OFF TO THE RACES: A COMPARISON OF MACHINE LEARNING AND ALTERNATIVE DATA FOR PREDICTING ECONOMIC INDICATORS

Jeffrey C. Chen Abe Dunn Kyle Hood

Alexander Driessen Andrea Batch*

July 5, 2019

Abstract

Timely alternative data sources such as credit card transactions and search query trends have become more readily available in recent years, while sophisticated machine learning (ML) techniques have enabled marked gains in predictive accuracy. These advances offer the benefit of revealing economic news earlier in the estimation cycle, reducing revisions, and improving estimate quality. But which combinations of data and ML techniques give the *most* accurate prediction of national economic activity? To answer this question, we conduct a prediction horse race using a one-step ahead model validation design to evaluate how each ML algorithm, data set, and variable selection method weighs on predictive accuracy. We test 73,884 model specifications, consider 1,180 variables drawn from both traditional and alternative sources, and predict 188 quarterly revenue and expenditure series for the services sector as published in the Quarterly Service Survey (QSS)—a key data set that accounts for nearly 80% of the revisions to Personal Consumption Expenditure for Services (PCE Services). Our results indicate that ensemble methods such as Random Forests afford the highest chance of reducing revisions. Relative to current national accounting methods, ensemble methods could reduce overall PCE revisions by 12% on average, with proportionally larger improvements among PCE sub-components. While alternative data are timelier, we find evidence that traditional data such as employment and lagged dependent variables contain relatively greater signaling power than alternative data; this finding demonstrates that more data does not necessarily translate into significantly better predictions.

* All authors are based at the U.S. Bureau of Economic Analysis. The paper has benefited from insightful comments with seminar participants at BEA Advisory Committee meeting, the Federal Economic Statistics Advisory Committee meeting, Strata Data Conference, and the BigNOMICS Workshop on Big Data and Economic Forecasting. We would like to also thank Pat Bajari, Francis Diebold, Brian Moyer, Sally Thompson, Dennis Fixler, Gary Cornwall, and Annabel Jouard for useful discussions.

1. Introduction

Gross Domestic Product (GDP) is one of the most widely used and cited measures of economic activity. Obtaining timely and accurate GDP estimates is essential for policymakers, the private sector, and individuals making a wide range of economic decisions. However, the Bureau of Economic Analysis (BEA), the agency responsible for producing GDP figures, must produce its initial estimates of GDP prior to when some critical source data are available. Thus, the reliability of advance estimates and the extent to which they capture news rather than noise hinges in part on successfully bridging data gaps.

One approach to bridging data gaps involves working with providers of source data to accelerate production of their estimates. For example, the U.S. Census Bureau accelerated publication of the Monthly Retail Trade and Sales Survey (MRTS) as an advance publication, which has translated into marked reductions in GDP revisions. While effective, this solution can be costly, may place undue burden on respondents, and may reduce the rate of response.

Alternatively, the breadth of timely proprietary data sources has expanded significantly in recent decades. The financial sector has relied on such data (including credit card transactions, e-mail receipts, search queries, etc.) to better forecast economic fundamentals and to anticipate financial performance of companies ahead of quarterly earnings reports. These data have the potential to do the same for official statistics. Nevertheless, these substitute data do suffer from some problems—non-traditional sampling, and large numbers of variables—that strain traditional statistical techniques. Instead, forecasters have developed sophisticated machine learning (ML) techniques in which non-parametric, non-linear, or otherwise computationally intensive algorithms yield predictions in just this type of environment. This combination of alternative data sources and contemporary ML techniques provides a possible bridge for the data availability gaps that producers of official statistics face.

These advancements are not without challenges and the transparency of ML is often called into question. Some view ML as a black box, especially because the techniques may not lend themselves to traditional modes of linear interpretation and because modeling decisions in non-parametric models may be too voluminous to efficiently evaluate. They also represent a philosophical shift: ML is aimed at producing predictions $\hat{y}_i$ rather than parameter estimates $\hat{\beta}$ (Mullainathan and Spiess 2017). Without being able to understand or interpret coefficients, there are some who argue that we can never fully understand the predictions given by ML models. Nonetheless, it is not the case that studies that use ML are devoid of economic intuition. In our case, the prediction target is of economic significance, and economic intuition will be preserved through the application of national economic accounting principles.

