Google Searches as Predictors of U.S. Recessions *

Gustavo Rojas-Matute †

Faculty Mentor: Xuguang Sheng

American University

May 10, 2019

Abstract

In this paper I explore the use of Google searches as predictors of the probability of recession in the U.S. economy. In particular, I evaluate several Google searches-based models and compare them, both individually and in groups, with two benchmarks: the yield spread model from Estrella and Mishkin (1998) and the smoothed U.S. recession probabilities from Chauvet and Piger (2008). I find that, while the yield spread is still a good indicator to anticipate a recession, Google searches-based models are more accurate for a threshold of 50 % of probabilities. I also find that Google searches-based models using principal components method and LASSO logistic regression outperform smoothed probabilities model in both accuracy (with a 99.43 % and 100 % of success, respectively) and signaling the recession in real-time.

Keywords— Google econometrics, Recession forecast, Principal components, Machine learning, LASSO.

1 Introduction

Predicting recessions is one of the most interesting challenge of the modern macroeconomic forecasting. Understanding the transition between expansions and recessions and the behavior of macroeconomic and financial variables commonly associated with the expectations of future economic events, has been the subject of research over the last century and has become more important since the Great Recession.

Macroeconomic models were subject of several critics as consequence of the their lack of ability to predict the global financial crisis. In this sense, Krugman (2009), says: "few economists saw our current crisis coming". Accordingly, there was nothing in the prevailing

*This paper was selected to receive the International Institute of Forecasters Student Award.

†E-mail: gr4422a@american.edu, AU ID 4324422

models suggesting the possibility of the kind of collapse that happened last year (Kugman, 2009). Kiley (2016), highlights that macroeconomic models in the last fifty years focus on the structural features of households, firm and government behavior that lead to cyclical fluctuations in employment and the roles of monetary and fiscal policy in reducing volatility in economic performance, but he also highlight the difficulties in modeling financial frictions and macroprudential policy in DGSE models.

In the empirical analysis, several variables have been tested in order to predict recession, but clearly, the yield curve and the coincident variables for monitoring U.S. recessions have been the most common used in the last two decades. Those models that use the yield curve have the advantage of being very informative in advance, however they also send some false alarms. On the other hand, those models that rely in coincident variables for monitoring U.S. recessions have the problem that the data are released one or two months later.

The emergence of Google data has become an innovative source of data for forecasting and predictive models. Google searches data reflects an aggregated micro behavior that could capture the behavior of macro variables in advance such as: unemployment and private consumption. For example, when a household is not doing well, he may try to search for coupons and discounts. When a person is unemployment he may try to search for unemployment benefits, and so on.

In this paper, I use data from Google Trends to evaluate their power in predicting U.S. recession in real time. In this sense, I try to address which Google-searches are good predictors of U.S. recessions. I find several Google searches that are good predictors individually. However, when using the first five principal components as independent variables of the probit model and a LASSO logistic regression, the models are successful in 99.43% and 100 % of the cases, respectively, outperforming the smoothed probabilities model from Chauvet and Pigger, not only in accuracy but also in signaling the recession in real time. The paper is structured as follows: in section 2, I review the most important contemporaneous literature; In session 3, I explain the methodology used for the empirical models including the data. In section 4, I show the results for individual predictors, the probit model using principal components and the LASSO logistic regression. In section 5, I evaluate the benchmark models against the Google-based models.

This paper contributes to the literature of Google econometrics and predicting recessions. To my knowledge, this is the first paper that evaluates Google searches as predictors of U.S. recessions.

2 Literature Review

In the last two decades, the literature offers an interesting survey of models and approaches to predict recessions. There are two kind of variables that are always considered. In one side, the financial variables, such as the yield curve. On the other hand, we have the real activity indicators, such as, the coincident variables used for monitoring the U.S. recession. When predicting a recession, the main goal must be an early warning indicator and a minimum of false alarms. One of the advantages of financial series is that they are released almost in real time while economic activity indicators are released with a delay of one or two months.

Estrella and Mishkin (1998) evaluate a series of financial variables in a probit model to predict the probability of recession. The idea behind this model is that when inflation is ex-

pected to fall the yield curve is inverted and the spread is negative, signaling that a recession is likely to happen. Their work show that the 10-year-3-month Treasury spread, lagged by 4 quarters, is the best fit to predict recession both in-sample and out-sample.

Chauvet and Potter (2002) also used the yield curve in a probit model with structural break. They used a standard Gibbs sampling method to evaluate the posterior of the probit with the yield curve as explanatory variable to address the probability of a recession state in December 2001. They found strong evidence of breakpoint but uncertainty over the value of this probability due to uncertainty over the location of the breakpoint.

Respect to economic activity variables, Chauvet and Potter (2009), examine several probit models with the coincident variables for monitoring U.S. recessions. The results show that the ability to correctly predict a recession improve when considering recurrent breakpoints.

Chauvet and Piger (2008) evaluate a non-parametric approach and a Markov-switching dynamic-factor model using "real-time" dataset of coincident monthly variables. They find that both approaches predict very well the NBER business cycle.

2.1 Google data

Since 2004, there is available a new set of data that reflects the interests or attention of the people in a specific term: Google Trends. This information reflects million of searches and returns a time series of search activity for a query term, date range and geography. According to Jun et al (2018), while Google search data does not provide simple data on Google Search usage, some estimate that searches on Google Trends reached 2 trillion in 2016.

