car manufacturer from 2004 to 2019.

We find that our proposed forecasting technique improves the out-of-sample prediction accuracy by up to 5% as compared to models based on the same Google Trends search data without systematic keyword selection. Furthermore, we find that our forecasting models improve the out-of-sample accuracy by up to 27% compared to well-accepted autoregressive benchmark models.

The remainder of this paper is organized as follows. The next section discusses related literature on forecasting using online search data. In section 3, we present our proposed forecasting technique. In section 4, we report the experimental evaluation and discuss our findings. We provide our conclusion in section 5.

II. Related work

Online search engines are frequently used as a starting point for the online research [17], [18]. With a market share of 88.5%, Google is by far the most frequently used online search service in the USA [19]. Due to the huge amount of daily search queries, Google represents a “Treasure House for web data mining” and previous research has focused on the predictive power of the search data [20]. Beside their popularity, search engines provide the benefit that the collected search data is less biased as compared to other user generated online data. In contrast to the use of social media platforms, online research is conducted in private and the personal activity is not revealed to others resulting in a less biased user behavior [21]. In recent years, several studies made use of Google data to improve forecasting as well as nowcasting accuracies. While forecasting is defined as the prediction of future events, nowcasting refers to the prediction of “the present, very near future and the very recent past” [22].

One of the first attempts to integrate search query data into a prediction model was made in the field of epidemiology [23]–[25]. Ginsberg et al. were able to predict the weekly influenza activity with a time lag of one day as they discovered a high correlation between influenza-related search queries and the percentage of daily physicians visits in which a patient had influenza-like symptoms. Further publications focus on the prediction of country-specific unemployment rates [13], [26]–[30], stock market movements and returns [31], [32], travel activities [33], and housing sales [13], [34]. During recent years, Google search data was employed in a wide range of different contexts, thus demonstrating the broad scope of possible application.

A. Prediction of car sales

The use of Google search data for the prediction of car sales or car registrations has raised significant attention in the literature. Chamberlin (2010), Seebach et al. (2011), Du and Kamakura (2012), and Choi and Varian (2012) were the first who examined the predictive power of Google search data in the context of car sales [35], [14], [36], [33]. They conclude that Google data reflect changes in the volume of car sales and appears to be an appropriate data source for prediction models.

Carrière-Swallow and Labbé (2013) propose an online search data index to improve nowcasting models, predicting automotive sales in Chile [15]. Although they observe a relatively low Internet usage among the Chilean population, the integration of Google data improved both in-sample and out-of-sample nowcasts. In the former case, the whole data sample is used to fit the model and the forecasted observations are part of this sample (in-sample). As an attempt to mimic real data constraints, in the latter case, only a subset of the data sample is used to fit the model of which the forecasted observations are not part of (out-of-sample) [37].

Barreira et al. (2013) examine the eligibility of Google search data as a predictor for car sales in four European countries (i.e., France, Italy, Portugal, Spain) [30]. In contrast to previous work, they find only little evidence that Google data improves the accuracy of the prediction models for the included countries.

Taking cars as an example of high-involvement products, Geva et al. (2015) aim to improve the accuracy of an out-of-sample forecast by combining forum data in form of social media mentions/sentiments and search data [21]. They find a significant improvement of the prediction accuracy if both data sources are included in the model as compared to forum data only. Moreover, they observe a stronger improvement of the prediction accuracy for value than for premium brands.

Benthaus and Skodda (2015) pursue a similar approach by combining search data with Twitter sentiment data [16]. The results are in line with the findings by Geva et al. as a combination of the two data sources leads to an improved accuracy, both in-sample and out-of-sample.

The findings of Wijnhoven and Plant (2017), however, indicate that social media sentiments only have a minor predictive power as compared to Google search data or social mentions [38]. Consequently, Wijnhoven and Plant propose to only incorporate Google data and social mentions in a prediction model.

Fantazzini and Toktamysova (2015) investigate the out-of-sample accuracy of multivariate models using Google search data and economic variables to predict monthly sales of several car brands in Germany [3]. They find that Google data-based prediction models outperform competing models especially for forecast horizons longer than 12 months.

Nymand-Andersen and Pantelidis (2018) investigate the usefulness of Google search data with respect to nowcasting of car sales in the euro area [39]. They highlight the predictive capabilities of online search data; however, they also underscore the need to further improve the data quality.

B. Motivation of the search engine user

Although the use of online search data for forecasting purposes has grown considerably, the search motivation of customers has so far received little attention. The impact of the search motive on the predictive quality of a search query index can be highlighted by the following example. The search term "Honda" comprises multiple search purposes such as gathering product information before purchase, gathering product information after purchase, and gathering news about the brand or the product. To use the Google Trends data as a predictor for new car sales, the search query index should only reflect search queries related to a purchase intention (i.e., pre-purchase search). Extracting pre-purchase searches from aggregated data, however, remains a challenge. One approach is to use appropriate search categories (e.g., vehicle shopping) to exclude searches unrelated to a purchase. Graevenitz et al. (2016) argue that the underlying algorithms might be altered over time or the customer search behavior changes with regard to the keyword use [40]. Instead, they develop a model that links distinct search motives to the search and sales data to estimate the effect of pre-purchase queries on car sales. Hu et al. (2014) pursue another approach and try to isolate pre-purchase searches by excluding terms associated to post-purchase and other non-new-car-shopping-related searches (e.g., "parts", "repair") [41]. Most of the studies discussed above use rather simple keyword combinations, which only comprise the brand and/or the model name (e.g., "Honda + Civic") depending on the level of aggregation.