On the data side, newer sources of data can be timelier, but the reliability and stability of alternative sources have yet to be proven for official statistical purposes as they are only a recent

phenomenon.[1] The universe captured in alternative sources are not typically disclosed, making it challenging to evaluate the properties of the data.

In this paper, we explore how ML and alternative data sources can play a role in stabilizing official national statistics when faced with publication lags. We focus on Personal Consumption Expenditures Services (PCE Services) that account for more than $9.8 trillion of the current-dollar estimate in 2018 (> 45% of GDP). Approximately $4.2 trillion of the PCE Services is based on the Quarterly Service Survey (QSS), which is only fully available 75 days after the end of each quarter and informs the third estimate of GDP.[2] The current estimate revision to quarterly GDP has averaged $27 billion since 2012 with an average revision of $14 billion attributable to PCE.[3] QSS-based estimates contribute the largest share to PCE revisions, averaging $11 billion. Thus, by predicting the QSS, estimates using ML and alternative indicators can deliver economic news earlier in the estimate cycle and improve data quality.

Our approach is not to apply an "off-the-shelf" ML algorithm, but rather to dedicate significant attention to the unique features of the problem at hand, while at the same time advocating broad principles that should apply to similar applications. For this purpose, forecasts must be both robust and stable, and we must carefully contemplate the way predictive accuracy should be defined in the national economic accounting context. More specifically, we evaluate potential revisions reductions, (a) for each PCE component across all modeling scenarios; (b) for each algorithm across all PCE components and other modeling choices (data set, inclusion criteria, etc.); and (c) for combinations of these concepts.

Predicting these types of official statistics presents a unique challenge which guides the approach that we favor. Surveys or censuses are not conducted at high frequencies, and the intersection between their observations and the observations contained in alternative datasets to which we have access yields a rather short time series. The ML paradigm prescribes partitioning data into multiple parts: one for estimation, one for model selection, and one for testing. We do not have enough observations to subset the data into these multiple parts, so we propose a unique approach. Specifically, we estimate thousands of potential models for every series where each model applies distinct methods and data. Rather than selecting just the "best" model, which may overstate the improved prediction, we report and analyze the full distribution of predictions across model scenarios for a large cross-section of series. This approach has two distinct advantages. First, using the cross-section of series allows one to evaluate and identify which

---

[1] Private sector data sources have been used for many components of the national accounts for decades. While the use of private sector data is not new, the availability and types of alternative data sources has changed dramatically (e.g., credit card data and search queries).

[2] The Census Bureau also publishes an advance estimate of QSS at 45 days; However, it is a limited subset of all series.

[3] The revision is calculated as the third estimate less the advance estimate.

modeling decisions result in poor predictions across many series. For example, we find that the method of using a 4-quarter moving average performs quite poorly across data series. A second advantage of this approach is that it avoids the over-fitting that might occur by selecting only the best model. Instead, using a distribution of many models for each series, we can determine which series show consistent improvement across a sample of model scenarios.[4]

The paper is organized as follows. Section 2 places this work in the landscape of forecasting and nowcasting literature for macroeconomics. Section 3 describes the process of a prediction horse race and criteria for identifying PCE components that can be reliably improved. Section 4 examines prediction results, placing an emphasis on producing rules of thumb for modeling and estimating the effects of PCE revisions.

2. Literature Review

Traditional forecasting typically employs linear time-series models wherein theory dictates the appropriate estimators, based, for example, on asymptotics and an assume class of data-generating processes. However, a major constraint especially of linear models is that the number of variables that can enter the forecast must be considerably less than the number of observations. This reduces the amount of data that can enter the models to help inform the prediction. The machine learning techniques applied in this paper are not bound by this constraint and allow for the consideration of a much larger number of variables.[5] The disadvantages associated with this approach are in the necessity to put one's faith in model validation and testing.

