Deutsches Institut für
Wirtschaftsforschung

DIW BERLIN

Discussion Papers

946

Konstantin A. Kholodilin • Maximilian Podstawski •
Boriss Siliverstovs • Constantin Bürgi

Google Searches as a Means of Improving the Nowcasts of Key Macroeconomic Variables

Berlin, November 2009

IMPRESSUM

DIW Berlin
German Institute for Economic Research
Mohrenstr. 58
10117 Berlin
Tel. +49 (30) 897 89-0
Fax +49 (30) 897 89-200
http://www.diw.de

ISSN print edition 1433-0210
ISSN electronic edition 1619-4535

Google searches as a means of improving the nowcasts of key macroeconomic variables‖

Konstantin A. Kholodilin* Maximilian Podstawski** Boriss Siliverstovs§

Constantin Bürgi¶

Abstract

The Google Insights data are a collection of recorded Internet searches for a huge number of the keywords, which are available since January 2004. These searches represent a kind of revealed perceptions of Internet users, which are a (possibly not entirely representative) sample of the general public. These data can be used to improve the short-term forecasts or nowcasts of various macroeconomic variables. In this paper, we compare the nowcasts of the growth rates of the real US private consumption based on both the conventional consumer confidence indicators and the Google indicators. The latter are extracted from the Google searches using the principal component analysis. It is shown that the Google indicators are especially successful at predicting private consumption in times of economic trouble, for they are 20% more accurate than the best alternative during the 2008m1-2009m5 forecast period. In addition, Google indicators are available at weekly frequency and not subject to revisions. This makes them an excellent source of information for the macroeconomic forecasting.

Keywords: Google indicators; forecasting; principal components; US private consumption.

JEL classification: C22, C53, C82.

‖The earlier version of this paper had been presented at a workshop "Neue Verfahren der Kurzfristprognose?", which took place in the German Ministry for Economy on July 30, 2009. We are grateful to the participants of the workshop for their insightful comments.

*DIW Berlin, Mohrenstrasse 58, 10117 Berlin, Germany, e-mail: kkholodilin@diw.de

**Universität Potsdam, Wirtschafts- und Sozialwissenschaftliche Fakultät, Potsdam, Germany, e-mail: mpod@gmx.net

§ETH Zurich, KOF Swiss Economic Institute, Weinbergstrasse 35, 8092 Zurich, Switzerland, e-mail: boriss.siliverstovs@kof.ethz.ch

¶London School of Economics, Department of Economics, London, United Kingdom, e-mail: corsbu@gmail.com

Contents

1 Introduction 1

2 Data 1

3 Construction of Google indicators 2

4 Forecasting models 3

5 Forecast evaluation 4

6 Conclusion 7

References 7

7 Appendix 8

List of Tables

1 Forecast accuracy (RMSFE) with Google indicators during "crisis period" (2008m1-2009m5) . . 8

2 Forecast accuracy (RMSFE) with sentiment indicators during "crisis period" (2008m1-2009m5) . 8

3 Forecast accuracy (RMSFE) of different combinations of forecasts during "crisis period" (2008m1-2009m5) 8

4 Forecast accuracy (RMSFE) with Google indicators during "non-crisis period" (2006m1-2007m12) 9

5 Forecast accuracy (RMSFE) with sentiment indicators during "non-crisis period" (2006m1-2007m12) 9

6 Forecast accuracy (RMSFE) of different combinations of forecasts during "non-crisis period" (2006m1-2007m12) 9

List of Figures

1 Actual growth rates of private consumption vs. the forecasts of two best models, 2008m1-2009m5 10

2 Actual growth rates of private consumption vs. the forecasts of two best models, 2006m1-2007m12 10

3 Histogram, distribution of the RMSFE during the "non-crisis" and the "crisis" period 11

1 Introduction

Recently, some attempts have been made to apply time series data on Google searches[1] in order to forecast labor market variables, such as the German unemployment rate (Askitas and Zimmermann (2009)) and the US unemployment rate by Google itself (Choi and Varian (2009)). The aim of this paper is, on the one hand, to improve the forecasts of another important macroeconomic variable, the US private consumption, using the Google data and, on the other hand, to contribute to the methodological development of the very young but promising field of Internet econometrics. A key property of Google data is their time advantage over other sentiment indicators. The Google search data are available in near to real time, that is, they have a time advantage of up to four weeks over the release of the sentiment indicators and even more if compared to the release of the national accounts data. Although conventional sentiment indicators, such as the well established consumer confidence indicators, are widely used and accepted, they perform poorly in times of high volatility and structural breaks. We suggest that our Google-based indicator provides similar and even higher forecast accuracy than conventional indicators, apart from being available at an earlier date.