III. Forecasting technique

Our proposed forecasting technique for new car sales using Google Trends search data for most relevant keywords is described as a five-step process depicted in Fig. 1. In a first step, relevant keywords and data are collected. To account for seasonality, the data is transformed to obtain deseasonalized time series. In a preliminary analysis, we detect the time lag of the Google Trends data and the economic variables with the car sales data. To identify the Google Trends data with the highest predictive power, we perform both an in-sample and an out-of-sample regression analysis. In the last step, we develop several multivariable regression models and determine the respective in-sample and out-of-sample performance.

A. Google Trends tool

In 2006, Google launched the search analysis website Google Trends. The publicly available tool provides information about aggregated individual searches expressed in a search volume index. Hence, Google does not report the data in absolute numbers but provides the relative popularity of a search term. The index is calculated by dividing the data points of a query by the total volume of searches of the geography and time range considered [42]. The query shares are normalized, such that 100 indicates the highest query share of the whole period. Since the search volume indices are proportionated to time and space, Google Trends allows to compare the relative popularity of a query across different geographic locations and time intervals.

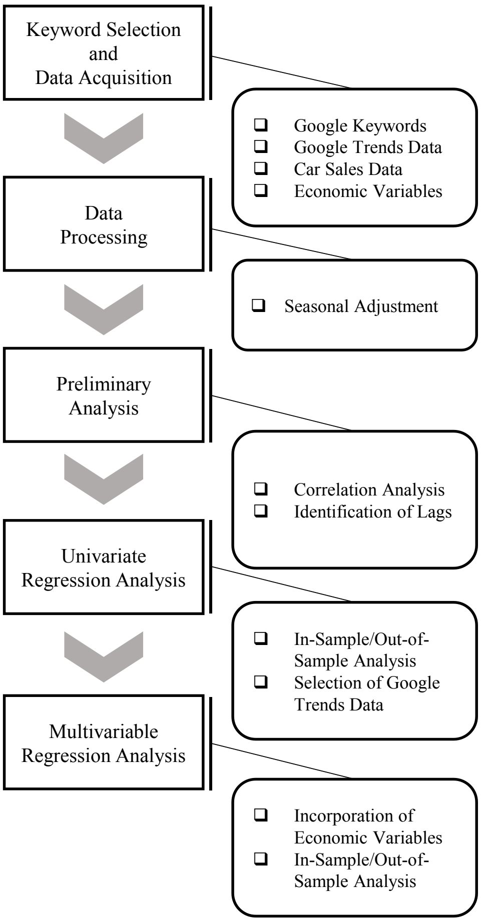
Keyword Selection and Data Acquisition
❑ Google Keywords
❑ Google Trends Data
❑ Car Sales Data
❑ Economic Variables
Data Processing
❑ Seasonal Adjustment
Preliminary Analysis
❑ Correlation Analysis
❑ Identification of Lags
Univariate Regression Analysis
❑ In-Sample/Out-of-Sample Analysis
❑ Selection of Google Trends Data
Multivariable Regression Analysis
❑ Incorporation of Economic Variables
❑ In-Sample/Out-of-Sample Analysis

Fig. 1 Forecasting technique for car sales using most relevant search data

Moreover, Google introduced different categories and subcategories to refine the search for terms with multiple meanings. In the context of the automotive industry, the search results for "beetle" can be narrowed down by the choice of an appropriate category to exclude queries regarding the insect and only obtain results for the car offered by Volkswagen.

B. Keyword selection

Selecting the most relevant keywords for Google Trends search is performed by using the online advertising platform Google Ads. Relevant keywords are identified following a backward induction approach [43]. The integrated Keyword planner tool suggests additional keywords based on keywords or groups of keywords entered by the user. The purpose of this process is to identify related keywords frequently employed by search engine users. We use the service to both identify top keywords that are commonly associated to new car purchase searches and keywords that relate to post-purchase or used car purchase searches.

C. Data processing

Some data may exhibit a strong seasonality. To account for the systematic seasonal variation, we perform a decomposition operation. The time series is decomposed into a seasonally adjusted times series and the corresponding seasonal factors. This process is an implementation of the ratio-to-moving-average method (census method I). Due to the same reporting granularity the periodicity has not been adjusted.

D. Preliminary analysis

This step encompasses a correlation analysis of the different Google Trends time series with the sales time series. As online information search is conducted in advance to new car shopping [44], we use cross-correlation to account for time lags. The incorporation of time lags is an essential prerequisite to obtain the optimal correlation between the data and to allow for forecasting the future instead of explaining the present. Cross-correlation has already been used to identify time lags in related previous work [14], [16]. The procedure is also applied to the selected economic variables. Moreover, the variables are checked for multicollinearity via bivariate Pearson correlation to prevent statistical and numerical issues in our subsequent regression analysis [45].

