DEPARTMENT OF ECONOMICS
AND BUSINESS ECONOMICS
AARHUS UNIVERSITY

CREATES
Center for Research in Econometric Analysis of Time Series

In Search of a Job: Forecasting Employment Growth in the US using Google Trends

Erik Christian Montes Schütte

CREATES Research Paper 2018-25

Department of Economics and Business Economics
Aarhus University
Fuglesangs Allé 4
DK-8210 Aarhus V
Denmark

Email: oekonomi@au.dk
Tel: +45 8716 5515

In Search of a Job: Forecasting Employment Growth in the US using Google Trends*

Erik Christian Montes Schütte†

August 29, 2018

Abstract

We show that Google search activity on relevant terms is a strong out-of-sample predictor of future employment growth in the US and that it greatly outperforms benchmark predictive models based on macroeconomic, financial, and sentiment variables. Using a subset of ten keywords, we construct a panel with 211 variables using Google's own algorithms to find related search queries. We use Elastic Net variable selection in combination with Partial Least Squares to extract the most important information from a large set of search terms. Our forecasting model, which can be constructed in real time and is free from revisions, delivers an out-of-sample R^2 statistic of 65% to 88% for horizons between one month and one year ahead over the period 2008-2017, which compares to between roughly 30% and 60% for the benchmark models.

JEL Classification: C22, C53, E24

Keywords: Forecast comparison, partial least squares, elastic net, complete subset regressions, bagging

*This research is supported by CREATES, Center for Research in Econometric Analysis of Time Series (DNRF78) and CONACYT, National Council for Science and Technology of Mexico.

†Department of Economics and Business Economics and CREATES, Aarhus University, Denmark, Email: christianms@econ.au.dk.

I am grateful to Thomas Quistgaard Pedersen, Jonas Nygaard Eriksen, Stig Vinther Møller, Daniel Borup, Giorgio Mirone, Eduardo Vera-Valdés, Benjamin Liengaard, and Jorge Hansen for the useful comments and discussions.

1 Introduction

Employment growth is a measure of economic expansion and regarded as a litmus test for US economic health. As such, it is a leading indicator that is important to policy makers, businesses and job seekers alike. For example, it is one of the key macroeconomic series looked at by the Federal Open Market Committee when determining the path of the federal funds rate, which is the primary tool of monetary policy used by the Fed. Additionally, job growth figures are carefully scrutinized by the media every time they are released. Thus, it is no coincidence that the word "jobs" was mentioned a total of 42 times during the 90 minutes long first presidential debate between candidates Hillary Clinton and Donald Trump in September 2016. Despite its significance, employment growth has historically been a relatively difficult macroeconomic series to forecast. A case in point is the period that covered the recession of 2008-9 and subsequent recovery, where it developed relatively different to projections made by the Bureau of Labor Statistics.

Given the salience of jobs and job growth in the minds of the US working-age population, it should not come as a surprise that latent labor market sentiment leaves a heavy footprint on internet search behavior, particularly from job seekers. A survey made by the Pew Research Center in 2015, found that 80% of the US population uses the internet when searching for a job, and 34% say that it is the most important resource available to them during the job search process (Smith, 2015). In a recent contribution, D'Amuri and Marcucci (2017) show that search volume for the term "jobs" is a strong predictor of the unemployment rate in the US. This predictability is also present in international markets, as evidenced by Askitas and Zimmermann (2009), D'Amuri (2009) and Fondeur & Karamé (2013) who find predictability for the unemployment rate in Germany, Italy, and France, respectively.[1] Nonetheless, these studies have solely focused on search volume for a single query or, at best, a small group of queries as predictors, failing to account for the inherent benefits of data rich environments. The potential for high-dimensional models to bring about significant improvements over classical univariate or low-dimensional forecasting models has been documented by several studies, among others, Stock and Watson (2002a; 2002b; 2006), Forni et al. (2005), De

[1] The evidence for the predictive power of internet search volume for macroeconomic series is not limited to the unemployment rate. Other macroeconomic variables for which there is evidence of predictability are private consumption (Vosen and Schmidt, 2011), initial claims (Choi and Varian, 2012), and building permits (Coble and Picheira, 2017).

Mol et al. (2008), Bai and Ng (2008), Kim and Swanson (2014), Groen and Kapetanious (2016).

The aim of this paper is to forecast employment using a data rich environment formed by Google search activity and as such the paper has two main contributions. The first one is to construct a real-time monitoring device for US employment growth using a broad spectrum of 211 internet search terms related to job-search activity, labor market sentiment and welfare policies. This index can be constructed instantaneously, is free from revisions, and displays much higher forecast accuracy than traditional macroeconomic, financial and sentiment variables. Our second contribution is to adapt state-of-the-art methods for forecasting with high-dimensional panels in a macroeconomic setting to Google search activity and show that this results in much higher predictive power than models based on a single keyword. By combining a large and heterogeneous set of Google search terms, we benefit in three important ways. First, each additional regressor has the potential of contributing with supplementary information. Second, the inclusion of different terms can possibly alleviate sample selection issues that arise due to variation in internet use across different groups by income and age since semantically related terms can potentially capture the same type of information but across distinctive demographical groups. Third, it minimizes the impact of noise in the data that arises due to changes in search terms or behavior across time.

Google Trends has several advantages over classical statistical measures used for macroeconomic forecasting. More specifically, official statistics are usually released with a lag and they are subject to substantial revisions. Household and business surveys can be more timely and they are relatively free from revisions but they are costly to obtain and might suffer from selection biases in response rates. Google Trends on the other hand, can be obtained in real time, be restricted to specific geographical areas, and can even be obtained at daily frequencies. Moreover, the ease with which you can download additional Google Trends series makes it easy to expand the panel of predictors.

Starting with a set of ten keywords, we use Google's own algorithms to find semantically linked search queries and thereby expand the panel to a high-dimensional setting. We then use soft thresholding variable selection based on Elastic Net, as proposed by Bai and Ng (2008), to choose the best thirty predictors within this large panel. We further reduce the dimensionality of these selected predictors into common components by using Partial Least Squares. This procedure, which we call targeted PLS, yields consistently superior performance to benchmarks models for horizons between one month and one year ahead, producing out-of-sample R^2 measures between 65% and 88%, which compares to 30% to 60% for models consisting of macro, financial, and sentiment variables. In contrast to the benchmark models, the targeted PLS model is particularly adept at forecasting employment growth during the latest recession and recovery that followed. We also compare the targeted PLS Google-based model to a PLS model that uses the same data but does not involve soft threshold variable selection. We find that the latter model delivers out-of-sample R^2 measures between 20% and 40%, implying that the pre-selection procedure is important since many of the search keywords harm performance by inducing noise in the estimated PLS factors.

The general superior performance of the model appears to arise from the combination of heterogeneous search queries with its flexibility to let the selected keywords vary over time. We investigate whether combining the Google Trends data set with the benchmark data improves on the forecasting performance of the targeted PLS model and find that the improvements are minimal, particularly at horizons below six months. This is noteworthy, since it is especially at these short horizons that official statistics are comparatively unsuitable due to publication lag and the possibility of future revisions. Finally, we show that our results are robust to the choice of search terms used to build the data set, estimation window, and to whether the data is detrended or differenced.

In Section 2 we present the methodology used to construct the panel of predictors for both the Google Trends data and the benchmark data set. Section 3 introduces the main models we use to forecast employment growth as well as the methods we use to draw inference on the results of the horse race. In Section 4 we present empirical findings, compare alternative models build, and discuss our results. We show the robustness of the results in Section 5. Finally, in Section 6 we present some concluding remarks.

2 Data

The sample that we use for this analysis spans from 2004:M1 to 2018:M1 and has a monthly frequency. The starting date is determined by the availability of Google Trends data. We obtain data for our target variable, seasonally adjusted employment growth in the US, from the Bureau of Labor Statistics.

[2] http://www.google.com/Trends

Our set of search volume data predictors are obtained from Google Trends, which provides a time series index on the proportion of queries for a search term in a given geographical area.[2] The proportion of queries for a particular keyword is normalized by the total amount of Google searches in the selected geographic area and time range. The resulting number is then scaled on a range between 0 and 100 such that the maximum volume for the particular query in the selected time period takes the value 100. Due to privacy concerns, Google Trends does not explicitly provide its users with the actual number of queries made for each keyword. Nonetheless, for the purpose of forecasting, this does not represent a problem since we are only interested in the time series dynamics of relative search activity. A very useful feature of Google Trends is that, for each keyword, the user is provided with a list of up to 25 *related terms* (also referred to as *related queries* henceforth).[3] According to Google, *related terms* are selected by looking at terms that are most frequently searched with the term you entered within the same search session. Although the precise algorithm that determines the *related terms* is proprietary, the output is generally intuitive. For example, querying for the term "*food stamps*" in the US for the period of interest returns a list of 25 *related terms* of which the top five are*: "food stamps apply", "apply for food stamps", "florida food stamps", "food stamps application", "food stamps online".* From a forecasting perspective, this functionality is appealing for three reasons: *i)* each semantically related keyword can potentially provide additional information about the target variable and thereby truly harness the predictive power of "big data"; *ii)* the algorithm performs a form of variable selection since it selects queries with high search volume that might be unknown to the researcher; *iii) related terms* are likely to be cross-sectionally correlated, which creates a natural factor structure between them, a feature that we can exploit to our advantage when estimating the forecasting model.

To construct the main set of predictors, which we denote $\boldsymbol{X}_g$, we start by selecting and downloading search volume series for ten queries: "*salary*"*, "jobs classifieds", "job listings", "companies hiring", "entry level jobs", "food stamps", "collect unemployment", "disability", "unemployment office", "welfare".* We follow Da et al. (2014) and call these words primitive queries (or alternatively primitive terms). Figure 1 shows the Google Trends for our primitive queries over the period

[3]Note that Google divides *related queries* into two main categories: *top* and *rising*, we use the top *related terms* in our analysis. The final number of related queries depends on the search volume of the original query, i.e. relatively low volume series will have fewer than 25 *related terms*.

2004:M1 to 2018:M1. Our criterion for selecting these primitive queries is based on a discontinued service by Google Finance called "Domestic Trends".[4] The idea behind "Domestic Trends" was to select a set of Google Trends that had predictive power within a certain category, i.e. advertising, air travel, luxury, jobs, unemployment etc., and construct an equal-weighted index that could act as a leading indicator within the category.[5] We select the first five words as those Google Finance used to construct the "Domestic jobs index" and the latter five as those used to construct the "Domestic unemployment index". Figure 1 also shows how some of the queries, i.e. *"jobs classified", "companies hiring", "collect unemployment"* and *"unemployment office"* increase steeply during the financial crisis as a result of the large surge in unemployment during this period which led an increasing amount people to look for job opportunities or unemployment benefits over the internet.

For each of the ten primitive queries, we add their *related terms* and remove duplicates, low volume series and series that are clearly unrelated to the employment sentiment.[6] This methodology follows Da et al. (2014), who start with a set of primitive queries and then add *related terms* (removing duplicates, low volume series and unrelated queries) to enrich the data set. Our raw data set (excluding duplicates) has 245 keywords that become 231 after removing low volume queries and 211 once economically unrelated terms are removed. As noted by D'Amuri and Marcucci (2017), Google Trends are created based on a sample of queries that change according to the time and IP address of the computer used to download the data. To account for sampling error, we compute the index for all Google Trends queries based on an average over 20 different days. The correlation across different samples is always above 0.99, hence, the results are, for all practical concerns, not sensitive to this precaution.

Following Da et al. (2011;2014) and Vozlyublennaia (2014), we start by converting the series to

[4]The domestic Trends service was removed from the Google Finance home page in January 2018. We are not aware of the reason behind this decision.

[5]The methodology that Google Finance used to select the keywords that composed each domestic trend index was not disclosed, but the queries are generally intuitive. For example the domestic "luxury index" was composed by an equal weighted index of *"jewelry","jewelers", "diamond", "rings"* and *"tiffany"*.

[6]We define low volume series as those for which less than 85% of the observations are larger than 0. Da et al. (2014), working with data at a daily frequency, define low volume series as those for which there is less than 1,000 positive observations in their sample. Economically unrelated terms are those which are clearly unrelated to the main query from an economic or sentiment perspective. For example, "*animal welfare*" and *"child welfare"* are among the related terms for the query *"welfare"* and we cannot expect these terms to have any predictive power for employment growth. Although the Elastic Net soft threshold we utilize for our main forecasting model is generally successful at removing these terms, they are not removed from the PLS model that does not involve variable selection. Thus, we remove them from the raw data set.

their natural logarithm. This is primarily done to account for the high volatility in some of the series. Looking at Figure 1, there are two other things that stand out from Google Trends data: *i)* they contain a strong yearly seasonal component, *ii)* the series appear to be relatively heterogeneous in terms of their order of integration and whether they contain deterministic Trends.[7] We account for the former by applying the Seasonal and Trend decomposition using Loess (STL) method proposed by Cleveland et al. (1990). To address the second point, we adopt a sequential testing strategy in the spirit of Ayat and Burridge (2000). The idea is to successively test for stationarity, linear trend stationarity and quadratic trend stationarity using an augmented Dickey-Fuller (ADF) test. Hence, the first test is an ADF test with a constant term. If the null of non-stationarity is rejected, we stop and use the series without any transformation; conversely, if the null is maintained, we use an ADF test that includes both a constant and a linear time trend. If the null of this second test is rejected, we linearly detrend the series by using the residuals of a regression of the series on a constant and a time trend, otherwise we run a final ADF test that includes a constant, a linear trend and a quadratic trend. If we reject the null of this test, we detrend the series by a similar methodology as before but including a quadratic trend in the regression, otherwise we take first differences. All ADF tests are performed with a maximum lag length of 4 with the optimal number of lags selected by the BIC. Following Ayat and Burridge, we conduct each sequential test at the 5% level.[8] To avoid look-ahead bias, we deseasonalize and perform the sequential testing for unit roots on a recursively expanding window, where the smallest window used matches our estimation window for the forecasting model. Hence, only information available at time t is used in both procedures.[9] Figure 2 shows four log deseasonalized queries exemplifying all the possible transformations that each series can undergo when applying the Ayat and Burridge procedure.

[7]There is indeed no consensus on the literature as to whether or not Google Trends data is best characterized by stationarity, trend stationarity or a unit root since this appears to be completely dependent on the query in question. Varian and Choi (2012), Vozlyublennaia (2014), Bijl et al.(2016) and D'Amuri and Marcucci (2017) do not perform any differencing or detrending of the series, which suggests that the Google Trends they use are stationary. Yu et al. (2018) use an ADF test on three Google Trends queries: *"oil inventory", "oil consumption"* and *"oil price"* and find evidence of stationarity at the 5% level (10% level) in *"oil inventory" ("oil consumption"),* but are not able to reject the null of a unit root for *"oil price".* Da et al. (2014) take log-differences on the series.

[8]Ayat and Burridge (2000) note that the procedure is able to retain relatively good size even though multiple tests are involved.

[9]Note that this can result in some discordance (across time) about the presence of a unit root or deterministic Trends in some series. In particular, due to the low power of unit root tests in small samples, some of the series might be initially characterized as having a unit root and later on, as more information becomes available, they will be characterized as stationary or trend stationary. In Section 5 we also show the results when using first differenced data.

For the series in the top panel, *"salary calculator"*, we find strong evidence of stationarity (no transformation). For the series in the second panel, *"career opportunities"*, we find evidence of linear-trend stationarity (linear detrending). For the series in the third panel, *"help wanted"* we find evidence of quadratic-trend stationarity (quadratic detrending), Finally, for the one in the bottom panel, *"jobs classifieds"*, we cannot reject the null of a unit root (first differences).[10] Note that the latter series is not a related term but a primitive term. Hence, the effect of taking the log transform and deseasonalizing can also be seen by comparing the raw series data, shown in the upper right panel of Figure 1 with the lower left panel in Figure 2, which is log transformed, deseasonalized and standardized.

2.1 Benchmark data set

Our benchmark data, which we denote $\boldsymbol{X_{mfs}}$, is composed of twenty predictors, where the first sixteen are inspired by the set of macroeconomic and financial leading indicators that Rapach and Strauss (2008;2010;2012) use to forecast employment growth rates in the US.[11] Table 1 displays the series used as well as the transformation applied to each of one them.

All nominal variables are converted to real by deflating them using the personal consumption deflator. The variables outlined above are mainly macroeconomic and financial. However, since Google Trends is arguably partially capturing sentiment it is conceivable that a more equitable comparison will incorporate sentiment variables. Consequently, we also include four sentiment covariates which are shown in Table 2 below.

The $UMICS$ has been shown to have predictive power for, among other variables, household spending (Carroll et al., 1994), GDP growth (Matsusaka and Sbordone, 1995), stock returns (Lemmon and Portniaguina, 2006) and recessions (Christiansen et al., 2014) in the US. D'Amuri and Marcucci (2017) report that the EPU, $EENM$ and EEM indices have a relatively good predictive power for the US unemployment rate.[12]

[10] The ADF test statistics over the whole sample are: -10.11 for *"salary calculator"*, -4.63 for *"career opportunities"*, and -5.46 for *"help wanted"*, all of which are significant at the 1% level, the trend coefficients for the latter two are also significant. The ADF test statistics with a constant (-0.519), linear trend (-2.48) and quadratic trend (-2.93) over the whole sample for *"jobs classifieds"*, are not significant at the 5% level.

[11] Note that some of the predictors used by Rapach and Strauss (2008;2010;2012) are from the Conference Board's Labor Force, Employment, and Unemployment database which is unavailable to the author.

