Irving Fisher Committee on
Central Bank Statistics

BANK FOR INTERNATIONAL SETTLEMENTS

Ninth IFC Conference on "Are post-crisis statistical initiatives completed?"

Basel, 30-31 August 2018

Developments in the residential mortgage market in Germany - what can Google data tell us?[1]

Simon Oehler,
Deutsche Bundesbank

[1] This paper was prepared for the meeting. The views expressed are those of the authors and do not necessarily reflect the views of the BIS, the IFC or the central banks and other institutions represented at the meeting.

Developments in the residential mortgage market in Germany – What can Google data tell us?

Session 5 – Big Data

Simon Oehler, Deutsche Bundesbank[1]

Abstract

This paper investigates the explanatory power of aggregate, publicly available Google Trends data for developments in the German residential mortgage market. For many consumption goods and services the internet serves as a means for households to acquire relevant information for example on prices, quality characteristics or legal and contractual conditions in advance of an actual purchase decision. Thus, households are also likely to rely largely on the internet (and in particular on search engines) in order to retrieve relevant information about potential providers of mortgage financing and the respective contractual conditions in the run-up to an actual loan agreement.
As households may subjectively choose different search terms in order to obtain (possibly the same) information on mortgage financing, the usefulness of several Google indicators, each representing the relative interest for a specific search term, is evaluated with respect to their predictive power for monthly changes in new mortgage business provided by banks to households in Germany. The performance of out-of-sample forecasts suggests that aggregate Google Trends data has the potential to serve as a valuable source of information for the prediction of mortgage market developments.[2]

Keywords: Forecasting, Internet search data, Google Trends, Google econometrics, residential mortgage markets, housing markets

JEL classification: C22, C52, C53

[1] Simon Oehler, Deutsche Bundesbank, Statistics Department, Email: simon.oehler@bundesbank.de

[2] The paper represents the author's personal opinions and does not necessarily reflect the views of the Deutsche Bundesbank or its staff.

Contents

1. Introduction 3

2. Data 4

3. Models 7

4. Results 8

5. Conclusion 10

References 12

Appendix 13

1. Introduction

Housing markets play an important role in most economies as a large share of overall economic activity is related to housing. Consequently, from a financial perspective, mortgage markets play an important role as mortgage debt often makes up for a large share of total outstanding aggregate debt in an economy. Thus, profound and timely analysis of mortgage market developments is of great importance for the whole economy and financial markets in general.

On a micro level, the housing market is of prime importance for many reasons. First, for many households dwelling accounts for a large share of their monthly expenditures, be they either for rental or mortgage rate payments. Second, for many households financing their property, mortgages make up the largest share of their total indebtedness. And third, the property at the same time often serves as a major part of a households' retirement provisions.

Therefore, and not least since the global financial crisis, analysts, policy makers and the wider public payed particular attention to mortgage market developments, resulting in an increased demand for detailed and timely information on this market segment.

One potential source of information on housing or mortgage markets in general is the internet. In most developed economies a large majority of households has access to the internet[3] which is used intensively for a vast amount of various economic activities. One of such is the acquisition of information in advance of purchase decisions. These can relate, for example, to consumption goods or (financial) services.

Most internet users worldwide rely on the Google search engine for information retrieval on the web. Accordingly, about 90% of all internet searches worldwide are currently processed by Google, with a similar share for Germany.[4] However, only few detailed official statistics exist on this topic. It is reasonable to assume that aggregate Google search data may contain useful signals on (changing) interests in certain topics, products, services etc. As these aggregate and anonymized data are published by Google Trends almost without any time lag, they could, in many cases, contain timely information exploitable e.g. for nowcasts or forecasts of demand for specific products & services.

In recent years, interest in internet search data has substantially increased and a number of research projects have made use of such data, in particular for forecasting exercises and the construction of novel indicators. For example, Choi and Varian (2009a,b) investigate the predictive power of Google data for nowcasting retail & automobile sales, home sales, travel activity and unemployment. At the Bank of England, McLaren and Shanbhogue (2011) have used internet search data for the analysis of labour and housing markets. Several papers have also investigated the potential of these data for the analysis of mortgage markets. Examples comprise Askitas and Zimmermann (2014) and Chauvet et al. (2016) who

[3] In Germany 87% of the overall population above the age of 10 years are internet users. For all individuals below the age of 45 years the share of internet users is almost 100%. See: https://www.destatis.de/EN/FactsFigures/SocietyState/IncomeConsumptionLivingConditions/UseInformationTechnologies/UseInformationTechnologies.html

[4] See: https://www.statista.com/statistics/216573/worldwide-market-share-of-search-engines/ and https://www.statista.com/statistics/445002/market-shares-leading-search-engines-germany/

seek to derive early indicators for mortgage delinquencies and default risk. Another example is Saxa (2014), who uses Google Trends data to forecast mortgage lending in the Czech Republic.

This paper contributes to the above mentioned literature as it evaluates the explanatory power of Google data for developments in the residential mortgage market in Germany, which, to my knowledge, has not been done so far. Moreover, this approach tries to strike a balance between pure data driven approaches (in the sense that the selection of potential predictors is solely based on the degree of correlation with a dependent variable of interest) and a "narrow" selection of search terms (i.e. selection of one or only very few search terms which are assumed to be thematically related to the object of interest). Thus, for this analysis, a set of search terms is selected from Google Trends, which is a priori restricted to the topic of mortgage lending. This approach significantly narrows down the number of potential predictors of actual mortgage business and is intended to help avoid finding possibly "spurious" correlations between the variable of interest and a very large set of Google Indicators. Consequently, this paper proposes a selection of specific search terms that could be of particular relevance for the construction of a mortgage market indicator for Germany. Further, and contrary to similar approaches, several "control" variables are included in the model specifications to account for other indicators that may have predictive power for mortgage market developments.

The remainder of the paper is structured as follows. In chapter 2 the Google data and some of their most important properties, including also potential drawbacks, are described. In chapter 3 model specifications and the variable selection procedure for the Google time series are explained. The respective results of in-sample as well as out-of-sample predictions are discussed in chapter 4. Chapter 5 concludes and provides an outlook on future work.

2. Data

The data on mortgage business are publicly available on the homepage of the Deutsche Bundesbank and comprise the volumes of new mortgages provided by banks to private households in Germany at a monthly frequency. Several breakdowns according to the respective agreed duration of a mortgage loan are available. In Germany mortgage contracts with durations longer than 5 years are largely prevailing. Nevertheless, as this project aims at evaluating German mortgage business as a whole, the aggregate over all agreed loan durations is taken into account. Thus, the respective interest rate, as far as considered, is a weighted average of interest rates recorded over all durations. As far as unemployment is considered as macroeconomic variable the respective time series contains the absolute number of unemployed persons according to the German law at a monthly frequency. The series is downloaded from the Bundesbank homepage.

All Google time series are downloaded from Google Trends' public website.[5] Each Google series obtained represents, in aggregate and anonymized form, the interest

[5] https://trends.google.de/trends/?geo=DE

of Google users in a specific search term over time.[6] More precisely, each series represents the number of queries for a specific search term relative to the overall number of all Google searches at a specific point in time, i.e. within a week or a month, and within a specific geographic region, i.e. at country or state level. Moreover, the data are provided as an index which means that the data is normalized to its maximum value equalling 100. This value corresponds to the highest relative "search intensity" over the sample period. As an example, the Google Index at period t for the German search term "Kredit" is computed as follows:

$$\boldsymbol{I(Kredit_t)} = \frac{\boldsymbol{R(Kredit_t)}}{\boldsymbol{\max\{R(Kredit_\tau)\}}} \times \mathbf{100} \text{ with } R(Kredit_t) = \frac{Kredit_t}{Google_t} \text{ and } \tau = 1, \ \ldots, T$$

In practice this index is computed by using a random sample of all Google searches rather than actual Google searches. The actual sample size used for the computation of the index is not reported. It is important to mention, that the sampling mechanism could threat the stability of some of the Google series. Practically, if a series corresponding to a specific search term is downloaded at different times, the results often vary slightly. Even though not every time series in this project is checked accordingly, it seems that most of the time series used here are rather stable in this sense.[7] However, McLaren and Shanbhogue (2011) report that this issue may occur in particular with less popular search terms. One possible cure for this shortcoming is to download each series several times and simply compute averages over all observations.