The popularity of big data and machine learning has been growing rapidly in the forecasting literature over the last decade. Our paper differs from many of these studies not so much in the techniques that are applied, but in the objects that we are forecasting. To our knowledge, forecasts using big data for incorporation into official statistics is a rather unique application. The closest application of these techniques in the recent literature has been to nowcast Macroeconomic aggregates.

---

[4] This approach is in the spirit of Leamer (1983) who advocated reporting a broad distribution of models as he was concerned that researchers searching for the "correct" specification may cause a high degree of bias and more recently Athey and Imbens (2015), who are concerned with misspecification uncertainty.

[5] It is not impossible to approach problems with more predictors than observations using a more traditional paradigm, and many of the important conventions of ML, such as validation and testing, are not unique thereto. Frequentist approaches applicable to such problems include model selection (for a review see Kadane and Lazar, 2004), model averaging (some recent examples include Hansen, 2007 and Hansen and Racine, 2012) and factor models (cf. Stock and Watson, 2006). Bayesian model averaging may also be applied to "wide" data sets, using a dimensionality-reduction techniques or stochastic searches (Fragoso *et al.* 2018). ML is thus one among many approaches that could be applied. Nevertheless, it is particularly well-suited to this problem based on the sheer number of right-hand-side-variable combinations that are possible.

A major benefit of writing a paper in a field that is growing in popularity is the existence of recent, high-quality review articles. Einav and Levin (2013) provide an overview of important concepts, data sources, and common forecasting techniques. They note that the larger scale, breadth of variables, lack of structure present new opportunities, but also new problems that must be dealt with by the researcher. In addition, they note the need for cross-validation—a technique that is rarely used by economists but is essential in this context. Varian (2014) also offers an overview and a sort of how-to guide in applying machine learning techniques to big data, while identifying where these techniques originated in the broader scientific literature. Kapetanios and Papailias (2018) provide an extensive review of very recent studies that have used these techniques, organized by prediction target (unemployment, inflation, output, and financial variables), as well as a detailed discussion of many important techniques.

Because in this paper we focus on near-term forecasts of the recent past, what we are doing can be called nowcasting. Nowcasting is a portmanteau of "now" and "forecasting," and was defined by Giannone, Reichlin, and Small (2008) to comprise forecasting of the recent past, present, or near future. However, we are not exposed to several problems that are particular to nowcasting: "ragged edges" in which because of real-time data flow, the forecaster does not have access to all data series at all points in time, and mixed-frequency data. As such, our application has more of a forecasting flavor.[6]

The constellation of big data, machine learning and nowcasting has spawned a literature that is somewhat distinct from the "traditional" nowcasting literature. This is precisely because these two approaches generally deal with a distinct collection of complications. Traditional approaches of regression and time series analysis have ready-made solutions to the ragged edge problem (that use, e.g., a Kalman filter), while the machine learning literature has generally ignored such considerations. As such, the types of "big data" that machine learning typically uses are somewhat different. Nevertheless, there is a recent and growing literature in this field summarized by Kapetanios and Papailias (2018). Biau and D'Elia (2013), for example, use survey data and a random forest algorithm to nowcast Euro-Area GDP, Nyman and Ormerod (2017) use a random forest algorithm to predict recessions, and Choi and Varian (2012) use Google Trends to nowcast several macroeconomic indicators such as auto sales and unemployment claims. Rajkumar (2017) compares various algorithms, including a Random Forest, to predict surprises in GDP growth.