E. Regression analysis

To determine the predictive power of Google Trends, search data (independent variable) and car sales (dependent variable) are used to estimate univariate linear regression models. We measure the in-sample and out-of-sample performance to identify the model with the best fit. Time lags detected during the preliminary analysis are taken into account for the model computation. We apply two performance criteria to evaluate the quality of the linear regression models. Both, Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are frequently employed for model evaluations [14], [16].

MAE measures the average magnitude of errors in a set of data regardless of the direction of the errors. As a linear score, all the individual differences are weighted the same. As shown in formula (1), the absolute difference of actual sales at time t (y_t) and the predicted sales at time t ($\hat{y}_t$) is divided by the number of observations (n).

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t| \qquad (1)$$

For the second regression metric, the error is also calculated as an average of the absolute differences between actual sales and predicted values, however, the individual deviations have been squared before. This leads to the fact that the RSME (see formula 2) is more sensitive to outliers.

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2} \qquad (2)$$

After identifying the Google Trends data with the best in-sample accuracy, we estimate additional multivariable linear regression models by including different combinations of economic variables.

As a benchmark, we use a seasonal autoregressive baseline model (see formula 3) previously applied in several studies [33], [39]. The model uses 12 months (S_{t-12}) and 1-month (S_{t-1}) lagged historic sales data and an error term ε_t to predict car sales S_t.

$$S_t = \beta_0 + \beta_1 S_{t-12} + \beta_2 S_{t-1} + \varepsilon_t \qquad (3)$$

IV. Evaluation

This section reports an experimental evaluation of our forecasting technique for car sales based on most relevant Google Trends data. We describe the setup, report the results, and discuss the findings

A. Experimental setup

We collected monthly search query indices for the respective car model and/or car brand in combination with the most relevant keywords selected via Google Ads. We focus on the car manufacturer Honda as a representative of a large seller in the US. To obtain Google Trends data for the brand Honda, we additionally include the model names of the four best-selling car models responsible for approximately 90% of the Honda car sales in the period considered. The intention is to achieve a high coverage of search queries for Honda cars by using the top sellers as a proxy. To exclude searches unrelated to the automotive industry the search query indices are generated within the category "Autos & Vehicles". The result data are limited to searches originating from the US in the period from January 2004 to February 2019.

Our evaluation is based on a unique dataset containing the monthly US car sales from January 2004 to January 2016. We obtained additional data from February 2015 to February 2019 from the automotive industry analysis website CarSaleBase [46], which has been used as a source for automotive sales information in prior research [47]. To ensure the consistency of the two datasets, we check that the car sales data are congruent in 2015. We obtained 182 observations for each Honda car model.

The economic variables have been systematically selected on the basis of relevant literature [3], [48]–[50]. The variables either reflect changes in the price paid by automobile consumers, affect the automobile sales demand, or describe the state of the US economy [50]. Table I shows the selected variables and the respective descriptions.

TABLE I.
ECONOMIC VARIABLES

Economic variable	Source	Description
Consumer confidence index (CCI)	OECD	The index provides a measure for the consumer confidence and indicates future developments regarding consumption and saving
Consumer price index for new vehicles (CPI)	BLS	The index reflects changes in the price level for new vehicles (base period 1982-1984=100)
Gasoline price	EIA	The monthly retail price of US regular all formulations gasoline price
Unemployment rate	BLS	US national unemployment rate
Standard & Poor's 500 Index (S&P 500)	Yahoo finance	US stock market benchmark

BLS: Bureau of Labor Statistics; EIA: Energy Information Administration; OECD: Organization for Economic Co-operation and Development

We used data from January 2004 to August 2017 to estimate the linear regression models. Consequently, this is also the period for the in-sample analysis. To evaluate the out-of-sample performance, we used the in-sample estimated models to predict car sales from September 2017 to February 2019.

As a linear relationship between the independent and the dependent variables is a fundamental prerequisite for a linear regression analysis, we verify linearity by using scatterplots. To ensure that the remaining assumptions are fulfilled, we analyzed the histogram of residuals and the P-P plot. To ensure homogeneity of variances, we examined a scatterplot of the predicted values and the residuals.

B. Results

We identified frequently employed keywords for searches relating to new car purchases and for searches not related to pre-purchase situations using the keyword planner tool of Google Ads. While pre-purchase keywords often relate to the procurement processes (e.g., search for car dealers), pre-purchase unrelated keywords predominantly cover attributes associated to used cars or car maintenance and repairs. Table II shows the different brand-related keyword sets, additional pre-purchase keywords, and the pre-purchase unrelated keywords that can be used for reducing search data results.

TABLE II.
KEYWORDS USED FOR RETRIEVING GOOGLE TRENDS SEARCH DATA

Brand-related keyword sets		Pre-purchase keywords	Pre-purchase unrelated keywords
1	honda	new + buy + dealers + dealerships + compare	repair -tires - mechanic - maintenance - inspection -old - used -owned - parts -lease
2	civic + accord + crv + odyssey		
3	honda + civic + accord + crv + odyssey		

Table III shows the in-sample performance of the different regression models. We conducted an in-sample cross-correlation analysis to detect the optimal time lag between the Google Trends data and the sales data. For each Google Trends time series, the highest correlation was identified without any time lag. Our results indicate a positive relationship between Google Trends search data and car sales for all univariate linear regression models. The correlation coefficients ranged from 0.69 to 0.83 and were significant at p<0.01. Search queries based on keywords for car models (set 2 in Table II) resulted in Google Trends data with the highest explanatory power in the in-sample analysis. The results also imply that specifying pre-purchase unrelated keywords to be excluded from search data further improves the model