The use of Google data has some advantages: 1) It is released in real time: Google data is available immediately at the end of the desired period to evaluate (week or month). 2) It reflects an aggregated micro behavior: for example when a person looses his or her job, a possible next strep is to search in Google how to fill a welfare unemployment form or he or she may try to find coupons for discounts when the personal economy is not doing well. In addition, the changes in searches for certain kind of categories of products such as apparel or vehicles may reflect the current situation of the economy before other indicators.

During the last decade, data from Google Trends have been widely used in different research including economics and, in particular, in forecasting. Jun et al (2018) say that Ginsberg et al. (2009) opened the door for the use of Google Trends in research studies by shown that Google Trends traced and predicted the spread of influenza earlier than the Centers for Disease Control and Prevention (CDC).

In the economics field, Della Penna & Huang (2009), constructed a Consumer Sentiment Index for the U.S. Economy using data from Google. They show that the Google search-based index is highly correlated with the Index of Consumer Sentiment form the University of Michigan. Kholodilin et al. show the predictive power of Google searches in nowcasting macro economic variables (2009) and nowcasting private consumption (2010). Choi and Varian (2012) show how can Google Trends be used in nowcasting sales of motor vehicles and parts, initial claims for unemployment benefits and travels.

More recently, D'Amuri and Marcucci (2017), use Google searches to forecast U.S. unemployment rate. They find that Google-based models outperform most of the competitors for predicting the US unemployment rate. Baker and Fradkin (2017) develop a job search activity index based on Google search data and study the effects of unemployment insurance policy changes.

Respect to the use of Google searches in predicting recessions, Tkacz, 2013, show that searches for "recession" and "jobs" are good predictors of the Canadian economy recession.

3 Methodology

3.1 Probit Model

In order to evaluate the predictive ability of Google-based models, I use probit model defined as:

$$P(Y_t^* \geqslant 0|\mathbf{X}_t, \beta) = \Phi(\beta_0 + \beta_1 \mathbf{X}_t) \tag{1}$$

Where $\mathbf{X}_t$ is a vector of independent variables, Y_{1+k}^* is a latent variable representing the state of the economy, and Φ is the cumulative distribution function of the standard normal distribution.

The observable recession indicator is defined by:

$$Y_t = \begin{cases} 1, & \text{if the economy is in recession} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

3.2 LASSO Logistic Regression

Following (2), a logistic regression can be expressed as:

$$P(Y_t^* \geqslant 0|\mathbf{X}_t, \beta) = \frac{1}{1 + e^{(\beta_0 + \beta_1 \mathbf{X}_t)}} \tag{3}$$

Let $p^*(x_i) = P(Y_t^* \geqslant 0|x_i)$ be the probability (2) for observation i at a particular value for the parameters (β_0, β), then the penalized log likelihood is maximized as follows:

$$\max_{\beta_0, \beta_1 \mathbf{X}_t} \frac{1}{N} \sum_{i=1}^{N} \{I(Y_t^* = 1) \log p^*(x_i) + I(Y_t^* = 0) \log(1 - p^*(x_i)) \lambda(\beta)\} \tag{4}$$

The value of λ is obtained with the cross validation (*cv.glmnet)* command of the package *glmnet* in R. More detail of the LASSO logistic regression can be seen in Friedman et al. (2009).

3.3 Empirical Models and Data

3.3.1 Benchmark models

In the first exercise I use the spread between the 10-year and 3-month U.S. Treasury yields, obtained from FRED. Following Estrella and Mishkin (1998), the spread is lagged 12 months.

The Smoothed U.S. recession probabilities from Chauvet and Piger (2008) Non-farm payroll employment, industrial production, real manufacturing and trade sales, and real personal income excluding transfer payments.

3.3.2 Google-Based Models

Selecting an ideal vector of Google searches is not an easy task since we can try with hundred, thousands or even millions of terms. Stephens-Davidowitz and Varian (2015), offer a helpful hands-on guide to Google data and how Google Correlate can facilitate the process of selection of the data. In the case of this paper, I would use Google Correlate to find searches highly correlated with the time series of the variables used by Estrella and Mishkin (1998) and Chauvert and Pigger (2008) and used them in the probit model. Unfortunately, at the moment in which I am conducting this research, the tool to use "my own data" is apparently broken.

As a result, I have to use a list of different variables used by the literature reviewed and other suggested by my classmates during my presentation at the forecasting course. I use the data in two ways: "search of terms" or categories. The advantage of using categories is that all different searches related are stored by Google. One possible disadvantage is that in the category there could be searches that behave in opposite directions. The list of variables used, their classification and expected sign respect to the probability of recession is as follows:

- Jobs. Search of term. (+)
- Coupons. Category. (+)
- Real Estate Listings (-)
- Bankruptcy. Category (+)
- Apparel. Category. (-)
- Machinery. Category. (-)
- Unemployment. Search of term (+)
- Welfare Unemployment. Search of term (+)
- Wholesale retailers. Category (-)
- Credit Cards. Search of term on the Shopping category (-)
- Mortgage. Search of term (+)
- Automotive Industry. Category (-)
- Commercial vehicles. Category (-)
- Alternative Energy. Category (+/-)
- Oil & Gas. Category (+).
- Construction. Category (-).
- Food retailers. Category (-).
- Car rental. Category (-).
- Suv & Trucks. Category (-)
- Recession. Search of term (+). Following Tckaz (2013).

The period of the data used is from January 2004 to July 2018. All variables, with the exception of "recession" are seasonal adjusted using the *stl* command in R.