[6] Earlier nowcasting work relied on regression-based methods, which include what is termed "bridging" or "bridge equations" and MIDAS regressions (cf. Bańbura, Giannone, and Reichlin (2011) for a review). Bridging uses time aggregation of monthly data combined with regression analysis to produce a nowcast, while in MIDAS models (Ghysels, Santa-Clara, and Valkanov 2004), variables of different frequencies to directly enter the regression equation. The ragged edge problem is solved with the application of "state-space" models in which variables that are used in the nowcast but are missing are themselves forecasted, a process typically implemented via a Kalman filter. Subsequent attempts to nowcast macroeconomic variables with large data sets involved the application of data-reduction techniques, e.g., dynamic factor models (Bańbura, Modugno, and Reichlin 2013). Bok et al. (2017) describe the New York Fed's nowcasting approach, which synthesizes many of these techniques. This summary does not cover the whole of the recent nowcasting literature, and so we refer the reader to Kapetanios and Papailias (2018) for a more detailed overview.

Finally, the adoption of any type of nowcasting technique for "filling in" series that are not yet available to be used in official statistics has few examples in the literature. Cavallo et al. (2018) use the "billion prices project" data (Cavallo and Rigobon 2016) to produce high-frequency purchasing power parities (PPPs), which could be used to bridge the period between releases of the World Bank's Penn World Table's International Comparisons Program's PPPs. Similar price indices might also be used to replace certain headline numbers such as Argentina's CPI, which is believed to be unreliable (Cavallo and Rigobon 2016). Finally, B. Chen and Hood (2018) use traditional nowcasting techniques (bridge equations, bridging with factors) combined with model selection to nowcast detailed components of personal consumption expenditures on services that go into the calculation of gross domestic product, showing the potential for significant reductions in revisions in many of these components.

3. Methods and Data

3.1. Modeling Considerations

The objective of this study is to reduce revisions to GDP by identifying predictive approaches that offer consistent improvements. There are challenges in this task, particularly in how we account for the properties of the data and in identifying where prediction can be reliably applied.

The properties of input data that are typically used for national economic accounts combined with the properties of alternative data present a unique forecasting challenge. Survey or census time series tend to be relatively coarse (e.g., monthly, quarterly, or annual). When used in conjunction with alternative data (which are a recent phenomenon, as mentioned above), the resulting time series tends to be short. The alternative data that we use, however, have a very broad cross-sectional dimension. As such, the number of variables, k, significantly exceeds the number of observations, n, a situation that is not a good fit for traditional statistical analysis. For this type of application, the problem with regression-based models is not that they are inaccurate (although they may be), but that they cannot even be estimated. One solution is to apply theory-driven methods that prune the input variables so that a model can be estimated, but this has proven to be ineffective for many applications (Stock and Watson 2014). Methods such as step-wise regression leave in inputs that are highly correlated with the series being predicted, but because pruning is based on in-sample correlations, estimation often results in overfitting and poor out-of-sample predictions.

In contrast, many ML techniques are designed for just this purpose, relying on a combination of model validation techniques and implicit variable selection. Traditional approaches often posit a "true model" that will obtain with enough observations, while ML focuses on producing generalizable predictions, using flexible non-linear approaches such as bootstrap aggregation or shrinkage to overcome overfitting, and relying nearly exclusively on partitioning to assess fit and

select models. As mentioned above, there are some trade-offs, but these types of models are needed to integrate into estimates the signal coming from these timely but high-dimensional data sets.

In this application, we are faced with a further problem that not only is the number of independent variables relatively large, but also the number of observations is small in absolute terms. Having a small sample size reduces power. Not only are a model's opportunities to learn economic patterns limited, but it is less likely to be resilient to structural instabilities that cause prediction accuracy to erode (Rossi 2013). Model selection also becomes challenging. When applying conventional forecast comparison techniques such as the Diebold-Mariano Test (Diebold and Mariano 1995), the lack of power prevents crowning a winning model. One can imagine a scenario in which a forecast model achieves lower error than its alternatives within sample, but the relative performance may not persist as the sample grows. This is particularly problematic if researchers estimate many forecasting models and then choose to report only their best-fitting estimate, which results in overfitting problems and poor out-of-sample performance.

