Forecasting US Birth Rates with Google Trends

Francesco Billari, Francesco D'Amuri, Juri Marcucci

Bocconi University, Bank of Italy, Bank of Italy

Bank of Italy's Workshop
on Harnessing Big Data & Machine Learning Techniques for Central Banks
Bank of Italy
Rome, 27 March 2018

Outline

- Data and Determinants for US fertility
- Short-Term emphasis
- New Leading indicators
- Forecasting models
- Out-of-sample evaluation
- Some robustness (state level)

Motivation

- A simple Supply-Side Decomposition
- Macro-based accounting framework

$$GDP_t = \underbrace{\frac{GDP_t}{Hours_t}}_{\text{Efficiency in production}} \times \underbrace{\frac{Hours_t}{Workers_t} \times \frac{Workers_t}{LaborForce_t}}_{\text{Labor market developments}} \times \underbrace{\frac{LaborForce_t}{Population_t} \times Population_t}_{\text{Demographic developments}}$$

- We concentrate on the last part (demographic developments)

Motivation

- Fertility is the major component of population dynamics
- The size and structure of population is entirely dependent on fertility
- Trends in fertility are the most difficult demographic variable to project
- Fertility rates represent the most important modeling variable in any population model
- These models are of critical importance
- Forecasts of births and birth rates are fundamental to forecasts of future population sizes (Keyfitz, 1972).
- Yet the forecasting of births and birth rates, even in highly developed countries has proven to be quite difficult to do

Introduction

- Demographers model long-run fertility (see for example Booth, IJF, 2006, for a review)
- However short-term perspective is useful to spot diverging trends (for example to assess the impact on births of a crisis)
- Our approach:
- Pure time series models with leading indicators:
- GDP, Unemployment rate dynamics (Goldstein et al., PDR, 2009)
- We also add Economic Policy Uncertainty (EPU) index by Baker, Bloom, and Davis (2016) as an additional leading factor affecting birth rates
- For a shorter sample (from 2004 onwards) we suggest using Google Index based on Google Trends: fertility-related web searches (in different contexts: Choi and Varian, 2009; D'Amuri and Marcucci, 2017; Ginsberg et al., 2009, Da, Engelberg, and Zha, 2011, etc.)

Forecasting the US Birth Rates Using Google - Preview

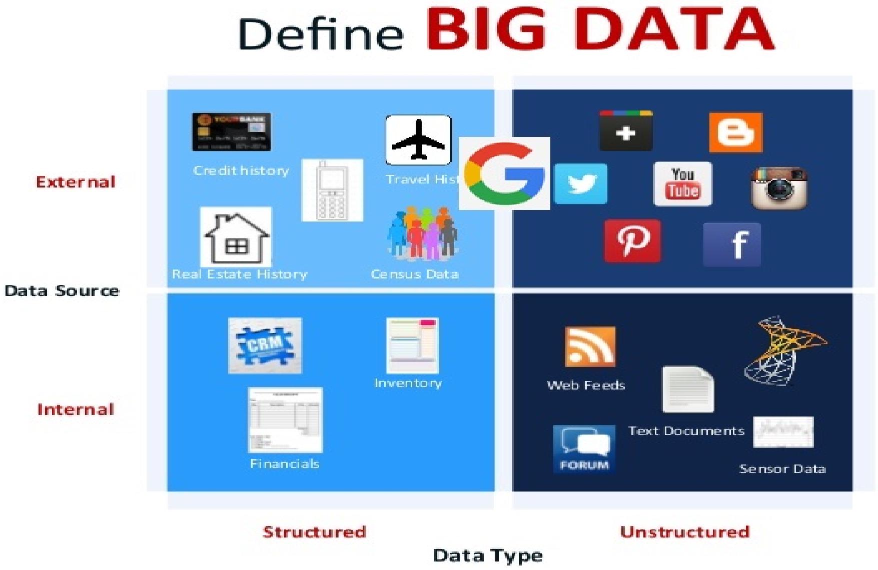
- We predict the US Birth Rates using:
  - Traditional macro indicators:
    - GDP
    - UR
  - New web-based indicators:
    - EPU (Economic Policy Uncertainty) Index by Baker, Bloom and Davis (2016)
    - Google Trends indicators
  - We find forecasting improvements with both web-based indicators