3.3.3 Principal Components Method and LASSO logistic regression

Since I am evaluating 20 variables, a plausible way to do it together is to reduce their dimensionality by applying the method of principal components. Kholodilin et al (2009 & 2010) used the principal component method to nowcast macroeconomic variables and private consumption in the U.S. with Google searches. In this case, the idea is to evaluate the principal components as independent variables of the probit regression.

Another method that has gained popularity among econometricians is the use of machine learning. Triffin (2016) explains in detail the use of ridge, LASSO and elastic net regressions in nowcasting GDP in Lebanon. According to him, dimension reduction is not enough: "the literature has generally found that factors extracted from fewer forecasts but more informative indicators can yield better forecasts that those obtained from larger datasets" (Triffin, 2016).

3.3.4 Control Variables

Because Google improved the geographical assignment on January 2011 and the data collection system on January 2016, I include two dummy control variables in the individual and principal component models to address this changes.

4 Results

4.1 Individual Predictors

To evaluate the results, I focus on the level of significance of the coefficients, the sign and the McFadden pseudo R-squared.

Tables 1-4 show the results for individual probit regressions. "Recession" is very significant, has a positive sign, as expected, and has the highest McFadden Pseudo R-squared of the sample (0.7394235). The coefficient of Coupons has the second highest McFadden Pseudo R-squared (0.4761417), is significant and positive as expected. Other variables with McFadden pseudo R-squared higher than 0.30 and significant coefficients are: jobs, real estate, machinery, apparel, wholesale, commercial vehicles, construction, car rental and suv & trucks.

The result for the search term "Recession" is particularly interesting because it reflects the attention on recession that Google users have at some time. In this sense, we may expect that since the recession announcement by the NBER was made in December 2008, the peak of Google searches for recession should have been after the announcement. However, Figure 1 shows that the peak of Google searches for "Recession" occurred in January 2008 just one month later of the beginning of the recession period. It is important to remember that, at this point, we did not know we were officially in recession. That means that we needed to wait until December 2008 to know that the economy was in recession. In addition, the interest remained over 50 (in an index between 0 and 100), until April 2009 and started to decline in May just one month before the end of the recession.

I made a quick search in the web page of the New York Times from November 2007 and February 2008 (click here: NYT-link). In November 2007 there is just one article titled with the word "recession". None in December 2007, but in January 2008 there were four articles including one from the Nobel laureated Paul Krugman. In February 2008 there were at least 5, including one note saying that "according to forecasters recession is unavoidable". In further

research I will explore more in depth, but the jump from zero articles in December 2007 to four in January 2008 and 5 in February 2008 may lead us to a plausible explanation. In section 4 I evaluate the predictive ability of searches for "recession" respect to the benchmark models.

Table 1: Probability of Recession: Google Searches

	Dependent variable:				
	Recession				
	(1)	(2)	(3)	(4)	(5)
jobs	0.093^{***}				
	(0.023)				
coupons		0.252^{***}			
		(0.063)			
real estate			-0.035^{***}		
			(0.011)		
bankruptcy				0.015	
				(0.013)	
machinery					-0.039^{***}
					(0.012)
con_1	-5.927	-13.030	-5.614	-4.790	-5.852
	(318.074)	(493.939)	(324.270)	(324.298)	(324.515)
con_2	0.223	2.741	0.655	0.075	0.410
	(550.628)	(941.413)	(557.389)	(558.166)	(556.851)
Constant	-6.186^{***}	-13.725^{***}	1.611^{**}	-1.784^{**}	1.689^{**}
	(1.403)	(3.259)	(0.778)	(0.831)	(0.776)
Observations	175	175	175	175	175
Log Likelihood	-35.279	-30.373	-38.909	-42.954	-38.175
Akaike Inf. Crit.	78.558	68.747	85.818	93.907	84.350
McFadden Pseudo R2	0.39153851	0.4761417	0.3289277	0.2591661	0.3415881
Note:					*p<0.1; **p<0.05; ***p<0.01

4.2 Principal Components

In this section I show the result of the model using the method of principal components. After evaluating different variables, the final vector includes: jobs, coupons, real estate, bankruptcy,

Table 2: Probability of Recession: Google Searches

	Dependent variable:				
	Recession				
	(1)	(2)	(3)	(4)	(5)
apparel	-0.043^{***}				
	(0.014)				
unemployment		0.012^{*}			
		(0.007)			
welfare unemployment			0.013		
			(0.012)		
wholesale				-0.031^{***}	
				(0.011)	
credit card					0.067^{**}
					(0.028)
con_1	-5.397	-5.263	-5.224	-5.446	-9.073
	(314.948)	(323.294)	(322.364)	(319.042)	(410.206)
con_2	0.550	0.309	0.228	0.648	-0.433
	(550.840)	(557.667)	(557.080)	(552.236)	(803.845)
Constant	2.050^{**}	-1.266^{***}	-1.513^{**}	1.266	-2.013^{***}
	(0.909)	(0.311)	(0.673)	(0.773)	(0.547)
Observations	175	175	175	175	175
Log Likelihood	-38.812	-42.113	-43.048	-40.260	-40.852
Akaike Inf. Crit.	85.624	92.226	94.095	88.520	89.704
McFadden Pseudo R2	0.33059683	0.2736692	0.2575448	0.3056265	0.2954144
Note:					*p<0.1; **p<0.05; ***p<0.01