2 Data

The dependent variable in this study is the year-on-year growth rate of monthly US private consumption. The source for monthly US private consumption data is FRED® database of the Federal Reserve Bank of St. Louis (http://research.stlouisfed.org/fred2/).

Two groups of leading indicators are used to forecast the US private consumption: 1) the data on Google searches and 2) the conventional leading indicators of consumer confidence, which are used as benchmarks to evaluate the forecast accuracy of the Google-based indicators.

Time series data on Google searches are available at weekly frequency from January 2004 onwards (http://www.google.com/insights/search/). The Google data are not subject to any revisions. Google normalizes each time series by dividing the count for each query by the total number of online search queries submitted during the week, resulting in a query fraction. A query fraction for the search query q is equivalent to the probability that a random search query submitted from a particular region at a particular time is exactly q (see

[1]In the following we will use the terms “Google data” and “Google searches” interchangeably.

Ginsberg et al. (2009)). Unlike Ginsberg et al. (2009) who had hundreds of computers at their disposal, our computational resources are quite limited. Therefore, we had to take a relatively small selection of the billions of Google searches. The data set used to predict the US private consumption was built in two steps. In a first step, a pool of search items provided by the top ten searches of each of the 27 main Google search categories and respective subcategories was collected. Then, we attempted to further eliminate the unrelated to the private consumption queries before fitting any models. Therefore, in the second step, we select only the search items from the entire pool of all top ten queries in the categories provided by Google Insights that are both statistically (have enough variability) and economically relevant from the viewpoint of private consumption. Following this algorithm we were able to construct a data set comprising of 220 consumption-relevant Google searches.

Sources for the two benchmark consumer confidence indicators are the Organization for Economic Co-operation and Development (OECD) and the US Conference Board. Both these indicators are published at monthly frequency.

3 Construction of Google indicators

Since Google Insights provides data at weekly frequency, while US private consumption data are published at monthly frequency, the Google searches time series must have to be aggregated. Due to the overlaps of weeks and months the data are interpolated to daily frequency by applying a spline methodology as a first step and subsequently aggregated to monthly frequency. Many Google search time series show a distinct seasonal pattern and thus make a seasonal adjustment necessary. This is done by applying an additive method based on the difference from the moving average. After testing for unit roots we find stationarity for a good portion of the data. For the remaining time series, stationarity is achieved by differencing the original series.

Given the large amount of the Google searches data we have collected, we need to reduce their dimensionality to a small and yet relevant number of regressors to be used in the forecasting the growth rate of US private consumption. This is achieved by applying the method of principal components, which allows us extracting a reduced number of common factors from the selected 220 time series. Based on information criteria, we retained the first three principal components.

4 Forecasting models

For all possible combinations of the principal components (henceforth "Google indicators") and for the two benchmark confidence indicators, five models are tested in two different subperiods, that is, before the crisis and during the crisis. Based on business cycle dates of the National Bureau of Economic Research (NBER), the former is defined as period from December 2001 to December 2007 and the latter as period from January 2008 to May 2009. The distinction between these two periods is supposed to allow us to evaluate the forecasting accuracy of the Google indicators both in normal and in turbulent times. The three baseline models are set up as follows, depending on what variables are included into the right hand side of equation as regressors:

1. contemporaneous leading indicators,
2. leading indicators and their lags, with the lag order being determined by the Akaike information criterion (AIC),
3. leading indicators, their lags, and autoregressive terms of the dependent variable, with the lag orders being determined by the AIC.

In addition, two special cases of Model 3 are tested: 3a) leading indicators, whose lag order is determined by the AIC plus an AR(1) term; and 3b) contemporaneous leading indicators plus an AR(1) term. In addition, a naive form of each of the above models is used, where the leading indicator is left out.