Further, Google Trends allows for the use of search operators which make it possible to further refine search results according to specific needs. For example, entering the term "Kredit" without apostrophe's yields results for all searches containing this specific term including related extensions of the word, for example, "Hauskredit". On the contrary, adding apostrophes to a search term would only yield results for searches exactly matching a specific term, thus not including any related or but slightly different words. Additionally it is possible to also use the operators "+" and "-" in order to explicitly include or exclude specific expressions from a particular search query. Such modifications are important as the respective results often differ substantially.

Moreover, Google Trends does not consider spelling errors. Therefore, in order to take account of such issues the explicit mentioning of a misspelled search term is necessary. Thus, if a search term were known to be often misspelled in a particular way or context, an explicit mentioning as part of a query would be a feasible way to account for potential bias due to such an error. However, if there is no knowledge of this sort, such an approach can cause biases by its own. Consequently, as for the relevant search terms used in this paper, typos and respective probabilities of this sort are not known, they are not accounted for. On the contrary, Google often suggests users corrected search terms if misspellings are detected. Thus, "following"

[6] For further details on how Google Trends data are adjusted, see the following link: https://support.google.com/trends/answer/4365533?hl=en&ref_topic=6248052

[7] To test for this particular issue, series representing identical search terms were downloaded at different times. The computed respective correlations were around 95% or above.

such a suggestion on Google's website corrects the search term and redirects the user to the results for the presumably intended query.
Further, Google Trends also allows to only consider searches within a specific category as for example "Finance". This can be of particular relevance if search terms have ambiguous meanings. By restricting search results to a specific category, "noisy" searches which, in essence, are unrelated to the topic of interest, are excluded.[8] Even though, most of the search terms in this approach do not have particular ambiguous meanings, all time series extracted from Google trends are restricted to the category "Finance". This assures however, that noise from unrelated terms is unintentionally included.
As a result of these properties, the Google data used in this paper are selected according to specific considerations and restrictions. First, the selection of search terms is not solely data driven. Rather, the selection is a priori restricted to terms which are in a logical sense connected to mortgage financing topics, as far as this can be assured. Additionally, both, rather general terms related to the topic of mortgage financing as well as more specific variations of such topics are considered. For example "Kredit" is taken into account together with related "extensions" as "Kreditvergleich", "Wohnungsbaukredit" etc. This means, however, that overlaps in searches can exist. For example, a specific search for "Wohnungsbaukredit" should also be comprised in a search results for the term "Kredit". Nevertheless, this approach allows investigating, whether narrowly defined searches can outperform broadly defined, generic searches.
Overall, 37 times series are obtained from Google Trends. By definition, these terms only account for searches in German language. By selection, only searches for the German geography are considered, i.e. searches in German language conducted in other countries, in particular Switzerland or Austria, are not included in the sample. The data is of monthly frequency, the sample ranging from January 2004 until April 2018.
Graphical representations of selected, unadjusted time series are depicted in Graph 5 in the appendix below.

All time series, except from the interest rates, show seasonal patterns of different sort. The series for unemployment, however, is already seasonally adjusted when downloaded from the Bundesbank homepage.
The series for new mortgage business shows a strong yearly seasonal pattern in form of a sharp rise in mortgage volumes in the mid of the year. Therefore, July often is the month with the largest volume of new mortgage business throughout the year. Therefore, the series is seasonally adjusted using the X-11 decomposition method.
Also the Google time series show a particular seasonality. The most striking seasonal pattern, which is common to all Google series in the sample, is a sharp decline of the Google index in December with a subsequent strong rise in January. This pattern, however, does not seem to be a peculiarity of search terms related to mortgage topics as also several other studies using Google Trends data report similar observations. Thus, it is reasonable to assume that this seasonality is not originating from actual seasonal patterns in Google users' interest for mortgages but from other, and in particular "economically" unrelated, sources. For example calendar effects could be one possible explanation. As it was discussed above, the

[8] For example the search term "Jaguar" could be either intended to find information about the animal or a car manufacturer.

Google data is essentially an index constructed as the ratio of searches for a particular term over all searches conducted at a given time and region. Thus, the reported seasonal pattern for the months December and January could be related to the Christmas season, as around this time, people simply use Google more intensely in preparation for various Christmas related activities. The result would be an inflated denominator of the Google index in December. On the contrary, in January after the Christmas season this effect disappears again, resulting in the above mentioned rise of the indicators. A similar explanation, however related to the nominator of the ratio, would be that people towards the end of the year tend to look for other than mortgage related topics thus the Google index would decline even if the total number of searches were constant over time. To avoid potential biases due to this type of seasonality all Google time series are seasonally adjusted using the X-11 decomposition method.

In order to check for potential unit roots in the time series, augmented Dickey-Fuller tests were performed for all series. After log-transforming and first-differencing[9] each series, the Dickey-Fuller test statistics strongly rejected the Null-hypothesis for the presence of unit roots, thus implying covariance stationarity for all series.

3. Models

In order to assess the usefulness of Google data with respect to its explanatory power for cyclical variation in the growth rate of mortgage volumes, different single equation linear models are estimated and consequently benchmarked against each other in terms of out-of-sample forecast performance. The models estimated can essentially be divided into three groups. First, an autoregressive model is estimated by simply regressing current values of the dependent variable on lagged terms of its own. Thus, the resulting model is of the form:

$$\Delta \boldsymbol{mortgages}_t = \boldsymbol{\beta}_m \boldsymbol{L}^m \Delta \boldsymbol{mortgages}_t$$

Second, this simple autoregressive model is augmented by current and lagged values of the series for the growth rate of the mortgage interest rate as well as the monthly growth rate of unemployment, as the latter is supposed to be an important indicator for households' willingness and ability to take up a loan for the financing of real estate.

$$\Delta \boldsymbol{mortgages}_t = \boldsymbol{\beta}_m \boldsymbol{L}^m \Delta \boldsymbol{mortgages}_t + \boldsymbol{\gamma}_m \boldsymbol{L}^m \Delta \boldsymbol{interest}_t + \boldsymbol{\theta}_m \boldsymbol{L}^m \Delta \boldsymbol{unempl}_t$$

Third, the model including the control variables is augmented by a set of Google terms representing the (lagged) growth rate of households' interest in mortgages, chosen according to the below mentioned procedure.

[9] None of the series was integrated of an order higher than one.

$$\Delta mortgages_t = \beta_m L^m \Delta mortgages_t + \gamma_m L^m \Delta interest_t + \theta_m L^m \Delta unempl_t + \delta_m L^m \Delta Google_t$$

For each of the above mentioned steps the model selection is performed with the help of a stepwise forward selection procedure in order to detect relevant regressors to be included in a model. The variables which are selected at a specific stage are kept in the model. Subsequently, additional potential regressors are added to the list of search regressors in the next stage and so forth. Thus at each stage, additional regressors potentially enter the model. The relatively large number of potential Google indicators results in a situation of a large number of potential models. Thus, for the Google indicators the procedure is twofold: First, the benchmark model, including the control variables, is estimated by adding a Google indicator (and its respective lags) one at a time to the list of search regressors in order to let the forward selection algorithm detect relevant lags. Those indicators, for which lags have been selected into the model, are kept aside and are further evaluated in the second stage. This procedure results in a reduced set of seven indicators. Subsequently, these preselected Google indicators are now used to find models for the forecasting exercise. For this purpose, the benchmark model including the controls is augmented by the preselected Google indicators which are allowed to be forward selected into the model one at a time. The resulting selected lags at a specific stage are then added to the model and additional indicators are again allowed to be forward selected. This procedure results in models including an increasing number of Google indicators as far as they prove to be stable and significant throughout the selection procedure.
This model selection and estimation procedure is performed on a subsample of the data ranging from January 2004 to December 2015 in order to treat the remaining part of the sample from January 2016 to April 2018 as "unseen" data, thus preserving it for out-of-sample forecast evaluation. Results are reported in section 4 as well as in the Appendix below.

