Nowcasting GDP with big data from internet

Luca Onorante

luca.onorante@ecb.int

Introduction

- Macro short term forecasting benefits from the availability of wide information.
- Internet-based information: potentially useful, but 'organic' and extremely noisy.
  - Conventional estimation techniques work well only for well-defined variables, require an important cleaning, and do not perform for wide macro aggregates.
- This paper nowcasts GDP for the four big countries in the euro area.

Literature using Google data

- Choi and Varian (2009): automotive sales, retail sales, travel.
- Vlastakis and Markellos (2010) and Moat and Stanley (2013) on financial markets using Google searches on individual stocks.
- Many papers nowcast car sales.
- Unemployment rate: Askitas and Zimmermann(2009) for DE, D' Amuri and Marcucci (2010, 2013) for US. Koop and Onorante on several European macro variables with "Google probabilities".
- Guzmin (2011) on inflation expectations.

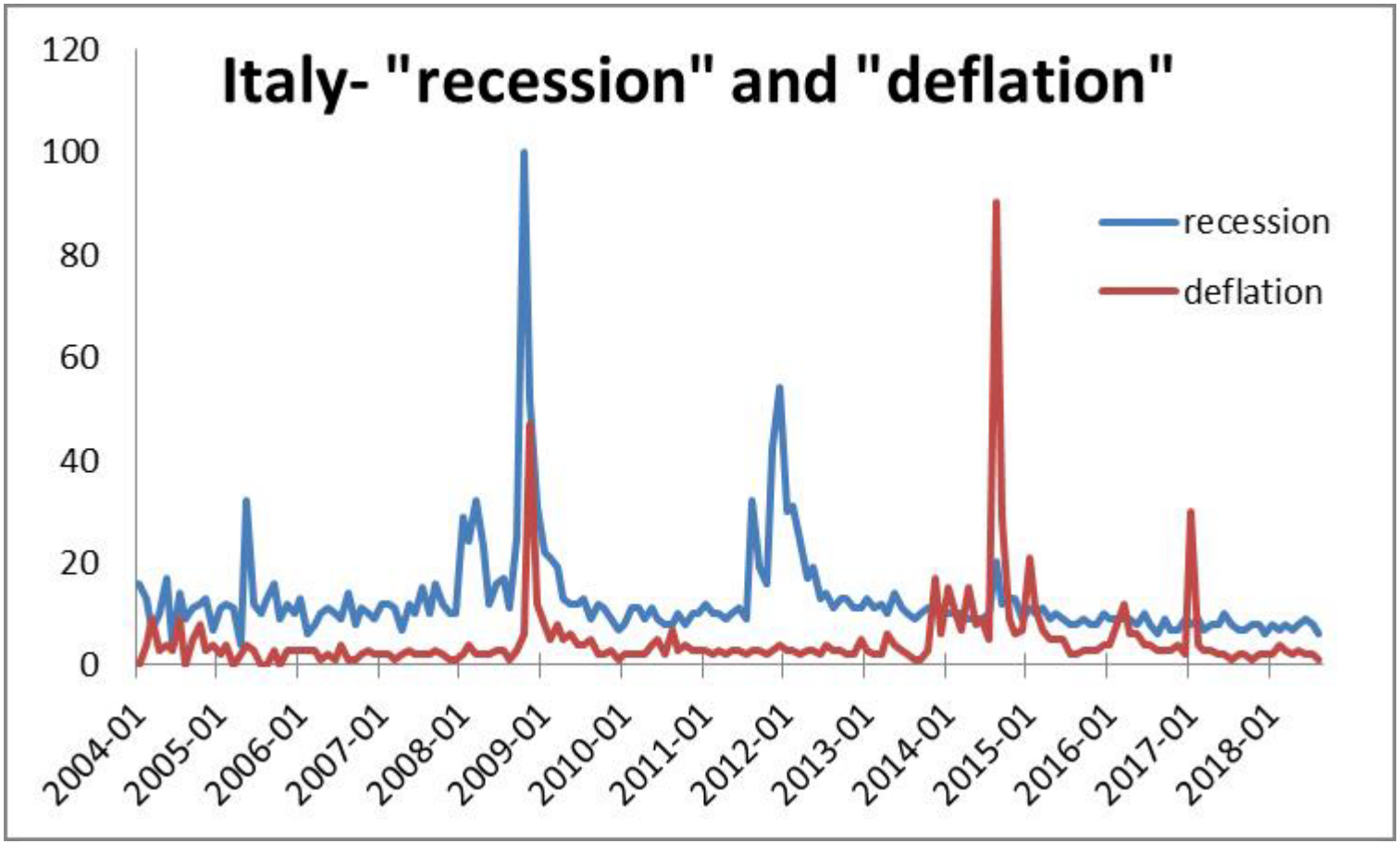
Contribution of the paper

- Nowcasting aggregate GDP using google search data.
  - Attacks the macro data directly, documents problems.
  - Augments a sizeable macro database with Search terms and Categories, “big data”.
  - Modifies standard BMA/BMS to deal with big, noisy data
  - Avoids hand picking, and introduces “data robustification” in Google Searches.

Google data

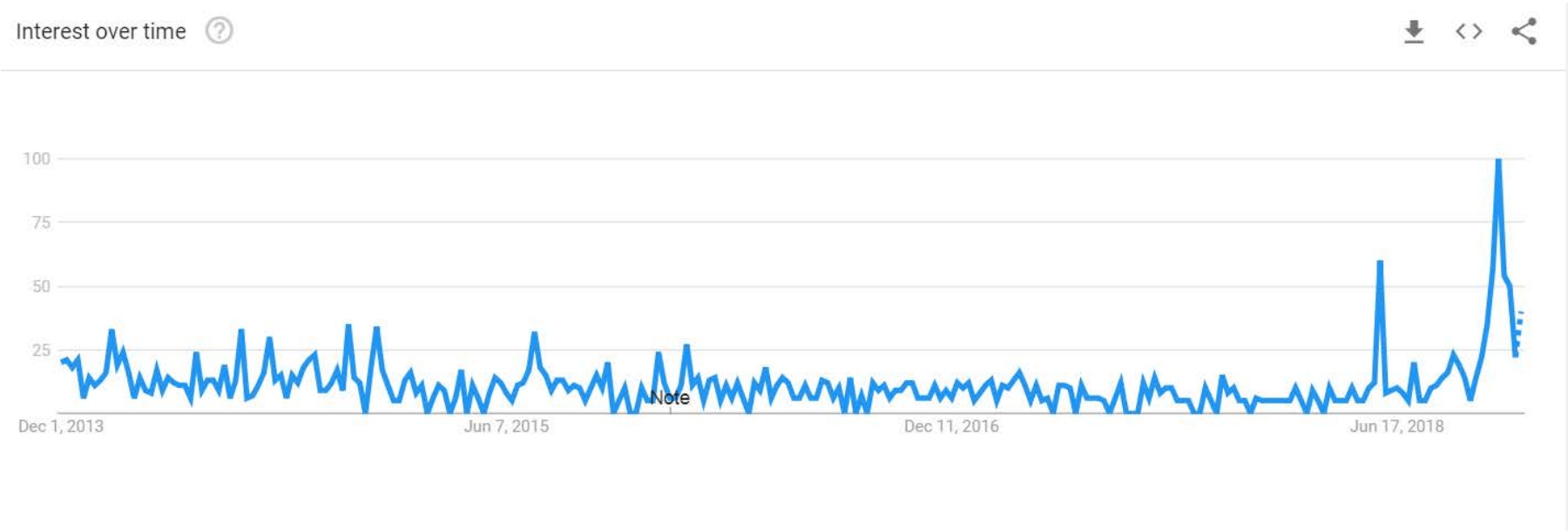
- Available in real-time and normally not revised. Cover *actual* behavior (e.g. purchase intentions), unlike e.g. surveys.
- Google trends: time series for different languages and countries.
- Added to macro database including: timely real indicators (unemployment rate), industrial production and retail sales, employment and trade data. Nominal variables (timely) : inflation flash estimate, inflation.