[12] The sources for each of the variables in $\boldsymbol{X_{mfs}}$ are shown in Table A1 in Appendix A.

3 Forecasting methodology and inference

In this section, we outline our empirical methodology and briefly describe the methods we use to draw inference on the performance of the models. When forecasting with Google Trends data, $\boldsymbol{X_g}$, we consider two models: targeted PLS and PLS since both frameworks can account for the large cross-sectional dimension of the data. In an effort to make a methodologically fair comparison that can keep model effects constant, we also use PLS on our benchmark data set, $\boldsymbol{X_{mfs}}$. Given the lower cross-sectional dimension of $\boldsymbol{X_{mfs}}$, it is possible that other forecasting methodologies are more appropriate. Thus, we also include bagging and the Complete Subset Regressions (CSR) method of Elliot et al. (2013). We include the former because Rapach and Strauss (2012) show that it can produce significant improvements in employment growth forecast accuracy over the autoregressive benchmark. Since the bagging model nests a (bootstrap) autoregressive model, we do not include AR specifications in our main set of benchmark models.[13] CSR is included because Elliot et al. (2013) report that this forecast combination approach shows strong performance when compared to alternative forecasting techniques such as ridge regression, bagging and LASSO.

Let our target variable, which is the h month ahead employment growth rate, be defined as $y^h_{t+h} = (1/h)\sum_{j=1}^{h} y_{t+j}$, where y_t is the log-difference of the seasonally adjusted employment growth at time t. Let us also define our $N \times 1$ vector of predictors at time t by $X_t = [X_{1,t,...,}X_{N,t}]^{'}$. Note that this should not be confused with the matrix of predictors, e.g. $\boldsymbol{X_g}$ or $\boldsymbol{X_{mfs}}$, which we denote with bold letters.

3.1 Targeted predictors

Targeted predictors, which combines shrinkage methods and factor models, was proposed by Bai and Ng (2008) to take into account the fact that not necessarily all series in X_t are important when forecasting the target variable. The idea is to first pre-select a subset $X^*_t \in X_t$ and then estimate a factor-based model using only X^*_t. Bai and Ng (2008) propose two methods for constructing X^*_t, soft and hard thresholding. For the sake of brevity, we focus only on soft thresholding, which is based on dropping uninformative regressors using penalized regressions. More specifically, we use the Elastic Net (EN) estimator of Zou and Hastie (2005) since it performs well when predictors are

[13] Consistent with the findings of Rapach and Strauss (2012) we find that the bagging forecasts outperform the AR(q) (with q chosen by the AIC) and hard threshold ADL models at all forecast horizons.

correlated.[14] If we let RSS be the residual sum of squares of a regression of y^h_{t+h} on X_t, Elastic Net solves the problem:

$$\hat{\beta}^{EN} = \underset{\beta}{argmin} \left[RSS + \lambda \left((1-\alpha)\frac{1}{2} \parallel \beta \parallel^2_{\ell_2} + \alpha \parallel \beta \parallel_{\ell_1} \right) \right] , \tag{1}$$

where $\alpha = (0, 1]$ selects a weight between the LASSO and ridge regression, λ is a tuning parameter and $\|\cdot\|_{\ell_i}$ denotes the ℓ_i norm for $i = \{1, 2\}$.[15] Thus, we can construct the soft threshold X^*_t by:

$$X^*_t = \left\{ X_i \in X_t \mid \beta^{EN}_i \neq 0 \right\} \text{ with } i = 1, ..., N. \tag{2}$$

We follow Bai and Ng (2008) and tune λ such that 30 predictors are selected. We set $\alpha = 0.5$, which means that ridge and LASSO regression get an equal weight.[16] Hence, the idea is to use the Elastic Net estimator to remove uninformative predictors from X_t and thereby improve on the forecast of the target variable.

3.2 PLS

Partial Least Squares (PLS) was originally proposed by Wold (1966) as a factor-based dimensionality reduction technique. The method is related to Principal Components, but instead of finding linear combinations of X_t that maximize explained variance, PLS finds linear combinations that maximize the covariance between the target and the explanatory variables. Our implementation is based on the generalization of PLS proposed by Kelly and Pruitt (2015). This estimation methodology is convenient since it relies only on OLS regressions to estimate the factors and forecast regression. The algorithm for the estimation and forecasting with PLS using L factors is as follows:

1. Initialize $Z_t = y^h_{t+h}$, for $t = 1, ..., T$, and set k (an indicator for the intermediate number of factors), equal to 1.
2. Run a time series regression of $X_{i,t}$ on $\boldsymbol{Z}_t$ for $i = 1, ..., N$, $X_{i,t} = \phi_{0,i} + \boldsymbol{Z}'_t\boldsymbol{\phi}_i + \epsilon_{i,t}$, and save

[14]The results for soft thresholding using the LASSO estimator of Tibshirani (1996) are presented in Table B1 in Appendix B. We find that using the LASSO instead of the Elastic Net does not alter the results in any significant way.

[15]Both the LASSO (Tibshirani, 1996) and ridge estimators work by regularizing the coefficients of unimportant or irrelevant predictors towards zero. The main difference is that ridge will only decrease the absolute size of the coefficients but it will never set them exactly equal to zero. In contrast, the LASSO is able to set the coefficients to zero and thus perform variable selection.

[16]Bai and Ng (2008) employ an alternative version of the Elastic Net estimator. Their implementation selects a second regularization parameter λ_2 for ridge instead of a weight between the two estimators.

the slope estimates $\hat{\boldsymbol{\phi}}_{\boldsymbol{i}}$.

3. Run a cross section regression of $X_{i,t}$ on $\hat{\boldsymbol{\phi}}_i$ for $t = 1, ..., T$, $X_{i,t} = \phi_{0,t} + \hat{\boldsymbol{\phi}}_{\boldsymbol{i}}' \boldsymbol{F_t} + \varepsilon_{i,t}$, and save the slope estimates $\hat{\boldsymbol{F}}_t$.

4. Run time series regressions of y^h_{t+h} on the predictive factor(s) $\hat{\boldsymbol{F}}_{\boldsymbol{t}}$, $y^h_{t+h} = \beta_0 + \hat{\boldsymbol{F}}_t' \boldsymbol{\beta} + \eta^h_{t+h}$. This delivers the PLS forecast with k factors $\hat{y}^{h,k}_{t+h}$. If $k = L$ stop, otherwise set $k = k+1$ and $\boldsymbol{Z} = \boldsymbol{y} - \hat{\boldsymbol{y}}^k$, where $\boldsymbol{y} = \left[y^h_{1+h}, y^h_{2+h}, ..., y^h_{T+h}\right]'$ and $\hat{\boldsymbol{y}}^k = \left[\hat{y}^{h,k}_{1+h}, \hat{y}^{h,k}_{2+h}, ..., \hat{y}^{h,k}_{T+h}\right]'$, and repeat step 2 through 4.

We consider PLS forecasts with $L = \{1, 2, 3\}$ factors. When the PLS model is estimated using the subset of variables selected by soft thresholding, we call it targeted PLS.

3.3 Bagging

Our implementation of bagging follows the lines of Inoue and Killian (2008) and Rapach and Strauss (2012). We start with the general autoregressive distributed lag (ADL) model that includes q_{max} autoregressive terms and N exogenous predictors, $y^h_{t+h} = \alpha + \sum_{j=1}^{q_{max}} \beta_{j,t} y_{t-j+1} + \sum_{i=1}^{N} \delta i, X_{i,t} + \varepsilon^h_{t+h}$. We then select the autoregressive lag by the AIC and apply a hard threshold on the variables in X, such that only the statistically significant variables remain.[17] Thus, the hard threshold ADL forecast for critical value t_c is given by:

$$
\begin{gathered}
\hat{y}^h_{t+h} = \hat{\alpha} + \textstyle\sum_{j=1}^{q} \hat{\beta}_j y_{t-j+1} + \sum_{i=1}^{N} \hat{\delta}_i X^*_{i,t} \\
q = \underset{1 \le q \le q_{max}}{argmin}\ AIC(q) \\
X^*_t = \{X_i \in X_t \mid |t_{X_i}| > t_c\} \text{ with } i = 1, ..., N.
\end{gathered}
\tag{3}
$$

The procedure is then augmented by using a moving block bootstrap to reduce variance coming from model uncertainty. More specifically, we generate B bootstrap samples by randomly drawing blocks of size m from the $\{y_{t+h:T}, X_{t:T-1}\}$ tuple. We then calculate (3) for each bootstrap sample using information until time t, and compute the hard threshold ADL bootstrap forecast, $y^{h,b}_{t+h}$ using the actual values of $y_{t-j+1:t}$ and bootstrap coefficients. The bagging forecast for $\hat{y}^h_{t+h}$ is then given

[17] Following Rapach and Strauss (2012) we use Newey and West (1987) standard errors to calculate the t-statistic. The lag truncation is set to $h - 1$.

as the average of the B hard threshold ADL bootstrap forecasts:

$$\hat{y}_{t+h}^{h,Bagg} = \frac{1}{B}\sum_{b=1}^{B}\hat{y}_{t+h}^{h,b}. \tag{4}$$

We maintain the autocorrelation structure of the target variable by applying the circular block bootstrap of Politis and Romano (1992) with block size chosen optimally according to Politis and White (2004).[18] We use $B = 500$, set $q_{max} = 4$ and $t_c = \{1.645,\ 1.96,\ 2.58\}$.

3.4 Complete subset regressions

The Complete Subset Regressions (CSR) method of Elliot et al. (2013) is based on the idea of taking all combinations of models restricted to use a fixed number of regressors $k < N$. Specifically, if we let $X_{m,t}$ denote the matrix of predictors containing k variables for each model $m = 1, ..., M$, the complete subset regression forecast is given by:

$$\begin{array}{c} \hat{y}_{t+h}^{h,m} = c + \hat{\beta}X_{t,m} \\ \hat{y}_{t+h}^{h,CSR} = \frac{1}{M}\sum_{m=1}^{M}\hat{y}_{t+h}^{h,m} \end{array}. \tag{5}$$

We select model combinations that include a maximum of $k = \{6, 9, 12\}$ variates. This choice is dictated by the number of variables in the benchmark data set, $\boldsymbol{X_{mfs}}$, which is 20.

3.5 Inference

We compare the performance of the competing models using the Campbell and Thomson (2007) out-of-sample R^2:

$$R_{OoS}^2 = 1 - \frac{\sum_t \left(y_t - \hat{y}_{m,t}\right)^2}{\sum_t \left(y_t - \bar{y}_t\right)^2}, \tag{6}$$

where $\bar{y}_t$ is the rolling-mean forecast, which is computed on a window that matches the model estimation window and $\hat{y}_{m,t}$ is the forecast of the model in question at time t. This measure lies in the range $(-\infty, 1]$, with negative numbers indicating that the model in question performs worse than the historical mean of the series. We conduct out-of-sample inference using the Diebold and

[18] For robustness we also used $m = h$ as Inoue and Killian (2008), but the results are insensitive to this alteration since the Politis and White (2004) method tends to choose an optimal block size close to h.

Mariano (1995), t_{DM}, test statistic. The null hypothesis of the Diebold and Mariano test used in this paper is that the model in question does not beat the rolling average of the series, while the alternative is that it does. Hence, it can be interpreted as the t-statistic of the R^2_{OoS}. When forecasting for horizons $h > 1$, we adjust for the moving average structure of the forecast errors by using Newey and West standard errors in the denominator of the test statistic with a bandwidth length equal to h. Since we are dealing with a large number of models, we use the model confidence set (MCS) approach developed by Hansen et al. (2011) to compare the performance of the models. This approach returns a confidence set that includes the best model with probability (1-α), i.e. small values of α will make the confidence set broader and thus include more models. We use squared forecast errors as a loss function and set the bootstrap block size equal to h when applying the MCS. We rely on the range statistic, T_R, to draw inference.

A positive R^2_{OoS} measure tells us that the model in question outperforms the rolling-mean benchmark by looking at the ratio of forecast errors over the whole out-of-sample period. However, it is possible that the model in question is only beating the rolling-mean during a subset of the evaluation period and it is underperforming during others. To formally look at the stability of forecast accuracy, we follow Goyal and Welch (2008) and compute and plot the cumulative sum of squared error difference (CSSED) between the model of interest and the rolling-mean model. The CSSED for model m at time t is computed as:

$$CSSED_{m,t} = \sum_{i=R}^{t} \left((y_t - \bar{y}_t)^2 - (y_t - \hat{y}_{m,t})^2 \right), \tag{7}$$

where R and t are the beginning and the end of the forecast evaluation period, respectively. For any t, a positive $CSSED_{m,t}$ means that model m is outperforming the rolling-mean. Positive changes in the slope mean that model m is improving against the rolling-mean benchmark and vice versa for negative changes.

4 Empirical results

This section presents the out-of-sample performance of our competing models. We then investigate where this predictive performance might arise from and finally we combine the Google Trends data set with our benchmark data set to see if combining the two results in improved forecasting power.

4.1 Employment growth predictability

Table 3 shows the R^2_{OoS}, t_{DM} test statistic, and an inclusion indicator for the 90% and 95% MCS for the competing models. All models are estimated using a rolling window scheme with 48 observations in the sample period 2004:M1 - 2018:M1. We forecast at horizons of $h = \{1, 2, 3, 4, 5, 6, 9, 12\}$ months ahead. Thus the first forecast for horizons between one and six months ahead occurs during the recession of 2008-09, allowing us to assess the performance of the models during a large contractionary period in the labor force. Panel A and Panel B show the targeted PLS, $t - PLS_{\boldsymbol{X_g}}(L)$, and PLS, $PLS_{\boldsymbol{X_g}}(L)$, models with Google Trends data, respectively. Panels C, D and E show the PLS, $PLS_{\boldsymbol{X_{mfs}}}(L)$, CSR, $CSR_{\boldsymbol{X_{mfs}}}(k)$, and bagging, $Bagg_{\boldsymbol{X_{mfs}}}(t_c)$, models with benchmark data, respectively.

The results in Panel A (Table 3) for $t-PLS$ models, which only include the thirty queries in $\boldsymbol{X_g}$ selected by soft thresholding, show that the method demonstrates a striking degree of forecasting power with R^2_{OoS} measures between 65.6% and 87.8%, depending on the forecast horizon and the number of factors in the model. All forecasts in Panel A beat the rolling-mean benchmark and the results are generally significant at the 5% level. The forecasting performance of the models increases in the number of factors, but the last two factors account for a decreasing marginal amount of predictive information. We do not show the results for a model with four factors since we find that out-of-sample predictive performance decreases with the inclusion of a fourth factor. The results in the table are not strictly making a comparison between models, however, the relative performance of the model is impressive, particularly at horizons of $h = 1$ and $h = 12$ where it achieves an R^2_{OoS} that is more than twice as large as the second best model in the table.

Panel B shows the results for the PLS model with all the series in $\boldsymbol{X_g}$. Although these models include a larger set of variables, and arguably more information since all the predictors in the $t - PLS_{\boldsymbol{X_g}}$ models are also included in the $PLS_{\boldsymbol{X_g}}$ models, taking a "kitchen-sink" approach significantly reduces performance. This implies that although some of the series in $\boldsymbol{X_g}$ are indeed containing valuable information, many of them are irrelevant and only induce noise. In fact, although $PLS_{\boldsymbol{X_g}}$ models generally achieve positive R^2_{OoS}, the best performing model, $PLS_{\boldsymbol{X_g}}(2)$ only produces statistically significant measures at the 10% level for horizons between one and four months ahead.

Before discussing the results, it is worth noting that the variables in the benchmark data, $\boldsymbol{X_{mfs}}$, are selected specifically because there is evidence in the literature of their predictive ability for either the unemployment rate or employment growth. Hence, a pre-selection procedure on a larger set of variables (similar to the one soft thresholding makes for $t-PLS$) has already been performed to arrive at this data set. When looking the benchmark models (Panels C, D and E), we find that the best $PLS_{\boldsymbol{X_{mfs}}}$ and $CSR_{\boldsymbol{X_{mfs}}}$ models perform similarly in terms of predictive power. In particular the $PLS_{\boldsymbol{X_{mfs}}}(1)$ and the $CSR_{\boldsymbol{X_{mfs}}}(9)$ models have R^2_{OoS} between 28.1% and 59.8% for the former, and 30.2% and 56.1% for the latter at horizons between one and nine months ahead. $PLS_{\boldsymbol{X_{mfs}}}$ and $CSR_{\boldsymbol{X_{mfs}}}$ models favor parsimony, with models with $L = 1$ and $k = \{6, 9\}$ generally performing best. Finally, if we look at Panel E we find that the $Bagg_{\boldsymbol{X_{mfs}}}$ models are also capable of generally beating the rolling-mean benchmark, with significant positive R^2_{OoS} measures at horizons between two and twelve months ahead. Despite the fact the bagging models are the only benchmark models that benefit from lagged values of the target variable, they tend to produce R^2_{OoS} below the best $PLS_{\boldsymbol{X_{mfs}}}$ and $CSR_{\boldsymbol{X_{mfs}}}$ models.