4. Results

A summary of the estimation results is depicted in table 1 below. The benchmark model includes only lagged values of the dependent variable, i.e. actual volumes of mortgages provided to households each month. The results show, that the strongest autoregressive predictors are the first and third lag of the mortgage series as well as a more distant seventh lag. Overall this model captures around one third of the series' variation, where trend and seasonal components have already been accounted for.[10] In a second step additional regressors are allowed to enter the model, potentially containing useful information with respect to the growth in volumes of mortgages. Thus, the model is again stepwise forward selected, allowing for current and lagged values of the monthly change in interest rates for mortgages and the monthly changes in unemployment in Germany. The interest rate enters the model with a lag of two month and is, as expected, indicating a negative relationship between the growth in mortgage volumes and the respective changes

[10] Regressing the mortgage series on a linear trend variable and on a set of deterministic monthly dummy variables, 79% of the series' variance is explained in terms of R^2.

in interest rates. Contrary to this finding, (lagged) changes in unemployment do not enter the model.
The benchmark model, including the lagged interest rate, is then stepwise augmented by the Google regressors according to the procedure described above in Section 3. The model selection procedure ultimately choses the model which is labelled as "Google augmented III" incorporating lagged values of different search terms. These are "Baufinanzierung", "Hypothek" and "Kreditvergleich" and "Bauzins". The Google regressors which are included in the models, as reported below, thus have all proven to enter the benchmark model individually and in combination with other Google indicators and control variables.
Additionally to this, robustness tests are performed by either including or removing search terms from the models and by estimating the selected models on different sub-periods of the sample. The models reported here, prove to be robust to these tests.

Following the in-sample model selection and estimation procedure, the out-of-sample performance of the selected models is evaluated. The results of this exercise are summarized in table 2 below. Notably, the specification "Google augmented I" incorporates the search terms "Baufinanzierung" as well as "Kreditvergleich". For this model all Google terms show the expected positive signs.
As mentioned before, the models are estimated on a subsample of the data. The estimation period ranges from 2004M9 to 2015M12 including 136 observations. The remainder of the sample is kept aside for performing out-of-sample forecast evaluation of the models. Thus out-of-sample forecasts are produced for the period 2016M1 to 2018M4, i.e. 28 out-of-sample estimates are obtained. Subsequently, forecasts are compared to actual observations of this period and standard measures of forecast accuracy are calculated to evaluate the performance of the models. Results are depicted in table 2 in the appendix. With respect to the forecast evaluation criteria, the model including lags of all three search terms clearly outperforms the other models, in particular those not incorporating any Google information. However, the model "Google augmented I" containing additional "hard" information does not outperform the "pure" benchmark model without additional independent variables. Concretely, for the out-of-sample period reported above, the model "Google augmented II", including three Google search terms, outperforms the benchmark model by about 19 % and the benchmark model including additional "hard" economic data, even by 27 % in terms of the reported Root Mean Squared Error. An additional improvement in terms of forecast accuracy can be obtained by including the term "Bauzins" as reported in the model "Google augmented III". Further results for other models as well as other forecast evaluation measures are reported as well in table 1 and table 2 below. A graphical representation is depicted in Graph 1 in the appendix.
Thus far, dynamic forecasts have been computed, essentially assuming, that all forecasts had been performed in December 2015 (i.e. 2015M12) based on information which was available at this time. Consequently, in order to compute forecasts over the whole sample from 2016M1 to 2018M4 forecasted values are used to compute out-of-sample forecasts for more distant future periods.
Additionally, static forecasts are computed relating to the case as if from December 2015 on, each month a one-step-ahead out-of-sample forecast exercise would have been performed using all information available until a specific month, i.e. actual past values of growth in mortgage volume rather than forecasted values. The respective results for several models are presented in Graph 2 to 4 in the appendix below. The

main results are similar to the previous exercise. The improvement of the forecasts of the Google augmented model II over the forecasting period is about 17% relative to the benchmark model. The graphs of the forecasted series and further evaluation criteria are depicted as well in Graphs 2 to 4 below.

5. Conclusion

In this paper the usefulness of aggregate, publicly available Google Trends data as an indicator for developments on the residential mortgage market is evaluated. Google search data is considered a proxy for consumers' interest in certain topics for example in the course of planning activities in relation to large household purchases. For this purpose, a number of time series are downloaded from Google Trends, each representing the interest of Google users in specific topics over time.
First, a purely autoregressive model for the mortgage time series is estimated which is then augmented by distributed lags of the mortgage interest rates and unemployment as a macroeconomic indicator. Subsequently, Google data is included into the benchmark autoregressive distributed lag model. In this respect, a stepwise forward-selection procedure is applied in order to detect relevant Google predictors out of a large set of potentially useful regressors and to choose from an accordingly large set of potential models.
The results indicate that Google data has the potential to improve out-of-sample forecast accuracy for the monthly growth rate in mortgage volumes. Moreover, the search terms "Baufinanzierung", "Bauzins" "Hypothek" and "Kreditvergleich" seem to be particularly useful in terms of improving out-of-sample forecast accuracy. Further, checks revealed that the indicators are robust to different model specifications and different sample periods on which the models are estimated. On the contrary, other variables can be excluded for almost all model specifications. Thus, the number of potential relevant search terms is narrowed down to only a small subset of four indicators containing useful information, constituting another result of the analyses conducted so far.
However, to some degree, the results at this stage need further robustness checks. One exercise to be done is augmenting the benchmark model with survey data on the mortgage market and subsequently comparing such model specifications to Google augmented models. In this respect also further "hard" economic data needs to be evaluated as potential predictors. Relating to the problem of "model uncertainty" there is variety of econometric methods related to "shrinkage" and dimensionality reduction techniques which likely hold potential for this "data rich" setting too. Thus, for example, Ridge and Lasso regressions seem "natural" candidates to be applied here. Another check at this stage would be to test if the results obtained for the aggregate mortgage business series as dependent variable are valid for the mortgage series broken down according to the different agreed interest rate durations.
A great advantage of Google data in general is their coverage of a wide variety of topics which are of interest for the public. Along with the real time availability of data, this makes Google Trends an interesting tool for economics and social sciences in general.
However, some drawbacks remain most of which have already been described by the literature in this field. As Google data does not incorporate any survey design, e.g. the representativeness of the data may be an issue. However, a large share of the population in many countries uses the internet, of which again the largest share

uses Google as the preferred search engine. Further Google is reported as an index, thus leaving open the question whether variance in the time series is attributed to changes in searches for a specific term (nominator) or overall search activity (denominator). Some of the drawbacks can potentially be cured by statistical methods; others need further investigation in order to clear out potential issues.

Overall, as this paper and the related literature shows, Google data incorporates useful information, for example in form of short term cyclicality, which can be exploited for nowcasting and forecasting. As Choi and Varian mention, this particularly works if consumers start planning purchases significantly in advance. Reasonably, this may specifically be the case for major household purchases, as mortgage lending is but one example. Thus, exploring the use and possibilities of Google data seems to be promising for the future, also in other fields of financial services. Given the flexibility and timely availability of this data source, it may serve researchers and analysts as a first insight into a specific topic of interest and beyond.

References

Choi, H., Varian, H. (2009). Predicting the Present with Google Trends. Google Research Blog http://googleresearch.blogspot.com/2009/04/predicting-present-with-google-trends.html.

Choi, H., Varian, H. (2009). Predicting Initial Claims for Unemployment Benefits. Available at SSRN: https://ssrn.com/abstract=1659307.

McLaren, N., Shanbhogue, R. (2011). Using internet search data as economic indicators. Bank of England Quarterly Bulletin Q2.

Askitas, N., Zimmermann, K. (2011). Detecting Mortgage Delinquencies with Google Trends. IZA Discussion Paper No. 5895.

Chauvet, M., Gabriel, S., Lutz, C. (2016). Mortgage default risk: New evidence from internet search queries. Journal of Urban Economics. Vol. 96. No. 91 – 111.

Saxa, B. (2014). Forecasting Mortgages: Internet Search Data as a Proxy for Mortgage Credit Demand. CNB Working Paper Series 14.