Google data – easy example



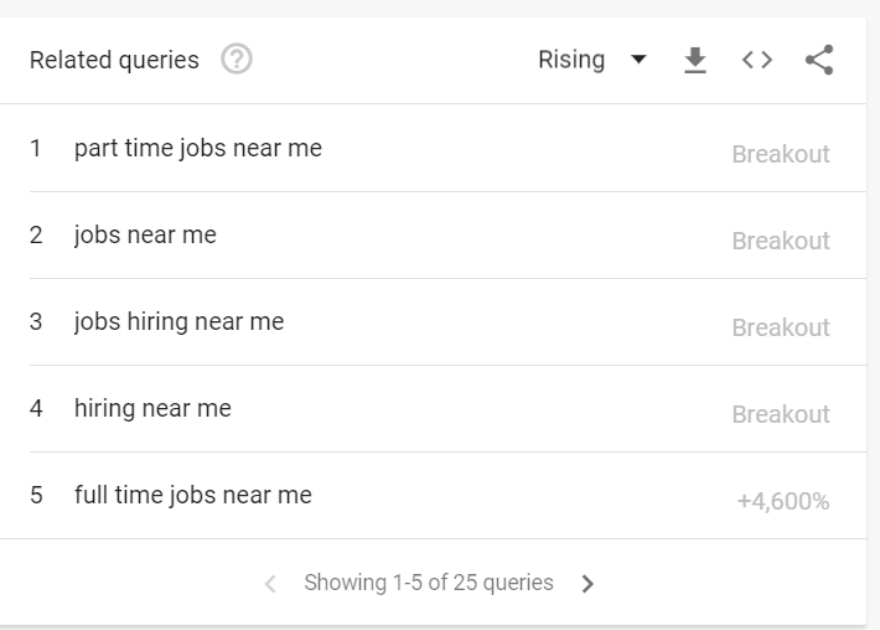
Google data – difficult example

- Italy – “open account in Switzerland”



Searches: “robustification”

- Start with “growth”, “jobs”
- Include related searches until 3rd level
- Delete duplicates
- Deals with:
  - different spelling of the same query (’jobs’ and ’job’),
  - common spelling mistakes,
  - wrong choice of the query by the researcher
  - account for related concepts (unemployment office near me)
  - time variation (macbook pro)
- Ensures that the concept is captured



Related queries | Rising

1	part time jobs near me	Breakout
2	jobs near me	Breakout
3	jobs hiring near me	Breakout
4	hiring near me	Breakout
5	full time jobs near me	+4,600%

Showing 1-5 of 25 queries

Categories

- Google groups Internet searches into general categories.
  - To advertisers: e.g. Apple.
- Directly from Google, not sampled
- Not noisy
- 296 categories

Dimensionality

	Searches	Macro	Gategories	Total
DE	807	57	296	1160
FR	723	16	296	1035
IT	728	22	296	1046
ES	394	18	296	708
NL	356	10	296	662

BMA-BMS

- BMA (Raftery), Sala-i-Martin (1994), Koop and Korobilis (2009) on inflation.
- BMA to deal with the short Google sample (starting in 2004), the very high number of potentially relevant search terms, and the high risk of misspecification due to the size and noisiness of the Google database.

$$Pr(M_i|y^{1:t-1}) = w_{i,t} \propto \frac{Pr\left(y_t|y^{1:t-1}, M_i\right)}{\sum_{j=1}^{K} Pr\left(y_t|y^{1:t-1}, M_j\right)}$$