Table 3: Probability of Recession: Google Searches

	Dependent variable:				
	Recession				
	(1)	(2)	(3)	(4)	(5)
mortgage	0.019				
	(0.014)				
automotive industry		0.013			
		(0.009)			
commercial vehicles			-0.045^{***}		
			(0.013)		
alternative energy				0.021^{*}	
				(0.013)	
oil & gas					-0.024^{**}
					(0.012)
con_1	-4.577	-4.763	-5.684	-4.485	-5.494
	(324.941)	(326.054)	(319.307)	(324.107)	(324.466)
con_2	0.120	0.010	0.732	0.065	0.196
	(558.269)	(559.206)	(553.558)	(557.687)	(556.729)
Constant	-2.188^{**}	-1.500^{***}	2.375^{***}	-2.305^{**}	0.680
	(1.004)	(0.542)	(0.879)	(0.922)	(0.748)
Observations	175	175	175	175	175
Log Likelihood	-42.585	-42.687	-36.777	-42.207	-41.734
Akaike Inf. Crit.	93.171	93.375	81.554	92.414	91.468
McFadden Pseudo R2	0.2655179	0.2637612	0.3657005	0.2720416	0.2802001
Note:					*p<0.1; **p<0.05; ***p<0.01

Table 4: Probability of Recession: Google Searches

	Dependent variable:				
	Recession				
	(1)	(2)	(3)	(4)	(5)
construction	-0.037^{***}				
	(0.011)				
food retailers		-0.013			
		(0.030)			
car rental			-0.041^{***}		
			(0.012)		
suv & trucks				-0.068^{***}	
				(0.018)	
recession					0.094^{***}
					(0.023)
con_1	-5.882	-5.088	-5.871	-5.430	-4.748
	(324.138)	(326.132)	(320.600)	(321.288)	(465.540)
con_2	0.437	0.300	0.251	1.475	0.609
	(557.178)	(558.174)	(553.738)	(544.990)	(852.668)
Constant	1.633^{**}	-0.037	1.459^{**}	3.783^{***}	-3.404^{***}
	(0.693)	(1.693)	(0.651)	(1.179)	(0.730)
Observations	175	175	175	175	175
Log Likelihood	-37.071	-43.553	-37.196	-34.054	-15.108
Akaike Inf. Crit.	82.143	95.107	82.393	76.108	38.217
McFadden Pseudo R2	0.3606224	0.2488262	0.3584651	0.4126659	0.7394235
Note:					*p<0.1; **p<0.05; ***p<0.01

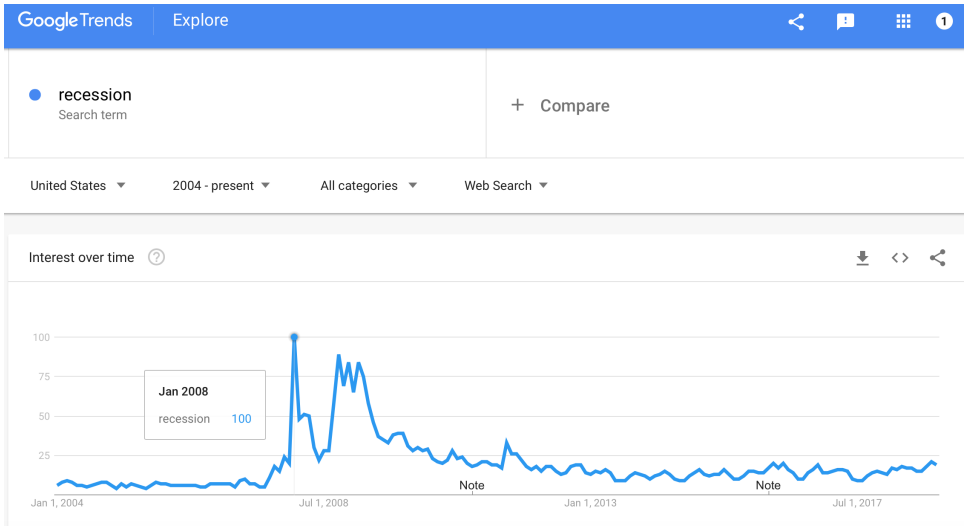

Figure 1: Google Searches for "Recession"

machinery, apparel, unemployment, welfare unemployment, wholesale, commercial vehicles, car rental, suv & trucks. Notice that some of these variables, like bankruptcy, were not significant respect to the probability of recession, however they are correlated with most of the rest of the variables. For example, bankruptcy has a correlation of -0.545 with jobs, 0.758 with machinery and 0.5303 with real estate listings. Welfare unemployment has a correlation of 0.439 with coupons, 0.462 with jobs and -0.492 with real estate listings.

Table 5 shows that principal components PC1, PC4 and PC5 are significant. I also evaluate the probit regression against PC1, PC4 and PC5 (Table 6). While the model in Table 5 shows a McFadden pseudo R-squared of 0.9172648, the model in Table 6 shows a McFadden pseudo R-squared of 0.9160916. Because of the high McFadden pseudo R-squared I decide to evaluate the model from Table 5 respect to the benchmark models.

It is not clear why PC1, PC4 and PC5 are significant while PC2 and PC3 are not. Figure 6 shows the variances of the principal components. In Figures 7-8 it can be observed that PC1, 2 and 3 have drastic changes at the beginning of 2011 and 2016, which coincide with the improvements in geographical assignment and data collection system described in the control variable subsection. Despite I control the models for this changes, my hypothesis is that PC2 and 3 are probably very sensitive to this changes and the dummy variables are not capturing them properly. It is possible that by increasing the number of variables both PC2 and 3 become significant. Because the model in Table 5 is very accurate and helpful for the purpose of this paper, I will address this issue more in depth in a further research.