"The Social Science (Big) Data Revolution" (Gary King)

- "Between the dawn of civilization and 2003, we only created five exabytes of information; now we're creating that amount every two days" (1 exabyte $= 2^6$ bytes $\approx 1.15 \times 10^{18}$ bytes, i.e. 10^9GB)
- "In 2010 human race created 800 exabytes of information" (around 800 billion gigabytes, 1GB $= 10^9$ bytes)
- 90% of the world's data was created in the last 2 years
- Most of the World's Data is **Unstructured**
  - 2009 HP survey: 70%
  - Gartner: 80%
  - Jerry Hill (Teradata), Anant Jhingran (IBM): 85%
- "There's a systemic gap between the low-frequency data employed by governments and the high-frequency data of business' (Hal Varian, Google)'
- "Data is like food. We used to be data poor, now the problem is data obesity' (Hal Varian, Google)'

Big Data

One of the 3 V's: **Variety**: Different sources, different types



Google Trends data

- *Google Trends* (previously separated from *Google Insights for Search*) tracks relative changes in Google search queries from January 2004
- Google search queries
  - web searches
  - news searches
  - image searches
  - product searches
  - YouTube searches
- Different Geographical areas and levels (national, state and metropolitan area level) based on the originating IP address
- Available to the public for free download online at www.google.com/trends/

Features of *Google Trends* data

- Google Trends data represent how many web searches are done for a particular *keyword*, relative to the total # of searches in certain geographical area over time
- indicate the likelihood of a random user to search for a particular keyword on Google from a certain location at a certain time on a relative basis
- are gathered using IP address information from Google logs and updated daily
- are gathered only if the number of searches exceeds a certain threshold of traffic
- are such that repeated queries from a single user/IP over a short period of time are eliminated
- are available world-wide (by country, by region, by city)
- are *normalized* (divided by the total website traffic in the geographical area) $\Rightarrow$ comparability issues $\Rightarrow$ Search Volume Index (SVI)
- are *scaled* (from 0 to 100) dividing each data point by the maximum
- available monthly, weekly, daily, and intra-daily (only on shorter samples)

Aggregation, Normalization and Scaling

- For region r, the SVI for week τ is constructed aggregating the daily data for each day t. Given the search volume on a term "V", ($V_{t,r}$) in region r on day t and the total search volume in that region $T_{t,r}$ we have the following for a total of T weeks
- Search Share for day t and week τ:

$$S_{t,r} = \frac{V_{t,r}}{T_{t,r}} \quad and \quad S_{\tau,r} = \frac{1}{7} \sum_{t=Sunday}^{Saturday} S_{t,r}$$

- Web Search Volume for week τ:

$$S^*_{\tau,r} = \frac{\mathbf{100}}{\max_\tau(\mathbf{S}_{\tau,\mathbf{r}})} \frac{1}{7} \sum_{t=Sunday}^{Saturday} S_{t,r}$$