A robust process for noisy data

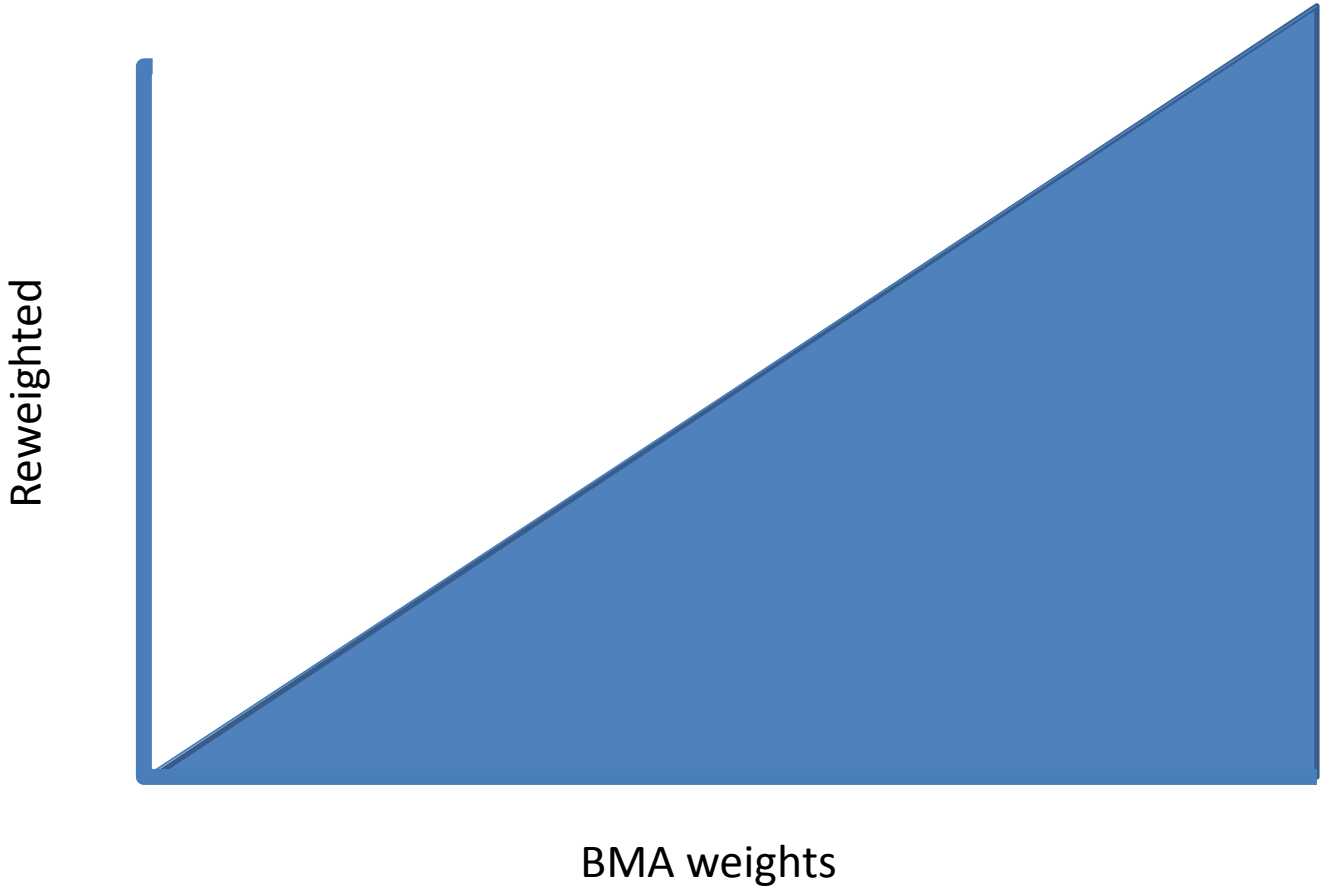
- Exponential reweighting of the posterior probabilities: (0 < α < inf)

$$w_{i,t} \propto \frac{L_i\left(y_t|y^{1:t-1}, M_i\right)^{\alpha}}{\sum_{j=1}^{K} L_j\left(y_t|y^{1:t-1}, M_j\right)^{\alpha}}$$