The results from the MCS confirm that the $t-PLS_{\boldsymbol{X_g}}$ models dominate all other models in predictive performance. In particular, the $t-PLS_{\boldsymbol{X_g}}(3)$ is the only model that is included in the 90% confidence set for all forecast horizons. The $t-PLS_{\boldsymbol{X_g}}(2)$ is also included in the 90% confidence set at horizons between one and five months ahead and at a horizon of nine months ahead. The $t-PLS_{\boldsymbol{X_g}}(1)$ model is only included in this set at a horizon of $h = 9$. When we consider the 95% confidence set, we also find that the $t-PLS_{\boldsymbol{X_g}}(3)$ and $t-PLS_{\boldsymbol{X_g}}(2)$ are included in the set for all horizons except $h = 12$, where the model with two factors is not included. This broader set also includes the $t-PLS_{\boldsymbol{X_g}}(1)$ at horizons of five, six and nine months ahead. Finally, it is worth mentioning that the only model that is not a targeted PLS model included in this set is the $CSR(6)$ model which is also in the 95% set at horizons of five and six months ahead.

Figure 3 shows actual employment growth and the forecast of the best performing model within each panel in Table 3 for horizons of three and six months ahead. The shaded period corresponds to the National Bureau of Economic Research (NBER) recession period. The most striking thing to note is that the $t-PLS_{\boldsymbol{X_g}}(3)$ model is the only model capable of accurately capturing the steep fall in employment growth during the recession. All other models either display a lag or are unable to capture the impact of the crisis. As with the R^2_{OoS} measures shown in Table 3, it is worth to notice the difference in performance between the $t - PLS_{\boldsymbol{X_g}}$ and the $PLS_{\boldsymbol{X_g}}$ models. Figure 4 shows the same comparison between actual values and forecasts as the preceding figure but for horizons of $h = 9$ and $h = 12$ months ahead. As with the preceding figure, the superior performance of the targeted PLS model is clear since this is the only model that can properly capture the increase in employment growth following the crisis without a substantial lag. Consistent with the R^2_{OoS} measures, we see that the $PLS_{\boldsymbol{X_g}}(2)$ model performs rather poorly when forecasting nine months and one year ahead.

Figure 5 plots the CSSED for the best model in each panel of Table 3 for horizons of $h = \{3, 6, 9, 12\}$ months ahead. We can easily see that all models have their greatest relative advantage over the rolling-mean model during the early part of the forecast evaluation period, i.e. the period during the recession and subsequent recovery. This is particularly the case for horizons of $h = 3$ and $h = 6$ where the CSSED lines are increasing steeply during the recession period to level off right after it ends. There is a second period of relative improvements for the majority of the models over the rolling-mean in the period between mid-2011 and 2012 where the US economy experienced an accelerated expansion on the number of jobs created. The relative local dominance of the $t - PLS_{\boldsymbol{X_g}}(3)$ model is also on display since the CSSED for the targeted PLS models is essentially above the CSSED values for other models. The CSSED line for $t - PLS_{\boldsymbol{X_g}}(3)$ is almost always either flat (performing as good as the rolling-mean model) or increasing (outperforming the rolling-mean model). It is also interesting to see that the $PLS_{\boldsymbol{X_g}}$ model is almost uniformly the worst performing model, implying that Elastic Net targeting prior to estimation is extremely important when dealing with search volume data. The overall conclusion is that the targeted PLS model is the best model throughout the whole evaluation sample.

4.2 Where is predictive power coming from?

In the preceding section, we show that Google Trends have a high degree of predictive power for future employment growth. We also show that including all the Google Trends series in the PLS model results in much poorer performance, implying that only some of the series in $\boldsymbol{X_g}$ have predictive power. These results lead us to the critical question: where is that predictive power coming from? The first thing we do to address this question is to calculate and show in Table 4 the R^2_{OoS} of a single factor PLS model where the explanatory variables are each of the primitive terms and their *related terms.* The number of variables in these explanatory sets, which we call primitive sets, vary between 17 and 25, depending on the primitive keyword they are related to.[19] Most primitive sets produce R^2_{OoS} measures that are positive, particularly at horizons of six months ahead or below, but they all fail to get a performance that is close to the performance of the $t-PLS_{\boldsymbol{X_g}}$models. The set of words related to *"collect unemployment"* and *"unemployment office"* stand out as relatively good predictors for horizons $h < 9$, having an average R^2_{OoS} (for $h = 1$ to $h = 6$) of 30% and 27.8%, respectively. The only primitive set that consistently beats the rolling-mean model across all forecast horizons is *"companies hiring",* which is able to produce an R^2_{OoS} of 19.1% and 20.2% at horizons of three quarters and one year ahead.

We also asses the forecasting power of each individual search term, X_i, using a univariate forecasting model:

$$y^h_{t+h} = \alpha + \beta_i X_{i,t} + \varepsilon^h_{t+h}, \tag{8}$$

where the parameters of the model are estimated using OLS. Figure 6 shows the R^2_{OoS} for univariate regressions using each of the top twenty predictors in $\boldsymbol{X_g}$ for $h = \{3, 6, 9, 12\}$ months ahead.[20] The first thing we notice in Figure 6 is that although some predictors achieve R^2_{OoS} above 30%, the majority of the search terms have predictive measures between 25% and 10%. Although, there is a large amount of heterogeneity in the search terms, some keywords appear relatively often. For example, queries that include "*food stamps*" appear in four and five of the top predictors for $h = 3$ and $h = 6$, respectively. It is interesting to see that the primitive set for *"food stamps"* does not seem to achieve higher R^2_{OoS} than the best keyword in the set *("food stamps md")* at any forecast horizon shown. This is because the words in the primitive sets are highly correlated and therefore seldomly add any additional information to the best individual predictor in each set. However, none of the primitive keywords are part of the top twenty predictors for any forecast horizon. Thus, most

[19]The number is below 26 (primitive keyword + 25 related terms) because we remove low volume queries and economically unrelated terms. We also note that some related terms are not unique to a single primitive keyword. Hence, there is some overlap in the data.

[20]For the sake of compactness we only include the top fifteen predictors for $\boldsymbol{X_{mfs}}$ in Figure 7. The last five predictors have large negative R^2_{OoS} values.

of the predictive power of the $t - PLS_{\boldsymbol{X_g}}$ must be arising from related keywords.

Other keywords often included in the top twenty predictors for horizons of three and six months ahead are *"unemployment benefits", "jobs"* and "*hiring*". These queries are related to *"collect unemployment",* "*jobs classified*" and *"companies hiring",* which unsurprisingly also have primitive sets that perform relatively well. For longer horizons (i.e. $h = 9$ and $h = 12$), we also find predictive power for terms related to *"food stamps"* and *"jobs",* but other terms such as "*salary"* and *"edd"* (Employment Development Department), which is a related term for *"unemployment office*", also begin to exhibit significance.[21] For example, the term *"salary calculator"* achieves an R^2_{OoS} of 34% at a horizon of nine months ahead, while *"edd employment"* attains 44% for a horizon of twelve months ahead. We do not show the bottom predictors in $\boldsymbol{X_g}$ for each forecast horizon, but it is worth mentioning that when forecasting at horizons of $h = 9$ and $h = 12$ months ahead, more than half of the predictors in $\boldsymbol{X_g}$ have a negative R^2_{OoS}. Thus, although PLS will asymptotically filter out irrelevant predictors (Kelly and Pruitt, 2015), on a finite and relatively short estimation window, these predictors are assigned non-zero weights in the construction of the factor, which explains the poor performance of the $PLS_{\boldsymbol{X_g}}$ models.

For the sake of comparison, we also show the R^2_{OoS} of the top fifteen individual predictors in the benchmark data set, $\boldsymbol{X_{mfs}}$, in Figure 7. Although some of the leading indicators in $\boldsymbol{X_{mfs}}$ are relatively good predictors, no single predictor achieves an R^2_{OoS} above 50% for any horizon. For horizons of three and six months ahead, we find that $EENM_t$ is a relatively good predictor, implying that managers' expectations on future employment growth are relatively accurate. In fact, for these forecast horizons $EENM_t$ produces R^2_{OoS} measures that are significantly higher than those achieved by the best individual predictor in $\boldsymbol{X_g}$**.** When we move over to a horizon of nine months ahead we find that $UNRATE_t$ stands out as the best predictor with an R^2_{OoS} around 38%, which is slightly better than the best predictor in $\boldsymbol{X_g}$ (*"salary calculator"*). For a horizon of one year ahead, we also find $UNRATE_t$ to be the best predictor in $\boldsymbol{X_{mfs}}$, but it performs slightly below its counterpart in $\boldsymbol{X_g}$ (*"edd employment"*). Generally, we find that the top three to five predictors at each horizon are capable of attaining R^2_{OoS} measures above 20% but moving lower down the

[21]The Employment Development Department is part of the Labor and Workforce Development Agency of the executive branch of the State of California. One of their main objectives is to help California job seekers obtain employment. Note that the state of California accounts for more than 12% of the total US population.

ranking does not result in particularly high predictive power. The best multivariate models based on $\boldsymbol{X_{mfs}}$ generally beat the best individual predictors in this data set. However, at a horizon of one year ahead $UNRATE_t$ performs better. One key difference between the benchmark data, $\boldsymbol{X_{mfs}}$, and the Google Trends data, $\boldsymbol{X_g}$, is that the predictive gain that we get for the former when using multivariate models is much smaller than the predictive gain that $t - PLS_{\boldsymbol{X_g}}$ models have over each of the best individual predictors in $\boldsymbol{X_g}$. Thus, variable selection and combination seems to work much better for Google Trends data.

The evidence presented until now implies that, although there is predictive information in $\boldsymbol{X_g}$ about future employment growth, no primitive set or individual predictor can account for the outstanding performance of the $t-PLS_{\boldsymbol{X_g}}$ models. Since the $PLS_{\boldsymbol{X_g}}$ models perform rather poorly in comparison, we can infer that soft thresholding is selecting a particularly useful set of Trends at each period. Our interest then shifts towards figuring out what those selected predictors are and whether or not they change over time. Figures 8 and 9 show the inclusion per period for the most often included predictors in $\boldsymbol{X_g}$ (ordered by inclusion frequency) as chosen by the Elastic Net soft threshold.

Several features of these figures are noteworthy. First, none of the series in $\boldsymbol{X_g}$ are included in the set for all periods at any forecast horizon. The most included predictors at each horizon, i.e. *"walmart career opportunities"* ($h = 3$), *"google job listings"* ($h = 6$ and $h = 12$) and *"entry level sales jobs"* ($h = 9$) are included in the set between 60% and 68% of the time. Second, there appears to be heterogeneity across the most frequent predictors, which means that they are generally related to different primitive keywords. This finding is not necessarily surprising since a good set of predictors would balance some correlation between predictors, which supports the use of a factor structure, with the inclusion of uncorrelated series that incorporate new information. Third, there is some overlap between the top predictors in Figure 6 and the most frequently included terms, but the overlap is far from complete, implying that the subset selected by soft thresholding is not necessarily composed of the best individual predictors. Finally, the Elastic Net estimator has some difficulties when choosing among highly correlated predictors. Hence there is some inherent instability in

[22] This intuitive fact is confirmed by unreported results in which we plot the inclusion frequency while setting $\alpha = 1$ in (1), which reduces the Elastic Net estimator to the LASSO estimator. Since the LASSO does not handle correlated predictors as well as the Elastic Net, inclusion frequencies for the soft thresholding based on the LASSO are much more fragmented across time periods.

inclusion across time periods.[22] Overall, it appears to be that the remarkable performance of the $t-PLS$ model does not arise from selecting a stable set of predictors, but from its flexibility to choose different predictors over time. These predictors appear to come from different primitive sets and therefore augment the amount of information in the latent factors. The key ingredient seems to be heterogeneity. Thus, the primitive words should cover a relatively broad spectrum of search queries. A disadvantage of our procedure is that factor interpretation becomes harder since the variables that go into each factor can vary over time. Hence, factor loadings will also vary over time. This trade-off between model interpretability and forecast accuracy is well known in the statistical learning literature and the choice of model will ultimately reflect the importance we give to this choice for the task at hand.

4.3 Combining data sets

The $t-PLS_{\boldsymbol{X_g}}$ models strongly outperform the models based on benchmark data. However, it is possible that $\boldsymbol{X_{mfs}}$ embodies useful information that is not contained in $\boldsymbol{X_g}$. To investigate this, we create a data set that combines $\boldsymbol{X_g}$ and $\boldsymbol{X_{mfs}}$, and denote it by: $\boldsymbol{X_c} = [\boldsymbol{X_g}\ \boldsymbol{X_{mfs}}]$. We then run the out-of-sample forecast with a $t-PLS$ model using this data, the results are presented in Table 5. For most forecast horizons, with the exception of $h = 9$ where the addition of $\boldsymbol{X_{mfs}}$ does result in increases in the R^2_{OoS} of approximately 8 percentage points, the three factor $t-PLS_{\boldsymbol{X_g}}$ model does not seem to improve with the addition of macroeconomic, financial and sentiment variables. We find that $UN27W_t$, $UNRATE_t$, EEM_t and $BUILDPERM_t$ are included in the set of predictors by soft thresholding between 50% and 65% of the time, depending on the predictor and horizon.[23] Although these inclusion frequencies are relatively high, the absence of large improvements in R^2_{OoS} when these variables are in the set implies that they are not providing useful additional information, particularly at horizons below six months ahead.

[23]For the interested reader, the inclusion frequencies for the combined data set, $\boldsymbol{X_c}$, are shown in Figures C1 and C2 of Appendix C.

5 Robustness checks

In this section we show that the forecasting power of targeted PLS, which is our main model, is not sensitive to the words we use to construct the data, alternative estimation windows or whether the data is detrended recursively or first simply differenced. Finally, we also show that the methodology we use does not result in spurious out-of-sample predictive power by running a placebo test.

5.1 Alternative keywords

The primitive set of words that we use to build the Google Trends data set, $\boldsymbol{X_g}$, are selected because they were part of two "domestic Trends" indices. This raises the possibility that the results depend entirely on that particular set of primitive queries. To address this issue and select a set of words in a manner that is as objective as possible, we use the Harvard and Lasswell Psychosociological Dictionary (the H4Lvd file).[24] More specifically, we start by selecting the 510 words in this dictionary that are classified as being related to economics.[25] From this set of words we manually select those that are unambiguously related to the labor market and/or employment. The keywords included in this alternative set of primitive words are the following 23 terms: "*career*", *"earn", "earner", "employee", "employer", "employment", "hire", "job", "occupation", "occupational", "payroll", "profession", "professional", "promotion", "unemployed", "unemployment", "wage"* and *"worker".* Note that this primitive set of keywords is larger than the 10 words used to build $\boldsymbol{X_g}$. We do this because we want to minimize any subjective choice in the selection.[26] Once this primitive set is constructed, we follow the same procedure as with the original primitive set. The resulting alternative Google Trends data set, which we denote by $\boldsymbol{X_g^*}$, has 381 terms. As with the other PLS based models, before estimating the targeted PLS model with this data, we remove seasonality and detrend on an expanding recursive window. The R^2_{OoS} and t_{DM} test statistic for this alternative Google Trends model, $t-PLS_{\boldsymbol{X_g^*}}$, are shown in Table 6 for a model with $L = \{1, 2, 3\}$ factors. We obtain an R^2_{OoS} between 83.4% and 92.6% for the three-factor model. The results are generally similar, albeit supe-

[24]http://www.wjh.harvard.edu/~inquirer/spreadsheet_guide.htm

[25]The Harvard general inquirer defines the "Econ@" category as those words with an economic, commercial, industrial, or business orientation, including roles, collectivities, acts, abstract ideas, and symbols, including references to money.

[26]The only exception is the word *"salary"*, which is included in the dictionary but explicitly excluded by us since it already forms part of the primitive set for $\boldsymbol{X_g}$.

rior, to the ones we obtain with our main data set, $\boldsymbol{X_g}$. This is possibly because soft thresholding is choosing among a larger set of predictors.

Figure 10 confirms that the forecasts are indeed relatively similar, especially at horizons of $h = 3$ and $h = 6$ suggesting that the predictive power in $\boldsymbol{X_g}$ can be recovered by an alternative set of primitive keywords related to the labor market. As we mentioned in Section 4.2, the important feature seems to be heterogeneity in the primitive terms, since this allows the model to cover different dimensions of job search or economic sentiment.

5.2 Alternative estimation windows

The CSSED analysis shows that the predictive performance of the $t - PLS_{\boldsymbol{X_g}}(3)$ has a stable performance over the evaluation window. To further check the robustness and stability of the results, we also perform the $t - PLS_{\boldsymbol{X_g}}$ forecast with three alternative estimation windows: a rolling window of 36 observations, a rolling window of 60 observations and an expanding window with 48 initial observations. Table 7 shows the R^2_{OoS} and t_{DM} of these alternative estimation windows for a $t - PLS_{\boldsymbol{X_g}}$ model with 3 factors. Decreasing the size of the rolling estimation window results in R^2_{OoS} measures that are between 65% and 80% (12 and 8 percentage points lower than the rolling window estimation with 48 observations). This decrease in performance is possibly arising for two reasons. First, a smaller estimation window will inevitably lead to noisier factor estimates. Second, the sample evaluation period is longer, which means that it covers most of the recession of 2008-2009, which is a period that is inherently more difficult to forecast. The results for the models that are estimated using a longer rolling window of 60 observations or an expanding window with an initial size of 48 observations are similar to the ones we obtain for our main model (with an estimation window of 48 observations), implying that the results are not sensitive to these changes. Overall, while decreasing the estimation window has a negative effect on predictability for $t - PLS_{\boldsymbol{X_g}}(3)$ model, the R^2_{OoS} measures are still far higher than those obtained by benchmark models. In unreported results, we find that decreasing the estimation window to 36 observations has an even larger (negative) effect on models based on $\boldsymbol{X_{mfs}}$.

