When are Google data useful to nowcast GDP? An approach via pre-selection and shrinkage[1]

Laurent Ferrara (Banque de France)
Anna Simoni (CREST, CNRS, ENSAE, École Polytechnique)

New analytic tools and techniques for economic policy-making
OECD, April 15, 2019

[1] The views expressed here are those of the authors and do not necessarily reflect those of the Banque de France.

Is there a *Big Data hubris?*

Hot debate within institutions (NSIs, CBs, IOs), only few evidence.
2 main issues:

Is there a *Big Data hubris?*

Hot debate within institutions (NSIs, CBs, IOs), only few evidence.
2 main issues:

1. Is there a gain from using *Big Data* in nowcasting/forecasting?
   - No significant gain when controlling by *official* data (hard, soft, financial data), see e.g. Choi and Varian (12) *vs* Li (16)/ Gotz and Knetsch (17, BBK WP)

Is there a *Big Data hubris?*

Hot debate within institutions (NSIs, CBs, IOs), only few evidence.
2 main issues:

1. Is there a gain from using *Big Data* in nowcasting/forecasting?
   - No significant gain when controlling by *official* data (hard, soft, financial data), see e.g. Choi and Varian (12) *vs* Li (16)/ Gotz and Knetsch (17, BBK WP)
2. When is there a gain?
   - Significant gain when there is no information or only fragmented

Is there a *Big Data hubris?*

Hot debate within institutions (NSIs, CBs, IOs), only few evidence.
2 main issues:

1. Is there a gain from using *Big Data* in nowcasting/forecasting?
   - No significant gain when controlling by *official* data (hard, soft, financial data), see e.g. Choi and Varian (12) *vs* Li (16)/ Gotz and Knetsch (17, BBK WP)
2. When is there a gain?
   - Significant gain when there is no information or only fragmented
     - Emerging and Low Income Countries (Carriere-Swallow and Labbe, 13 JoF, Assessing real-time inflation in Venezuela and Argentina by Cavallo and Rigobon at MIT "The Billion Prices Project" ...)
     - Low frequency information: Annual World GDP (Ferrara and Marsilli, 17 World Eco.), Annual National Accounts in LICs ...
     - Lagged information (QNA nowcasts, flash estimates ...)
     - Measuring unobserved variables, e.g. Economic uncertainty (Baker et al., 16 QJE)

What do we do in this paper?

Main questions: Are Google data useful to nowcast EA GDP? And when?

What we do:

- Nowcast EA GDP on a weekly basis accounting for data releases
- Estimate a linear regression model with a shrinkage method (Ridge regression) using weekly Google data as explanatory variables, in addition to official macro variables (IPI and Survey)
- Assess the gain from using Google data when controlling from official data
- Extend the shrinkage method by pre-selecting data
- Implement a true real-time analysis

Main results

What we get:

1. We point out the usefulness of Google data in nowcasting euro area GDP for the first four weeks of the quarter when there is no information about the state of the economy
2. As soon as official data become available, that is starting from week 5 (Survey), then the relative nowcasting power of Google data progressively vanishes
3. We show that pre-selecting Google data before entering the nowcasting models appears to be a pertinent strategy in terms of accuracy (strong reduction of RMSFEs).
4. We show that the 3 previous results hold when we carry out a true real-time analysis (ie: using data vintages)

The nowcasting approach

Our goal is to nowcast EA GDP (Y) using a bridge regression involving 3 types of predictors :

$$Y_t = \beta_0 + \beta_g' x_{t,g} + \beta_s' x_{t,s} + \beta_h' x_{t,h} + \varepsilon_t \quad (1)$$

where:

- $x_{t,g}$ is the N_g-vector of Google variables
- $x_{t,s}$ is the N_s-vector containing *soft* variables (e.g. DG EcFin surveys)
- $x_{t,h}$ is the N_h-vector containing *hard* variables (e.g. IPI)

Issue: $N_g >> T$, thus we introduce a shrinkage of the OLS estimator.

Ridge regression

We use a standard approach referred to as *Ridge regression* :

$$\hat{\beta}^R = Arg\min\{\frac{1}{T}\sum_{t=1}^{T}(Y_t - \beta_0 - \beta_g' x_{t,g} - \beta_s' x_{t,s} - \beta_h' x_{t,h})^2 + \alpha||\beta||^2\} \quad (2)$$

where α is the shrinkage parameter.