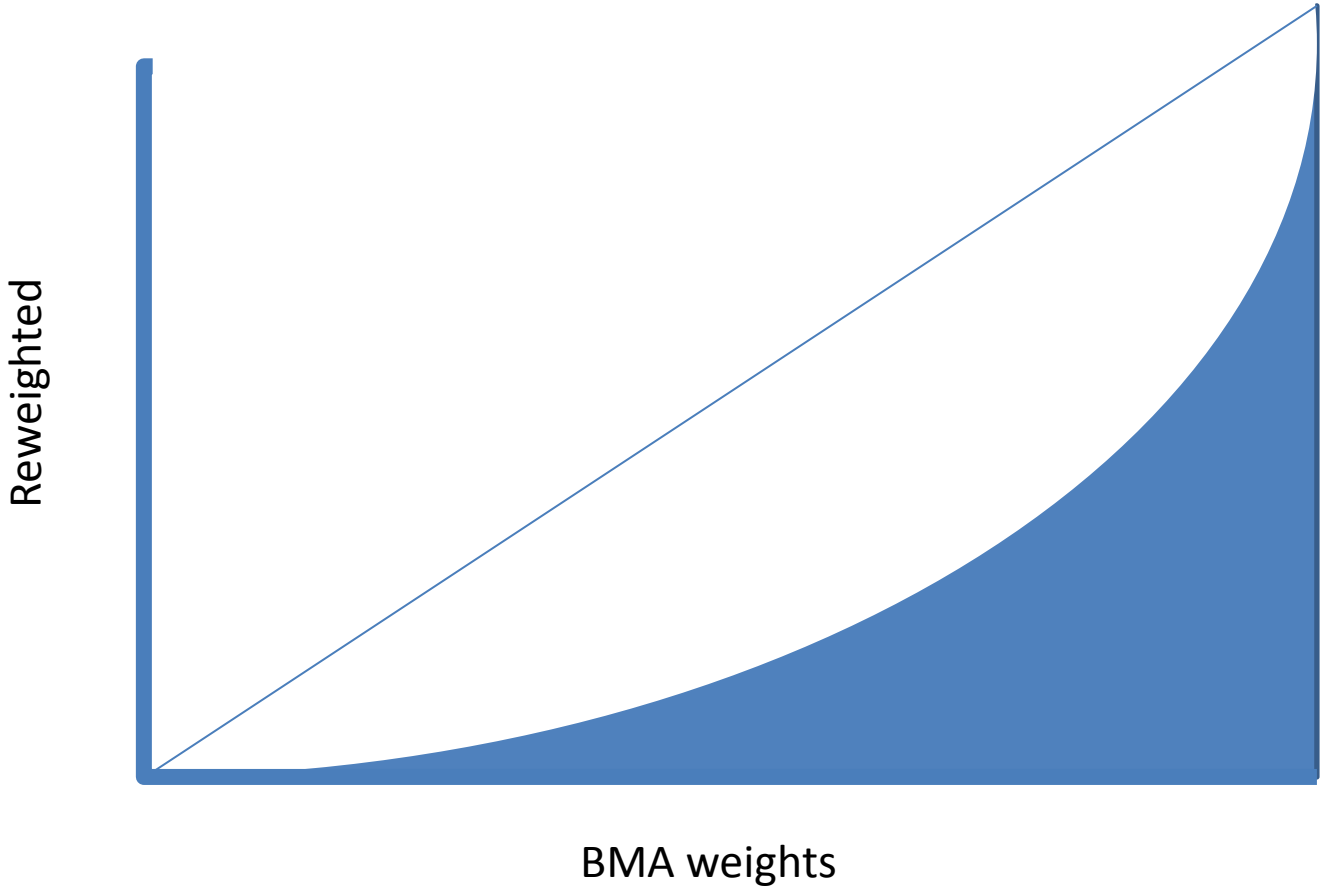
- Eliminates (rather than just downplaying) badly specified models. Trimming.