Appendix

Table 1

Estimation Results

Dependent variable – new mortgage business, volumes, growth rates

	Benchmark	Incl. Controls	Google augmented I	Google augmented II	Google Augmented III
$\Delta mortgages_{t-1}$	-0.34***	-0.57***	-0.56***	-0.50***	-0.52***
	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)
$\Delta mortgages_{t-2}$		-0.34***	-0.35***	-0.39***	-0.38***
		(0.07)	(0.07)	(0.07)	(0.07)
$\Delta mortgages_{t-3}$	0.29***				
	(0.08)				
$\Delta mortgages_{t-7}$	-0.18**	-0.21***	-0.20***	-0.27***	-0.27***
	(0.08)	(0.07)	(0.07)	(0.06)	(0.06)
$\Delta interest_{t-2}$		-1.11***	-1.02***	-0.99***	-1.02***
		(0.23)	(0.22)	(0.20)	(0.20)
$\Delta Baufinanzierung_{t-1}$			0.16***	0.16***	0.13***
			(0.05)	(0.04)	(0.04)
$\Delta Baufinanzierung_{t-3}$			0.12***	0.12***	0.11***
			(0.05)	(0.04)	(0.04)
$\Delta Hypothek_{t-1}$				-0.04**	-0.06***
				(0.02)	(0.02)
$\Delta Hypothek_{t-3}$				-0.08***	-0.09***
				(0.02)	(0.02)
$\Delta Kreditvergleich_{t-3}$			0.06***	0.08***	0.08***
			(0.03)	(0.03)	(0.03)
$\Delta Bauzins_{t-1}$					0.03***
					(0.01)
R^2	0.32	0.41	0.49	0.55	0.58
$Adj.R^2$	0.31	0.40	0.47	0.53	0.55
$DW\ stat$	2.09	1.91	1.96	1.99	2.03
$Sample\ Period$	2004M09	2004M09	2004M09	2004M09	2004M09
	- 2015M12	- 2015M12	- 2015M12	- 2015M12	- 2015M12
$Num.Obs.$	136	136	136	136	136

Note: All time series are log-linearized and differenced.

Forecast Evaluation Table 2

Sample: 2016M01 2018M04
Included observations: 28
Evaluation sample: 2016M01 2018M04
Number of forecasts: 7

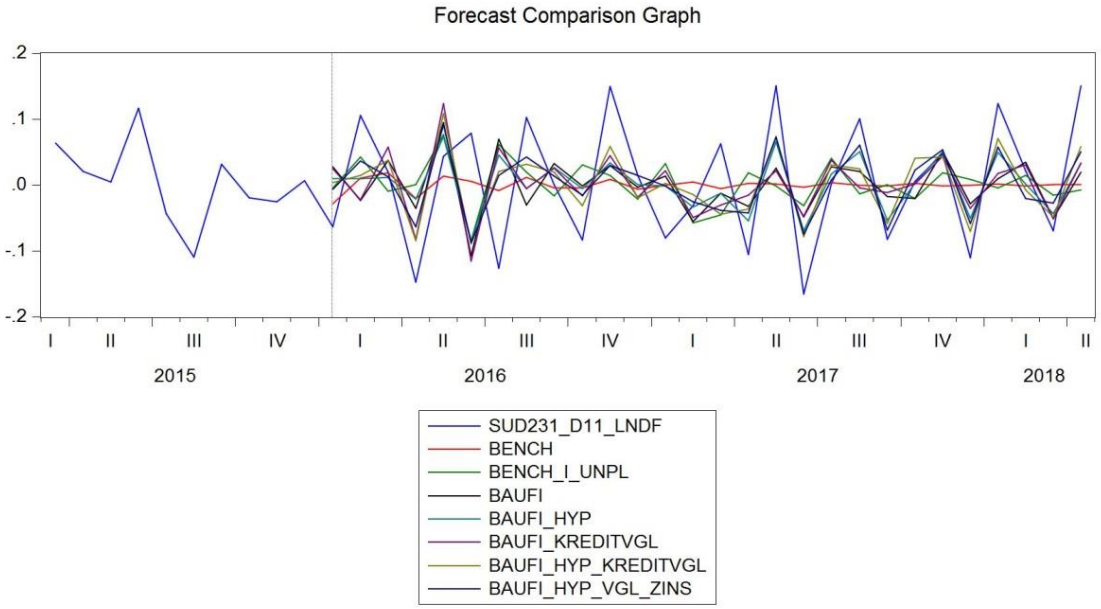
Combination tests
Null hypothesis: Forecast i includes all information contained in others

Equation	F-stat	F-prob
BENCH	7.933748	0.0001
BENCH_I_UNPL	5.851299	0.0010
BAUFI	10.54317	0.0000
BAUFI_HYP	6.456631	0.0006
BAUFI_KREDITVGL	10.02877	0.0000
BAUFI_HYP_KREDITVGL	4.755206	0.0033
BAUFI_HYP_VGL_ZINS	4.376260	0.0051

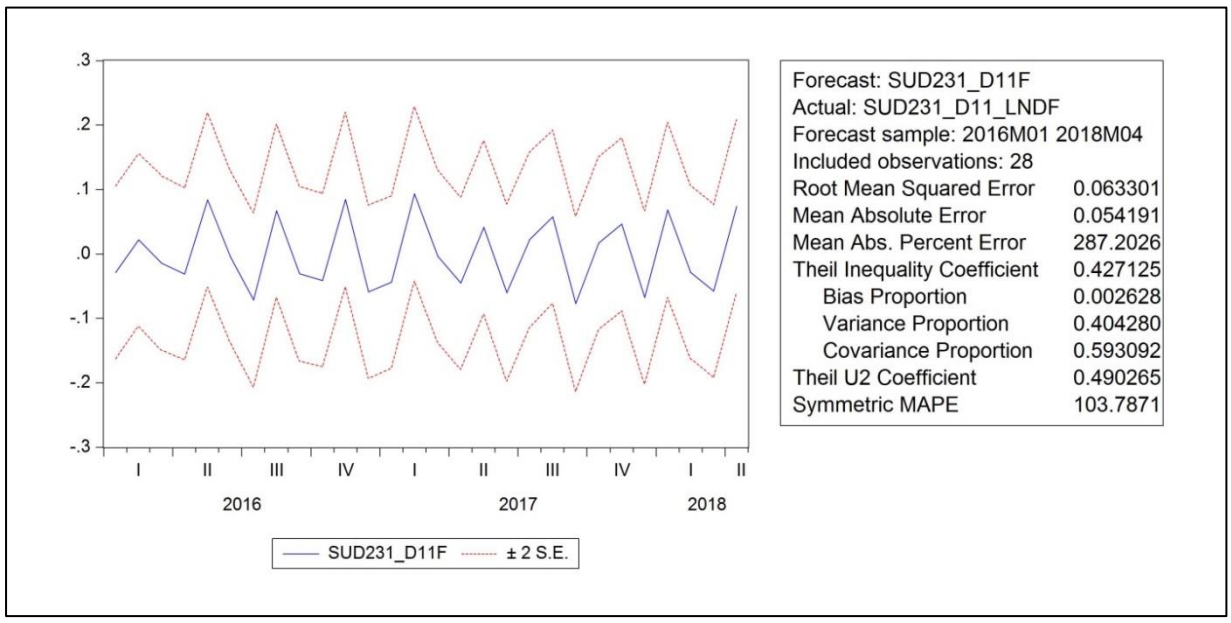
Evaluation statistics

Forecast	RMSE	MAE	MAPE	SMAPE	Theil U1	Theil U2
BENCH	0.090353	0.076160	112.5150	167.2447	0.874054	0.957484
BENCH_I_UNPL	0.102884	0.088681	282.8204	166.0138	0.803323	1.298619
BAUFI	0.095041	0.079783	306.7584	141.2235	0.698204	0.927898
BAUFI_HYP	0.079904	0.064154	222.1042	120.8835	0.590645	0.839572
BAUFI_KREDITVGL	0.093064	0.078871	290.7800	136.7531	0.656967	0.857488
BAUFI_HYP_KREDITVGL	0.072731	0.061808	252.6928	116.4031	0.498878	0.658610
BAUFI_HYP_VGL_ZINS	0.072015	0.058441	149.0011	102.9245	0.509604	0.726737