These models are summarized as follows:

$$\text{Model 1:} \quad y_t = \alpha + \beta x_t + \varepsilon_t \tag{1}$$

$$\text{Model 2:} \quad y_t = \alpha + \sum_{p=0}^{P} \beta_p x_{t-p} + \varepsilon_t \tag{2}$$

$$\text{Model 3:} \quad y_t = \alpha + \sum_{p=0}^{P} \beta_p x_{t-p} + \sum_{q=1}^{Q} \gamma_q y_{t-q} + \varepsilon_t \tag{3}$$

$$\text{Model 3a:} \quad y_t = \alpha + \sum_{p=0}^{P} \beta_p x_{t-p} + \gamma y_{t-1} + \varepsilon_t \tag{4}$$

$$\text{Model 3b:} \quad y_t = \alpha + \beta x_t + \gamma y_{t-1} + \varepsilon_t \tag{5}$$

where y_t is the year-on-year growth rate of the real private consumption; x_t is an indicator (Google indicator or one of the sentiment indicators); and ε_t is the error term.

In order to assure comparability of the results two different lengths of the confidence indicator series are used. Firstly, for the sake of comparison the estimation sample of all indicators is reduced to the period, for which the Google indicator series are available, that is, from 2004m1 onwards. Secondly, to account for the fact that the sentiment indicators are available for a much longer period than the data on Google searches[2] the forecasts involving the confidence indicators are performed using the entire available series.

The above models are used to make quasi-real-time nowcasts, i.e., forecasts for the current month using only information available up to that month. They are called "quasi-real-time", since the final (revised) data on the private consumption and not the real-time data are used in the estimation and forecast evaluation. The model parameters used for the out-of-sample nowcasts are estimated using an expanding window. That is, the whole period is split into an estimation subperiod (2004m1-2005m12) and forecasting subperiod (2006m1-2009m5). After the first forecast is made, the estimation period is extended by one observation — 2004m1-2006m1 — and the next forecast is made for 2006m2. The forecast accuracy of different models is evaluated using the root mean squared forecast error (RMSFE) measure.

5 Forecast evaluation

In the "crisis period", the most accurate out-of-sample forecast as measured by the RMSFE is obtained using the second and third principal components extracted from the Google searches as the leading indicator in the specification of Model 3 — see Tables 1 and 2. Its RMSFE is 0.71, which is 20% lower than that of its best competitor, that is, the model using the long series of the OECD consumer confidence indicator under the specification of Model 3b. The superiority of the Google forecasts becomes even more evident when the models involving the sentiment indicators are estimated using the shorter estimation window. In this setting, the forecast of the Google indicator is nearly 29% more accurate than the best performing sentiment indicator, the year-on-year growth rate of the OECD indicator, as measured by the RMSFE. Figure 1 compares the year-on-year growth rate of US private consumption and the two best forecasts: Google indicator under specification of

[2] The OECD consumer confidence indicator for the US is available and used from 1978m1 onwards, whereas the Conference Board consumer confidence indicator is used from 1997m6, although it is available from 1967 onwards.

Model 2 and OECD indicator under specification of Model 1. This plot reveals a further auspicious property of our indicator. The forecast based on the Google indicator consistently anticipates changes of the dependent variable, while the forecast based on the sentiment indicator only follows the movements with a lag. At the actual edge the Google indicator clearly points to a recovery of US private consumption, whereas the nowcast based on the sentiment indicator keeps falling. Nevertheless, it has to be stated that the autoregressive term contributes heavily to the quality of the forecast. Models 1 and 2, which are based exclusively on the indicators, perform poorly in the crisis period.

In the pre-crisis period, evidence is not as clear cut. In fact, the best forecast is obtained using Model 3 without any indicator, but is closely followed by the forecast based on the long series of the year-on-year growth rates of the Conference Board indicator in the specification of Model 3b, see Tables 4 and 5. The simple autoregressive process performs about 2% better than the best sentiment indicator and 15% better than the best forecast based on the Google indicator. Figure 2 compares the actual growth rates with the best Google-based forecast and the overall best forecast (in this case it is a purely autoregressive process). It can be seen that the AR model replicates the actual values much closer, while Google-based forecast produces especially large errors in the period 2006m8-2006m12. However, when creating similar conditions for the Google indicators and the alternative indicators by reducing the estimation window to the period, for which the Google series are available, both the Google indicators and the sentiment indicators produce forecasts with a similar accuracy. Thus, when using the short estimation sample, the best model based on Google indicators has a RMSFE of 0.50, which is slightly better than that of the best alternative, i.e., Conference Board confidence indicator (RMSFE=0.52).