TABLE III.
IN-SAMPLE PERFORMANCE OF GOOGLE TRENDS BASED LINEAR REGRESSION MODELS

Keyword set (brand)	Keywords (pre-purchase)	Keywords (pre-purchase unrelated)	Correlation coefficient	Root mean squared error	Mean absolute error
1			0.70**	14845.5	11830.2
	●		0.69**	15133.2	11591.5
		●	0.72**	14603.9	11751.8
	●	●	0.69**	15150.8	11631.4
2			0.82**	11873.5	8864.5
	●		0.71**	14787.2	11687.4
		●	0.83**	11815.0	8856.0
3			0.71**	14819.7	11829.4
	●		0.69**	15228.5	11796.4
		●	0.79**	12952.9	10335.0

**p<0.01

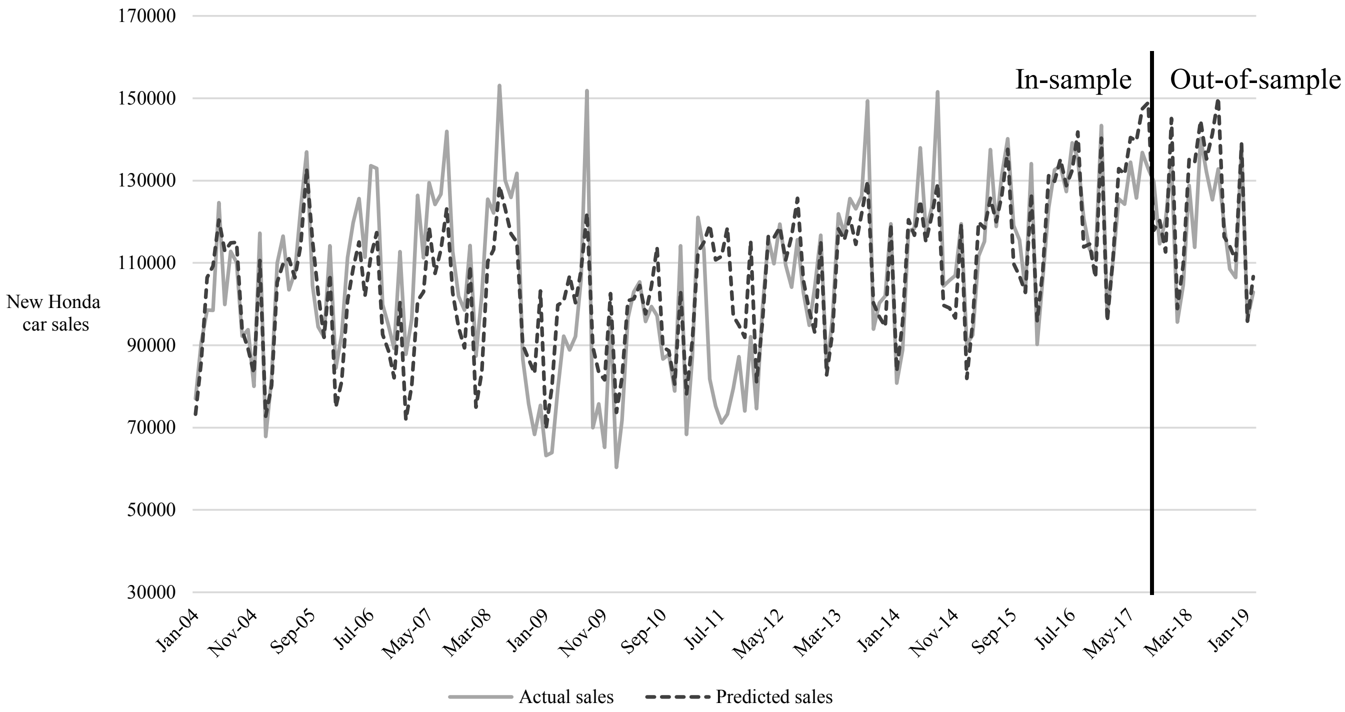

Fig. 2 Actual car sales and predicted sales based on most predictive Google Trends data

performance. That is, using car model keywords and specifying pre-purchase unrelated keywords for exclusion leads to lowest error measures among all regression models in the in-sample analysis (RMSE=11815; MAE=8856).

Excluding pre-purchase unrelated keywords from Google Trends data on car model keywords (set 2 in Table II) reduced the out-of-sample MAE by 5% from 7564.2 to 7183.8. Compared to Google Trends data on brand name (set 1 in Table II) without considering further keywords, the out-of-sample error (MAE=16796.8) is reduced by more than half. However, including keywords related to new car purchases do not reduce the prediction error. Fig. 2 illustrates actual sales and predicted sales using Google Trends data with the highest in-sample and out-of-sample accuracy. The figure demonstrates face validity of our approach.

After selecting the Google Trends data with the lowest prediction error, we conducted an out-of-sample analysis with a time horizon of 18 months. We included a set of economic variables to test for further reducing the prediction error. Since most of the economic variables are known to be leading or lagging indicators, we first identified the most predictive time lags via cross-correlation with the car sales data. Time lags were restricted to -12 to 0 months. If positive time lags for the variables were detected (i.e., economic variable from January 2016 has the highest correlation with car sales from December 2015), we incorporated no time lag. Table IV shows the chosen time lags for the economic variables and the corresponding correlation matrix.