4.3 LASSO Logistic Regression

Coefficients of the LASSO logistic regression are shown in table 7, but the key measure of effectiveness of the model are shown in the next section. The signs of credit cards, bankruptcy,

Table 5: Principal Components

	Dependent variable:
	rec
PC1	1.313^{**}
	(0.639)
PC2	0.159
	(1.169)
PC3	0.166
	(1.676)
PC4	-5.006^{*}
	(3.035)
PC5	-10.030^{**}
	(4.730)
con_1	-17.958
	$(1{,}008.974)$
con_2	11.541
	$(1{,}950.172)$
Constant	-3.421
	(2.504)
Observations	175
Log Likelihood	-4.797
Akaike Inf. Crit.	25.594
McFadden Pseud R2	0.9172648
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 6: Principal Components

	Dependent variable:
	rec
PC1	1.321^{**}
	(0.617)
PC4	-5.366^{*}
	(2.975)
PC5	-10.219^{*}
	(5.274)
con_1	-18.755
	(1,116.533)
con_2	12.430
	(1,998.432)
Constant	-3.620^{*}
	(2.054)
Observations	175
Log Likelihood	-4.865
Akaike Inf. Crit.	21.730
McFadden Pseud R2	0.9160916
Note:	*p<0.1; **p<0.05; ***p<0.01

coupons apparel, jobs, machinery, commercial vehicles and SUV are the expected. These coefficients, however, are not necessarily easy to interpret. As comented by Triffin (2016), machine learning has focused on prediction, with emphasis on accuracy rather than its interpretability.

Table 7: LASSO Logistic Regression

	Coef
Intercept	93.2267340
construction	.
oil & gas	-0.2588968
credit card	-0.1101311
Bankruptcy	0.8939866
Welfare Unemployment	.
Unemployment	-0.8064770
Coupons	0.5816892
Apparel	-0.4837614
Jobs	0.2909882
Real Estate	1.1156784
Machinery	-0.1791458
Wholesale	.
Commercial vehicles	-0.2818964
Car rental	.
SUV	-2.4725278

5 Model Evaluations

In this section I evaluate the Google-based models obtained in the results against the benchmarks. Figure 2 shows the behavior of the predicted probability of recession based on the yield spread. It can be seen that the model signals the recession of the 1990's and the Great Recession of 2008-2009 in advance, but also give us two false alarms between 1995 and 2000 and failed to signal the recession of 2001 in advance. In the three events observed in the figure, the probability of recession barely exceeds 50 % in two of them.

Figure 3 shows the comparison between the probit regression using "recession" (Google 3) and the smoothed probabilities from Chauvet and Pigger (2008). The probability predicted by "recession" reached 1 in January 2008, just one month after the beginning of the period, while the smoothed probabilities jumps from less than 50% to more than 50% between March and April 2008. However, the probabilities from Chauvet and Pigger remain consistently over 80% until June 2009 while the "recession" model falls to low probabilities between June and August 2008. This Google search-based model also sends signals of recession in July, September and October 2009 despite the recession ended in June 2009.

Figure 4 shows the predicted probabilities of the Google-based model using the method of principal components. The results are surprising. The probabilities jump from zero to 1