Small sample size is not a problem that ML is specifically designed for. Standard application of ML algorithms might involve splitting the data into three sets: One for training (estimation), one for validation (in this case, model selection), and one for testing (assessment of fit). Fit (i.e., accuracy) cannot be assessed using any part of the sample on which estimation or model selection is done, and model selection cannot be done using the sample from which the models were estimated. If we were to divide all 30-some quarters into three distinct sets, no inferences could be made with reasonable statistical power.

For this reason, we propose to run a prediction "horse-race," in which we estimate a large collection of models for each series. We vary these models along several dimensions: algorithm, data, and variable selection. By varying the conditions and comparing their results through a prediction horse race, we can determine which dimensions drive accuracy for each industry. If one modeling choice seems to produce inaccurate predictions in most series that are being forecasted, or if one modeling choice seems to do the best on average, we can decide as to which modeling choices can be included or excluded from the final ensemble. In our analysis, model performance is gauged by pooling the estimates of fit (root mean squared revision, or RMSR) of all models and series into a single data set. A statistical analysis is then performed to assess the effect of each modeling choice on the expected revision.

In the subsequent subsections, we describe the process of constructing thousands of models that are trained under a multitude of modeling scenarios (e.g., combinations of algorithms, data, and variable selection procedures). We then construct measures of revision reductions and a simple framework to rate PCE components that are well-suited for this prediction approach.

3.2. Prediction Models

We conduct a horse race between different algorithms, data, and variable selection procedures (each individual combination of the three we call a "model") the results of which are compared with current BEA methodologies to evaluate the improvement. Each model can be expressed as:

$$y_{it} = f_m[g_k(X_t)]$$

where y_{it} is the not seasonally adjusted (NSA) quarterly growth in percentages of a QSS industry i in time t, f_m is any one of nine ML algorithms (see Section 3.1.1), X_t is a matrix of input variables and dependent lags in the form of quarterly growths at time t, and g_k is the procedure k for variable selection that guide how input variables are included (see Section 3.3.3).[7]

3.2.1. Algorithms

A diverse array of algorithms is selected that interact with the data in different ways. Some are commonly employed in the social sciences, whereas others are used in sectors that rely more heavily on data science techniques. We categories these techniques into two broad buckets: *linear methods* and *non-parametric methods*.

To represent techniques that overlap with the traditional econometric toolkit, we consider four linear methods:

Four-Quarter Moving Average (4QMA). The simplest of the linear methods is the 4QMA that smooths the univariate series using a one-year sliding window:

$$\hat{y}_{it} = \frac{1}{4}\sum_{j=1}^{4} \frac{y_{i,t-j}}{y_{i,t-j-1}}$$

where j is an index of prior quarters. The effect is an extrapolation that appears to be seasonally adjusted. Its simplicity is also its weakness, producing predictions with the risk of carrying forward momentum from prior periods and ignore contemporaneous information.

Forward Stepwise Regression (Stepwise). Forward stepwise regression is an automated variable selection procedure built around linear regression. The process adds variables to a regression one at a time, doing so based on partial F-tests. Each step of the process is computationally intensive, starting by estimating a null model without predictors, then adding one variable at a time starting

[7] We model growth rates rather than trends or levels because growth rates in the QSS are applied to update PCE estimates, not the levels. Moreover, through the benchmarking and revision process, levels will eventually be replaced with data from more reliable sources.

with the lowest partial F-test that is below a pre-defined threshold α. This requires that a set of candidate models is estimated prior to adding new variables (Efroymson 1960). We set $\alpha = 0.05$ requiring additional variables to yield partial F-test values below the threshold. In addition, given the small sample constraints, we place a cap on the number of parameters at $k = \sqrt{n}$. The technique has drawbacks, particularly that it conducts variable selection in-sample that results in predictions that are not generalizable (Copas 1983). In addition, the estimate is constructed on unconstrained least squares, so that ill-posed problems where $k > n$ are non-invertible.

Ridge Regression *and* ***Least Absolute Shrinkage and Selection Operator (LASSO)***. Several challenges with Stepwise method are addressed through regularized least squares methods, which introduces a constraint that forces sparse solutions in the regression coefficients. We consider two varieties: *Ridge Regression* (Hoerl and Kennard 1970) and *Least Absolute Shrinkage and Selection Operator (LASSO) regression* (Tibshirani 1996).