$$w'_{i,t} \propto w_i * I(w_i < Percentile(w, x))$$

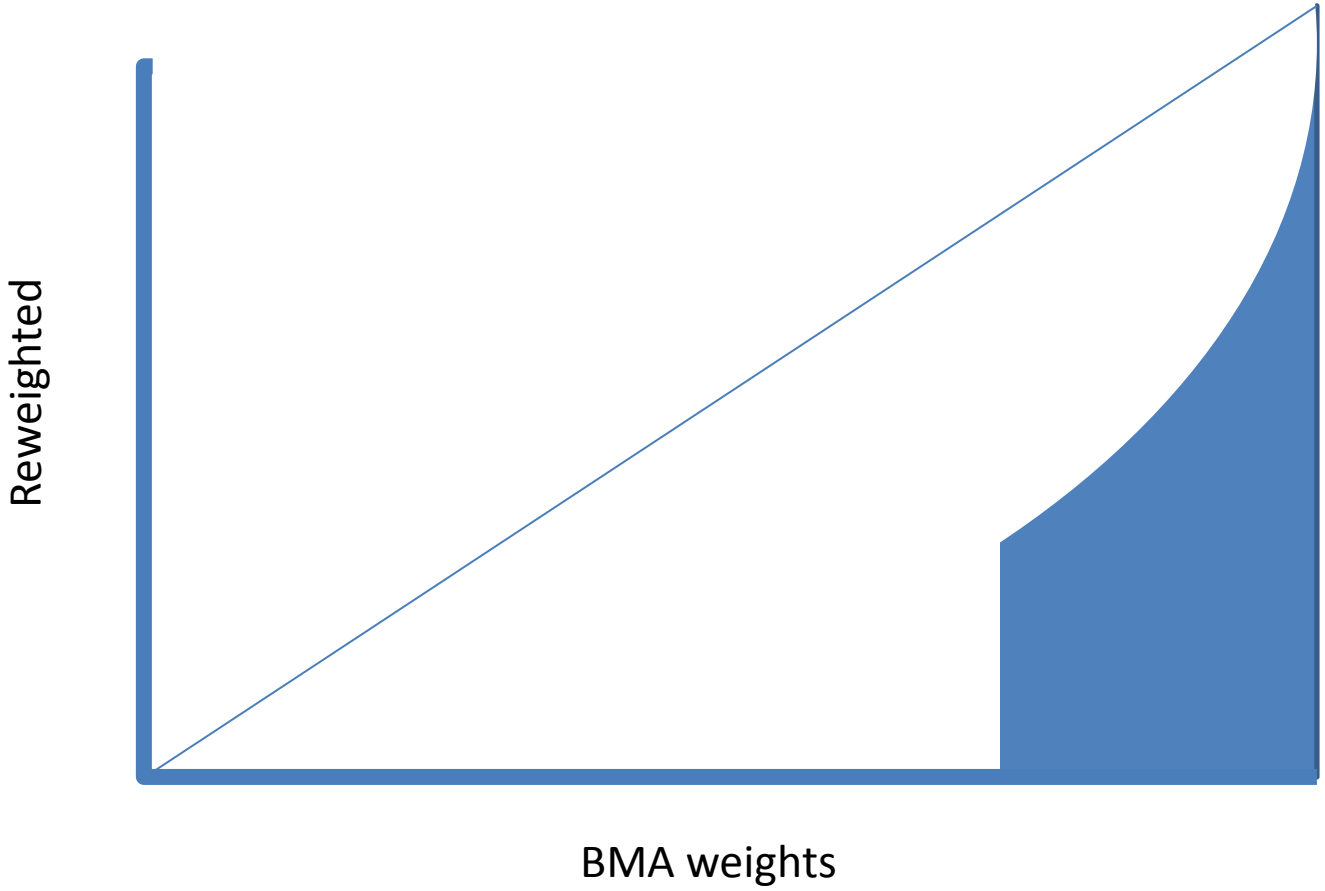
A robust process for noisy data



A robust process for noisy data



A robust process for noisy data



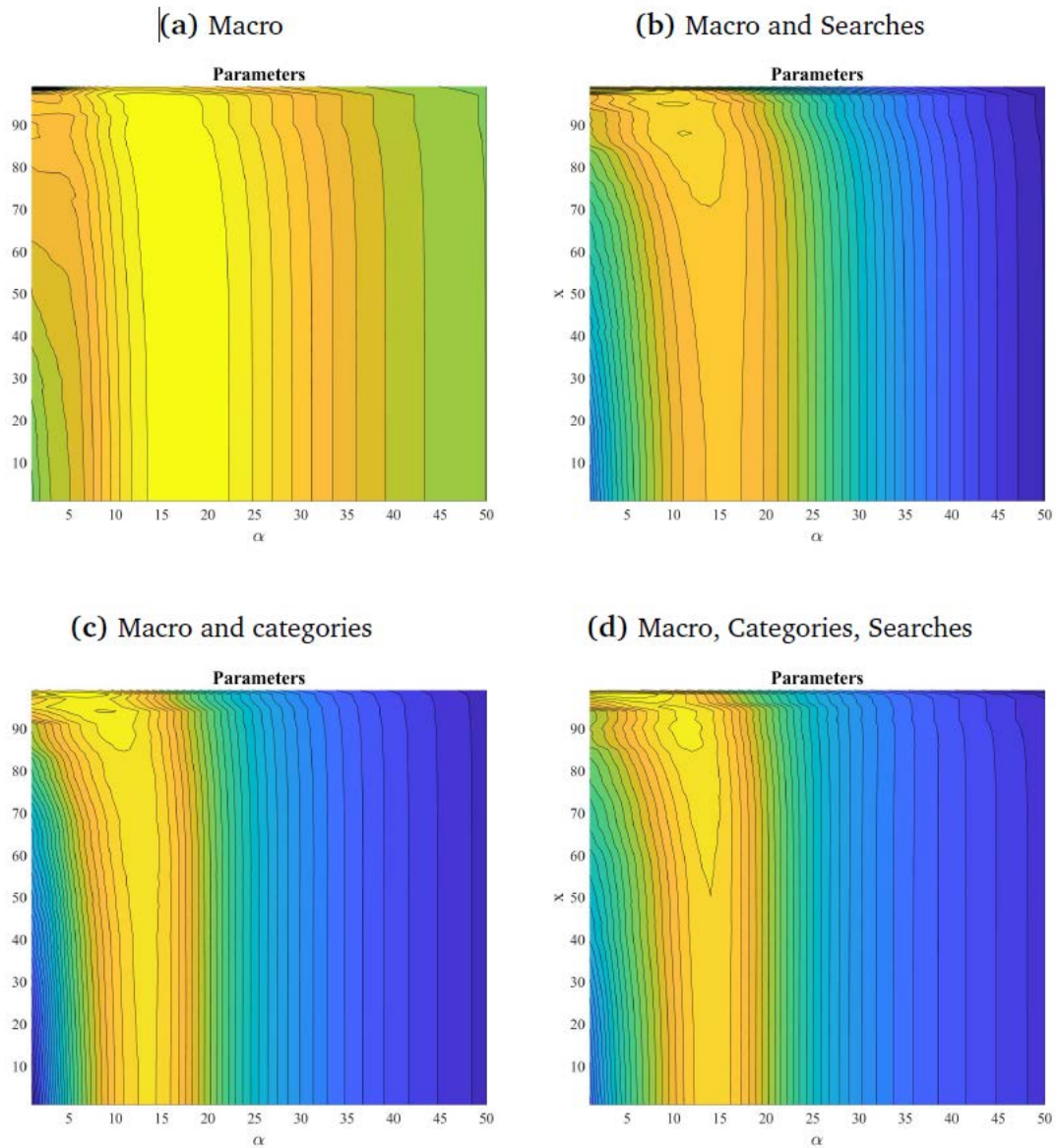
Results with additional shrinkage

Table 1: RMSFE of different estimators and corresponding shrinkage

	AR4	AR1	BMA	BMS	BMA-a*	a*	BMA-x*	x*	Optimal	a*	x*
DE - Macro and Categories	0.26	0.31	0.26	0.20	**0.09**	13	**0.09**	98	**0.08**	6	98
DE - Macro, Searches, Categories	0.26	0.31	0.28	0.21	**0.09**	14	**0.06**	98	**0.06**	4	98
DE - Macro and Searches	0.26	0.31	0.25	0.20	**0.11**	15	**0.07**	98	**0.07**	1	98
DE - Macro	0.26	0.31	0.26	0.21	**0.09**	18	**0.15**	93	**0.09**	17	63
ES - Macro and Categories	0.51	0.82	0.46	0.43	**0.28**	16	**0.28**	97	**0.28**	14	76