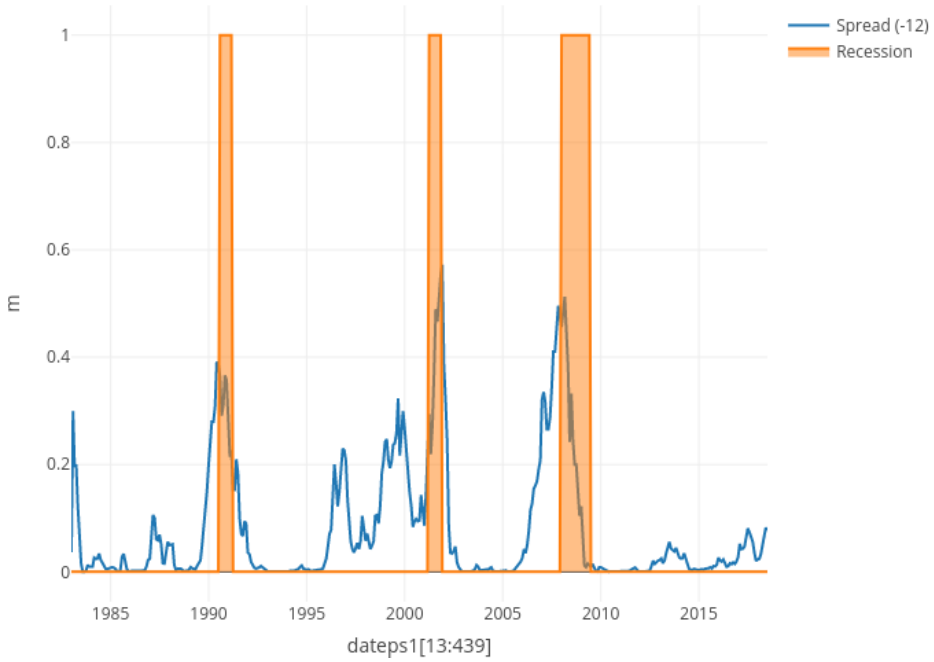

Figure 2: Probability of Recession: Yield Spread

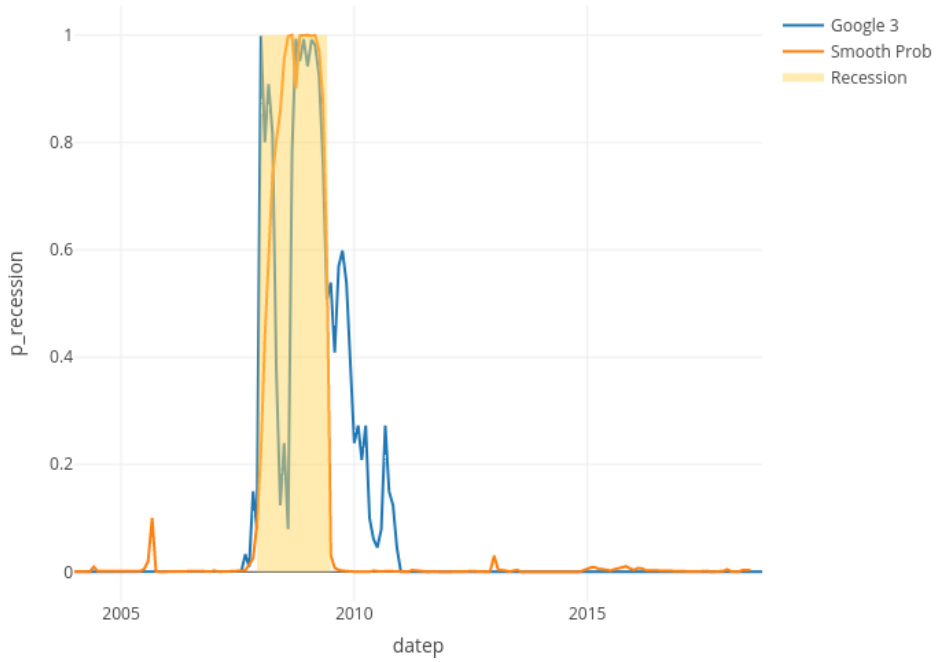

Figure 3: Probability of Recession: Google Searches (”recession”) vs Smoothed Probabilities

in December 2007 at the beginning of the recession and remain consistently over 0.60 during the whole period of recession. The model only fails by signaling recession in July 2009, one month after the end of the recession. However, it immediately falls to zero in August 2009.

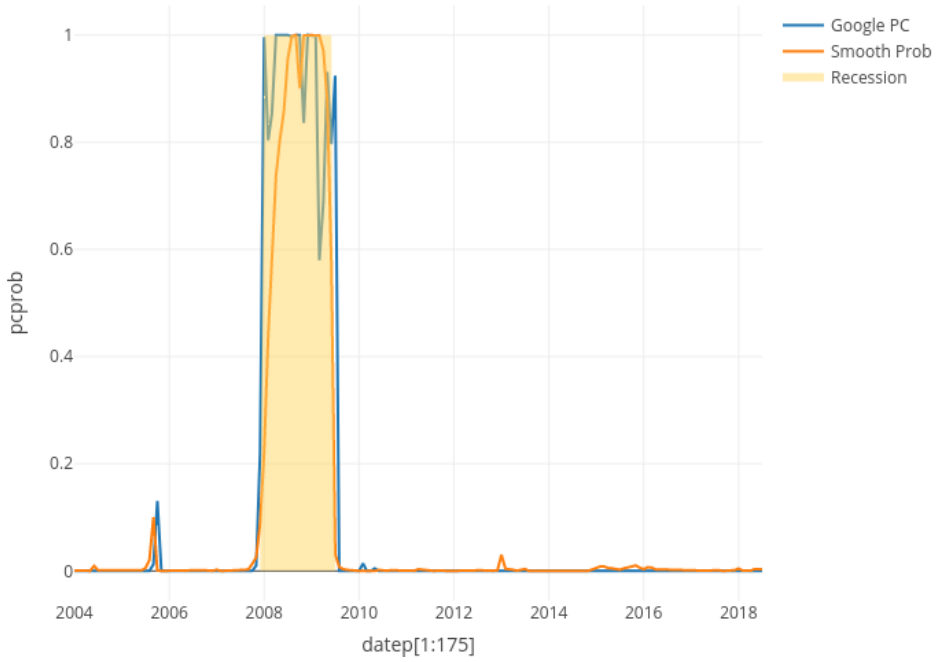

Figure 4: Probability of Recession: Google Searches - PC vs Smoothed Probabilities

Table 8 shows the predictive ability of the models using a threshold of 50%. The yield curve spread model is evaluated from 1983 to July 2018, while the rest is evaluated from January 2004 to July 2018. Google Searches - PC model is accurate 99.43 % of the cases and LASSO logistic regression is accurate 100 %, outperforming the smoothed probabilities model from Chauvet and Pigger, not only in accuracy but also in signaling the recession in real time. One caveat of this Google Searches models is that it was evaluated in just one event of recession (the Great Recession from December 2007 to June 2009), because Google data is only available since 2004. Therefore we will need to continue evaluating the model in future events.

Figure 5 includes the predicted probabilities of the Google-based model using the LASSO regression. As shown in Table 8, the percentage of successes using the machine learning is 100 % of success.

