Report on the project:

"Using Google search data for nowcasting macroeconomic and financial aggregates in OECD countries", Laurent Ferrara (Skema Business School) and Anna Simoni (CREST, CNRS)

In this project we analyze the importance of Google Search data for short-term macroeconomic and financial forecasting and nowcasting purposes. The main research questions that we want to study are: (A) when such data improve nowcasting accuracy, and (B) whether they are useful even after controlling for official variables, such as opinion surveys or production, generally used by forecasters. Answering these questions requires the use of appropriate Machine Learning/econometrics methods. Recent macroeconomic empirical literature has seen an explosion of various methods to account for the specificity of alternative data, but for most of them we do not know their out-of-sample (OOS hereafter) theoretical properties, which matter the most for forecasting purposes. Our project puts forward a new methodology to deal with Google Search data (GSD in the remaining) for nowcasting purposes and establishes OOS large-sample properties for the proposed method.

So far we have made the following contributions:

1) We have developed a nowcasting method, which we call *Ridge after Model Selection*, made of the following two-steps. (i) First, GSD variables are preselected, conditionally on the official variables, by targeting the macroeconomic aggregate to be nowcast. Preselection in the first step is based on the t-statistics associated with each GSD variable in a regression that includes the official predictors as well. (ii) Second, a Ridge regularization is applied to those preselected GSD and official variables. The Ridge tuning parameter is chosen by Generalized Cross Validation (GCV, in the remaining). Past literature has proposed forecasting approaches based on Ridge regression to deal with dense models with a large number of predictors. We go beyond this literature by considering models where the dimension can be ultra-high and that can be either sparse or dense.
2) We have provided the following theoretical contributions. First, we have proved that our targeted preselection retains all the variables in the true model with probability approaching one (Sure Screening property). Second, we have established an upper bound for both the in-sample and OOS prediction error associated with the Ridge after model selection estimator. This upper bound is a function of the number of predictors N, the number of time-series observations T and the Ridge regularization parameter α. Third, we evaluate optimality of GCV to choose the regularization parameter α for OOS prediction. To the best of our knowledge, previous literature has established in-sample optimality of the GCV in the setting of Ridge regularization but not OOS optimality.
3) We have studied finite sample properties of our procedure through a Monte Carlo exercise. Our study analyzes how the dimension of the problem, N and T, the degree of sparsity s in the model and the correlation among the predictors affect the

performance of our method compared with other widely used methods in macroeconomic nowcast, like Lasso, Ridge without preselection and Principal Component Analysis estimators. We show that when the true data generating process is sparse with a large number of active predictors our *Ridge after Model Selection* procedure outperforms all the considered competitors for OOS prediction.

4) We have conducted an empirical study to answer questions (A) and (B) stated above for GSD with respect to GDP growth nowcast for three countries/areas: the euro area, the U.S. and Germany. Usual GDP nowcasting tools integrate standard official macroeconomic information stemming, for instance, from national statistical institutes, central banks and international organizations. Typically, two sources of official data are considered: (i) hard data (production, sales, employment …) and (ii) opinion surveys (households or companies are asked about their view on current and future economic conditions). Sometimes, financial markets information, generally available on high frequency basis, is also integrated into the information set. In our study we have also considered financial market information for robustness check. In addition to these official data, we include the alternative GSD into our information set. The challenging feature of GSD as a whole is their high dimension. GSD differ from Google Trends mainly because GSD are volume variations of Google queries with respect to the first value while Google Trends provides the ratio between the search shares for a particular keyword/category over a given sub-period and the maximum search share for the same keyword/category over a chosen larger period.

   We have analyzed three different periods: a period of cyclical stability (2014q1-2016q1), a period that exhibits a sharp downturn in GDP (2017q1-2018q4) and a period of recession (the *Great Recession* period from 2008q1 to 2009q2). Overall, empirical results show that GSD are useful when trying to nowcast GDP growth. At the beginning of the quarter, when there is no official information available about the current state of the economy, we show that using only Google data leads to very reasonable Mean Squared Forecasting Errors (MSFEs), sometimes only slightly higher than those obtained at the end of the quarter when the information set is complete. As soon as we integrate official macroeconomic information, starting from the fifth week of the quarter, MSFEs decrease reflecting the importance of this type of data in nowcasting. Overall, combining macroeconomic variables and GSD variables in the same model appears to be generally fruitful.

   A striking result coming out from our empirical analysis is that, on the one hand, the preselection step is crucial in the first two periods considered as it generates better outcomes compared to nowcasting procedures without any preselection. On the other hand, recession periods present specific patterns as a model that only contains GSD, without any preselection step, tends to be preferred in terms of nowcasting accuracy. This result is quite robust over the three countries/areas that we consider in the study.

We are currently working on the following extensions (which have been suggested by two referees):

1) So far the results have been established for i.i.d. covariates. We want to extend the i.i.d. setting and allow for time dependence for some (or all) covariates. We plan to do this by assuming that some covariates are strong mixing stationary processes.
2) The upper bounds that we have established for the in-sample and OOS prediction error associated with our *Ridge after model selection* estimator are random in the sense that they depend on the model selected in the first step of our procedure. We are trying to establish a uniform bound which is independent of the selected model and so it is nonrandom.

Publications related to the Project

The paper related to the project has been submitted to *Journal of Business and Economics Statistics* in June 2020. We got a "Rejection with Resubmission". We have resubmitted a thoroughly revision of the paper in June 2021 and the journal asked for a "Revision". We are currently working on this revision.

A working paper version of the paper is available on ArXiv: https://arxiv.org/abs/2007.00273