where $\tau = 1, \ldots, T$

Matching options for "fertility"-related web queries in Google Trends

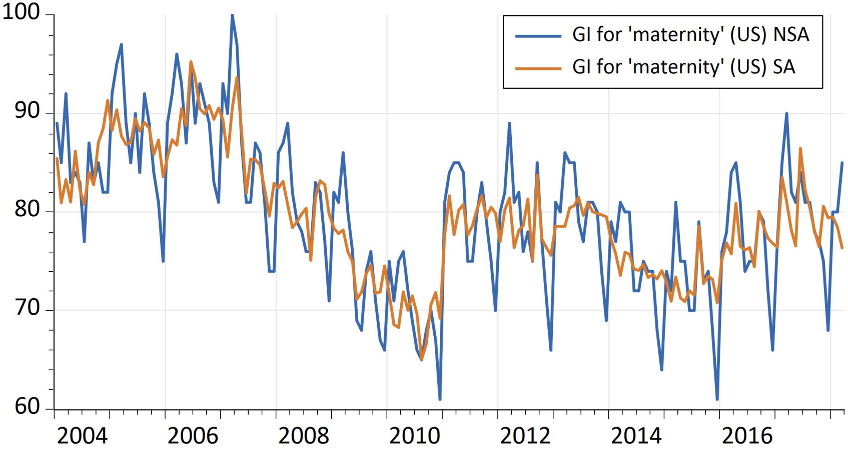
- typing `maternity leave`: GI includes searches containing both `maternity` and `leave` in any oder and along with additional terms before (e.g. `using short term disability for maternity leave`) and after (e.g. `maternity leave replacement`)
- typing `"pregnancy test"`: GI includes searches with that specific order in quotes along with additional terms before and after (e.g. `"pregnancy test" calculator`)
- typing `"ovulation" + "pregnancy test"`: GI includes searches with either `"ovulation"` or `"pregnancy test"`, but not both

Which **keyword(s)** to forecast birth rates?

- We tried to imagine what an average American internet user who wanted to have children would type in the Google bar. For example:
  - `‘maternity’`
  - `‘pregnancy’`
  - `‘ovulation’`

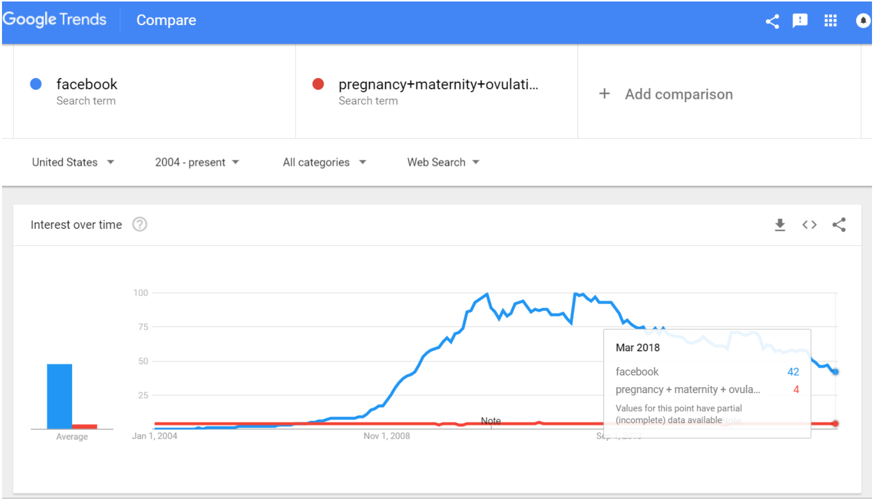
Peculiar seasonality

Monthly Google index for "maternity" - Sample: Jan. 2004 - Mar. 2018



Relevance of our preferred keyword?

Incidence of keywords "maternity" + "pregnancy" + "ovluation" vs other popular keywords ‘‘facebook" (**highest incidence**)



Economic Policy Uncertainty

How is it built by Baker, Bloom and Davis (2016)?

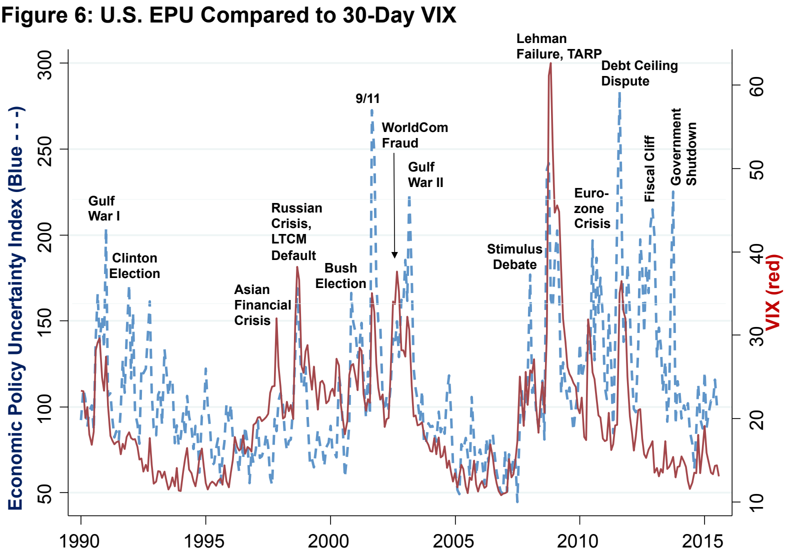
- Counting articles from newspapers containing (E)conomic, (P)olicy and (U)ncertainty words