TABLE IV.
CORRELATIONS BETWEEN CAR SALES AND ECONOMIC VARIABLES

Variable	**Optimal time lag in months**	**Sales**	**CPI**	**CCI**	**S&P 500**	**Unempl.**
Sales						
CPI	-12	0.60**				
CCI	-10	0.57**	0.40**			
S&P 500	0	0.67**	0.91**	0.44**		
Unempl.	0	-0.63**	-0.38**	-0.89**	-0.50**	
Gasol. p.	0	0.04	0.69	-0.44**	0.04	0.39**

**p<0.01; Unempl.: Unemployment; Gasol. p.: Gasoline price

All economic variables except the gasoline price showed a statistically significant correlation with car sales at p<0.01. The strongest correlation with car sales was observed for S&P 500 without time lag. Based on this preliminary analysis, we systematically generated univariate and multivariable linear regression models. The combination of predictors was restricted by the prevention of multicollinearity effects. Multicollinearity refers to a state of very high intercorrelation among the independent variables, which potentially impairs the unbiased estimation of the regression coefficients.

Because the gasoline price did not correlate with car sales in our analysis, the variable was not incorporated in any model. In addition to combinations of Google Trends data and 1-month lagged Google Trends data with each economic variable, we included all eligible combinations of economic variables.

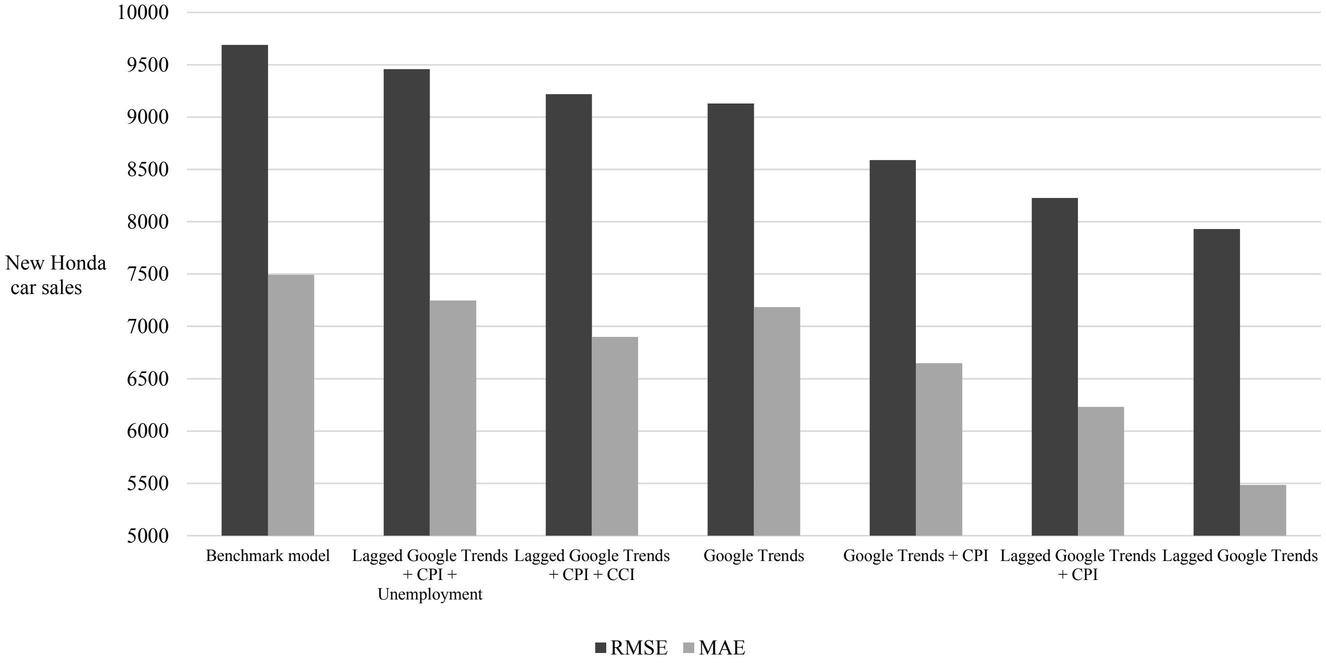

Fig. 3 Out-of-sample car sales prediction error of linear regression models with different predictive variables

Fig. 3 shows the results of the out-of-sample evaluation on a prediction horizon of 18 months. The figure only includes prediction models that outperformed the benchmark model. Both univariate Google Trends models achieved higher prediction accuracies compared to the benchmark model. A combination of Google Trends with economic variables did not necessarily improve the out-of-sample performance. A 12-months lagged CPI appeared to be the only predictor that decreases the forecasting error for unlagged Google Trends. All other multivariable regression models failed to improve the performance of the respective univariate Google Trends model. Although the in-sample error of Google Trends data with a time lag of one month (MAE=9536.5) is higher than that of Google Trends without time lag (MAE=8811.8), lagged Google Trends data achieved the smallest out-of-sample prediction error.