The estimated coefficients $\hat{\beta}^R$ are shrunk towards zero.

How to choose the shrinkage parameter α? Generalized cross-validation by minimizing the RMSFEs over the estimation period.

Preselection of data using SIS approach

- Main idea: Pre-select Google data by targeting GDP (Bai-Ng, 2008 JoE, Schumacher, 2010, EcoLet, when dealing with dense models).

Preselection of data using SIS approach

- Main idea: Pre-select Google data by targeting GDP (Bai-Ng, 2008 JoE, Schumacher, 2010, EcoLet, when dealing with dense models).

- We follow Fan and Lv (2008, JRSS): **Sure Independence Screening** *"all important variables survive after applying a variable screening procedure, with probability tending to 1"*

Preselection of data using SIS approach

- Main idea: Pre-select Google data by targeting GDP (Bai-Ng, 2008 JoE, Schumacher, 2010, EcoLet, when dealing with dense models).

- We follow Fan and Lv (2008, JRSS): **Sure Independence Screening** *"all important variables survive after applying a variable screening procedure, with probability tending to 1"*

- Basic idea of SIS: only the variables with the highest absolute correlation with the dependent variables should be used in modelling (correlation learning, hard thresholding).

Preselection of data using SIS approach

- Only consider $\overline{X}_g$ the $T \times N_g$ matrix of average Google variables as explanatory variables.
- Let M^* be the true model with shrinkage
- Compute $\omega = (\omega_1, \ldots, \omega_{N_g})'$ the vector of absolute marginal correlations of predictors with the response variable Y,
- For any given $\lambda \in]0, 1[$, the N_g componentwise magnitudes of the vector ω are sorted in a decreasing order and we define a submodel M_λ such as: $M_\lambda = \{1 \leq i \leq N_g : |\omega_i|$ is among the first $[\lambda N_g]$ largest of all $\}$,
- Under some conditions the sure screening property holds, namely for a given λ:

$$P(M^* \subset M_\lambda) \to 1, \qquad N_g \to \infty$$

Preselection of data using SIS approach

- Ridge estimator with pre-selection is $\widehat{\beta}_R^{(w)} = (\widehat{\beta}_{R,M_\lambda}^{(w)\prime}, \widehat{\beta}_{R,M_\lambda^c}^{(w)\prime})'$

$$\widehat{\beta}_{R,M_\lambda}^{(w)} = \left(\frac{1}{T} \sum_{t=1}^{T} X_{t,M_\lambda} X_{t,M_\lambda}' + \alpha I \right)^{-1} \frac{1}{T} \sum_{t=1}^{T} X_{t,M_\lambda}' Y_t, \qquad \widehat{\beta}_{R,M_\lambda^c}^{(w)} = 0$$

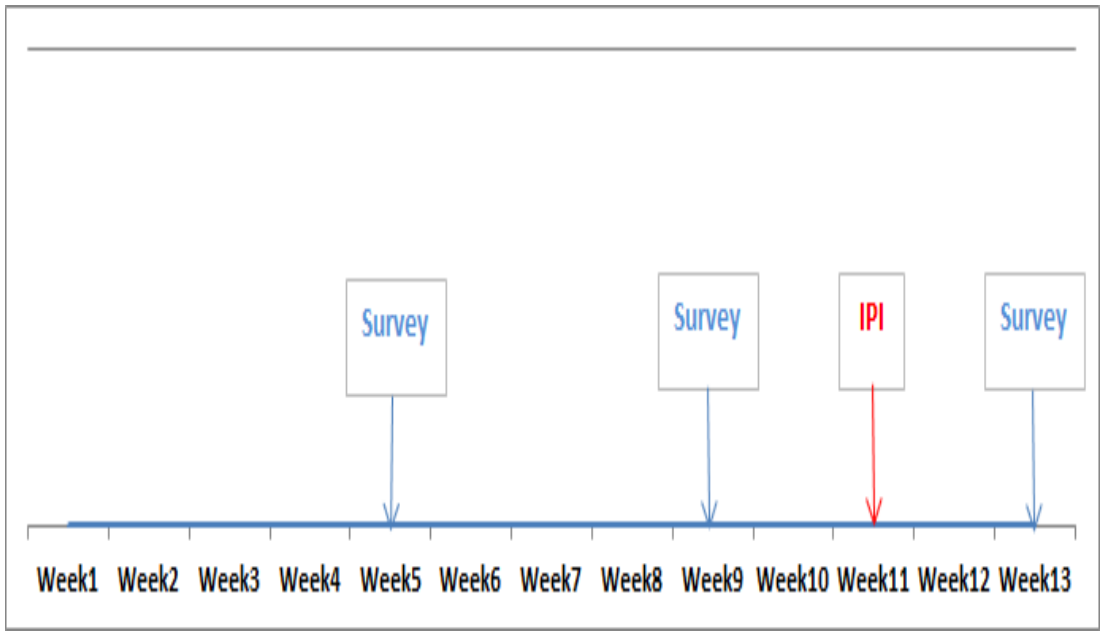
- Empirical issue: How to choose λ?
- Again, cross-validation by minimizing the RMSFEs over the estimation period