5.3 Differencing data

As a final check for robustness, we present the results of the targeted PLS model using our Google Trends data set, but instead of applying the Ayat and Burridge (2000) sequential test, we simply take first differences of the deseasonalized Google Trends data. Due to first differencing, the first observation of the first differenced panel, $\Delta \boldsymbol{X_g}$ is set to the unconditional mean of each series. We do this in an effort to keep the evaluation period constant and ease comparison since the alternative is to reduce the forecast window by one observation.[27] Table 8, which presents the results of this robustness check, shows that forecasting power decreases when all series are first differenced, particularly at long forecast horizons ($h > 6$). Since there is strong evidence that many of the series are stationary (or stationary around a deterministic trend), taking first differences implies over-differencing and thus lower predictive power. We should mention, however, that the first differenced data is still able to produce higher R^2_{OoS} than the models based on $\boldsymbol{X_{mfs}}$.

5.4 Placebo test

To show that the targeted PLS methodology does not result in spurious out-of-sample predictive power, we follow Kelly and Pruitt (2013) and run a placebo test. If the methodology results in a mechanical bias, simulated placebo data that is similar to the data in $\boldsymbol{X_g}$, but unrelated to our target variable, will also display out-of-sample predictability. For each time period, we generate thirty AR(1) series that have the same mean, variance and autoregressive coefficient as the series selected by soft thresholding at this time period. Innovations in the series are generated using an i.i.d. normal distribution that has zero covariance with our target variable. Thus, they are independent of y_t. Table 9 shows the 95% and 99% quantile of the null region for 1,000 simulations. Although we can expect a result of zero asymptotically, in finite samples the results are mostly negative due to small sample bias. The placebo test shows that the probability of actually getting a positive R^2_{OoS} by chance is virtually null. This is especially the case for the model with three factors which has negative values even for the 99% quantile of the null region, making the results of the $t - PLS_{\boldsymbol{X_g}}(3)$ model even more remarkable.

[27]The results are unaffected if the first observation is set to zero or if the evaluation period is pushed one period ahead.

6 Concluding remarks

Employment growth is a leading indicator that has important implications for both policy makers and the private sector. Therefore, the need for accurate and timely predictions is relatively self-evident. In this paper, we show that there is plenty of relevant information about future employment growth in internet search volume. Our findings imply that Google-based forecasting models can be a particularly valuable tool for obtaining accurate real-time information on future employment growth and labor market conditions. We also show that individual Google Trends series do not appear to embed enough information to be better predictors than the classical macroeconomic, financial or opinion survey series. However, the combination of many Google Trends series can substantially increase the forecasting power and substantially improve upon our models based on classical series. A caveat is that including all series in the large panel usually results in relatively poor forecast results in comparison to models based on the usual series. This can be solved by using a variable pre-selection procedure such as soft thresholding. Overall, our contribution shows that the high predictive power of Google Trends implies that it should be added to the toolbox of practitioners and policy makers interested in forecasting employment growth. Our results also suggest that internet search volume should be further investigated to forecast other macroeconomic variables.

References

ASKITAS, N. AND K. F. ZIMMERMANN (2009): "Google econometrics and unemployment forecasting." *Applied Economics Quarterly*, 55, 107–120.

AYAT, L. AND P. BURRIDGE (2000): "Unit root tests in the presence of uncertainty about the non-stochastic trend." *Journal of Econometrics*, 95, 71–96.

BAI, J. AND S. NG (2008): "Forecasting economic time series using targeted predictors," *Journal of Econometrics*, 146, 304–317.

BAKER, S. R., N. BLOOM, AND S. J. DAVIS (2016): "Measuring economic policy uncertainty," *The Quarterly Journal of Economics*, 131, 1593–1636.

BIJL, L., G. KRINGHAUG, P. MOLNÁR, AND E. SANDVIK (2016): "Google searches and stock returns," *International Review of Financial Analysis*, 45, 150–156.

CAMPBELL, J. Y. AND S. B. THOMPSON (2007): "Predicting excess stock returns out of sample: Can anything beat the historical average?" *The Review of Financial Studies*, 21, 1509–1531.

CARROLL, C. D., J. C. FUHRER, AND D. W. WILCOX (1994): "Does consumer sentiment forecast household spending? If so, why?" *The American Economic Review*, 84, 1397–1408.

CHOI, H. AND H. VARIAN (2012): "Predicting the present with Google Trends," *Economic Record*, 88, 2–9.

CHRISTIANSEN, C., J. N. ERIKSEN, AND S. V. MØLLER (2014): "Forecasting US recessions: The role of sentiment," *Journal of Banking & Finance*, 49, 459–468.

CLEVELAND, R. B., W. S. CLEVELAND, J. E. MCRAE, AND I. TERPENNING (1990): "STL: A Seasonal-Trend Decomposition," *Journal of Official Statistics*, 6, 3–73.

COBLE, D. AND P. PINCHEIRA (2017): "Nowcasting Building Permits with Google Trends." *MPRA Working Paper No. 76514, University Library of Munich, Germany.*

DA, Z., J. ENGELBERG, AND P. GAO (2011): "In search of attention." *Journal of Finance*, 66, 1461–1499.

Da, Z., J. Engelberg, and P. Gao (2014): “The sum of All FEARS: Investor sentiment and asset prices,” *Review of Financial Studies*, 28, 1–32.

D’Amuri, F. and J. Marcucci (2017): “The predictive power of Google searches in forecasting US unemployment,” *International Journal of Forecasting*, 33, 801–816.

De Mol, C., D. Giannone, and L. Reichlin (2008): “Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components?” *Journal of Econometrics*, 146, 318–328.

Diebold, F. X. and R. S. Mariano (1995): “Comparing predictive accuracy.” *Journal of Business & Economic Statistics*, 13, 253–263.

Elliott, G., A. Gargano, and A. Timmermann (2013): “Complete subset regressions.” *Journal of Econometrics*, 177, 357–373.

Fondeur, Y. and F. Karamé (2013): “Can Google data help predict French youth unemployment?” *Economic Modelling*, 30, 117–125.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2005): “The generalized dynamic factor model: One-sided estimation and forecasting,” *Journal of the American Statistical Association*, 100, 830–840.

Francesco, D. (2009): “Predicting unemployment in short samples with internet job search query data.” *MPRA Working Paper No. 18403. University Library of Munich, Germany.*

Groen, J. J. and G. Kapetanios (2016): “Revisiting useful approaches to data-rich macroeconomic forecasting.” *Computational Statistics & Data Analysis*, 100, 221–239.

Hansen, P. R., A. Lunde, and J. M. Nason (2011): “The model confidence set,” *Econometrica*, 79, 453–497.

Inoue, A. and L. Kilian (2008): “How useful is bagging in forecasting economic time series? A case study of US consumer price inflation,” *Journal of the American Statistical Association*, 103, 511–522.

Kelly, B. and S. Pruitt (2013): "Market expectations in the cross-section of present values," *The Journal of Finance*, 68, 1721–1756.

Kelly, B. and S. Pruitt (2015): "The three-pass regression filter: A new approach to forecasting using many predictors," *Journal of Econometrics*, 186, 294–316.

Kim, H. H. and N. R. Swanson (2014): "Forecasting financial and macroeconomic variables using data reduction methods: New empirical evidence," *Journal of Econometrics*, 178, 352–367.

Lemmon, M. and E. Portniaguina (2006): "Consumer confidence and asset prices: Some empirical evidence," *The Review of Financial Studies*, 19, 1499–1529.

Matsusaka, J. G. and A. M. Sbordone (1995): "Consumer confidence and economic fluctuations," *Economic Inquiry*, 33, 296–318.

Newey, W. K. and K. D. West (1987): "A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix." *Econometrica*, 55, 703–708.

Politis, D. N. and J. P. Romano (1992): *A circular block-resampling procedure for stationary data.*, Wiley, 263–270.

Politis, D. N. and H. White (2004): "Automatic block-length selection for the dependent bootstrap." *Econometric Reviews*, 23, 53–70.

Rapach, D. E. and J. K. Strauss (2008): "Forecasting US employment growth using forecast combining methods." *Journal of Forecasting*, 27, 75–93.

Rapach, D. E. and J. K. Strauss (2010): "Bagging or combining (or both)? An analysis based on forecasting US employment growth," *Econometric Reviews*, 29, 511–533.

Rapach, D. E. and J. K. Strauss (2012): "Forecasting US state-level employment growth: An amalgamation approach," *International Journal of Forecasting*, 28, 315–327.

Smith, A. (2015): "Searching for work in the digital era." *Pew Research Center: Internet, Science & Tech*, 19.

Stock, J. H. and M. W. Watson (2002a): “Forecasting using principal components from a large number of predictors.” *Journal of the American Statistical Association*, 97, 1167–1179.

Stock, J. H. and M. W. Watson (2002b): “Macroeconomic forecasting using diffusion indexes.” *Journal of Business & Economic Statistics*, 20, 147–162.

Stock, J. H. and M. W. Watson (2006): “Forecasting with many predictors.” *In G. Elliott, C. Granger, & A. Timmermann (Eds.) Handbook of economic forecasting*, 1, 515–554.

Tibshirani, R. (1996): “Regression shrinkage and selection via the lasso.” *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 58, 267–288.

Vosen, S. and T. Schmidt (2011): “Forecasting private consumption: Survey-based indicators vs. Google trends,” *Journal of Forecasting*, 30, 565–578.

Vozlyublennaia, N. (2014): “Investor attention, index performance, and return predictability,” *Journal of Banking & Finance*, 41, 17–35.

Welch, I. and A. Goyal (2008): “A comprehensive look at the empirical performance of equity premium prediction.” *The Review of Financial Studies*, 21, 1455–1508.

Wold, H. (1966): “Estimation of principal components and related models by iterative least squares.” *In: Krishnaiaah, P. (Ed.), Multivariate Analysis, Academic Press, New York, 391-420.*

Yu, L., Y. Zhao, L. Tang, and Z. Yang (2018): “Online big data-driven oil consumption forecasting with Google trends,” *International Journal of Forecasting*, In press.

Zou, H. and T. Hastie (2005): “Regularization and variable selection via the elastic net,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67, 301–320.

Appendices

Appendix A

Table A1 shows the source of each predictor variable in $\boldsymbol{X_{mfs}}$.

Table A1: Predictor variables sources

Predictor	**Source**
Unemployment rate*	Bureau of Labor Statistics
Unemployment claims*	Bureau of Labor Statistics
Average mean duration of employment*	Bureau of Labor Statistics
Number of civilians unemployed for less than 5 weeks*	Bureau of Labor Statistics
Number of civilians unemployed for 5-14 weeks*	Bureau of Labor Statistics
Number of civilians unemployed for 14-26 weeks*	Bureau of Labor Statistics
Number of civilians unemployed for 27 weeks and over*	Bureau of Labor Statistics
Average weekly manufacturing hours*	Bureau of Labor Statistics
Ratio of help wanted advertising to number of unemployed*	National Bureau of Economic Research
Real disposable personal income*	Bureau of Economic Analysis
Real house price index*	Federal Housing Finance Agency
Building permits for new private housing units*	Bureau of Census
Real new consumer goods orders in durable goods*	Bureau of Census
10 year Treasury bond yield	Board of Governors from the US Federal Reserve
Real stock returns	Center for Research in Security Prices
Real oil price growth, WTI spot price per barrel	US Energy Information Administration
University of Michigan's index of consumer sentiment	http://www.sca.isr.umich.edu/
Economic policy indicator of Baker et al. (2016)	http://www.policyuncertainty.com/
Employment expectations in the non-manufacturing sector	Institute for Supply Management
Employment expectations in the manufacturing sector	Institute for Supply Management

Predictors marked with an asterix * are seasonally adjusted.

Appendix B

Table B1 shows the results of the $t - PLS_{\boldsymbol{X_g}}$ model when $\alpha = 1$ in equation (1), which reduces the Elastic Net estimator to the LASSO estimator. As above we tune λ such that 30 Google Trends are selected at each point in time.

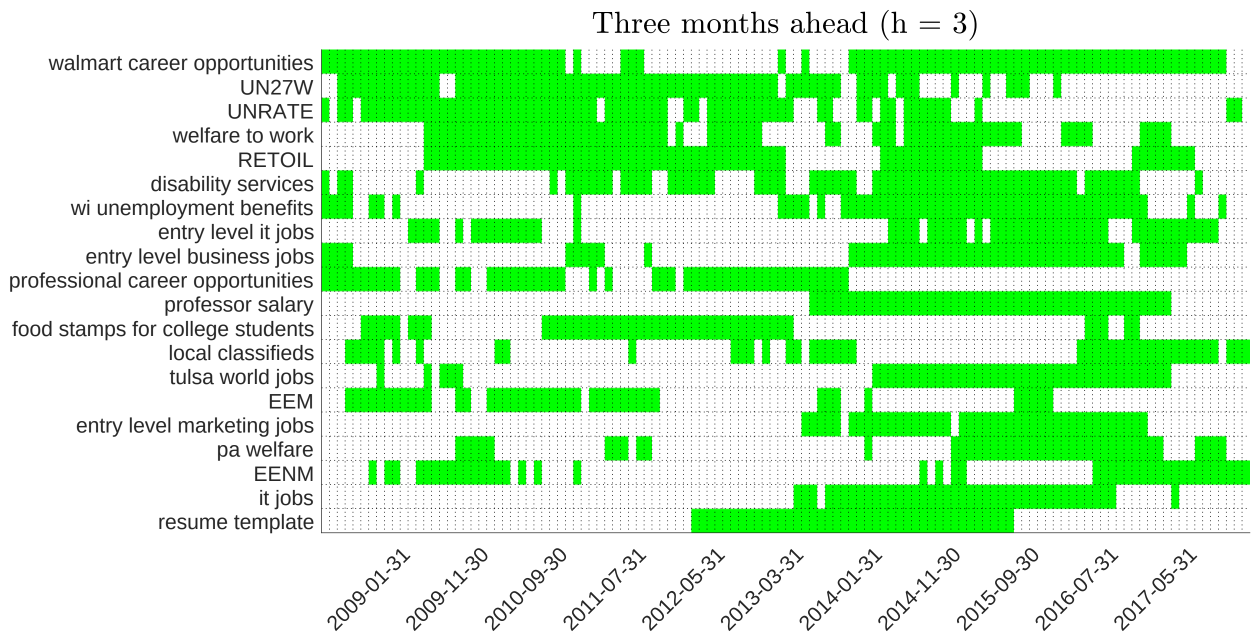
Table B1: Out-of-sample predictive power of the $t - PLS_{\boldsymbol{X_g}}$ using LASSO-based soft thresholding

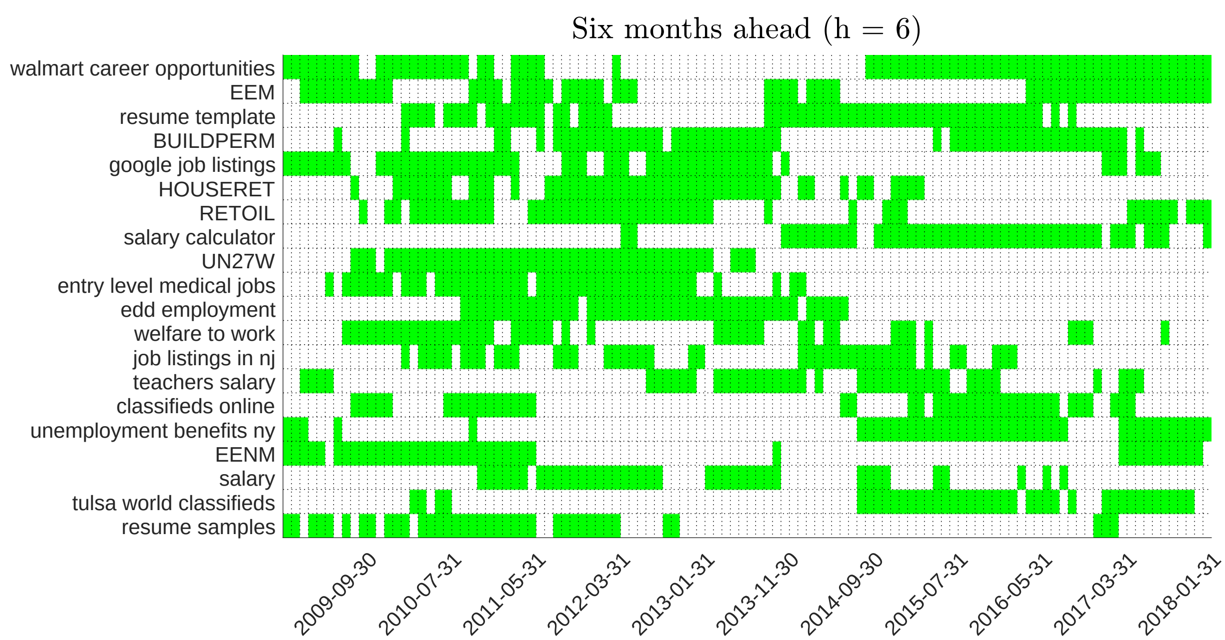
	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=9$	$h=12$
$t - PLS_{\boldsymbol{X_g}}(1)$ $LASSO$	68.16	74.38	67.84	73.16	73.95	74.33	66.74	67.18
	[4.27]	[2.72]	[2.06]	[1.87]	[1.72]	[1.66]	[1.68]	[1.45]
$t - PLS_{\boldsymbol{X_g}}(2)$ $LASSO$	78.40	82.69	83.59	86.25	86.17	86.19	82.67	74.89
	[4.97]	[2.92]	[2.13]	[1.91]	[1.73]	[1.66]	[1.63]	[1.54]
$t - PLS_{\boldsymbol{X_g}}(3)$ $LASSO$	79.77	83.07	84.57	86.95	87.55	88.23	86.01	81.79
	[4.98]	[2.95]	[2.16]	[1.91]	[1.72]	[1.66]	[1.66]	[1.59]

The table shows the R^2_{OoS} and Diebold-Mariano test statistic, $[t_{DM}]$, for the $t - PLS_{\boldsymbol{X_g}}$ model with LASSO-based soft thresholding. The numbers inside the parenthesis indicate the number of factors in the model. The models are estimated using a rolling window of 48 observations.