ES - Macro, Searches, Categories	0.51	0.82	0.47	0.44	**0.28**	7	**0.27**	99	**0.27**	1	99
ES - Macro and Searches	0.51	0.82	0.47	0.45	**0.32**	11	**0.32**	97	**0.32**	1	97
ES - Macro	0.51	0.82	0.48	0.45	**0.31**	21	**0.33**	97	**0.31**	21	7
FR - Macro and Categories	0.58	0.76	0.60	0.58	**0.58**	5	**0.54**	93	**0.54**	1	93
FR - Macro, Searches, Categories	0.58	0.76	0.60	0.58	**0.54**	7	**0.50**	97	**0.50**	1	97
FR - Macro and Searches	0.58	0.76	0.61	0.58	**0.57**	6	**0.55**	94	**0.55**	1	94
FR - Macro	0.58	0.76	0.59	0.57	**0.28**	50	**0.38**	99	**0.28**	50	98
IT - Macro and Categories	1.48	1.88	1.36	1.26	**0.71**	7	**0.85**	98	**0.71**	7	5
IT - Macro, Searches, Categories	1.48	1.88	1.41	1.34	**0.72**	9	**0.85**	99	**0.72**	9	6
IT - Macro and Searches	1.48	1.88	1.34	1.26	**0.59**	8	**0.80**	97	**0.59**	8	10
IT - Macro	1.48	1.88	1.32	1.21	**0.91**	5	**0.79**	96	**0.79**	1	96