(E) "economic" or "economy";

(U) "uncertain" or "uncertainty";

(P) "congress", "deficit", "Federal Reserve", "legislation", "regulation" or "White House"

US Economic Policy Uncertainty (EPU) index vs VIX



Notes: The figure shows the U.S. EPU Index from Figure 1 and the monthly average of daily values for the 30-day VIX.

The setup of the forecasting horse-race

- **Timing:** $T = R + P$ observations
  - In the '**Long sample**' (**1990.1-2013.12**) we have $T = 288$
  - In the '**Short sample**' (**2004.1-2013.12**) we have $T = 120$
- The first R are used to estimate the models (**in-sample**) while the last P are used for **out-of-sample** evaluation.
- Want to predict u_t using linear AR models w/ and w/o exogenous leading indicators x_t:
  - $\boxed{x_t = \{GI_t, ..., GI_{t-k}\}}$
  - $x_t = \{GDP_t, ..., GDP_{t-k}\}$
  - $x_t = \{UR_t, ..., UR_{t-k}\}$
  - $x_t = \{EPU_t, ..., EPU_{t-k}\}$
- $GI1_t$= '*Maternity*', $GI2_t$= '*Pregnancy*', and $GI3_t$= '*Ovulation*'.

The setup of the forecasting horse-race

- Forecasting scheme: we use a **rolling** scheme.
  - '**In Sample**' (**2004.2-2008.12**) w/$R = 60$
  - '**Out-Of-Sample**' (**2009.1-2013.12**) w/ $P = 60$
- We use **direct** forecasts.
  - Benchmark $AR(p)$ with p selected by BIC recursively ex-ante at each forecast origin

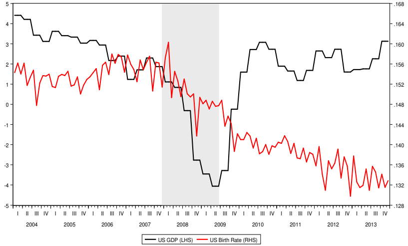
$$\boxed{y_{t+h}^{h} = \beta_0 + \beta_1(L)y_t + \eta_{t+h}, \qquad t = 1, 2, \ldots, T} \tag{1}$$

  - versus AR-$X(p)$ model w/ LI x_t with lags p and q selected by BIC recursively and sequentially $(p_{max} = q_{max} = 4)$

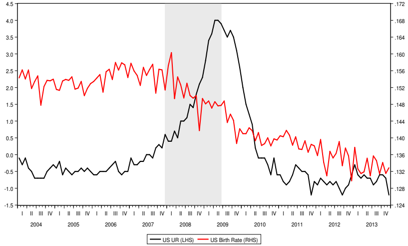
$$\boxed{y_{t+h}^{h} = \beta_0 + \beta_1(L)y_t + \beta_2(L)x_t + \varepsilon_{t+h}, \qquad t = 1, 2, \ldots, T} \tag{2}$$

Google Web Searches for US Birth-related Keywords

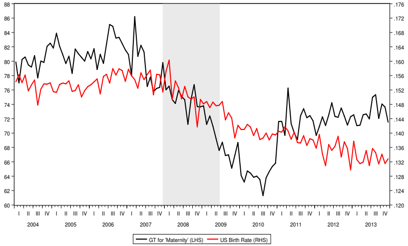
GDP



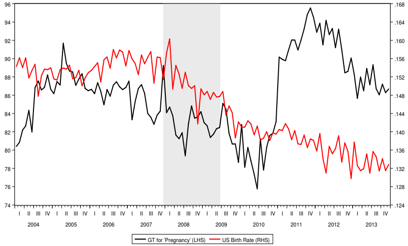
UR



GI1 - ‘Maternity’