Appendix C

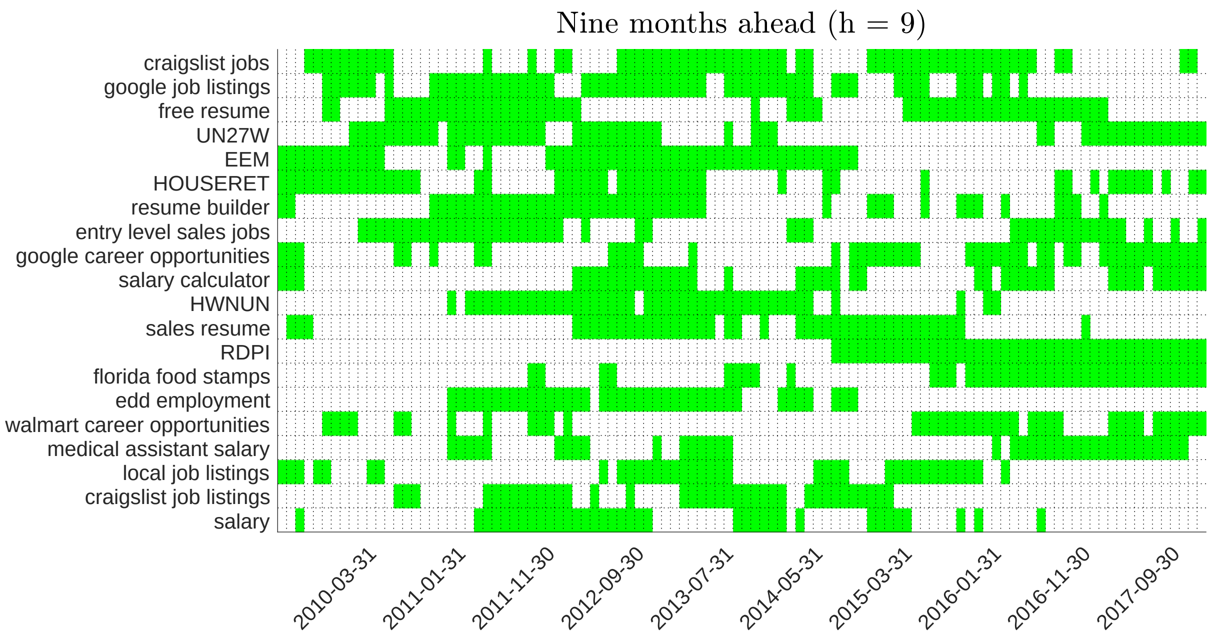
Figure C1: Soft thresholding inclusion for individual predictors in the combined data set $\boldsymbol{X_c} = [\boldsymbol{X_g}\ \boldsymbol{X_{mfs}}]$ at horizons of three and six months ahead

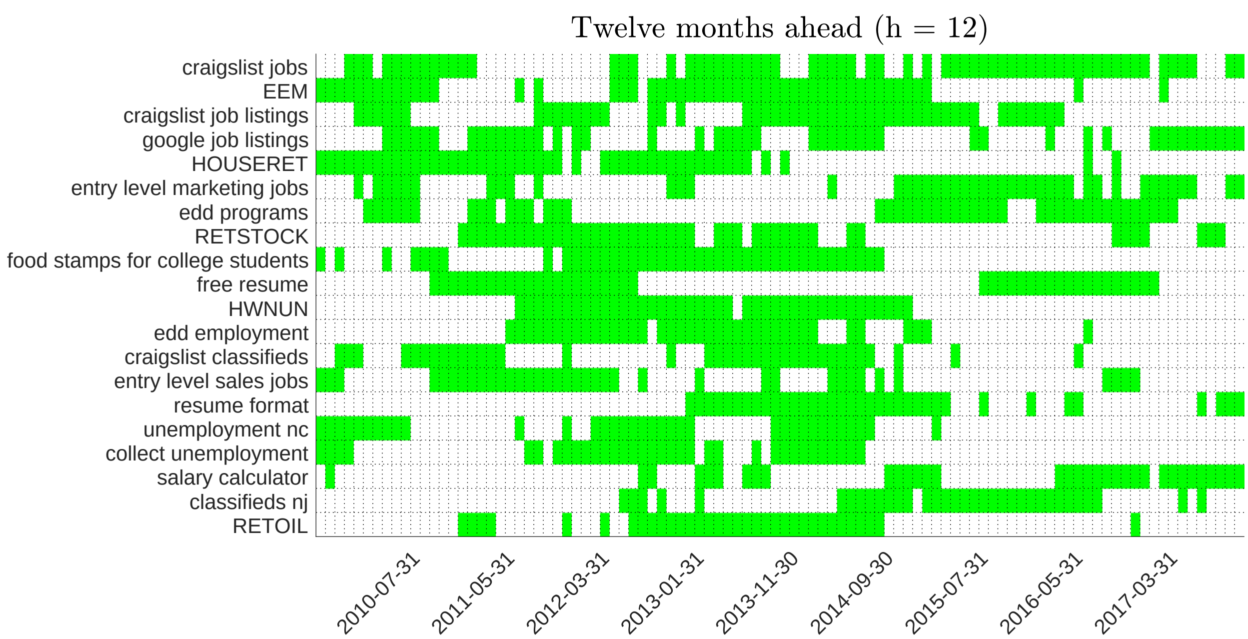




The figure shows the soft thresholding inclusion for the predictors in $\boldsymbol{X_c} = [\boldsymbol{X_g}\ \boldsymbol{X_{mfs}}]$ during the out-of-sample evaluation period for the top twenty predictors at horizons of three and six months. Note that predictors from $\boldsymbol{X_{mfs}}$ are capitalized. The predictors are ordered from top to bottom according to their inclusion frequency.

Figure C2: Soft thresholding inclusion for individual predictors in the combined data set $\boldsymbol{X_c} = [\boldsymbol{X_g}\ \boldsymbol{X_{mfs}}]$ at horizons of three and six months ahead

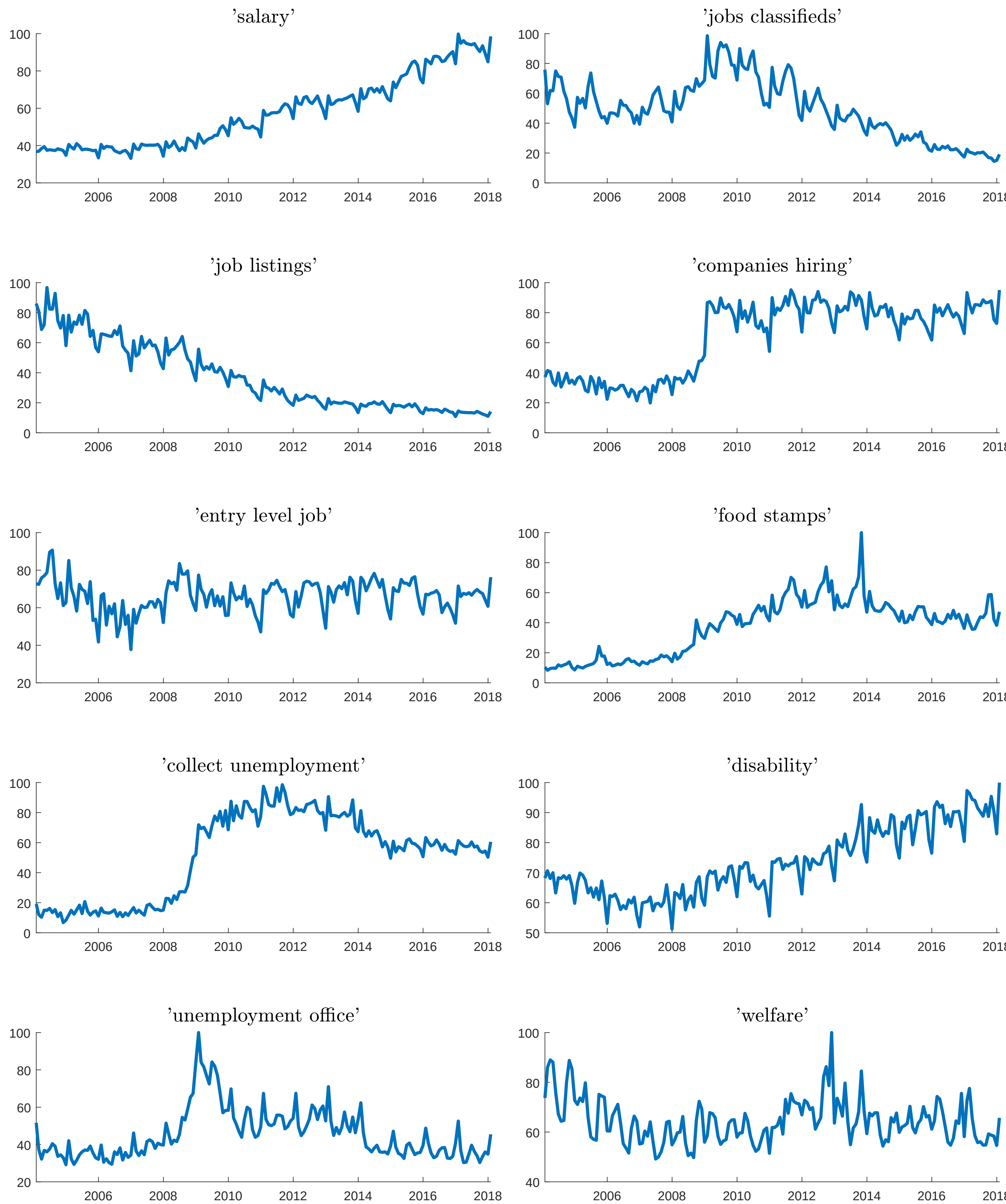




The figure shows the soft thresholding inclusion for the predictors in $\boldsymbol{X_c} = [\boldsymbol{X_g}\ \boldsymbol{X_{mfs}}]$ during the out-of-sample evaluation period for the top twenty predictors at horizons of nine and twelve months. Note that predictors from $\boldsymbol{X_{mfs}}$ are capitalized. The predictors are ordered from top to bottom according to their inclusion frequency.

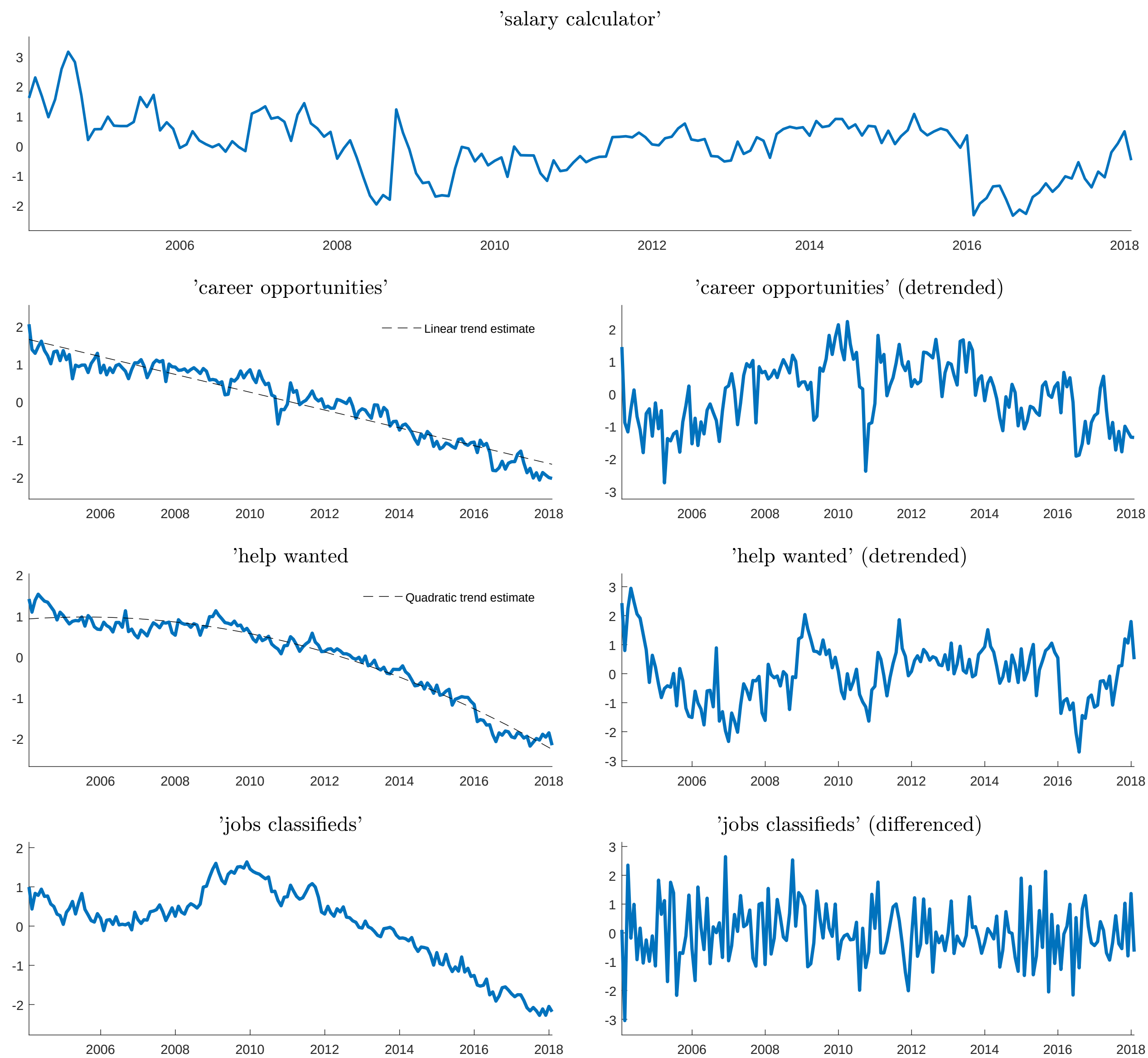
Figures and Tables

Figure 1: Primitive queries



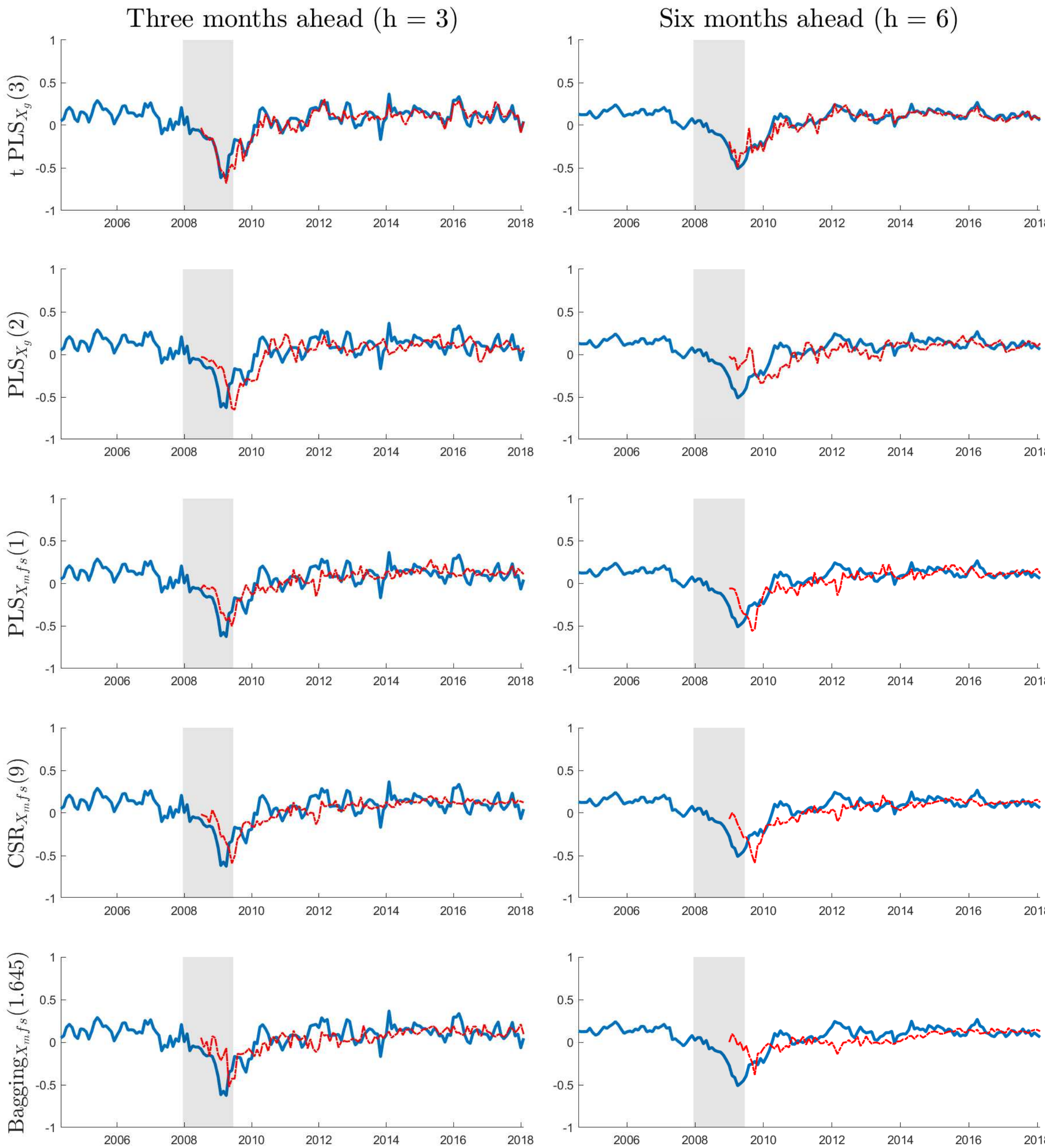
The figure shows the ten primitive Google Trends queries in the period 2004:M1 - 2018:M1. Note that the index is calculated as a simple of average of the index for each word over twenty different days. This averaging across different series is the reason why *"entry level job"* does not reach a 100 as its maximum.