Ridge regression modifies least squares by adding a pre-selected constant λ into the coefficient estimator:

$$\hat{\beta} = (X'X + \lambda I)^{-1}X'Y$$

The parameter estimates are obtained by minimizing the penalized sum of squares with a l_2 norm penalty:

$$PSS = \sum_{i=1}^{n}(y_i - \sum_{j=1}^{m} x_{ij}\,\beta_j)^2 + \lambda \sum_{j=1}^{m} \beta_j^2$$

By adding the penalty, we can see that as coefficients β_j grows, the cost function is penalized and places greater preferences for smaller coefficients. The value of λ is tuned through k-fold cross-validation to minimize the cost function. A more recent innovation to this method is the LASSO model, that makes a simple modification to the penalty—replacing the $l2$ norm with a $l1$ norm:

$$PSS = \sum_{i=1}^{n}(y_i - \sum_{j=1}^{m} x_{ij}\,\beta_j)^2 + \lambda \sum_{j=1}^{m} |\,\beta_j|$$

Whereas the Ridge regression forces smaller parameter estimates, LASSO conducts variable selection by forcing some parameters to the edge case of exactly zero. While regularized least squares methods is an improvement on least squares, linear methods may not capture non-linearities and interactions that non-parametric algorithms can. We thus also consider five non-parametric techniques that are more flexible.

Regression Trees (CART). The building block for a number of these non-parametric techniques is Classification and Regression Trees (CART), more specifically the *regression tree* (Breiman

et al. 1984). The objective of CART is to recursively split a sample into smaller, more homogeneous partitions known as nodes. Each split yields two child nodes that are defined by a threshold θ along variable x_j:

$$I^- = \{i: x_j < \theta\}$$

$$I^+ = \{i: x_j \geq \theta\}$$

where I^- and I^+ are sets of observations that are below and above θ. As multiple values of θ are considered, the best θ minimizes the sum of squares:

$$SS = \sum_{i \in I^-} (y_i - \overline{y}^-)^2 + \sum_{i \in I^+} (y_i - \overline{y}^+)^2$$

in which y^- and y^+ are the mean of y_i for candidate partitions above and below θ. Each resulting child node $(X_i, y_i)_{i \in I^-}$ and $(X_i, y_i)_{i \in I^+}$ is further partitioned until it cannot be split any further or when additional splits do not improve the model fit. Each terminal node is referred to as a *leaf* c. A fully-grown tree minimizes the sum of squares of tree f:

$$SS = \sum_{c=1}^{C} \sum_{i=1}^{n} (y_i - \hat{y}_c)^2$$

where C are all leaves in the tree, n is the number of observations within a leaf c, and $\hat{y}_c = \frac{1}{n} \sum_{i=1}^{n} y_i$.

While we can see that CART implicitly conducts variable selection by selecting split thresholds along variables, each node could in theory be split until all leaves are $n = 1$. An overgrown, overly complex CART thus may overfit the data and introduce unnecessary variance into predictions. One remedy is to *prune* the tree to reduce the complexity, choosing a level of complexity that minimizes out-of-sample error. In small samples, however, these tuning strategies may have minimal effect on the quality of predictions as each leaf is an average of a small cell of observations that lend little statistically meaningful support.

Random Forests. Regression trees can be improved upon by an ensemble method known as Random Forests (Breiman 2001). The algorithm process is simple:

1. Construct B number of samples with replacement with n observations and m randomly drawn variables from X.
2. Train regression tree f_b on the sample b.
3. Average the predictions from each f_b to obtain $\hat{y}_i$

$$\hat{y}_i = \frac{1}{B}\sum_{b=1}^{B} f_{bm}(x_i)$$

where $B = 500$ in this study and the number of variables m per tree is determined through tuning.