GI2 - ‘Pregnancy’



Forecasts of US Birth Rates

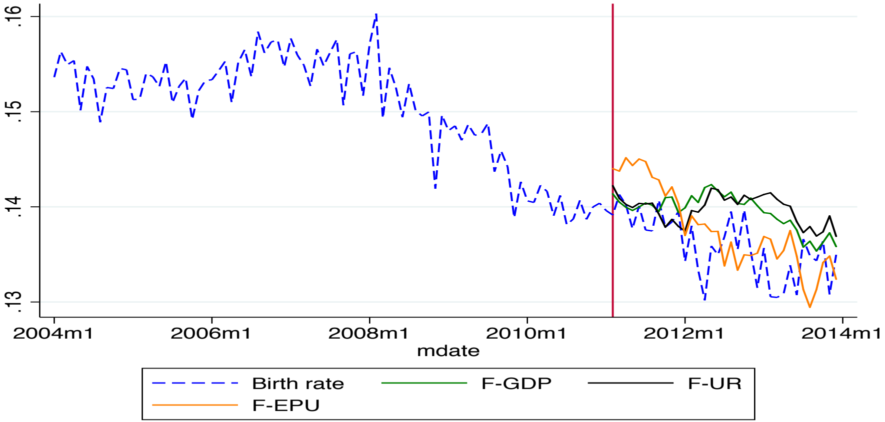
Short sample: IS: 2004M1-2008M12 - OOS: 2009M1-2013M12[1]				
$t + \ldots$	6	12	18	24
$AR(p)$ (RMSE)	3.634	3.566	3.655	4.611
$DGDP_t$	0.956	1.005	1.278	1.384
DUR_t	1.019	1.175	1.466++	1.529++
EPU_t	0.942	0.949	0.822	1.002
$GI1_t$	0.972	0.988	0.955	0.848
$GI2_t$	0.999	1.032++	1.017+	1.011
$GI3_t$	1.03	1.152	1.16	1.138
Long sample: IS: 1990M1-2008M12 - OOS: 2009M1-2013M12				
$t + \ldots$	6	12	18	24
$AR(p)$ (RMSE)	0.982	1.100	1.247	1.162
$DGDP_t$	0.890	0.968	1.103	1.063
DUR_t	0.944	1.097	1.221	1.139
EPU_t	0.859**	0.988	0.867***	0.897**

[1] *,**,*** indicate significance at 10%, 5%, and 1% respectively of the Diebold and Mariano's (1995) test of equal forecast accuracy, when competing models beats the benchmark. +,++, and +++ are defined in the same way when the benchmarks outperforms.

RMSE in red and ratios w.r.t. benchmark in black.

Long sample forecasts

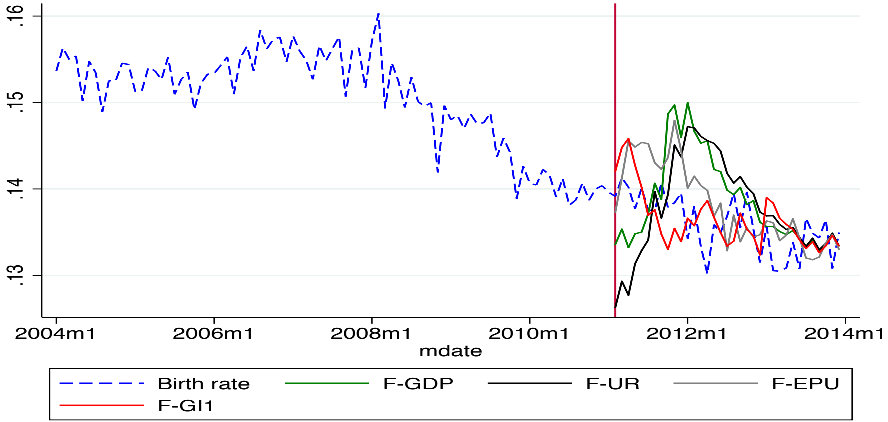
24-month-ahead - Sample: 1990M1-2013M12[2]