All (multiple) linear regressions met the criteria for linear regressions. For some of the linear regressions, we observed a slight bunching of the residuals, resulting in not perfectly identically distributed values. However, we considered the homoscedasticity assumption as fulfilled.

C. Discussion

Our experiments demonstrate the effectiveness of carefully selected customer online search data from Google for accurately predicting automotive sales. Our findings provide evidence that our proposed forecasting technique benefits from the predictive power of Google Trends data. In the following paragraphs, we discuss the insights that can be obtained from our research.

Although most customers engage in online information search prior to the purchase of a new car, our results imply that Google Trends search data without any time lag yield the highest correlation with car sales. That is, Google Trends data is most effective for predicting car sales of the current month. This finding is consistent with the results of prior research that identifies only few to no month(s) time lag [15], [16].

In the 18 months out-of-sample analysis, however, we find the highest performance using Google Trends data with a time lag of one month. Prior research suggests that, high in-sample prediction accuracy does not necessarily lead to high accuracy in an out-of-sample analysis and vice versa [14].

Our technique achieved the highest in-sample and out-of-sample accuracy for Google Trends data based on car model names combined with an exclusion of search queries containing keywords unrelated to pre-purchase situations. This finding becomes particularly evident in the out-of-sample analysis. Here, the prediction error was reduced by approximately 5% as compared to Google Trends data without keyword exclusion. Although adding pre-purchase associated keywords did not improve the model performance, systematic keyword use improved the predictive power of the Google Trends data in general.

While we find, with one exception, that incorporating the selected economic variables does not reduce the out-of-sample error, the in-sample performance was generally improved by adding the economic variables. While for the basic Google Trends data, combinations with both CCI and unemployment rate reduce the in-sample error, any two-variable combination of Google Trends data with a time lag of one month with one of the economic variables (CPI, CCI, S&P 500, unemployment rate) improves the in-sample performance. Moreover, any tested three-variable combination (Google Trends + economic variable 1 + economic variable 2) outperformed the respective univariate Google Trends regression model in the in-sample analysis. As depicted in Fig. 3, several multivariable regression models attained smaller prediction errors in the out-of-sample analysis as compared to the benchmark model. A combination of Google Trends data with economic variables, however, did not always improve the accuracy of the corresponding univariate Google Trends model.

Future research might be pursued in at least two directions. First, while we focus on top keywords proposed by the keyword planner tool of Google Ads in this work, integrating additional keywords and keyword combinations could further improve the accuracy of the prediction. These additional keywords could be obtained by empirical studies that focus on customer search behavior. Second, although our experimental setup appears to be sufficient for our research purpose, more sophisticated methods for sales forecasting are available. Hence, our approach might be extended to machine learning methods such as Neural Networks.

V. Conclusion

Our findings imply that predictions based on most relevant Google Trends search data that exclude pre-purchase unrelated searches improve the out-of-sample accuracy by up to 5% as compared to Google Trends data without systematic keyword selection. Moreover, we combine Google Trends data with relevant economic variables commonly employed for new car sales forecasting. In the performance evaluation of our linear regression models against a common seasonal autoregressive benchmark model, we find an improvement of the out-of-sample accuracy of up to 27%. Our findings help car manufacturers to obtain better forecasts and to make more informed decisions concerning their business strategies for production planning.

Acknowledgment

We thank Hansjörg Tutsch and Joerg Leukel for their valuable comments on earlier versions of this paper.

References

[1] J.-H. Thun and D. Hoenig, “An empirical analysis of supply chain risk management in the German automotive industry,” *International Journal of Production Economics*, vol. 131, no. 1, pp. 242–249, 2011, http://dx.doi.org/10.1016/j.ijpe.2009.10.010.

[2] J. Roehrich, G. Parry, and A. Graves, “Implementing build-to-order strategies: enablers and barriers in the European automotive industry,” *International Journal of Automotive Technology and Management*, vol. 11, no. 3, pp. 221–235, 2011, http://dx.doi.org/10.1504/IJATM.2011.040869.

[3] D. Fantazzini and Z. Toktamysova, “Forecasting German car sales using Google data and multivariate models,” *International Journal of Production Economics*, vol. 170, pp. 97–135, 2015, http://dx.doi.org/10.1016/j.ijpe.2015.09.010.

[4] J. Leukel, A. Jacob, P. Karaenke, S. Kirn, and A. Klein, “Individualization of goods and services: towards a logistics knowledge infrastructure for agile supply chains,” in *Proceedings of the 2011 AAAI Spring Symposium on AI for Business Agility*, Stanford, CA, USA, 2011, pp. 36–49.

[5] T. Widmer, A. Klein, P. Wachter, and S. Meyl, “Predicting Material Requirements in the Automotive Industry Using Data Mining,” in *Business Information Systems*, Seville, Spain, 2019, pp. 147–161.

[6] G. Nunnari and V. Nunnari, “Forecasting Monthly Sales Retail Time Series: A Case Study,” in *2017 IEEE 19th Conference on Business Informatics (CBI)*, Thessaloniki, Greece, 2017, pp. 1–6.

[7] K. Akalamkam and J. K. Mitra, “Consumer Pre-purchase Search in Online Shopping: Role of Offline and Online Information Sources,” *Business Perspectives and Research*, vol. 6, no. 1, pp. 42–60, 2018, http://dx.doi.org/10.1177/2278533717730448.