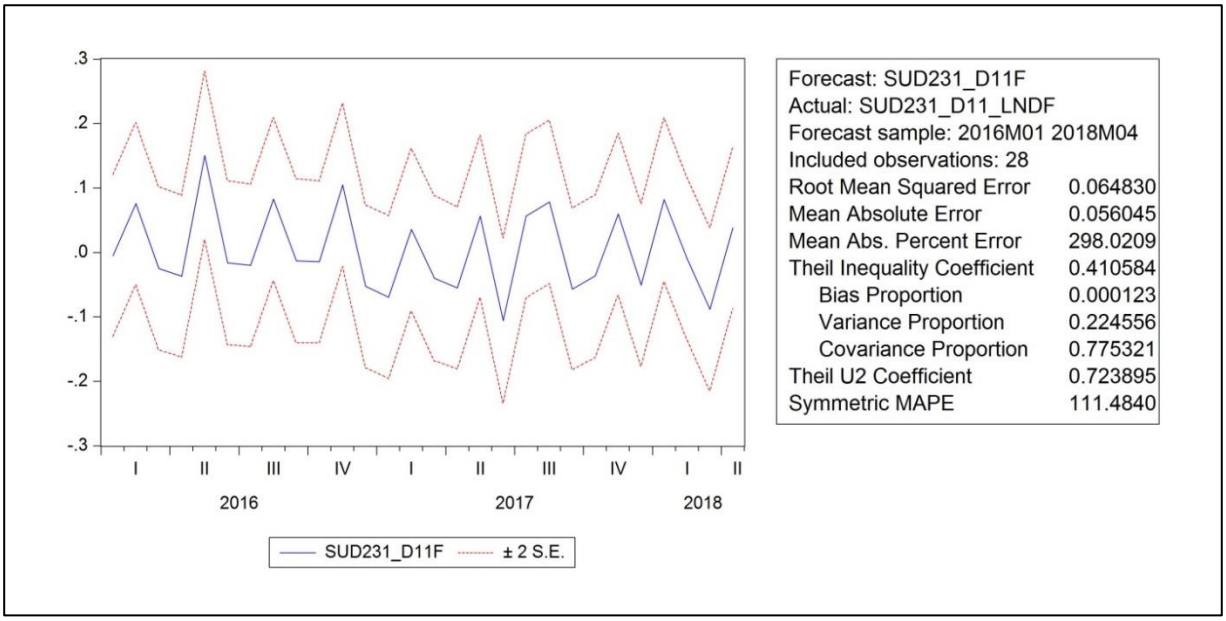
Graph 1 – dynamic out-of-sample forecasts



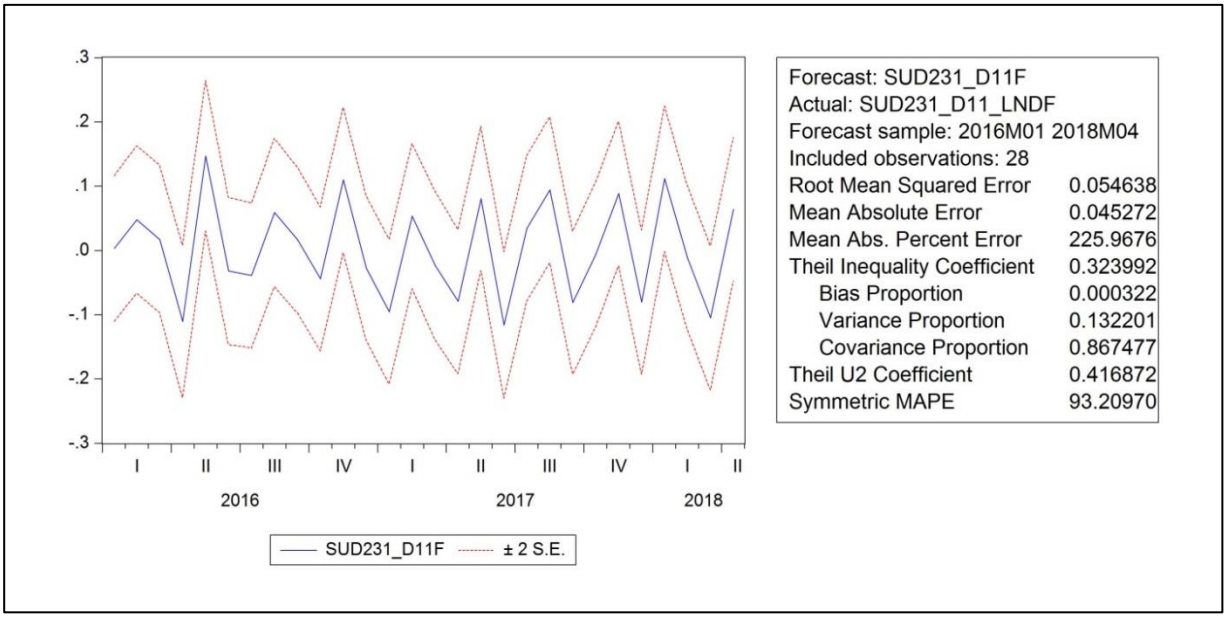
Graph 2 – Benchmark model – one step ahead (static) out-of-sample forecasts of monthly mortgage growth



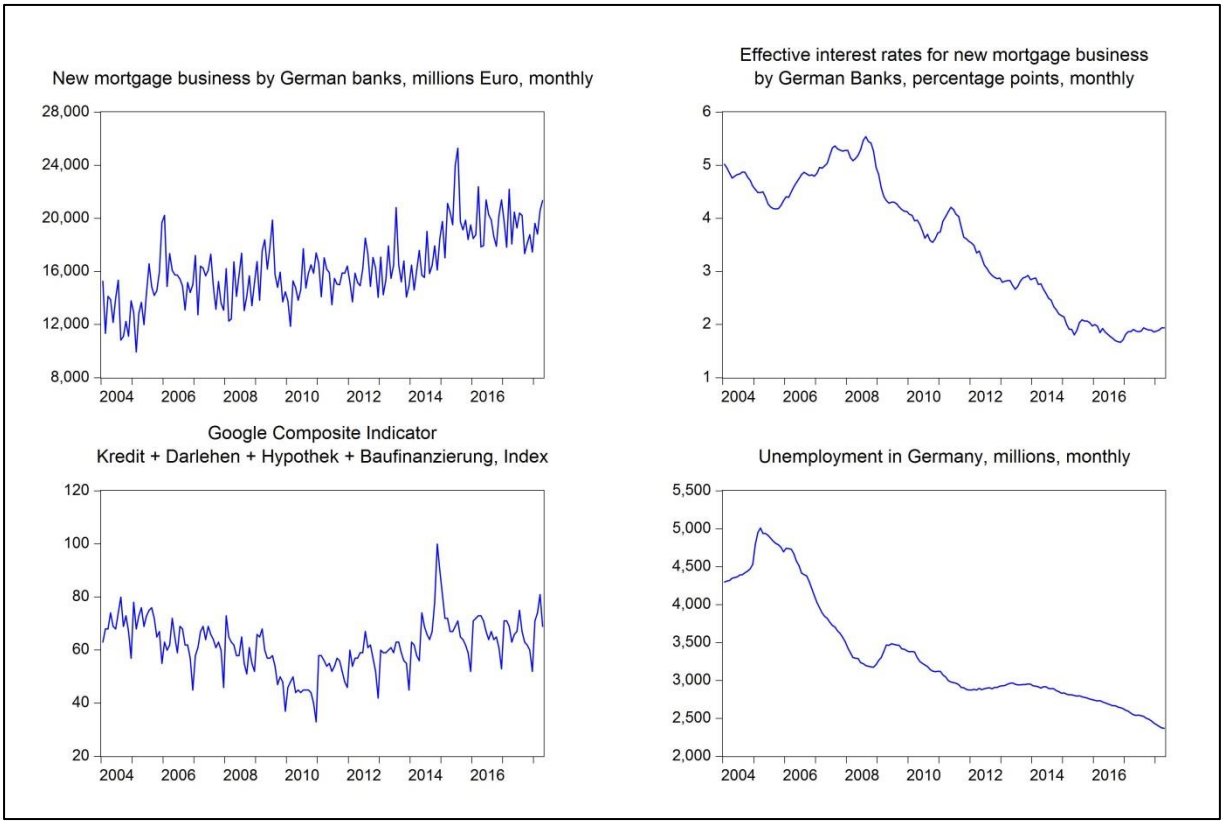
Graph 3 - Benchmark model with controls – one step ahead (static) out-of-sample forecasts of monthly mortgage growth



Graph 4 – Google Augmented II - one step ahead (static) out-of-sample forecasts of monthly mortgage growth



Graph 5 – graphical representation of selected time series, levels



Irving Fisher Committee on
Central Bank Statistics

BANK FOR INTERNATIONAL SETTLEMENTS

Ninth IFC Conference on "Are post-crisis statistical initiatives completed?"