[2] $R = 228$, $P = 60$, In-sample = 1990:M1-2008:M12; Out-of-sample = 2009:M1-2013:M12

Short sample forecasts

24-month-ahead - Sample: 2004M1-2013M12[3]



[3] $R = 60$, $P = 60$, In-sample = 2004:M1-2008:M12; Out-of-sample = 2009:M1-2013:M12

CSSED

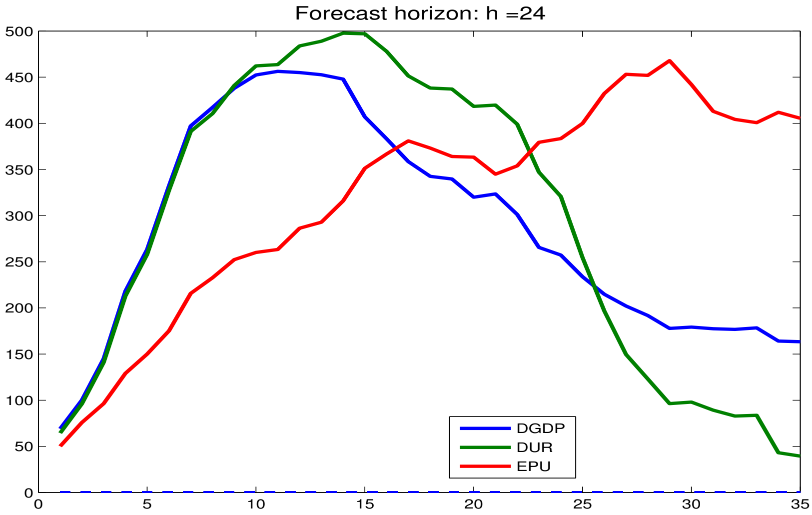
- $$\boxed{CSSED_{m,\tau} = \sum_{\tau=R}^{T}(\hat{e}^2_{bm,\tau} - \hat{e}^2_{m,\tau})} \tag{3}$$

$$\hat{e}_{k,\tau} = u_\tau - \hat{u}_{k,\tau|t} \tag{4}$$

- where $\hat{e}^2_{bm,\tau}$ is the squared forecast error of the AR benchmark model and $\hat{e}^2_{m,\tau}$ denotes the same for the competing model
- What happens if the benchmark model (bm) outperforms the competing model (m)?
- $\hat{e}^2_{bm,\tau} < \hat{e}^2_{m,\tau} \quad \Rightarrow \quad CSSED_{m,\tau} < 0$
- And if the competing model m beats the benchmark bm?
- $\hat{e}^2_{bm,\tau} > \hat{e}^2_{m,\tau} \quad \Rightarrow \quad CSSED_{m,\tau} > 0$

Predictive Relative Performance with CSSED - Long sample

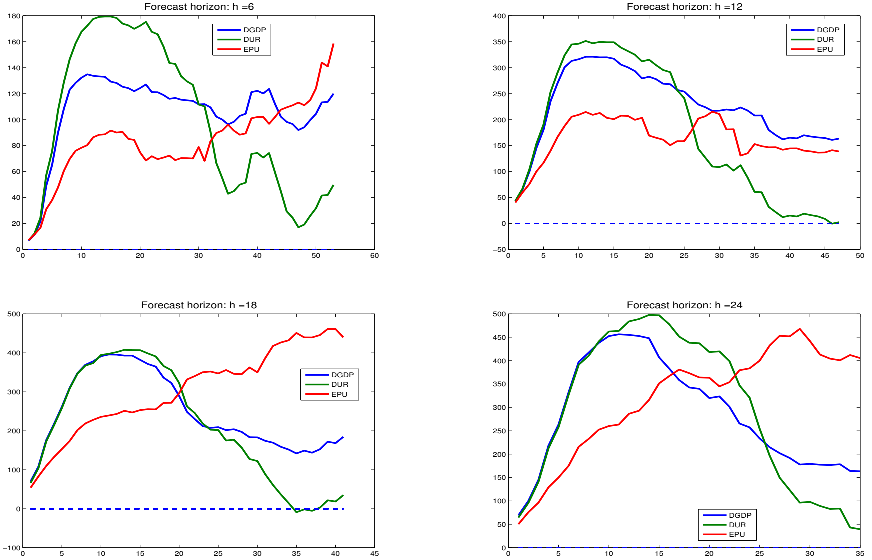
$CSSED_{m,\tau} = \sum_{\tau=R}^{T}(\hat{e}^2_{bm,\tau} - \hat{e}^2_{m,\tau})$ 'best' competing model w.r.t. the $AR(P)$ benchmark[4]