[8] K. Kandaswam and A. Tiwar. "Driving through the consumer's mind: Steps in the buying process." https://www2.deloitte.com/content/dam/Deloitte/in/Documents/manufacturing/in-mfg-dtcm-steps-in-the-buying-process-noexp.pdf (accessed Apr. 18, 2019).

[9] EY. "Future of automotive retail Shifting from transactional to customer-centric." https://www.ey.com/Publication/vwLUAssets/EY-future-of-automotive-retail/%24FILE/EY-future-of-automotive-retail.pdf (accessed Apr. 18, 2019).

[10] S. Shahabuddin, “Forecasting automobile sales,” *Management Research News*, vol. 32, no. 7, pp. 670–682, 2009, http://dx.doi.org/10.1108/01409170910965260.

[11] R.M.J. Heuts and J.H.J.M. Bronckers, “Forecasting the Dutch heavy truck market,” *International Journal of Forecasting*, vol. 4, no. 1, pp. 57–79, 1988, http://dx.doi.org/10.1016/0169-2070(88)90010-6.

[12] F.-K. Wang, K.-K. Chang, and C.-W. Tzeng, “Using adaptive network-based fuzzy inference system to forecast automobile sales,” *Expert Systems with Applications*, vol. 38, no. 8, pp. 10587–10593, 2011, http://dx.doi.org/10.1016/j.eswa.2011.02.100.

[13] H. Choi and H. Varian, “Predicting the Present with Google Trends,” *Google Inc*, 2009.

[14] C. Seebach, I. Pahlke, and R. Beck, “Tracking the Digital Footprints of Customers: How Firms can Improve their Sensing Abilities to Achieve Business Agility,” *Proceedings of the 19th European Conference on Information Systems (ecis))*, 2011.

[15] Y. Carrière-Swallow and F. Labbé, “Nowcasting with Google Trends in an Emerging Market,” *Journal of Forecasting*, vol. 32, no. 4, pp. 289–298, 2013, http://dx.doi.org/10.1002/for.1252.

[16] J. Benthaus and C. Skodda, “Investigating consumer information search behavior and consumer emotions to improve sales forecasting,” in *Proceedings of the 21 st Americas Conference on Information Systems*, Puerto Rico, 2015.

[17] J. Otterbacher, “Searching for product experience attributes in online information sources,” in *Proceedings of the International Conference on Information Systems (ICIS 2008)*, 2008, paper 207.

[18] N. Kumar, K. R. Lang, and Q. Peng, “Consumer Search Behavior in Online Shopping Environments,” in *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, Big Island, HI, USA, Jan. 2005, 175b-175b.

[19] statcounter. "Search Engine Market Share United States Of America." http://gs.statcounter.com/search-engine-market-share/all/united-states-of-america (accessed Apr. 18, 2019).

[20] L. Vaughan and Y. Chen, “Data mining from web search queries: A comparison of google trends and baidu index,” *Journal of the Association for Information Science and Technology*, vol. 66, no. 1, pp. 13–22, 2015, http://dx.doi.org/10.1002/asi.23201.

[21] T. Geva, G. Oestreicher-Singer, N. Efron, and Y. Shimshoni, “Using forum and search data for sales prediction of high-involvement products,” *MIS Quarterly*, vol. 41, no. 1, pp. 65–82, 2017, http://dx.doi.org/10.25300/MISQ/2017/41.1.04.

[22] M. Banbura, D. Giannone, and L. Reichlin, “Nowcasting,” *ECB Working Paper No. 1275*, 2010.

[23] J. Ginsberg *et al.,* “Detecting influenza epidemics using search engine query data,” *Nature*, vol. 457, no. 7232, pp. 1012–1014, 2009, http://dx.doi.org/10.1038/nature07634.

[24] P. M. Polgreen, Y. Chen, D. M. Pennock, and F. D. Nelson, “Using internet searches for influenza surveillance,” (eng), *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, vol. 47, no. 11, pp. 1443–1448, 2008, http://dx.doi.org/10.1086/593098.

[25] A. F. Dugas *et al.,* “Influenza forecasting with Google Flu Trends,” (eng), *PloS one*, vol. 8, no. 2, e56176, 2013, http://dx.doi.org/10.1371/journal.pone.0056176.

[26] J. Pavlicek and L. Kristoufek, “Nowcasting unemployment rates with google searches: Evidence from the visegrad group countries,” *PloS one*, vol. 10, no. 5, e0127084, 2015, http://dx.doi.org/10.1371/journal.pone.0127084.

[27] Y. Fondeur and F. Karamé, “Can Google data help predict French youth unemployment?,” *Economic Modelling*, vol. 30, no. C, pp. 117–125, 2013, http://dx.doi.org/10.1016/j.econmod.2012.07.017.

[28] N. Askitas and K. F. Zimmermann, “Google econometrics and unemployment forecasting,” *Applied Economics Quarterly*, vol. 55, no. 2, pp. 107–120, 2009, http://dx.doi.org/10.2139/ssrn.1465341.

[29] F. D’Amuri and J. Marcucci, “The predictive power of Google searches in forecasting US unemployment,” *International Journal of Forecasting*, vol. 33, no. 4, pp. 801–816, 2017, http://dx.doi.org/10.1016/j.ijforecast.2017.03.004.