Figure 2: Data transformation to construct X_g



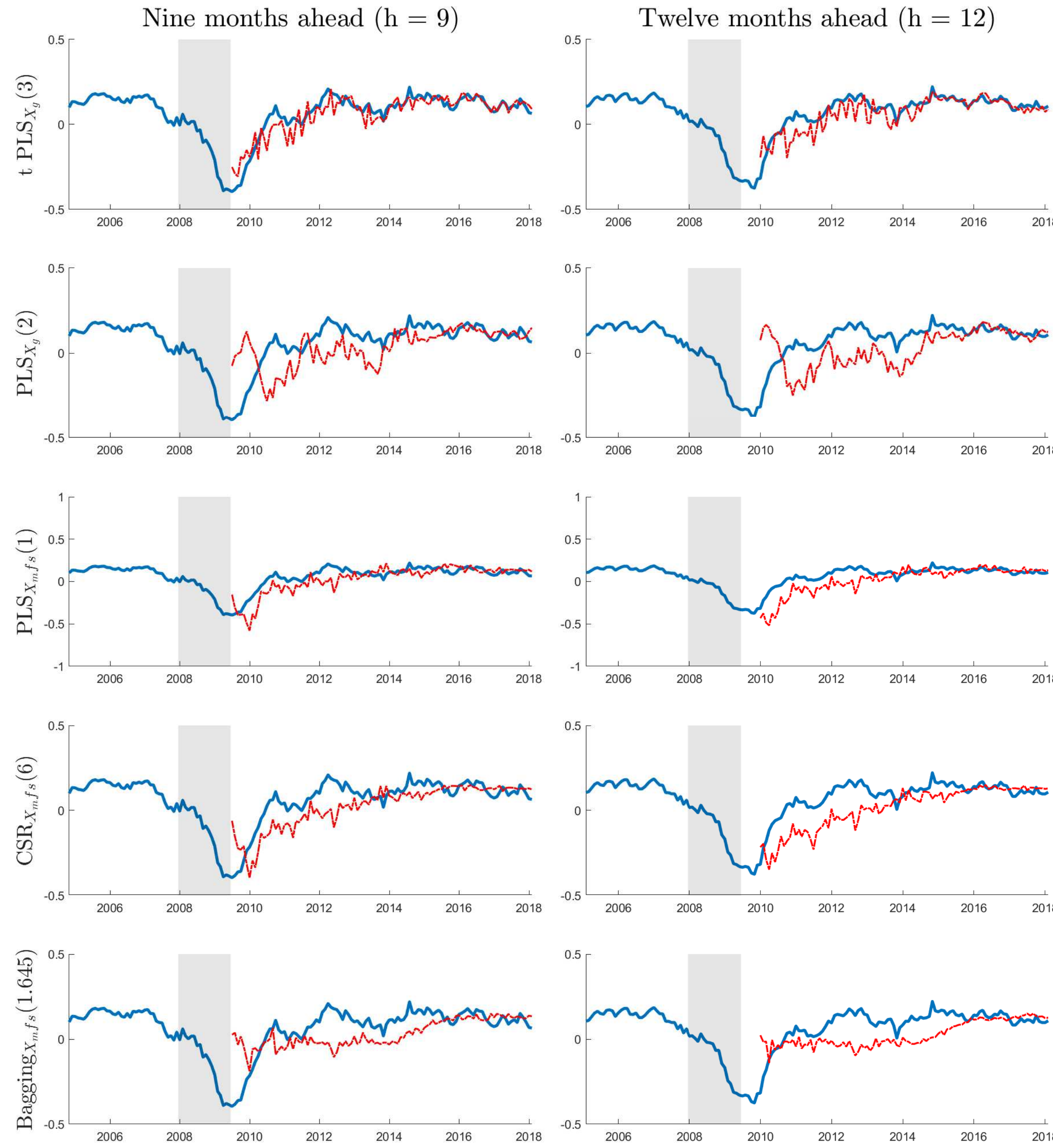
The figure shows the natural logarithm of four deseasonalized Google Trends queries in the period 2004:M1 - 2018:M1. The panel on top shows an example of a stationary query*: "salary calculator"*, for which we do not perform any detrending or diferencing. The second panel on the left shows a linear trend-stationary query: "*career opportunities*" (solid line) and its linear trend estimate (dashed line) while the panel on the right shows deviations from this trend. The third panel on the left shows shows a quadratic trend stationary query (*"help wanted"*) and its trend estimate (dashed line) while the panel on the right shows deviations from this trend. The panel at the bottom right shows a series for which we could not reject the null of a unit root (*"jobs classifieds"*) and the panel on the bottom right shows the same series in differences. For ease of comparison the series have been standardized to have mean zero and standard deviation of one. Note that the only series in the panel that is a primitive querie is *"jobs classifieds",* hence, this series can be compared to the raw series shown in Figure 1. The differences between the two stem from the log transformation and the deseasonalizing.

Figure 3: Employment growth: actual value vs. forecasts for best model



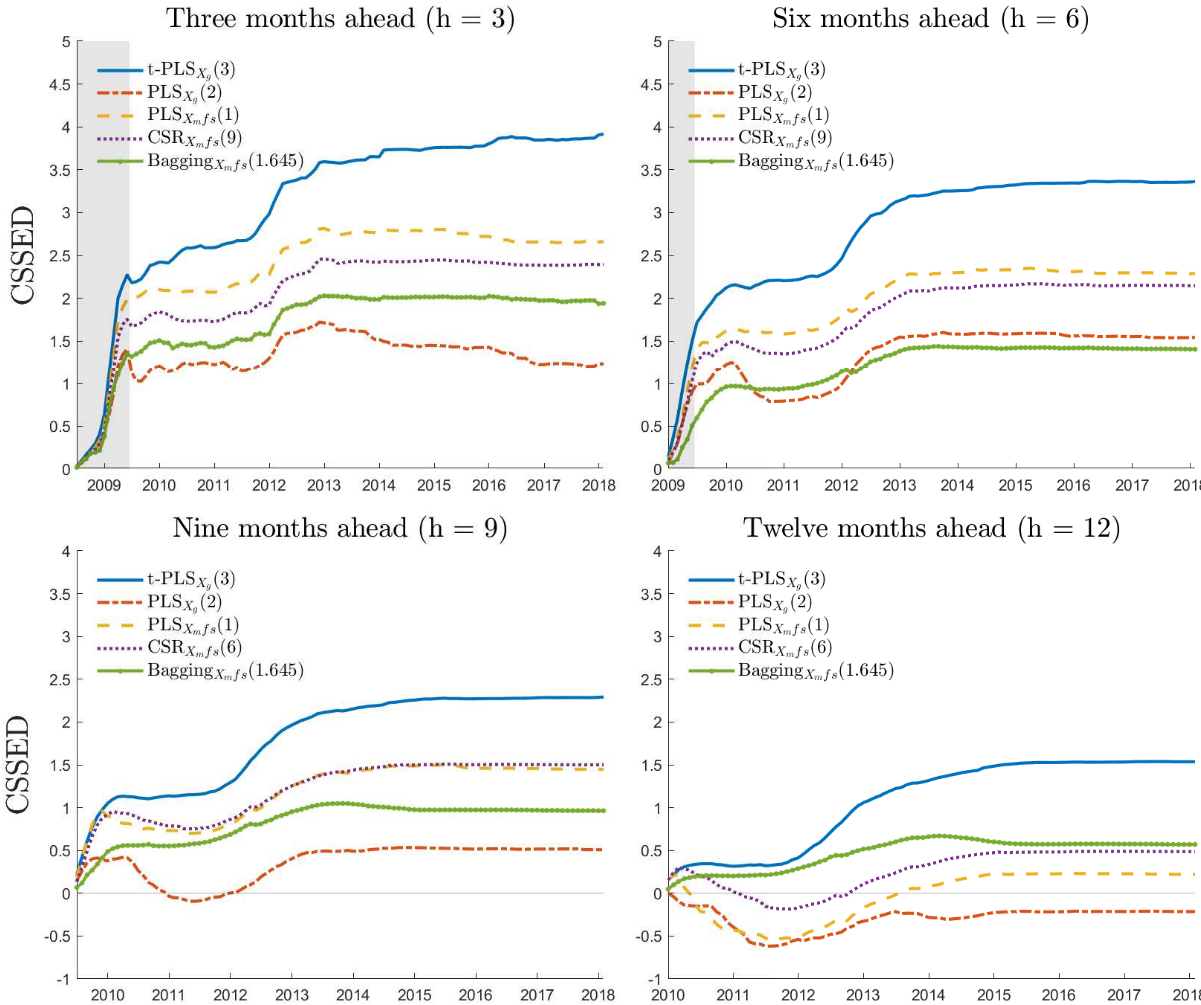
The figure shows the three and six months ahead forecast for the competing models. The blue (solid) line is the actual value of employment growth and the red (dashed) line is the forecast. The number of factors for the PLS based models, variates for the CSR models and hard threshold value for bagging models are selected such that the best performing model in terms of R^2_{OoS} is selected.

Figure 4: Employment growth: actual value vs. forecasts for best model



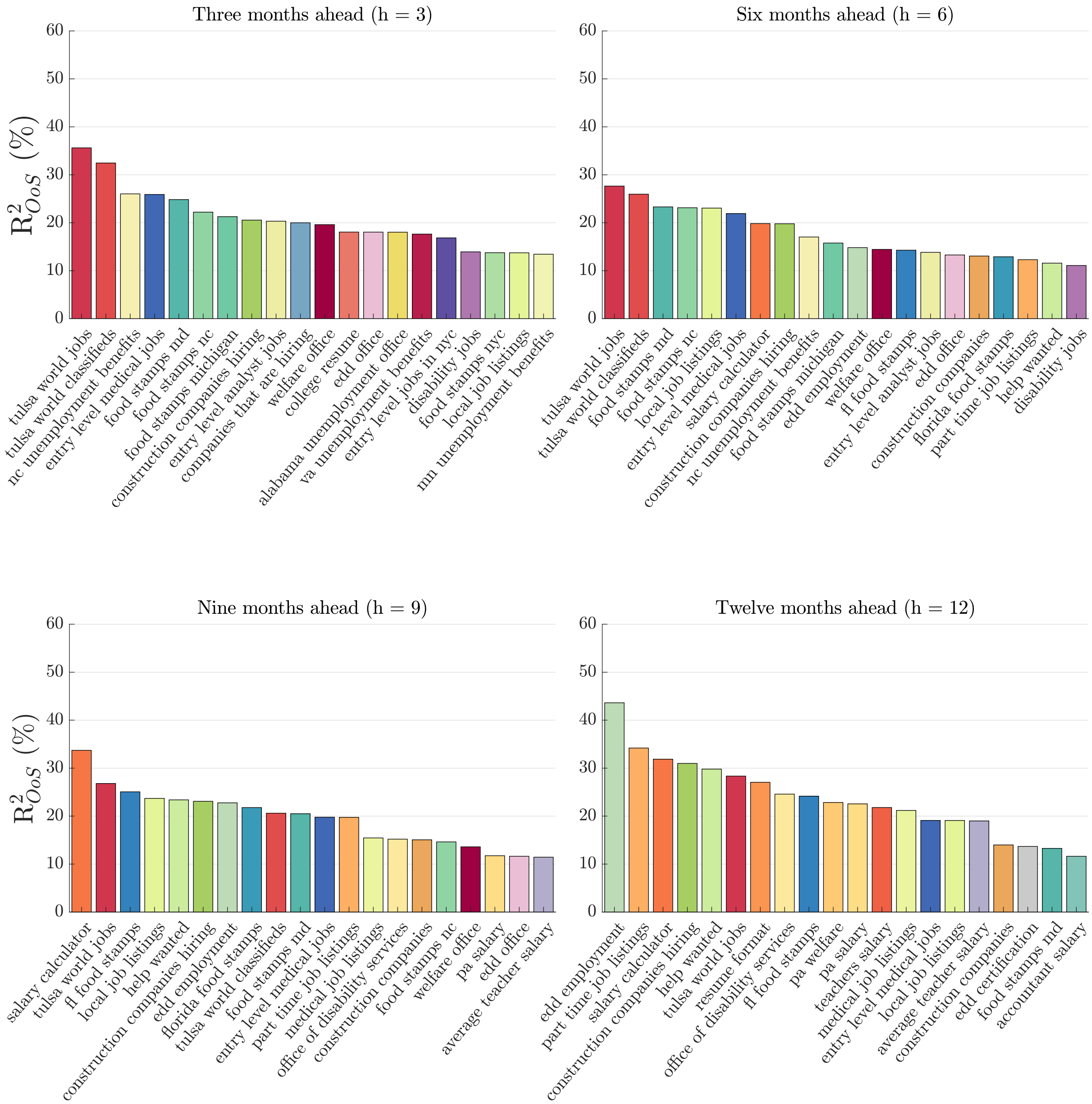
The figure shows the nine and twelve months ahead forecast for the competing models. The blue (solid) line is the actual value of employment growth and the red (dashed) line is the forecast. The number of factors for the PLS models, variates for the CSR models and hard threshold value for bagging models are selected such that the best performing model in terms of R^2_{OoS} is selected. Note that the forecasts for $h = \{9, 12\}$ do not overlap with the recession period since $(2h) - 1$ observations are lost in the estimation and forecast. Hence, the first forecast for $h = 9$ is 2009:M6 and for $h = 12$ it is 2009:M12. For $h = 12$, the R^2_{OoS} of the $PLS_{\mathbf{X_g}}(1)$ (not shown in the figure) is slightly less negative than the $PLS_{\mathbf{X_g}}(2)$, but the forecast of the latter is shown due to better overall performance.

Figure 5: Cumulative sum of squared error difference (CSSED) for best models



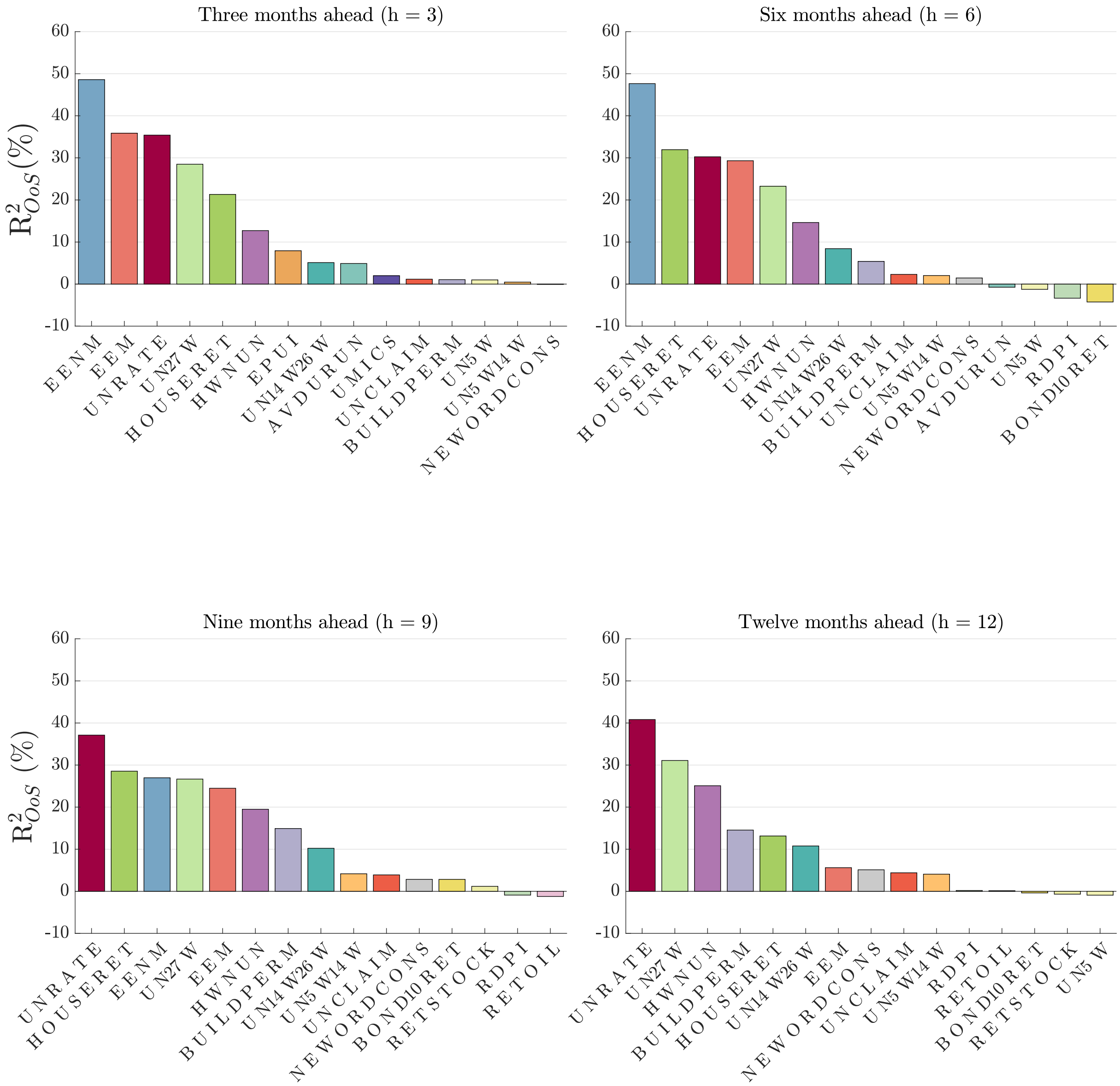
The figure shows the cumulative sum of squared errors difference ($CSSED_{t,m}$) for the best models for $h = \{3, 6, 9, 12\}$ months ahead. The number of factors for the PLS based models, variates for the CSR models and hard threshold value for bagging models are selected such that the best performing model in terms of R^2_{OoS} is selected. Note that the $CSSED_{t,m}$ measures for $h = \{9, 12\}$ do not overlap with the NBER recession period.