[4] $R = 228$, $P = 60$, In-sample = 1990:M1-2008:M12; Out-of-sample = 2009:M1-2013:M12

Predictive Relative Performance with CSSED - Long sample

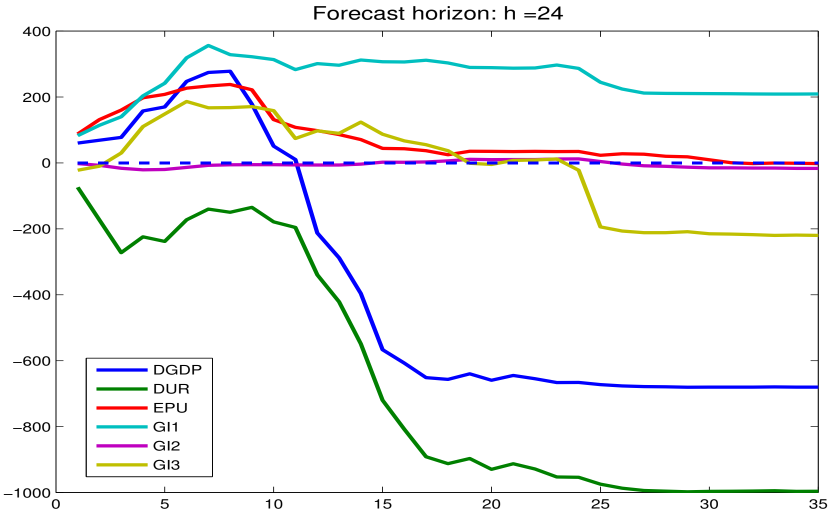
$CSSED_{m,\tau} = \sum_{\tau=R}^{T}(\hat{e}^2_{bm,\tau} - \hat{e}^2_{m,\tau})$ 'best' competing model w.r.t. the $AR(P)$ benchmark[5]



[5] $R = 228,\ P = 60$, In-sample = 1990:M1-2008:M12; Out-of-sample = 2009:M1-2013:M12

Predictive Relative Performance with CSSED - Short sample

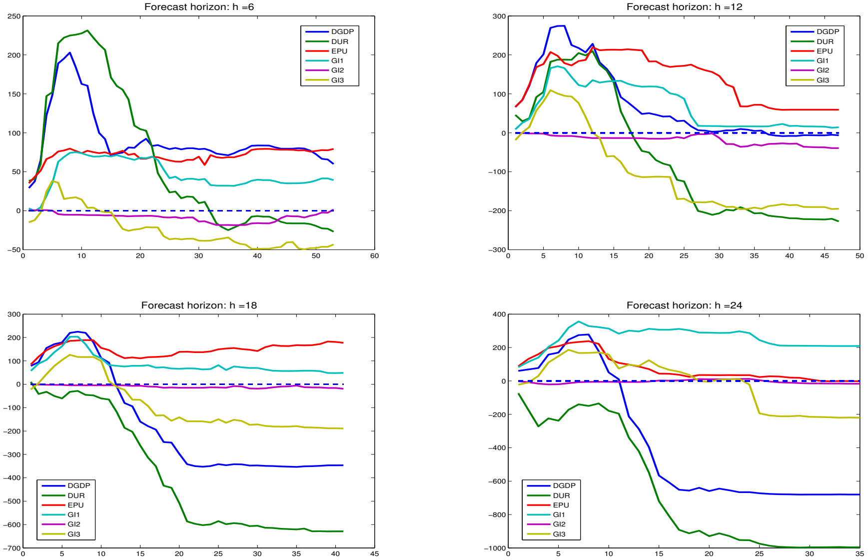
$CSSED_{m,\tau} = \sum_{\tau=R}^{T}(\hat{e}^2_{bm,\tau} - \hat{e}^2_{m,\tau})$ 'best' competing model w.r.t. the $AR(P)$ benchmark[6]