Table 8: Evaluation of Predictive Ability

		SP			*R*			*PC*			*S*			*L*		
	Y=0	Y=1	*Total*	Y=0	Y=1	*Total*	Y=0	Y=1	*Total*	Y=0	Y=1	*Total*	Y=0	Y=1	*Total*
$P \leqslant 0.5$	157	2	159	153	4	157	156	0	156	392	31	423	157	0	157
$P > 0.5$	0	16	16	4	14	18	1	18	19	13	3	16	0	18	18
Total	157	18	175	157	18	175	157	18	175	405	34	439	157	18	175
%Correct	100.00	88.90	98.86	97.45	77.78	95.40	99.36	100.00	99.43	96.80	8.80	90.00	100.00	100.00	100.00

S: spread, SP: smoothed probabilities, R: recession, PC: Principal Components, L: LASSO

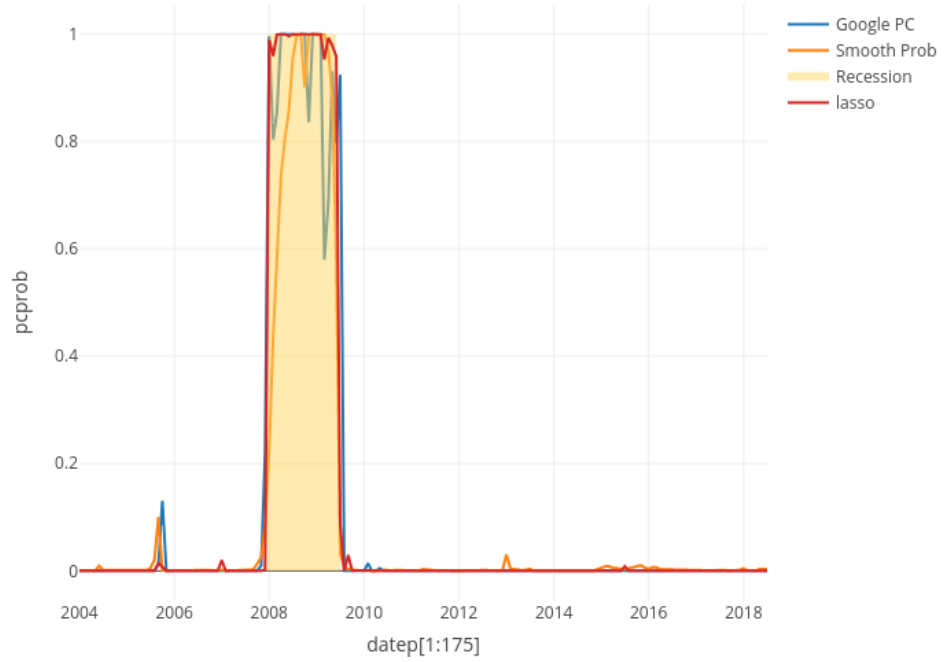

Figure 5: Probability of Recession: Google Searches - LASSO vs PC vs Smoothed Probabilities

6 Conclusions

In this paper I evaluate the use of Google searches as predictors of the probability of U.S. recessions. Applying the principal components method to reduce de dimensionality of the different time series used, I find that the probit model using the first five principal components as predictors is successful in 99.43% of the cases evaluated, outperforming the benchmark models. The model using LASSO logistic regression, however, is 100 % accurate, outperforming the other models.

I also find that the search for "recession" is a good predictor and could be subject of further research as empirical measure of attention.

The empirical findings in this paper have significant implications: a whole Google-based model could be implemented as a complement of the yield curve-based and smoothed probabilities model in order to have an early warning indicator system to face recessions. In this sense, the main claim of this paper is that using Google data and methods such as principal components and machine learning is a great combination that needs more room in forecasting and nowcasting economic variables.

There is still room for more research. We can increase the number of Google searches. The use of machine learning is still a work in progress.

References

[1] Baker, Scott & Fradkin, Andrey (2017) "The Impact of Unemployment Insurance on Job Search: Evidence from Google Search Data". *The Review of Economics and Statistics*, December 2017, 99(5): 756-768

[2] Chauvet, M., & Pigger, J. (2008) "A Comparison of the Real-Time Performance of Business Cycle Dating Methods". American Statistical Association. *Journal of Business & Economic Statistics*. January 2008, Vol. 26, No. 1

[3] Chauvet, M., & Potter, S. (2002) "Predicting a recession: evidence from the yield curve in the presence of structural breaks". *Economic Letters* 77 (2002) 245-253.

[4] Chauvet, M., & Potter, S. (2009) "Business Cycle Monitoring with Structural Change". *International Journal of Forecasting*. Volume 26, Issue 4, October?December 2010, Pages 777-793

[5] Choi, H., & Varian, H. (2012) "Predicting the Present with Google Trends". *THE ECONOMIC RECORD*, VOL. 88, SPECIAL ISSUE, JUNE, 2012, 2?9

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

[7] Estrella, A. & Mishkin, F. (1998) "Predicting U.S. Recessions: Financial Variables as Leading Indicators". *The Review of Economics and Statistics.* Vol. 80, No. 1 (Feb., 1998), pp. 45-61

[8] Friedman, J., Hastie, T & Tibshirani, R (2009) "Regularization Paths For Generalized Linear Models via Coordinate Descent". *Stanford University.* April 2009.

[9] Jun, S., Yoo, H., and 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 (2018) 69-87

[10] Kholodilin, K., Podstawski, M., Siliverstovs, B., & Burgi, C.,(2009) "Google searches as a means of imprioving the nowcasts of key macroeconomic variables". DIW, Working Paper, November 2009.

[11] Kholodilin, K., Podstawski, M., Siliverstovs, B., & Burgi, C.,(2009) "Google searches as a means of imprioving the nowcasts of key macroeconomic variables". DIW, Working Paper, November 2009.

[12] Kholodilin, K., Podstawski, M., & Siliverstovs, B., (2010) "Do Google Searches Help in Nowcasting Private Consumption". KOW Swiss Economic Institute, No. 256, April 2010.

[13] Stephen-Davidowiitz, S., & Varian, H. (2015) "A Hands-on Guide to Google Data". Google.