As Tables 1 and 4 show, the forecast accuracy of the models based on the Google indicators (principal components of Google searches time series) is not always superior compared to alternative models and rather highly dispersed. Therefore, to overcome this drawback different combinations of forecasts are evaluated in order to provide a robust and applicable workhorse instrument for forecasters. The combinations are constructed as averages of the forecasts produced within each of the five models (equation (1) to (5)). Following combinations are constructed, listed in the order of the columns in Tables 3 and 6: (1) combination of the 220 individual Google search time series; (2) combination of the seven forecasts made by the Google indicators; (3) combination of the average of the forecasts made by the seven Google indicators and the average of the individual forecasts based on the 220 individual Google search time series, both with equal weights; (4) combination of the forecasts based on the short series of the sentiment indicators; (5) combination of the forecasts based on the long series of the sentiment indicators; (6) combination of the forecasts based on the short and the long series of the sentiment indicators; (7) combination of (1) to (4)[3]. The results are consistent with the evaluation of the underlying, not combined forecasts.

During the "crisis period" (see Table 3), the combination of forecasts made by Google indicators in the specification of Model 3 outperforms all other forecasts. The corresponding RMSFE of 0.81 is 16% lower than the lowest RMSFE of the combined sentiment-indicator-based forecasts. Depending on the specification of the model the average of individual forecasts based on single Google search series contributes only little to the quality of the combination forecasts or — as in majority of cases — even deteriorates it. Furthermore, except for Models 1 and 2 that are very poor anyway, a combination including Google-based forecasts always is more accurate than any of the combinations of the sentiment indicators. In the "crisis period", all the forecast combinations are less accurate than the best underlying, not combined forecast.

As opposed to the "crisis period", during the "non-crisis period" (see Table 6), the combining of forecasts leads to improvement of the forecasting accuracy in some cases. Thus, the best combined Google-based forecast, that is a combination of Google indicators and individual Google forecasts, is improved by 3% from an RMSFE of 0.50 to 0.48, while the best combination of sentiment indicators leads to an improvement of 5% from an RMSFE of 0.42 to 0.40. Again, combining all available forecasts based on Google and sentiment indicators as well as the individual Google search series, does not pay off.

In general, as one would have expected, the forecast accuracy of all the models is much worse in the "crisis period" than in the "non-crisis period" — see Figure 3, where the distributions of the RMSFEs of Google-based models and alternative models, including all the forecast combinations, are plotted for the two differing periods. For the "non-crisis period", the density of the distribution of the RMSFEs of forecasts based on Google searches is concentrated at RSMFE values lower than one, while the median is similar to the one of the benchmark forecasts. With regards to the "crisis period", the biggest mass (about 60% of the RMSFEs) in the distribution of Google-based RSMFEs is found in the lower part of the distribution, that is, around an $RMSFE = 1$. Notice

[3] With a weight of one for (1) and weights proportional to the underlying number of forecasts for (2), (3) and (4), that is 7, 5, and 5.

that for the Google-based forecasts 35% of the RMSFEs are below one, whereas only 25% of the RMSFEs of the alternative models are below one. The higher RMSFEs of the two distributions, grouping around a value of 4 are entirely due to the bad performance of of Models 1 and 2.

6 Conclusion

As was demonstrated, whereas in the normal, or "non-crisis", times the forecast accuracy of all the leading indicators (both Google indicators and sentiment indicators) is more or less the same, in the turbulent times the Google indicator is clearly superior to all its competitors. It allows predicting the changes in the growth rate of US private consumption more accurately. This, coupled with such an advantage of the Google data as their availability at high (weekly) frequency, makes them an excellent indicator for nowcasting the year-on-year growth rate of US private consumption and possibly other key macroeconomic variables in other countries.

References

Askitas, N. and K. F. Zimmermann (2009). Google econometrics and unemployment forecasting. IZA Discussion Papers 4201, Institute for the Study of Labor (IZA).

Choi, H. and H. Varian (2009). Predicting the present with Google trends. Technical report, Google Inc.

Ginsberg, J., M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant (2009). Detecting influenza epidemics using search engine query data. *Nature 457*, 1012–1014.

7 Appendix

Table 1: Forecast accuracy (RMSFE) with Google indicators during "crisis period" (2008m1-2009m5)

	(pc1)	(pc2)	(pc3)	(pc1 pc2)	(pc1 pc3)	(pc2 pc3)	(pc1 pc2 pc3)
Model 1:	4.39	3.38	4.20	3.59	4.33	**3.35**	3.51
Model 2:	5.20	3.25	4.33	4.15	5.25	**3.19**	3.87
Model 3:	1.04	1.12	1.01	1.09	1.04	**0.71**	0.89
Model 3a:	1.01	1.11	1.10	1.09	1.07	**0.97**	1.05
Model 3b:	1.00	1.01	0.97	0.97	0.95	0.87	**0.82**

Note: (pc1) is the model, which only includes the first principal component; (pc1 pc2) is the model, which includes the first two principal components.