[6] $R = 60$, $P = 60$, In-sample = 2004:M1-2008:M12; Out-of-sample = 2009:M1-2013:M12

Predictive Relative Performance with CSSED - Short sample

$CSSED_{m,\tau} = \sum_{\tau=R}^{T}(\hat{e}^2_{bm,\tau} - \hat{e}^2_{m,\tau})$ 'best' competing model w.r.t. the $AR(P)$ benchmark[7]



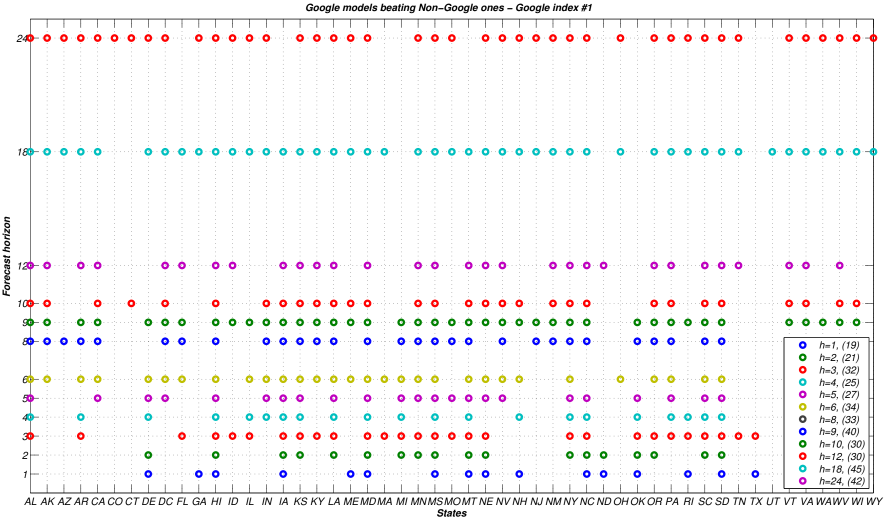
[7] $R = 60$, $P = 60$, In-sample = 2004:M1-2008:M12; Out-of-sample = 2009:M1-2013:M12

Robustness - Forecasting birth rates across states

- We repeat the same forecasting exercise for each one of the 50 US states plus DC.
- We ran the same horse race between a benchmark $AR(p)$ and an $ARX(p)$ where the leading indicator could be
  - GDP
  - UR
  - EPU (federal level)
  - GI i.e. the Google Index for '*pregnancy*'[8]
- Good results even at the state level for Google-based models
  - At 12-month ahead, Google-based models are better for 59% of states
  - At 18-month ahead, Google-based models are better for 88% of states
  - At 24-month ahead, Google-based models are better for 82% of states

[8] The GI for '*maternity*' was not populated for some states, i.e. below the Google threshold.

Robustness - Forecasting birth rates across states[9]



[9] A circle indicates the Google-based model outperforms for state on x axis at forecast horizon on y axis. States in alphabetical order.

Conclusion

- EPU index seems useful in forecasting US birth rates, at least in the long sample
- Google Trends data seems even more useful in forecasting birth rates at 12, 18 and 24 month ahead
- Google-based model outperform over the short sample
- Only caveat: Google Trends data available only from January 2004
- Google-based model tend to outperform even at the state level

Thank you!

E-mail: francesco.damuri@bancaditalia.it
E-mail: juri.marcucci@bancaditalia.it