Data

- **GDP:** EA quarterly GDP growth rate (1995q1 - 2016q3) as available on 24 April 2017 from Eurostat
- **Hard data:** EA monthly IPI growth rate released by Eurostat
- **Soft data:** EA monthly composite index of various sectors : EA Sentiment Index released by DG EcFin
- **Google data:** Google search data are weekly data related to queries performed with Google data
  Data are available for the 6 largest countries (Germany, France, Italy, Spain, Netherlands, Belgium) for 296 sub-categories by countries, that is a total of $N_g = 1776$ variables.

Timeline of data releases within the quarter

Google information is available for the 13 weeks of the quarter
Survey data is available at weeks 5, 9 and 13, and IPI at week 11



Bridge models for the 13 weeks

For weeks $w = 1, \ldots, 4$, the models are of the form:

$$Y_t = \beta_{0,w} + \beta'_{1,w} x^{(w)}_{t,g} + \varepsilon_{t,w} \tag{3}$$

From week $w = 5$ to $w = 10$, the models are of the form:

$$Y_t = \beta_{0,w} + \beta'_{1,w} x^{(w)}_{t,g} + \beta_{2,w} S_t + \varepsilon_{t,w} \tag{4}$$

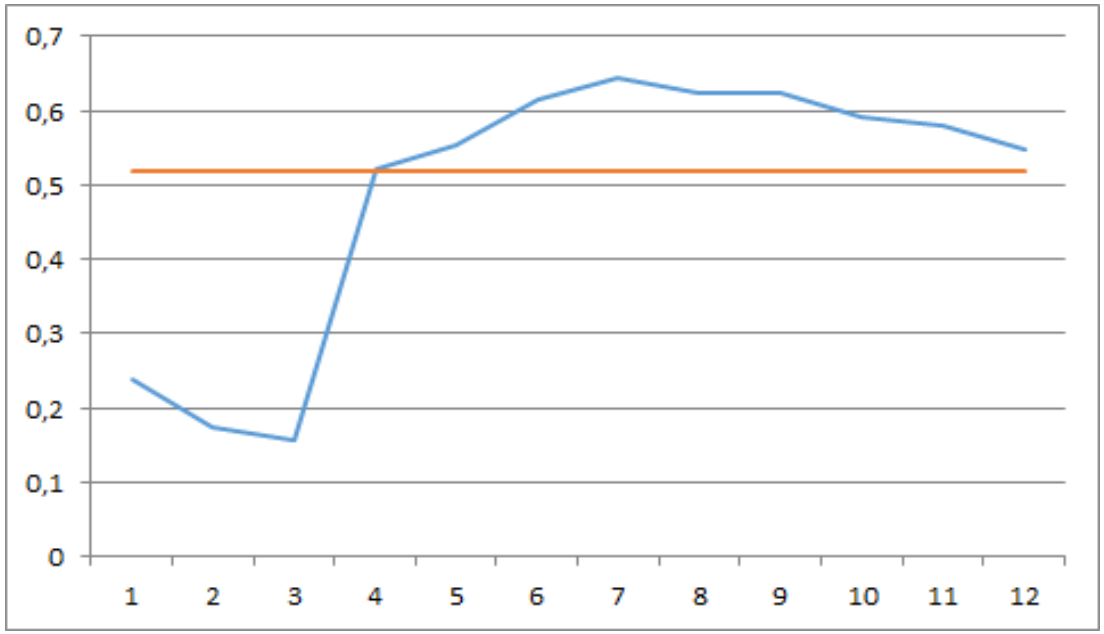
For weeks $w = 11, \ldots, 13$, the models are of the form:

$$Y_t = \beta_{0,w} + \beta'_{1,w} x^{(w)}_{t,g} + \alpha_{2,w} S_t + \beta_{3,w} IP_t + \varepsilon_{t,w} \tag{5}$$