[14] Tiffin, Andrew, (2016) "Seeing in the Dark: A Machine-Learning Approach to Nowcasting in Lebanon". *IMF Working Paper* WP/16/56. March 2016.

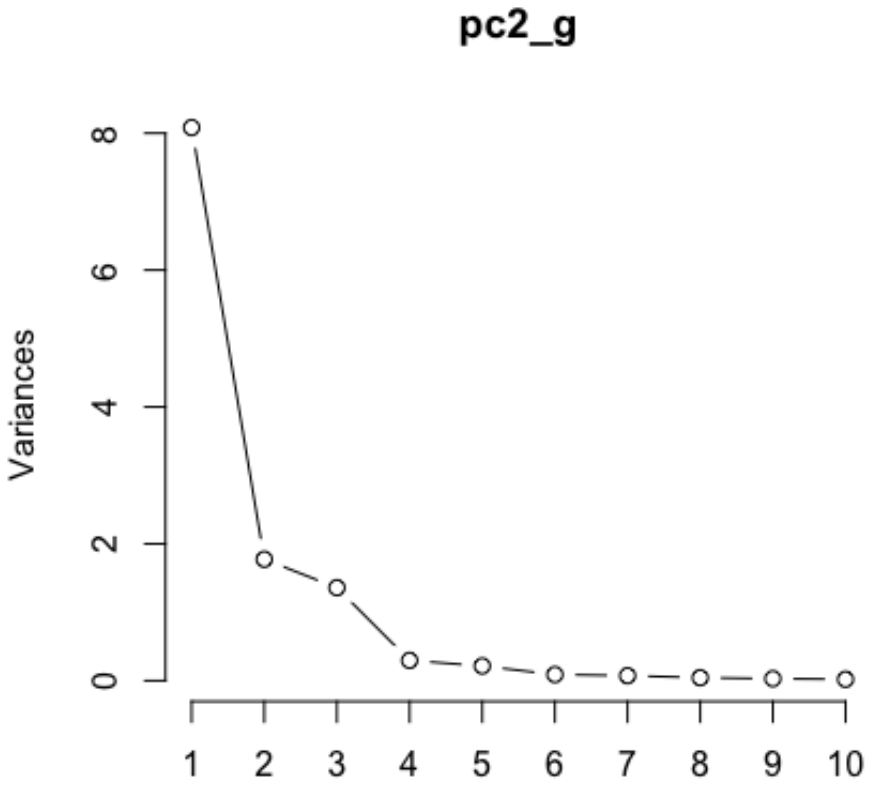

Figure 6: Principal Components Variances

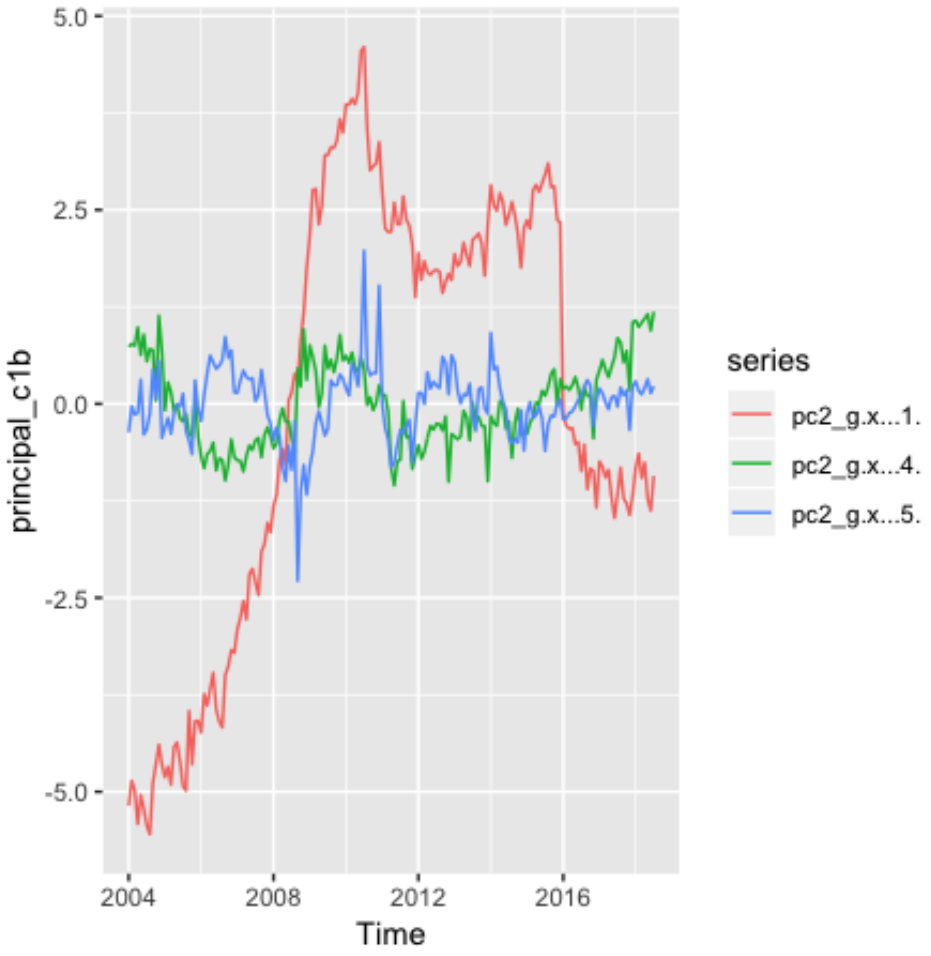

Figure 7: Principal Components 1, 4 and 5

Figure 8: Principal Components 2 and 3