[30] N. Barreira, P. Godinho, and P. Melo, “Nowcasting unemployment rate and new car sales in south-western Europe with Google Trends,” *NETNOMICS: Economic Research and Electronic Networking*, vol. 14, no. 3, pp. 129–165, 2013, http://dx.doi.org/10.1007/s11066-013-9082-8.

[31] T. Preis, H. S. Moat, and H. E. Stanley, “Quantifying trading behavior in financial markets using Google Trends,” *Scientific reports*, vol. 3, p. 1684, 2013, http://dx.doi.org/10.1038/srep01684.

[32] L. Bijl, G. Kringhaug, P. Molnár, and E. Sandvik, “Google searches and stock returns,” *International Review of Financial Analysis*, vol. 45, no. C, pp. 150–156, 2016, http://dx.doi.org/10.1016/j.irfa.2016.03.015.

[33] H. Choi and H. Varian, “Predicting the Present with Google Trends,” *Economic Record*, vol. 88, no. 1, pp. 2–9, 2012, http://dx.doi.org/10.1111/j.1475-4932.2012.00809.x.

[34] L. Wu and E. Brynjolfsson, “Chapter 3 - The Future of Prediction,” in *Economic Analysis of the Digital Economy*, A. Goldfarb, S. M. Greenstein, and C. E. Tucker, Eds.: University of Chicago Press, 2015, pp. 89–118.

[35] G. Chamberlin, “Googling the present,” *Economic & Labour Market Review*, vol. 4, no. 12, pp. 59–95, 2010, http://dx.doi.org/10.1057/elmr.2010.166.

[36] R. Y. Du and W. A. Kamakura, “Quantitative trendspotting,” *Journal of Marketing Research*, vol. 49, no. 4, pp. 514–536, 2012, http://dx.doi.org/10.1509/jmr.10.0167.

[37] A. Inoue and L. Kilian, “In-Sample or Out-of-Sample Tests of Predictability: Which One Should We Use?,” *Econometric Reviews*, vol. 23, no. 4, pp. 371–402, 2005, http://dx.doi.org/10.1081/ETC-200040785.

[38] F. Wijnhoven and O. Plant, “Sentiment analysis and Google trends data for predicting car sales,” in *38th International Conference on Information Systems*, 2017.

[39] P. Nymand-Andersen and E. Pantelidis, “Google econometrics: nowcasting euro area car sales and big data quality requirements,” ECB Statistics Paper, 2018.

[40] G. von Graevenitz, C. Helmers, V. Millot, and O. Turnbull, “Does Online Search Predict Sales? Evidence from Big Data for Car Markets in Germany and the UK,” *CGR Working Paper*, 2016, http://dx.doi.org/10.2139/ssrn.2832004.

[41] Y. Hu, R. Y. Du, and S. Damangir, “Decomposing the Impact of Advertising: Augmenting Sales with Online Search Data,” *Journal of Marketing Research*, vol. 51, no. 3, pp. 300–319, 2014, http://dx.doi.org/10.1509/jmr.12.0215.

[42] Google. "How Trends data is adjusted." https://support.google.com/trends/answer/4365533?hl=en&ref_topic=6248052 (accessed Apr. 18, 2019).

[43] A. Ross, “Nowcasting with Google Trends: a keyword selection method,” *Fraser of Allander Economic Commentary*, vol. 37, no. 2, pp. 54–64, 2013.

[44] L. R. Klein and G. T. Ford, “Consumer search for information in the digital age: An empirical study of prepurchase search for automobiles,” *Journal of Interactive Marketing*, vol. 17, no. 3, pp. 29–49, 2003, http://dx.doi.org/10.1002/dir.10058.

[45] A. F. Siegel, “Multiple Regression,” in *Practical Business Statistics*: Elsevier, 2016, pp. 355–418.

[46] Carsalebase. "Automotive Industry analysis, opinions and data." carsalesbase.com/ (accessed Apr. 18, 2019).

[47] K. Afrin, B. Nepal, and L. Monplaisir, “A data-driven framework to new product demand prediction: Integrating product differentiation and transfer learning approach,” *Expert Systems with Applications*, vol. 108, pp. 246–257, 2018, http://dx.doi.org/10.1016/j.eswa.2018.04.032.

[48] M. Hülsmann, D. Borscheid, C. M. Friedrich, and D. Reith, “General Sales Forecast Models for Automobile Markets and their Analysis,” *Transactions on Machine Learning and Data Mining*, vol. 5, no. 2, pp. 65–86, 2012.

[49] A. Sa-ngasoongsong, S. T.S. Bukkapatnam, J. Kim, P. S. Iyer, and R. P. Suresh, “Multi-step sales forecasting in automotive industry based on structural relationship identification,” *International Journal of Production Economics*, vol. 140, no. 2, pp. 875–887, 2012, http://dx.doi.org/10.1016/j.ijpe.2012.07.009.

[50] J. Gao, Y. Xie, X. Cui, H. Yu, and F. Gu, “Chinese automobile sales forecasting using economic indicators and typical domestic brand automobile sales data: A method based on econometric model,” *Advances in Mechanical Engineering*, vol. 10, no. 2, 1-11, 2018, http://dx.doi.org/10.1177/1687814017749325.