This technique offers a couple of gains over regression trees. First, by constructing many trees under similar but randomly drawn conditions minimizes model variance while keeping bias uniform. Second, the bootstrapping builds in a natural validation sample to calculate the *out-of-bag* (OOB) error for evaluating generalizability of predictions.[8] Parameter tuning can also take advantage of the OOB error by training Random Forest algorithms under varying conditions such as variables per tree, then *comparing* the average OOB error between models.

Gradient Boosting (XG Boost). Another ensemble technique that has gain in popularity is *gradient boosting*. As developed in Friedman (2001), gradient boosting generates *m*-number of base learners $f_m(x)$ that are trained to correct errors made by prior iterations. Each base learner $f_m(x)$ is a *weak learner*—a model that may only have slightly better than random predictive power. In this case, we rely on a decision stump, which is a regression tree with only one split. Each base learner is generated sequentially and added to produce a prediction $F_M(x)$

$$F_M(x) = \sum_{m=1}^{M} \eta f_m(x)$$

where η is a shrinkage parameter between 0 and 1 that controls the rate in which the boosting model converges and has been shown to be an effective way to mitigate overfitting. As η decreases, the number of iterations *M* required to converge needs to be increased—these parameters are tuned together.

At some iteration m, the loss will have effectively converged, meaning that the addition of subsequent base learners may add noise to estimates and use unnecessary computational resources. For simplicity, we set the $M = 300$ with a learning rate of $\eta = 0.05$, but specify an early stopping rule that ends training if 15 consecutive iterations fail improve the model. We rely on XGBoost implementation of the technique as described in T. Chen and Guestrin (2016).

Support Vector Regression (SVR). SVR fits a linear regression on input data that has been mapped using a non-linear function. The non-linear function can take on various functions $k(x_i, x_j)$, such as a Gaussian radial basis function kernel:

[8] The out-of-bag error is the error based on the observations left out of the bootstrap draw, which is a commonly applied in ML models.

$$k(x_i, x_j) = exp(|x_i - x_j|^2)$$

$k(x_i, x_j)$ transforms the input variables *x* into a higher-dimensional space to better model patterns in the data. The linear regression yields a hyperplane in which each y_i resides within a hard margin of error ϵ: $(\hat{y}_i - \epsilon) \leq y_i \leq (\hat{y}_i + \epsilon)$. Each prediction $\hat{y}_i$ is found on this hyperplane. This constrained optimization problem can be infeasible as some observations may lie beyond the margin, thus a cost parameter *C* can "soften" the margins. A soft margin of error allows some observations to reside beyond the margin but penalizes those observations by their distance from the margin (i.e., the amount of "slack" they are permitted), thereby regularizing the model to reduce the incidence of overfitting (Drucker et al. 1996).

SVR may require more time to train than other algorithms, thus for cost efficiency, we tune the cost parameter C along a grid for a sample of industry targets, then fix values of the parameters for all other industries based on the optimum.

Multi-Adaptive Regression Splines (MARS). MARS fits k-number of basis functions that are combined to produce a prediction (Friedman 1991):

$$\hat{f}(x) = \sum_{i=1}^{k} \alpha_i \, b_i(x)$$

where each basis function b_i is weighted by a coefficient α_i learned by minimizing the sum of squared errors. Each basis function $b_i(x)$ can take on one of three forms: a constant term—or intercept, a hinge function, or the interaction of hinge functions. Hinge functions fit splines to the data—allowing a regression line to bend at a threshold along *x* so that the slopes may vary on either side. By taking advantage of a potentially large number of splines, MARS molds to the non-linearities and discontinuities in even highly dimensional data sets, but a potentially large number of basis functions may overfit the data. The technique thus unfolds as a two-step process: a forward stage and a backward stage. The forward stage fits and weights candidate pairs of hinge functions, choosing only to add the pair to the overall model if it reduces training error by the largest margin. The *backward pass* mitigates overfitting by removing least effective terms subject to generalized cross validation.