Figure 6: R^2_{OoS} for the best twenty individual predictors in X_g



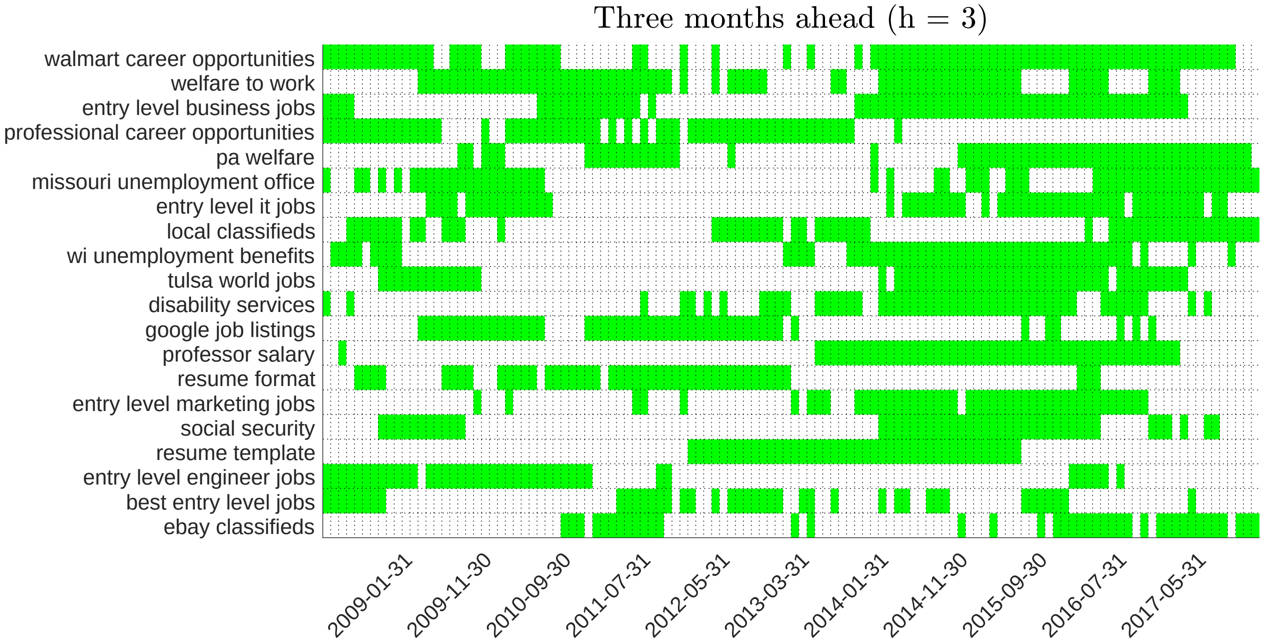
The figure shows the R^2_{OoS} for univariate regressions, $y^h_{t+h} = \alpha + \beta_i X_{i,t} + \varepsilon^h_{t+h}$, for the top twenty predictors in the Google Trends panel, $\boldsymbol{X_g}$, at horizons of $h = \{3, 6, 9, 12\}$ months ahead.

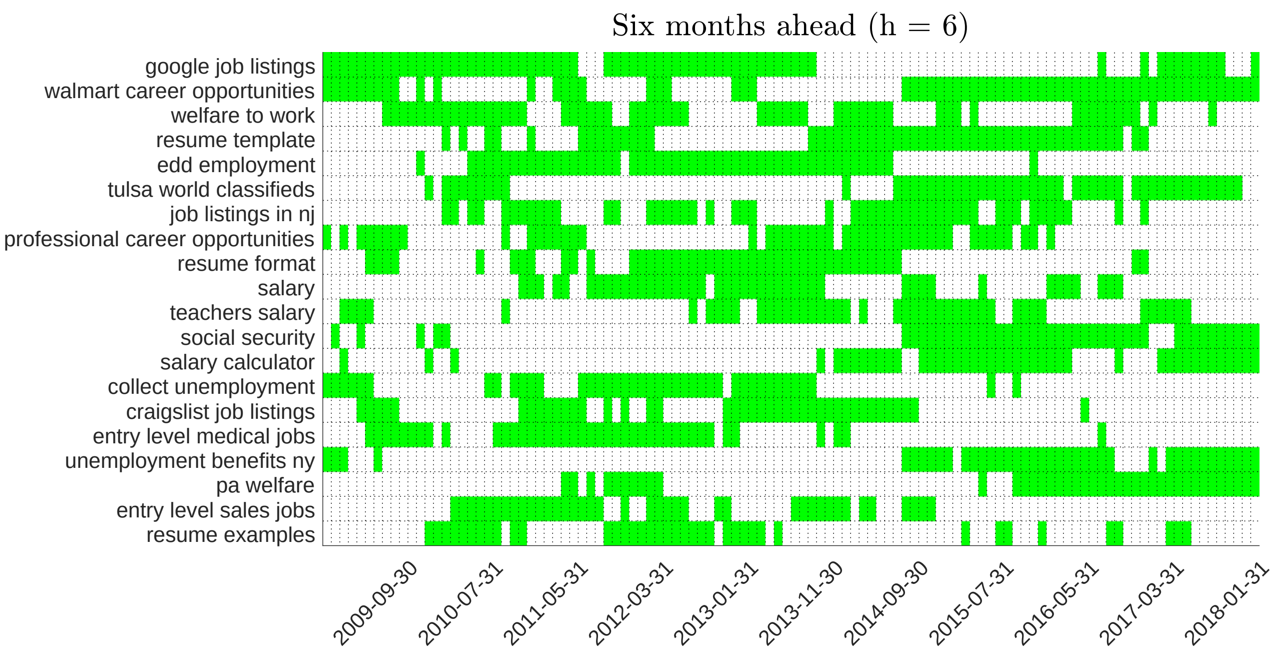
Figure 7: R^2_{OoS} for the best fifteen individual predictors in X_{mfs}



The figure shows the R^2_{OoS} for univariate regressions, $y^h_{t+h} = \alpha + \beta_i X_{i,t} + \varepsilon^h_{t+h}$, for the top fifteen predictors in the benchmark panel, $\boldsymbol{X_{mfs}}$, at horizons of $h = \{3, 6, 9, 12\}$ months ahead. The last five variables in the ranking, for each forecast horizon, have large negative R^2_{OoS} measures and are thus excluded from the figure.

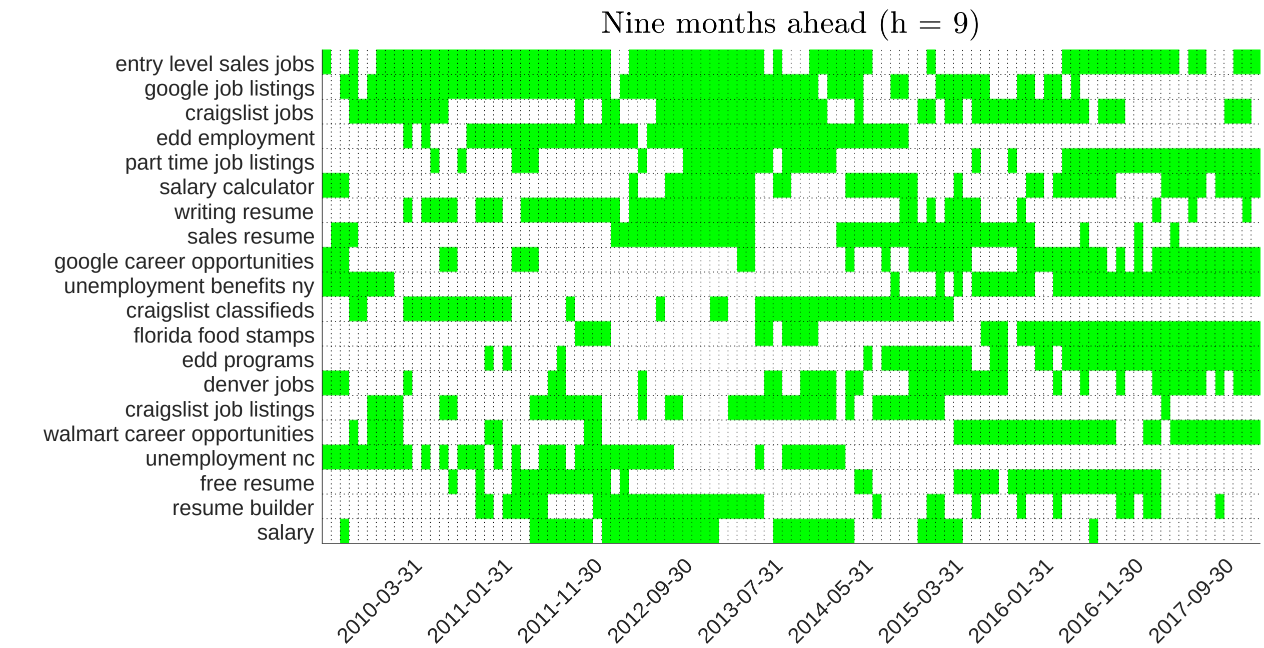
Figure 8: Soft thresholding inclusion for individual predictors in $t - PLS_{\boldsymbol{X_g}}$ models at horizons of three and six months ahead

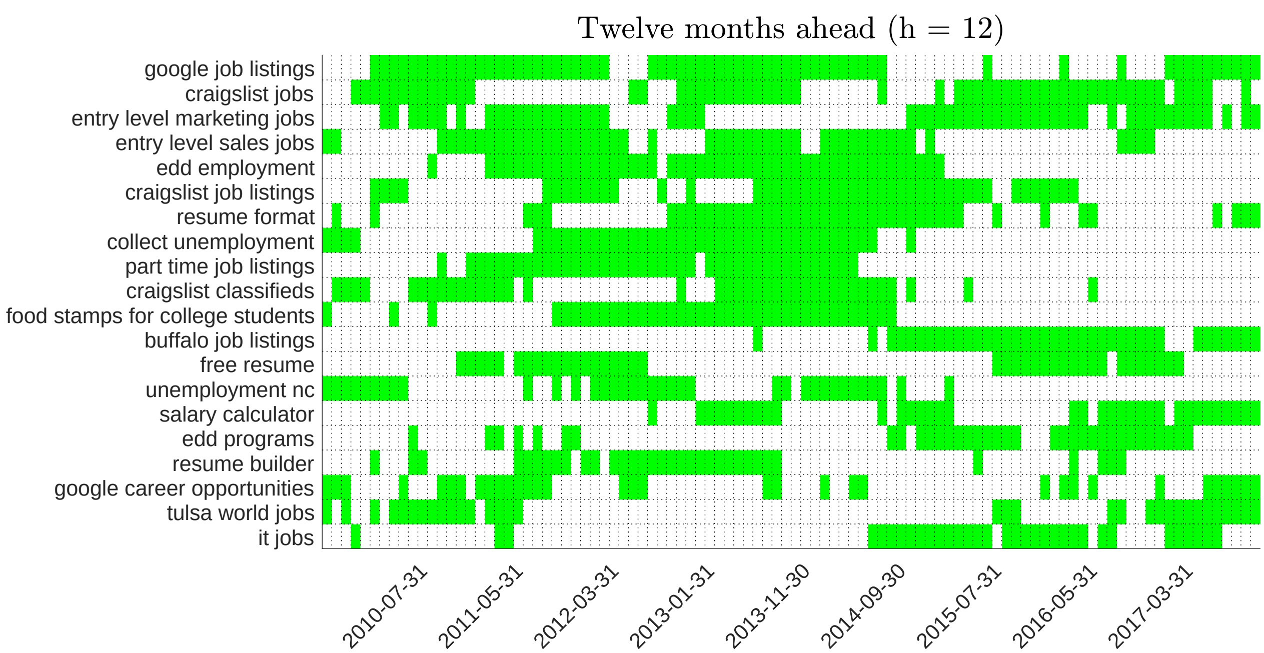




The figure shows the soft thresholding inclusion for the predictors in $\boldsymbol{X_g}$ during the out-of-sample evaluation period for the top twenty predictors in the $t - PLS_{\boldsymbol{X_g}}$ model at horizons of three and six months. The predictors are ordered from top to bottom according to their inclusion frequency.

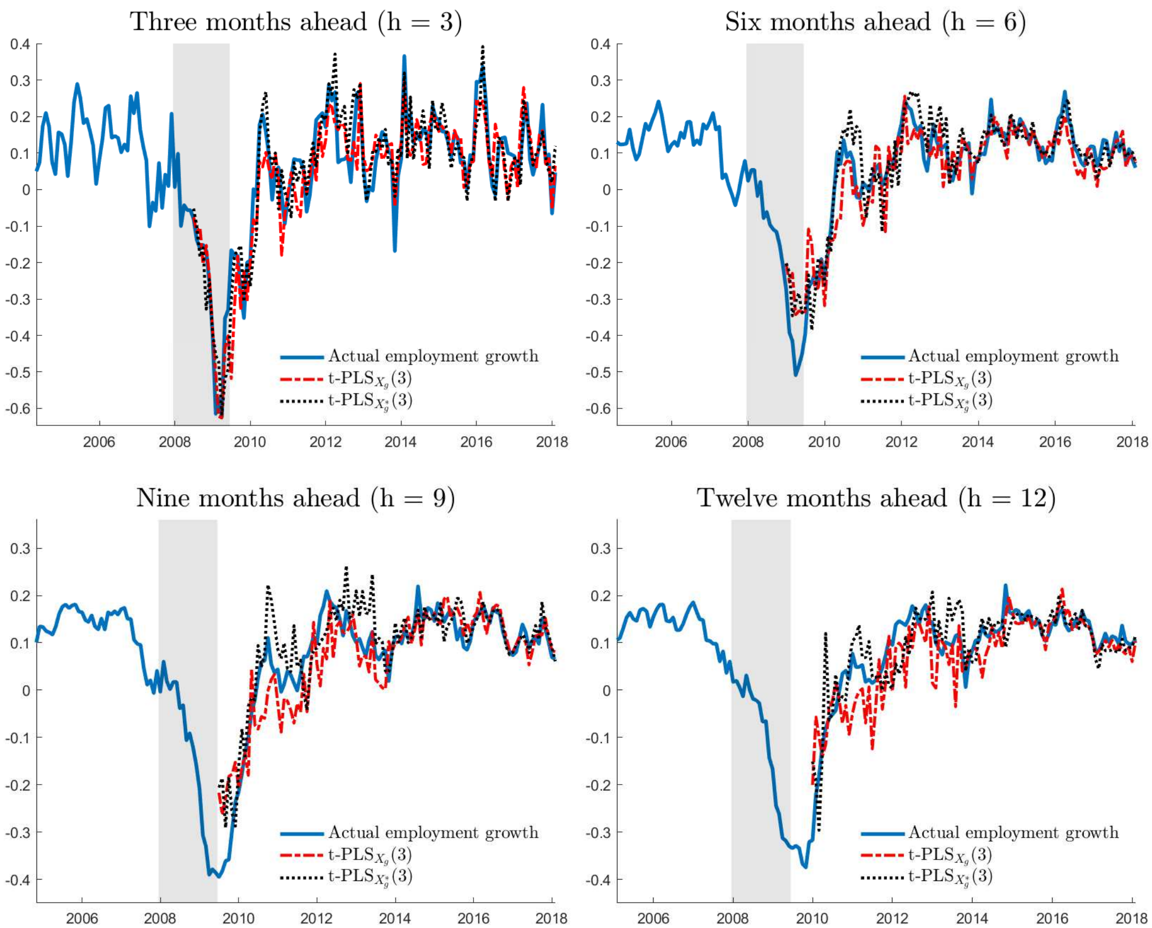
Figure 9: Soft thresholding inclusion for individual predictors in $t - PLS$ models at horizons of nine and twelve months ahead





The figure shows the soft thresholding inclusion for the predictors in $\boldsymbol{X_g}$ during the out-of-sample evaluation period for the top twenty predictors in the $t - PLS_{\boldsymbol{X_g}}$ model at horizons of nine and twelve months. The predictors are ordered from top to bottom according to their inclusion frequency.