Basel, 30-31 August 2018

Developments in the residential mortgage market in Germany - what can Google data tell us?[1]

Simon Oehler,
Deutsche Bundesbank

[1] This presentation was prepared for the meeting. The views expressed are those of the authors and do not necessarily reflect the views of the BIS, the IFC or the central banks and other institutions represented at the meeting.

Developments in the residential mortgage market in Germany – What can Google data tell us?

9th IFC Conference , „Are post-crisis statistical initiatives completed?“, Session 5 – Big Data

Simon Oehler, Deutsche Bundesbank

Agenda

1. Motivation & Literature Review
2. Google Data
3. Econometric Approach
4. Results
5. Conclusion

1. Motivation & Literature Review

- **In recent years interest in internet search data has increased & research has started to investigate the potential of this new data source.**
- **Examples comprise:**
  - Choi, Varian (2011); Predicting the present with Google Trends
  - Schmidt, Vosen (2009); Forecasting Private Consumption, Survey-based Indicators vs. Google Trends
  - McLaren, Shanbhogue (2011); Using internet search data as economic indicators, BoE Quarterly Bulletin, Q2
  - Askitas, Zimmermann (2014); Detecting Mortgage Delinquencies with Google Trends
  - Chauvet, Gabriel, Lutz (2016); Mortgage default risk: New evidence from internet search queries
  - Saxa (2014); Forecasting Mortgages, CNB Working Paper

1. Motivation & Literature Review

Why Google search data ?

“An individual's **interest in certain documents** (and not in others) is a **function of the individual's state** and so are search queries which are used to locate them. These queries are therefore utterances worth being investigated […]” - Askitas, Zimmermann (2014)

“We have found that [search] queries can be useful leading indicators for subsequent consumer purchases in situations where **consumers start planning purchases significantly in advance of their actual purchase decision**.” - Choi, Varian (2011), Predicting the Present with Google Trends

- Real estate & the financing thereof should meet this condition

Research question:

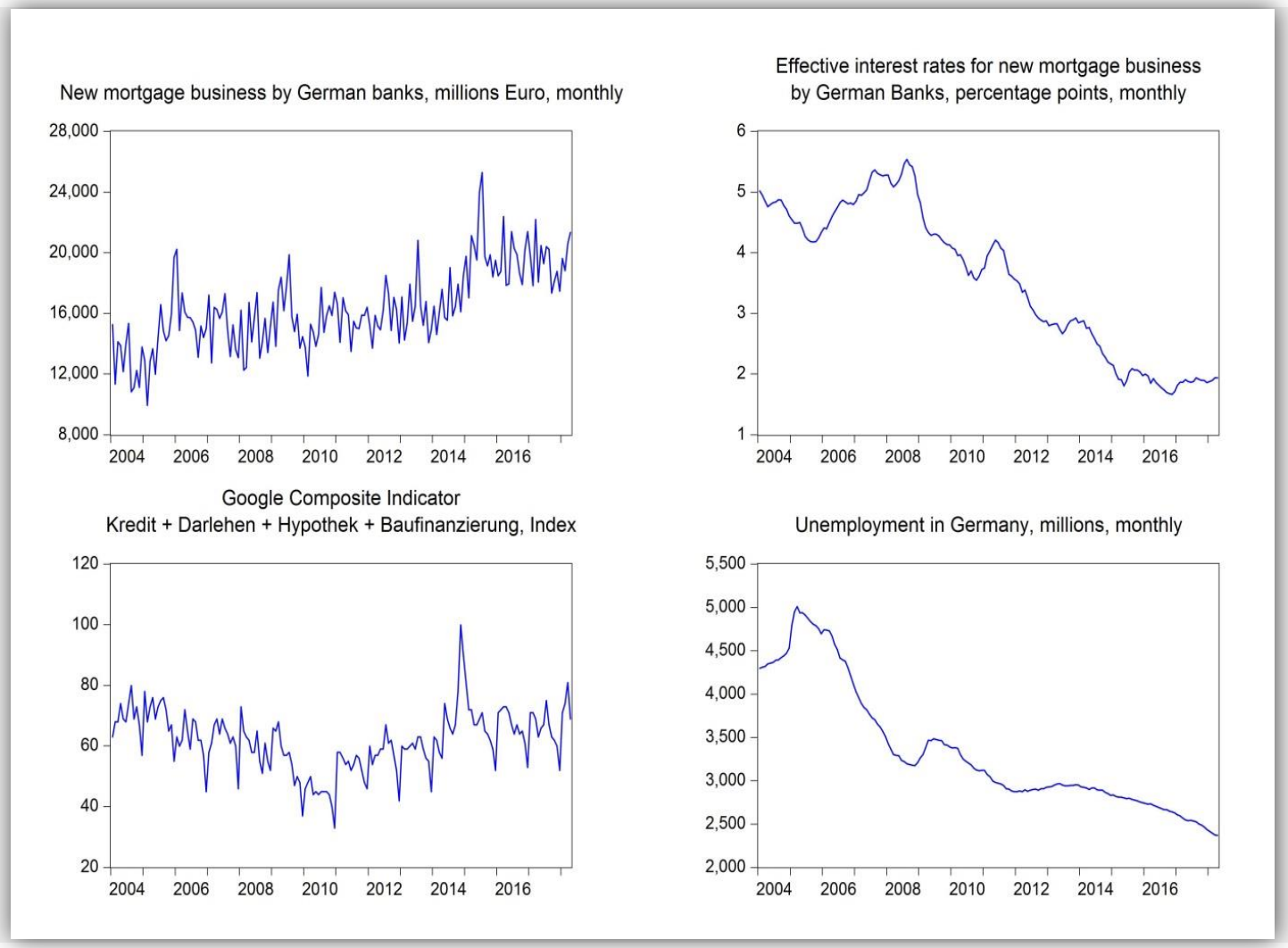
In how far can Google search data explain the variation in volumes of mortgage transactions at the federal level in Germany?

2. Google Data

- **37 Google series** are downloaded from https://trends.google.de/trends
- Selection is not solely "data driven". A priori **"economic/human reasoning"** involved as selection of time series is restricted to search terms relating to "mortgage" or "housing".
- Geography: Germany
- Language: German
- Frequency: Monthly
- Period: 2004 – April 2018
- Sampling: random sample of total searches is drawn by Google
- **Index:** no information about actual volumes or query shares

$$\boldsymbol{I(Kredit_t) = \frac{R(Kredit_t)}{\max\{R(Kredit_\tau)\}} \times 100} \quad \text{with } R(Kredit_t) = \frac{Kredit_t}{Google_t} \text{ and } \tau = 1, \dots, T$$

2. Google Data



3. Econometric approach

- **All time series are log-transformed and first differenced.**

- **Seasonal adjustment:**
  - **Response**: New mortgage business with seasonal patterns, particularly in July
  - **Controls**:
    - Effective Interest rate: no seasonality
    - Unemployment: seasonally adjusted
  - **Google**:
    - Almost all Google series with (strong) seasonal pattern around the end of the year: large drop in December and sharp rise in January of the subsequent year.

- **Modeling approach: Benchmark augmented by controls and Google data (stepwise forward selection procedure)**

  - $\Delta mortgages_t = \beta_m L^m \Delta mortgages_t$
  - $\Delta mortgages_t = \beta_m L^m \Delta mortgages_t + \gamma_m L^m \Delta interest_t + \theta_m L^m \Delta unempl_t$
  - $\Delta mortgages_t = \beta_m L^m \Delta mortgages_t + \gamma_m L^m \Delta interest_t + \theta_m L^m \Delta unempl_t + \delta_m L^m \Delta Google_t$

4. Results
Out-of-sample forecasts

Forecast Evaluation
Date: 10/31/18 Time: 16:20
Sample: 2016M01 2018M04
Included observations: 28
Evaluation sample: 2016M01 2018M04
Number of forecasts: 7