Table 2: Forecast accuracy (RMSFE) with sentiment indicators during "crisis period" (2008m1-2009m5)

	Short sample					Long sample				
	OECD	Δ^{12}OECD	CB	Δ^{12}CB	AR	OECD	Δ^{12}OECD	CB	Δ^{12}CB	AR
Model 1:	2.90	4.14	**2.53**	3.18	4.24	6.21	4.78	**2.58**	3.10	5.41
Model 2:	**2.25**	3.14	2.39	2.32	4.24	6.26	4.23	**2.54**	2.39	5.41
Model 3:	1.28	1.14	1.29	1.28	**1.08**	**1.01**	1.12	1.12	1.15	1.03
Model 3a:	1.11	1.08	1.20	1.15	**1.04**	**0.91**	1.01	1.05	1.05	0.91
Model 3b:	1.11	**1.00**	1.07	1.16	1.04	**0.89**	1.00	1.06	1.00	0.91

Note: OECD (CB) is the log of the confidence indicator of the OECD (US Conference Board) confidence indicator; Δ^{12}OECD (CB) is the year-on-year monthly growth rate of the OECD (US Conference Board) confidence indicator; AR is the purely autoregressive model.

Table 3: Forecast accuracy (RMSFE) of different combinations of forecasts during "crisis period" (2008m1-2009m5)

	Google individual	Google indicators	Google indicators and Google individual	Sentiment indicators (short)	Sentiment indicators (long)	Sentiment indicators (all)	All indicators
Model 1:	4.22	3.80	4.00	**3.37**	4.31	3.84	3.83
Model 2:	4.24	4.09	4.16	**2.83**	4.07	3.44	3.73
Model 3:	1.01	**0.81**	0.88	1.15	1.07	1.06	0.93
Model 3a:	0.99	0.98	0.97	1.08	0.97	0.98	**0.96**
Model 3b:	1.01	**0.91**	0.95	1.05	0.96	0.97	0.94

Table 4: Forecast accuracy (RMSFE) with Google indicators during "non-crisis period" (2006m1-2007m12)

	(pc1)	(pc2)	(pc3)	(pc1 pc2)	(pc1 pc3)	(pc2 pc3)	(pc1 pc2 pc3)
Model 1:	0.69	0.69	**0.60**	0.71	0.62	0.61	0.63
Model 2:	0.71	0.69	0.58	0.69	**0.53**	0.63	0.55
Model 3:	0.59	0.60	**0.55**	0.82	0.76	0.83	0.90
Model 3a:	0.57	0.57	**0.54**	0.56	**0.54**	0.60	0.69
Model 3b:	0.55	0.56	**0.50**	0.57	0.51	**0.50**	0.53

Note: (pc1) is the model, which only includes the first principal component; (pc1 pc2) is the model, which includes the first two principal components.

Table 5: Forecast accuracy (RMSFE) with sentiment indicators during "non-crisis period" (2006m1-2007m12)

	Short sample					Long sample				
	OECD	Δ^{12}OECD	CB	Δ^{12}CB	AR	OECD	Δ^{12}OECD	CB	Δ^{12}CB	AR
Model 1:	0.66	0.66	**0.61**	0.68	0.67	1.56	1.37	**0.61**	0.77	1.48
Model 2:	0.65	0.73	**0.61**	0.68	0.67	1.52	1.33	**0.61**	0.76	1.48
Model 3:	0.56	0.63	**0.55**	0.56	0.57	0.46	0.50	0.43	0.90	**0.42**
Model 3a:	0.55	0.64	**0.52**	0.61	0.55	0.46	0.50	**0.45**	0.47	0.46
Model 3b:	0.55	0.55	**0.52**	0.56	0.55	0.46	0.50	0.45	**0.43**	0.46

Note: OECD (CB) is the log of the confidence indicator of the OECD (US Conference Board) confidence indicator; Δ^{12}OECD (CB) is the year-on-year monthly growth rate of the OECD (US Conference Board) confidence indicator; AR is the purely autoregressive model.