Implementation

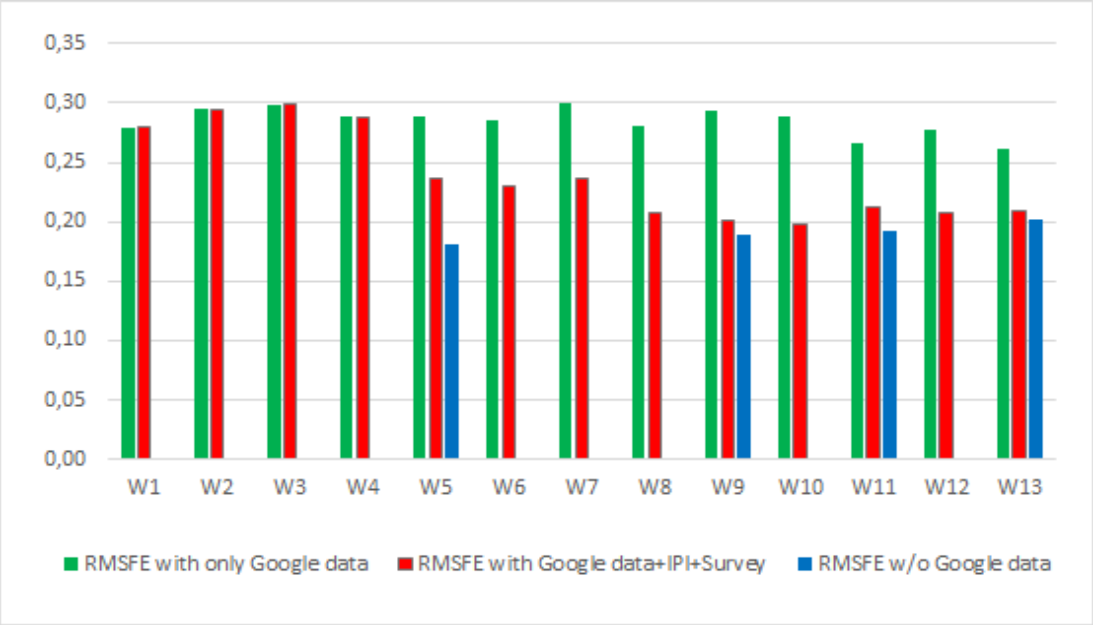
- Estimation sample: 1995q1 - 2013q4
- Nowcasting sample: 2014q1 - 2016q3
- Recursive approach over the nowcasting sample
- Parameter estimation is done for each new data, including cross-validation for hyper-parameters (α, γ)
- Mixed-frequencies issue is managed by time averaging

Example: EA GDP nowcasts for 2016q1



Are Google data useful and when?

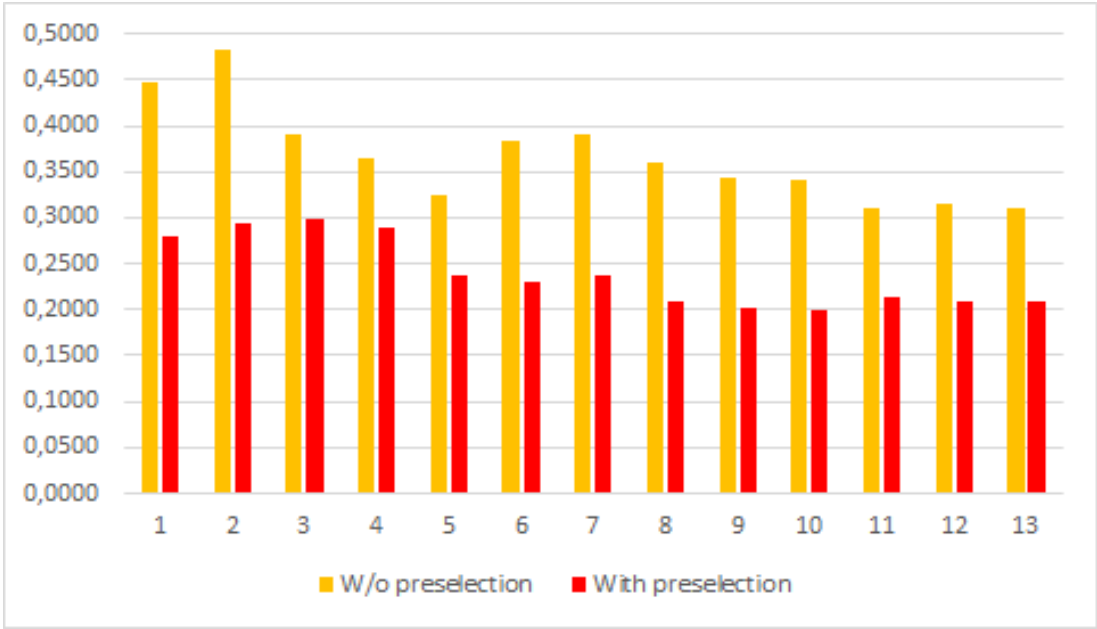
Evolution of RMSFEs within the quarter in a pseudo-real time analysis with pre-selection:



1. Decreasing pattern overtime
2. Reasonable RMSFEs until W4, ie with only Google information
3. Improvement as soon as the first survey arrives in W4

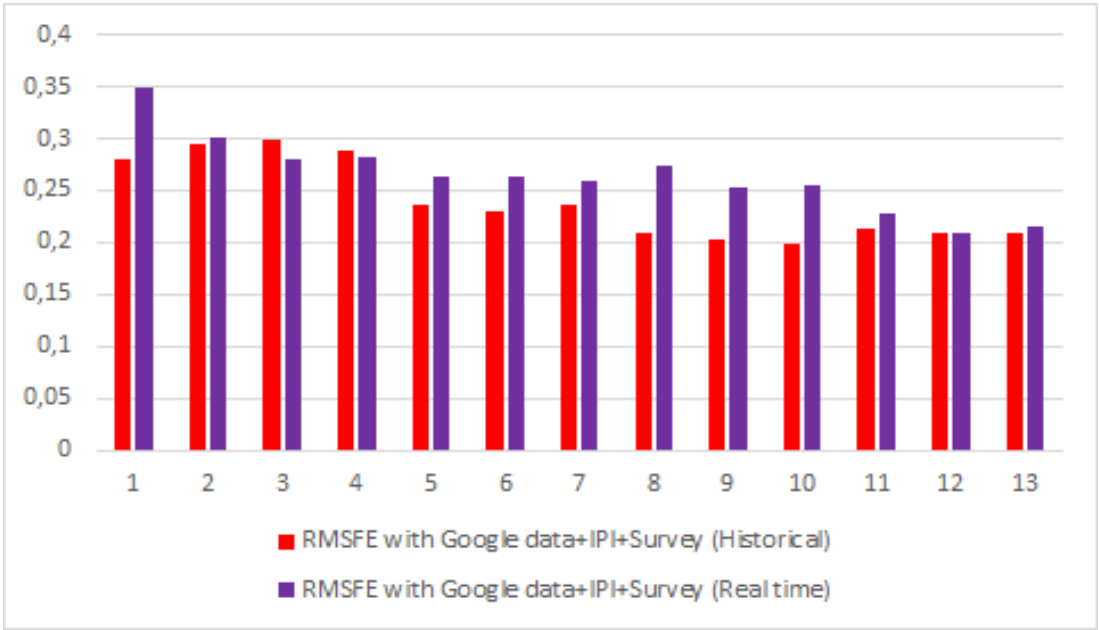
Is it worth to preselect Google data?

Compared evolution of the RMSFEs of the Ridge model with and without data pre-selection: A large gain from pre-selecting



Real-time analysis

Real-Time: RMSFEs (obtained with pre-selection) are slightly higher in real-time (expected result) but remain reasonable for the first 4 weeks



Conclusion = Main results

1. We point out the usefulness of Google data in nowcasting euro area GDP for the first four weeks of the quarter when there is no information about the state of the economy
2. As soon as official data become available, that is starting from week 5 (Survey), then the relative nowcasting power of Google data progressively vanishes
3. We show that pre-selecting Google data before entering the nowcasting models appears to be a pertinent strategy in terms of accuracy (strong reduction of RMSFEs).
4. We show that the 3 previous results hold when we carry out a true real-time analysis (ie: using data vintages)