Figure 10: Employment growth forecasts for $t - PLS_{\boldsymbol{X_g}}$ and $t - PLS_{\boldsymbol{X_g^*}}$ models



The figure shows the forecast comparison between the $t - PLS_{\boldsymbol{X_g}}(3)$ and the $t - PLS_{\boldsymbol{X_g^*}}(3)$ models for horizons of $h = \{3, 6, 9, 12\}$ months ahead. The blue solid line is the actual employment growth rate, the red stippled line is the $t - PLS_{\boldsymbol{X_g}}(3)$ forecast and the black dotted line is the $t - PLS_{\boldsymbol{X_g^*}}(3)$ forecast.

Table 1: Macroeconomic and financial predictors

Abbreviation	Predictor	Transformation
UNRATE	Unemployment rate*	Differences
UNCLAIM	Unemployment claims*	Log levels
AVDURUN	Average mean duration of employment*	Differences
UN5W	Number of civilians unemployed for less than 5 weeks*	Log differences
UN5W14W	Number of civilians unemployed for 5-14 weeks *	Log differences
UN14W26W	Number of civilians unemployed for 14-26 weeks *	Log differences
UN27W	Number of civilians unemployed for 27 weeks and over*	Log differences
AVMANHRS	Average weekly manufacturing hours*	Differences
HWNUN	Ratio of help wanted advertising to number of unemployed*	Differences
RDPI	Real disposable personal income*	Log differences
HOUSERET	Real house price index*	Log returns
BUILDPERM	Building permits for new private housing units*	Log differences
NEWORDCONS	Real new consumer goods orders in durable goods*	Log differences
BOND10RET	10 year Treasury bond yield	Log returns
STOCKRET	Real stock returns	Log returns
ROIL	Real oil price growth, WTI spot price per barrel	Log returns

Predictors marked with an * are seasonally adjusted.

Table 2: Sentiment predictors

Abbreviation	Predictor	Transformation
UNMICS	University of Michigan's index of consumer sentiment	Levels
EPUI	Economic policy indicator of Baker et al. (2016)	Levels
EENM	Employment expectations in the non-manufacturing sector	Levels
EEM	Employment expectations in the manufacturing sector	Levels

Table 3: Out-of-sample predictive power for employment growth rates

	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=9$	$h=12$
Panel A: targeted PLS with X_g								
$t-PLS_{\boldsymbol{X_g}}(1)$	65.56	74.17	68.66	73.95	73.54†	73.49†	66.59‡	69.38
	[4.18]	[2.46]	[2.08]	[1.87]	[1.69]	[1.65]	[1.60]	[1.51]
$t-PLS_{\boldsymbol{X_g}}(2)$	76.46‡	82.52‡	83.40‡	86.28‡	85.90‡	85.92†	81.96‡	76.91
	[4.91]	[2.85]	[2.17	[1.88]	[1.71]	[1.63]	[1.60]	[1.53]
$t-PLS_{\boldsymbol{X_g}}(3)$	77.99‡	83.04‡	84.74‡	86.83‡	86.66‡	87.80‡	85.62‡	83.69‡
	[4.77]	[2.88]	[2.14]	[1.86]	[1.70]	[1.65]	[1.62]	[1.59]
Panel B: PLS with X_g								
$PLS_{\boldsymbol{X_g}}(1)$	16.21	29.75	33.68	35.13	28.40	25.18	6.36	-9.45
	[1.03]	[1.23]	[1.36]	[1.36]	[1.15]	[1.08]	[0.25]	[-0.29]
$PLS_{\boldsymbol{X_g}}(2)$	19.59	32.40	36.51	41.78	40.09	40.31	13.52	-11.66
	[1.50]	[1.35]	[1.27]	[1.30]	[1.25]	[1.22]	[0.45]	[-0.34]
$PLS_{\boldsymbol{X_g}}(3)$	9.75	22.38	26.65	38.54	35.43	40.07	18.98	-16.05
	[0.71]	[0.93]	[1.08]	[1.36]	[1.29]	[1.34]	[0.64]	[-0.48]
Panel C: PLS with X_{mfs}								
$PLS_{\boldsymbol{X_{mfs}}}(1)$	28.07	46.52	57.59	57.07	59.58	59.78	54.10	11.99
	[2.11]	[1.81]	[1.69]	[1.56]	[1.53]	[1.48]	[1.45]	[0.28]
$PLS_{\boldsymbol{Xmf_s}}(2)$	14.38	36.07	45.57	44.20	46.00	48.39	30.85	-52.09
	[1.00]	[1.50]	[1.41]	[1.33]	[1.35]	[1.32]	[0.78]	[-0.60]
$PLS_{\boldsymbol{Xmf_s}}(3)$	-0.87	28.96	38.17	37.25	37.64	41.74	18.77	-88.75
	[-0.06]	[1.28]	[1.26]	[1.24]	[1.24]	[1.22]	[0.45]	[-0.79]
Panel D: CSR with X_{mfs}								
$CSR_{\boldsymbol{X_{mfs}}}(6)$	30.17	46.20	51.03	50.10	52.63†	54.98†	56.06	26.52
	[2.87]	[2.15]	[1.82]	[1.70]	[1.64]	[1.56]	[1.42]	[0.80]
$CSR_{\boldsymbol{X_{mfs}}}(9)$	28.71	46.66	51.85	50.86	53.30	56.07	50.56	2.03
	[2.40]	[2.02]	[1.71]	[1.61]	[1.57]	[1.51]	[1.27]	[0.04]
$CSR_{\boldsymbol{X_{mfs}}}(12)$	22.81	42.45	47.82	46.53	48.43	51.38	39.94	-23.11
	[1.76]	[1.81]	[1.53]	[1.46]	[1.45]	[1.40]	[0.98]	[-0.33]
Panel E: Bagging with X_{mfs}								
$Bagg_{\boldsymbol{X_{mfs}}}(1.645)$	-16.76	30.30	42.36	38.56	36.65	36.72	35.97	31.05
	[-0.94]	[1.67]	[1.85]	[1.63]	[1.57]	[1.55]	[1.44]	[1.33]
$Bagg_{\boldsymbol{X_{mfs}}}(1.96)$	-0.60	37.49	42.26	38.72	38.09	34.57	33.38	27.30
	[-0.04]	[2.08]	[2.00]	[1.67]	[1.60]	[1.60]	[1.47]	[1.25]
$Bagg_{\boldsymbol{X_{mfs}}}(2.58)$	11.29	40.05	41.90	37.51	35.37	33.70	31.20	28.22
	[0.82]	[2.17]	[2.01]	[1.70]	[1.67]	[1.59]	[1.48]	[1.27]

The table shows the R^2_{OoS} and Diebold-Mariano test statistic, $[t_{DM}]$, for all models using a rolling window of 48 observations. The ‡ symbol indicates that the model is included in the 90% confidence set ($\alpha = 10\%$) and † indicates that the model is in the 95% confidence set ($\alpha = 5\%$). The numbers in the parentheses indicate the number of factors (for PLS models), number of variates (for CSR models) and hard threshold critical value (for bagging models).

Table 4: R^2_{OoS} for a PLS(1) model with each primitive keyword + related terms

Primitive keyword	n	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=9$	$h=12$
"salary"	25	11.77	18.49	19.58	21.44	17.22	9.72	-16.46	-57.43
"job classifieds"	18	14.65	32.91	27.55	18.95	8.70	3.16	-10.90	9.01
"job listings"	18	5.33	8.25	6.98	1.62	-2.15	-6.16	-25.15	-41.57
"companies hiring"	24	16.85	30.15	26.93	24.94	21.21	17.76	19.10	20.22
"entry level jobs"	19	5.62	21.79	19.78	14.95	8.21	3.32	-36.80	-63.73
"food stamps"	25	7.91	21.21	22.04	20.92	19.90	20.21	-8.18	-92.16
"collect unemployment"	19	28.32	38.48	39.99	40.15	24.92	8.21	-18.79	-37.29
"disability"	19	10.72	17.73	14.22	12.40	10.34	5.97	-7.96	-35.37
"unemployment office"	25	22.51	34.90	34.03	33.29	25.04	17.14	-2.37	-8.78
"welfare"	18	-3.59	2.74	6.29	8.27	-4.42	-7.61	-31.95	-53.47

The table shows R^2_{OoS} for a PLS model with a single factor using each of the primitive keywords and its related terms as explanatory variables. The number of Google Trends, n, included in each set is shown in the second column, this number varies since low volume and economically unrelated terms are removed. Some related terms can also appear as related terms for more than one primitive keyword.

Table 5: Out-of-sample predictive power of the $t-PLS$ with the combined data set

	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=9$	$h=12$
$t-PLS_{X_c}(1)$	67.09	76.93	75.51	78.29	80.40	81.86	83.32	79.89
	[4.40]	[2.73]	[2.04]	[1.78]	[1.70]	[1.67]	[1.69]	[1.54]
$t-PLS_{X_c}(2)$	78.42	84.36	84.06	86.31	87.38	89.96	89.48	78.12
	[5.00]	[2.99]	[2.17]	[1.87]	[1.72]	[1.64]	[1.66]	[1.50]
$t-PLS_{X_c}(3)$	77.90	83.34	83.49	86.27	88.81	89.85	93.35	81.44
	[4.80]	[2.98]	[2.15]	[1.87]	[1.72]	[1.63]	[1.67]	[1.53]

The table shows the R^2_{OoS} and Diebold-Mariano test statistic, $[t_{DM}]$, for the $t-PLS_{\boldsymbol{X_c}}$, where $\boldsymbol{X_c}$ denotes a data set that combines the Google Trends panel, $\boldsymbol{X_g}$, and the benchmark panel, $\boldsymbol{X_{mfs}}$. The numbers inside the parenthesis indicate the number of factors in the model. The models are estimated using a rolling window of 48 observations.

Table 6: Out-of-sample predictive power of targeted PLS models using $\boldsymbol{X_g^*}$

	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=9$	$h=12$
$t-PLS_{\boldsymbol{X}_g^*}(1)$	70.40	80.44	78.88	82.19	79.70	79.86	76.94	82.42
	[4.35]	[2.85]	[2.17]	[2.01]	[1.85]	[1.73]	[1.66]	[1.57]
$t-PLS_{\boldsymbol{X}_g^*}(2)$	80.16	85.59	85.69	91.41	88.21	85.90	86.46	86.38
	[4.97]	[2.93]	[2.21]	[1.96]	[1.80]	[1.68]	[1.67]	[1.61]
$t-PLS_{\boldsymbol{X}_g^*}(3)$	83.38	86.17	87.39	92.63	89.59	88.37	86.63	86.21
	[5.09]	[2.97]	[2.25]	[1.98]	[1.78]	[1.66]	[1.67]	[1.60]

The table shows the R^2_{OoS} and Diebold-Mariano test statistic, $[t_{DM}]$, for targeted PLS models with the alternative Google Trends panel, $\boldsymbol{X_g^*}$. The numbers inside the parenthesis indicate the number of factors in the model. The models are estimated using a rolling window of 48 observations.

Table 7: Out-of-sample predictive power of $t-PLS_{\boldsymbol{X_g}}(3)$ with alternative estimation windows

R^2_{OoS}/t_{DM}

	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=9$	$h=12$
Panel A: Rolling window of 36 observations								
$t-PLS_{\boldsymbol{X}_g}(3)$	65.63	69.11	69.55	77.15	73.58	79.97	74.03	72.71
	[4.18]	[2.46]	[2.08]	[1.87]	[1.69]	[1.65]	[1.60]	[1.51]
Panel B: Rolling window of 60 observations								
$t-PLS_{\boldsymbol{X}_g}(3)$	77.82	81.10	83.17	86.55	85.00	87.46	83.05	76.39
	[4.91]	[2.85]	[2.17]	[1.88]	[1.71]	[1.63]	[1.60]	[1.53]
Panel C: Expanding window with an initial size of 48 observations								
$t-PLS_{\boldsymbol{X}_g}(3)$	77.87	81.53	82.91	86.04	85.96	90.57	86.62	82.04
	[4.77]	[2.88]	[2.14]	[1.86]	[1.70]	[1.65]	[1.62]	[1.59]

The table shows the R^2_{OoS} and Diebold-Mariano test statistic, $[t_{DM}]$, for the $t-PLS_{\boldsymbol{X_g}}(3)$ model with alternative estimation windows.

Table 8: Out-of-sample predictive power of targeted PLS models using $\Delta \boldsymbol{X_g}$

	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=9$	$h=12$
$t-PLS_{\boldsymbol{\Delta X_g}}(1)$	54.50	39.96	42.32	42.44	43.21	36.89	38.87	44.00
	[4.56]	[2.66]	[2.10]	[1.85]	[1.79]	[1.68]	[1.79]	[1.53]
$t-PLS_{\boldsymbol{\Delta X_g}}(2)$	65.74	65.39	61.76	68.95	61.56	53.27	57.76	53.71
	[4.49]	[2.82]	[1.96]	[1.79]	[1.65]	[1.60]	[1.62]	[1.52]
$t-PLS_{\boldsymbol{\Delta X_g}}(3)$	70.20	73.29	67.52	72.62	65.53	58.11	61.22	58.22
	[4.53]	[2.71]	[1.86]	[1.76]	[1.60]	[1.54]	[1.58]	[1.52]

The table shows the R^2_{OoS} and Diebold-Mariano test statistic, $[t_{DM}]$, for targeted PLS models with first differenced Google Trends, $\Delta \boldsymbol{X_g}$. The numbers inside the parenthesis indicate the number of factors in the model. The models are estimated using a rolling window of 48 observations.

Table 9: Placebo test R^2_{OoS}

		$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=9$	$h=12$
$t-PLS_{\boldsymbol{X_{placebo}}}(1)$	95% quantile	2.38	-46.63	-70.64	-88.44	-93.61	-98.33	-177.17	-251.47
	99% quantile	10.43	-39.69	-62.94	-75.80	-84.41	-88.48	-162.40	-229.57
$t-PL\boldsymbol{S}_{\boldsymbol{X_{placebo}}}(2)$	95% quantile	-11.65	-69.51	-97.89	-117.05	-124.04	-128.36	-216.87	-304.84
	99% quantile	-2.49	-57.63	-84.36	-100.62	-108.50	-113.83	-196.65	-276.68
$t-PLS_{\boldsymbol{X_{placebo}}}(3)$	95% quantile	-27.21	-91.63	-124.41	-144.49	-152.25	-159.98	-254.01	-356.49
	99% quantile	-17.02	-76.12	-107.13	-126.33	-137.36	-137.45	-228.37	-326.34

The table shows the R^2_{OoS} for the 95% and 99% quantile of the null region for the targeted PLS model with placebo data. The data, $\boldsymbol{X_{placebo}}$, is constructed to have the same mean, variance and AR(1) coefficient as the series selected by soft thresholding but with no true predictive power for employment growth. The models are estimated using a rolling window of 48 observations.

Research Papers
2018

CREATES
Center for Research in Econometric Analysis of Time Series

2018-08: Torben G. Andersen, Nicola Fusari and Viktor Todorov: Short-Term Market Risks Implied by Weekly Options

2018-09: Torben G. Andersen and Rasmus T. Varneskov: Consistent Inference for Predictive Regressions in Persistent VAR Economies

2018-10: Isabel Casas, Xiuping Mao and Helena Veiga: Reexamining financial and economic predictability with new estimators of realized variance and variance risk premium

2018-11: Yunus Emre Ergemen and Carlos Velasco: Persistence Heterogeneity Testing in Panels with Interactive Fixed Effects

2018-12: Hossein Asgharian, Charlotte Christiansen and Ai Jun Hou: Economic Policy Uncertainty and Long-Run Stock Market Volatility and Correlation

2018-13: Emilio Zanetti Chini: Forecasting dynamically asymmetric fluctuations of the U.S. business cycle

2018-14: Cristina Amado, Annastiina Silvennoinen and Timo Teräsvirta: Models with Multiplicative Decomposition of Conditional Variances and Correlations

2018-15: Changli He, Jian Kang, Timo Teräsvirta and Shuhua Zhang: The Shifting Seasonal Mean Autoregressive Model and Seasonality in the Central England Monthly Temperature Series, 1772-2016

2018-16: Ulrich Hounyo and Rasmus T. Varneskov: Inference for Local Distributions at High Sampling Frequencies: A Bootstrap Approach

2018-17: Søren Johansen and Morten Ørregaard Nielsen: Nonstationary cointegration in the fractionally cointegrated VAR model

2018-18: Giorgio Mirone: Cross-sectional noise reduction and more efficient estimation of Integrated Variance

2018-19: Kim Christensen, Martin Thyrsgaard and Bezirgen Veliyev: The realized empirical distribution function of stochastic variance with application to goodness-of-fit testing

2018-20: Ruijun Bu, Kaddour Hadri and Dennis Kristensen: Diffusion Copulas: Identification and Estimation

2018-21: Kim Christensen, Roel Oomen and Roberto Renò: The drift burst hypothesis

2018-22: Russell Davidson and Niels S. Grønborg: Time-varying parameters: New test tailored to applications in finance and macroeconomics

2018-23: Emilio Zanetti Chini: Forecasters' utility and forecast coherence

2018-24: Tom Engsted and Thomas Q. Pedersen: Disappearing money illusion

2018-25: Erik Christian Montes Schütte: In Search of a Job: Forecasting Employment Growth in the US using Google Trends