Table 6: Forecast accuracy (RMSFE) of different combinations of forecasts during "non-crisis period" (2006m1-2007m12)

	Google individual	Google indicators	Google indicators and Google individual	Sentiment indicators (short)	Sentiment indicators (long)	Sentiment indicators (all)	All indicators
Model 1:	0.66	**0.64**	0.65	0.65	1.02	0.83	0.74
Model 2:	0.65	**0.54**	0.58	0.64	1.02	0.82	0.69
Model 3:	0.51	0.57	0.53	0.55	**0.40**	0.44	0.48
Model 3a:	0.51	0.49	0.48	0.55	0.45	0.47	**0.45**
Model 3b:	0.54	0.52	0.53	0.54	**0.45**	0.46	0.48

Figure 1: Actual growth rates of private consumption vs. the forecasts of two best models, 2008m1-2009m5

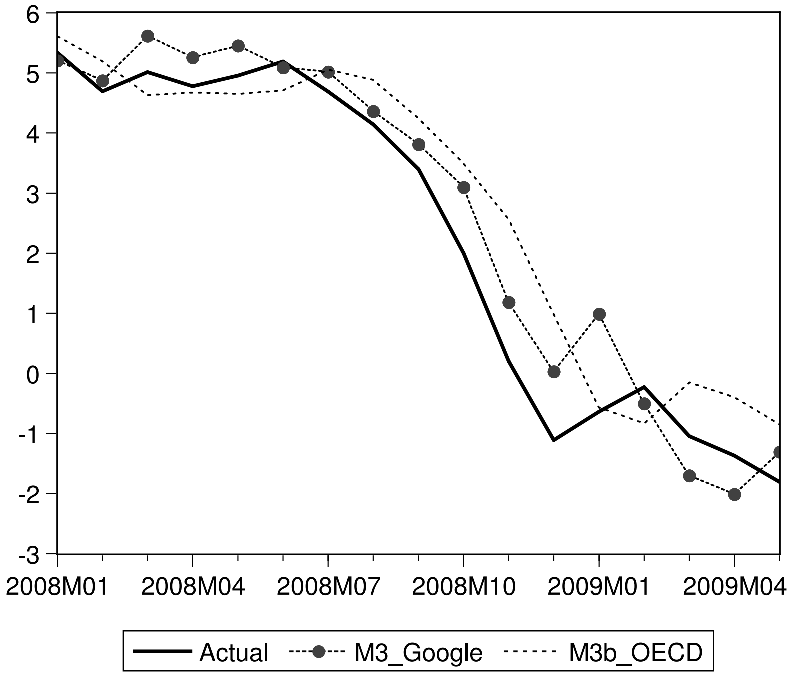

Figure 2: Actual growth rates of private consumption vs. the forecasts of two best models, 2006m1-2007m12

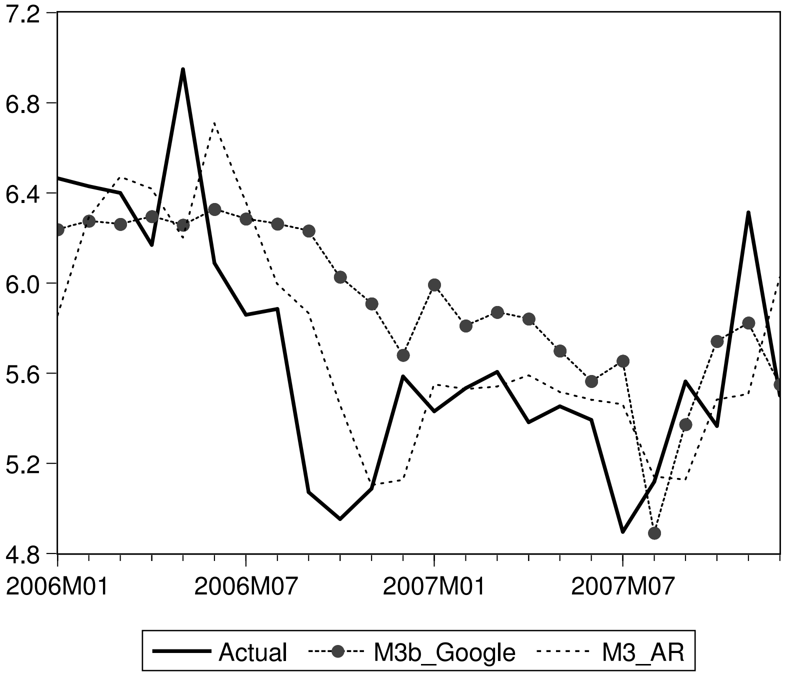

Figure 3: Histogram, distribution of the RMSFE during the "non-crisis" and the "crisis" period