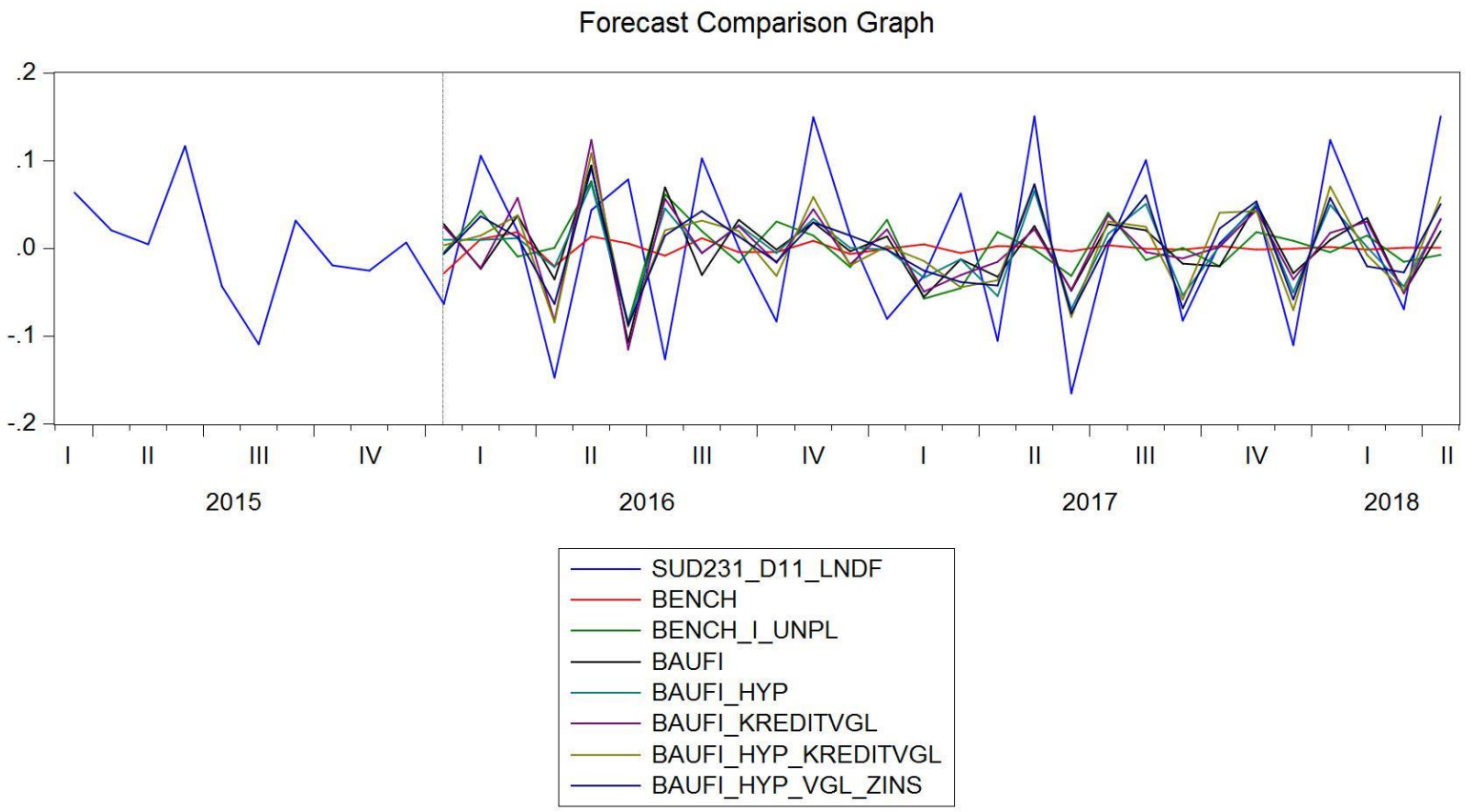
Combination tests
Null hypothesis: Forecast i includes all information contained in others

Equation	F-stat	F-prob
BENCH	7.933748	0.0001
BENCH_I_UNPL	5.851299	0.0010
BAUFI	10.54317	0.0000
BAUFI_HYP	6.456631	0.0006
BAUFI_KREDITVGL	10.02877	0.0000
BAUFI_HYP_KREDIT	4.755206	0.0033
BAUFI_HYP_VGL_ZI	4.376260	0.0051

Evaluation statistics

Forecast	RMSE	MAE	MAPE	SMAPE	Theil U1	Theil U2
BENCH	0.090353	0.076160	112.5150	167.2447	0.874054	0.957484
BENCH_I_UNPL	0.102884	0.088681	282.8204	166.0138	0.803323	1.298619
BAUFI	0.095041	0.079783	306.7584	141.2235	0.698204	0.927898
BAUFI_HYP	0.079904	0.064154	222.1042	120.8835	0.590645	0.839572
BAUFI_KREDITVGL	0.093064	0.078871	290.7800	136.7531	0.656967	0.857488
BAUFI_HYP_KREDIT	0.072731	0.061808	252.6928	116.4031	0.498878	0.658610
BAUFI_HYP_VGL_ZI	0.072015	0.058441	149.0011	102.9245	0.509604	0.726737

4. Results
Out-of-sample forecasts



5. Conclusion

- **Results suggest that Google data contain (short term) cyclicality** which can be exploited for forecasting/nowcasting.

- In particular the search terms **„Baufinanzierung“, „Hypothek“, „Kreditvergleich“** and **„Bauzins“** proved to be significant and relevant indicators for the change in growth rates of mortgage business in Germany under the tested model specifications.

- Thus far, the models presented here control for mortgage market interest rates and unemployment as a macroeconomic indicator.

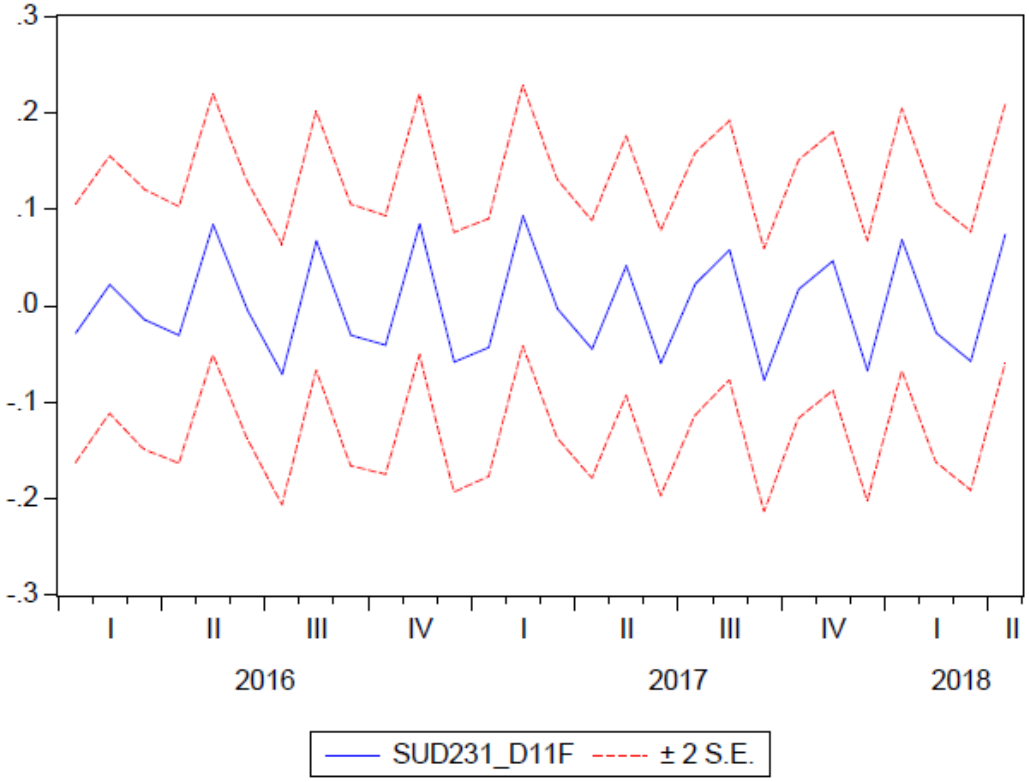
- Further robustness checks are needed. In particular:

  → Evaluate GoogleTrends relative to survey indicators
  → Further variable selection procedures to be applied

Thank you for your attention!