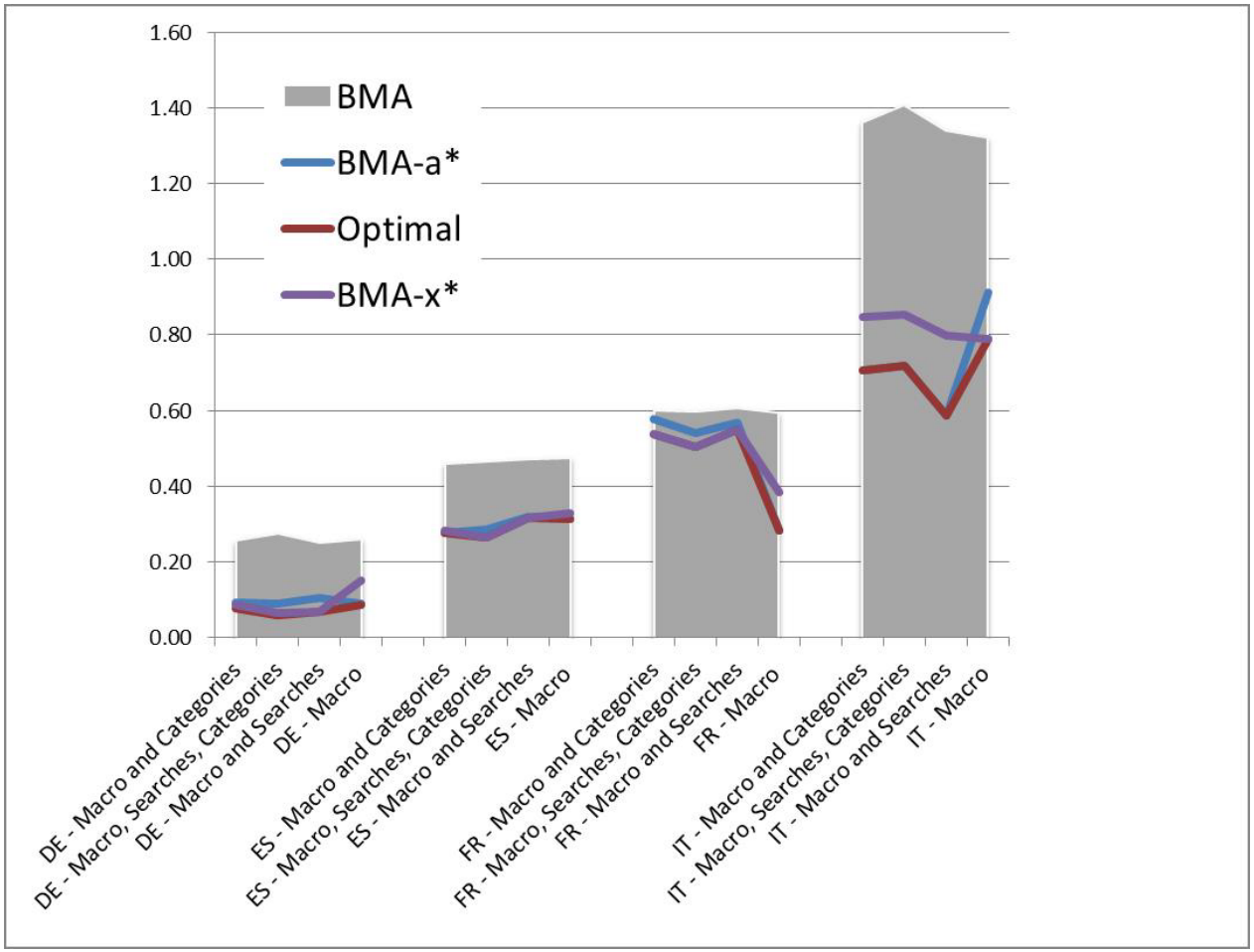
Conclusions

- Shrink!
- Trim!



Conclusions 2

- Google data: help but lots of mining needed.



Conclusions and way ahead

- Conventional result: standard MA does not perform well
- DMS performs only slightly better.
- Nowcasting and short-term forecasting can work with big and noisy data sets.
  - A flexible but high degree of shrinkage is needed.
  - Trimming also improves performance (as a median is more robust)
  - “Robustification” of the data is often necessary.

Next steps

- Dynamic Model Averaging to assess whether different groups of variables are relevant at different times.
- Experiment with different, big data sets.
- We want to extend the range of evaluation criteria
- Evaluate short-term forecasting in addition to now-casting.
- A toolbox to perform this analysis in an easy manner.

Thank you!

Luca Onorante

luca.onorante@ecb.int