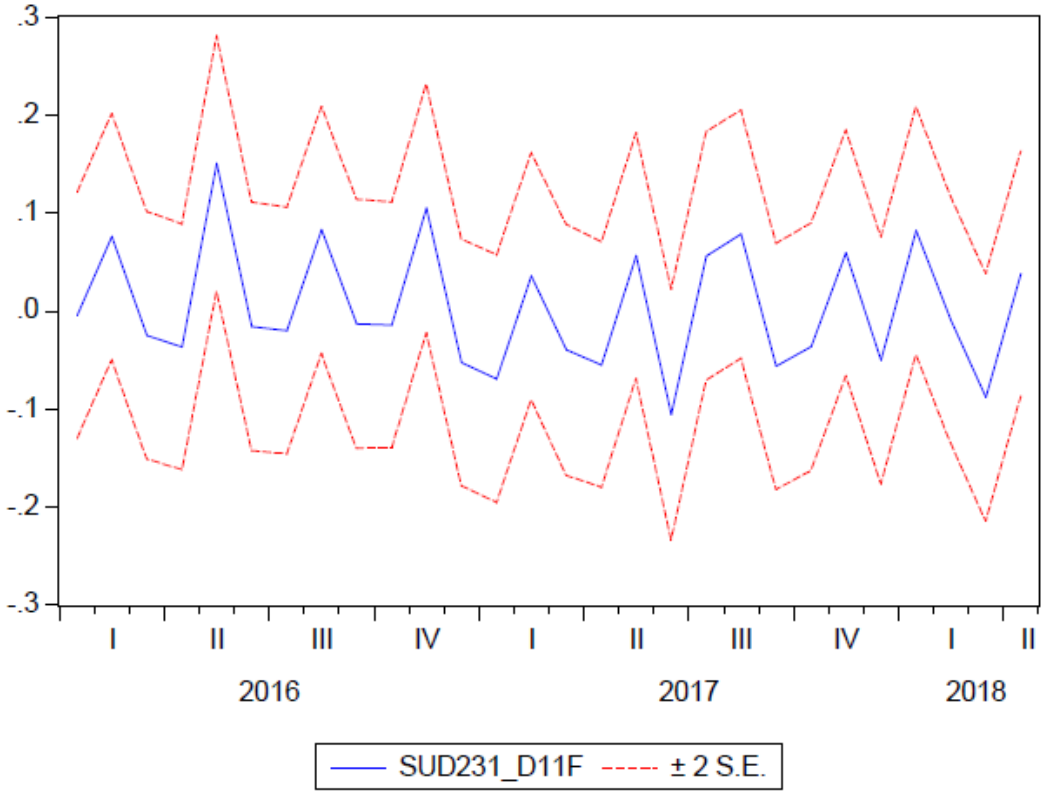
E-mail: simon.oehler@bundesbank.de

Backup I
Static Forecasts Benchmark



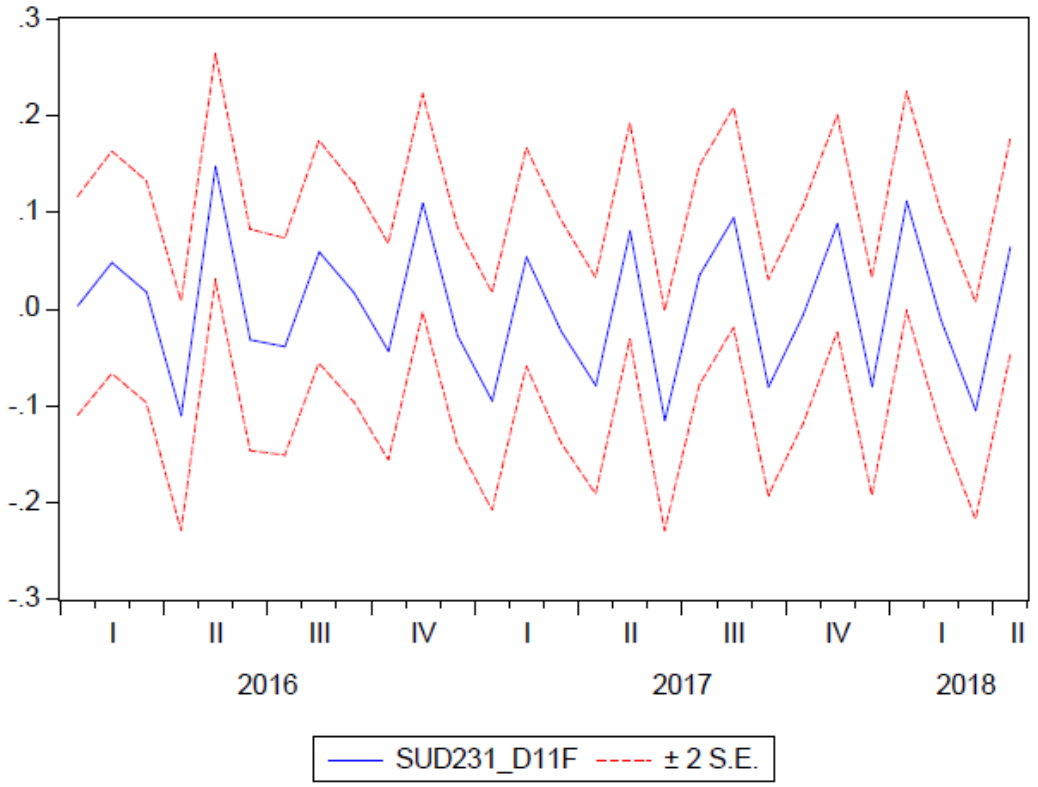
Forecast: SUD231_D11F	
Actual: SUD231_D11_LNDF	
Forecast sample: 2016M01 2018M04	
Included observations: 28	
Root Mean Squared Error	0.063301
Mean Absolute Error	0.054191
Mean Abs. Percent Error	287.2026
Theil Inequality Coefficient	0.427125
Bias Proportion	0.002628
Variance Proportion	0.404280
Covariance Proportion	0.593092
Theil U2 Coefficient	0.490265
Symmetric MAPE	103.7871

Backup II
Static Forecasts Benchmark with interest rate



Forecast: SUD231_D11F	
Actual: SUD231_D11_LNDF	
Forecast sample: 2016M01 2018M04	
Included observations: 28	
Root Mean Squared Error	0.064830
Mean Absolute Error	0.056045
Mean Abs. Percent Error	298.0209
Theil Inequality Coefficient	0.410584
Bias Proportion	0.000123
Variance Proportion	0.224556
Covariance Proportion	0.775321
Theil U2 Coefficient	0.723895
Symmetric MAPE	111.4840

Backup III
Static Forecast Google augmented II



Forecast: SUD231_D11F	
Actual: SUD231_D11_LNDF	
Forecast sample: 2016M01 2018M04	
Included observations: 28	
Root Mean Squared Error	0.054638
Mean Absolute Error	0.045272
Mean Abs. Percent Error	225.9676
Theil Inequality Coefficient	0.323992
Bias Proportion	0.000322
Variance Proportion	0.132201
Covariance Proportion	0.867477
Theil U2 Coefficient	0.416872
Symmetric MAPE	93.20970

Backup IV
Regression Output BAUFI_HYP_KREDITVGL

Dependent Variable: SUD231_D11_LNDF
Method: Least Squares
Date: 08/10/18 Time: 17:43
Sample (adjusted): 2004M09 2015M12
Included observations: 136 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
SUD231_D11_LNDF(-2)	-0.387154	0.067556	-5.730879	0.0000
SUD231_D11_LNDF(-1)	-0.503458	0.068513	-7.348352	0.0000
SUD231_D11_LNDF(-7)	-0.270081	0.063962	-4.222515	0.0000
SUD131_LNDF(-2)	-0.991662	0.204519	-4.848749	0.0000
GOOGLE_BAUFI_D11_LNDF(-3)	0.119307	0.044107	2.704970	0.0078
GOOGLE_HYP_D11_LNDF(-3)	-0.083612	0.020631	-4.052762	0.0001
GOOGLE_BAUFI_D11_LNDF(-1)	0.155329	0.043875	3.540303	0.0006
GOOGLE_KREDITVGL_D11_LNDF(-	0.086669	0.027745	3.123732	0.0022
GOOGLE_HYP_D11_LNDF(-1)	-0.042178	0.019847	-2.125220	0.0355

R-squared	0.558962	Mean dependent var	0.004244
Adjusted R-squared	0.531180	S.D. dependent var	0.080372
S.E. of regression	0.055031	Akaike info criterion	-2.897948
Sum squared resid	0.384611	Schwarz criterion	-2.705199
Log likelihood	206.0605	Hannan-Quinn criter.	-2.819619
Durbin-Watson stat	1.996244		