We apply MARS using an open source implementation called *earth* (Milborrow 2018). The forward pass requires tuning the degree of interaction effects among basis functions. The backward pass is also tuned based on the number of terms to retain. We conduct a grid search by considering all combinations of interaction effects for degrees 1 through 3 and number of retained terms (5, 10, 15).

A summary list of the different methods is shown in Table 1 for reference.

TABLE 1: Algorithms consider for the prediction horse race.

Technique	Training and Tuning Procedure
Linear Methods	
4-Quarter Moving Average (4QMA)	Calculate 4-quarter moving average.
Forward Stepwise Regression	Set max number of parameters k to the square root of the sample size.
LASSO Regression	Leave-one-out cross validation to find value of lambda that minimizes mean squared error.
Ridge Regression	Leave-one-out cross validation to find value of lambda that minimizes mean squared error.
Nonparametric Methods	
Regression Trees	Grow tree to full depth and cross validate error in each step, then select tree complexity that minimizes MSE.
Random Forests (RF)	Number of trees set to 500. Select the number of variables per tree along a grid of possible values choosing the lowest OOB error.
Gradient Boosting	Set maximum iterations to 300, η = 0.05, early stopping if model error does not improve after 15 rounds.
Multi-Adaptive Regression Splines (MARS)	Tune over a search grid of degree of interaction effects (1 to 3) and number of terms to retain during pruning pass (5, 10, 15).
Support Vector Regression (SVR) with Radial Basis Function (RBF)	Search hyperparameter C along a grid for a sample of industry targets, then fix values of the parameters for all other industries based on the optimum.

3.2.2. Variable Selection

Models are only as good as their inputs. Too much information may lead to an overfit model and highly variable predictions. Too little information place disproportionate weight on a few variables, thereby introducing bias into predictions. In machine learning, a happy medium involves conducting dimensionality reduction to reduce the number of variables considered, while still extracting the key information from the variables. Sample size constraints may limit the effectiveness of more sophisticated variable selection techniques.

We instead consider two contrasting approaches that represent the extremes of variable selection: *cherry picking* and *kitchen sink*. Economic intuition tends towards parsimonious specifications, including only variables that capture economic and behavioral forces. Thus, cherry picking in this context is defined as the inclusion of input variables that are conceptually like the left-hand side variable. For example, if physician offices revenue (NAICS 6211) is the target, then only

medical-related factors are included as input variables. However, if important information is omitted, then models are underfit and can miss the trend.

Alternatively, kitchen sink models include all available data, placing no assumption on which variables should be included. This implies that the algorithms have the capacity to conduct implicit variable selection and can incorporate information without introducing excess noise.

3.2.3. Data Sources

National accounts are an amalgam of public and private sources. In fact, private source data are incorporated in various areas of economic measurement such as motor vehicle production and Value Put In Place (VPIP) estimates for construction. Alternative private data offer the possibility of capturing news that may otherwise be overlooked by indicator series or projections, though recognizing that private administrative data are collected with a goal other than national statistics (e.g., profit maximization). Thus, our proposed machine learning-alternative data hybrid should not be viewed as a replacement for current projection methods, but rather a supplemental source that is run in parallel that assesses the validity of current projections.

The target series are 188 industry time series published in the QSS, available in time for the third estimate of GDP. To ensure predictions produce an output that are useful for estimate production, we target NSA percentage quarterly growths for both revenue and expenditure series for a 31-quarter period—between the second quarter of 2010 and the first quarter of 2018.

We assemble a variety of input data from traditional and alternative sources. Among traditional sources are NSA aggregates from the Bureau of Labor Statistics's (BLS) Current Employment Survey (CES) and Consumer Price Indices (CPI). These sources are currently used in estimating national indicators, are publicly available and are constructed on probability samples—in other words, these are generalizable samples with known universes and quantifiable biases.

Two alternative data sources are considered. First, credit card transactions are acquired from First Data, which offers credit card processing services for a network of merchants across the United States. The data are available daily within the first 10 days after the end of a month and are processed by Palantir using a methodology developed by the Federal Reserve Board of Governors (Aladangady et al. Forthcoming). To minimize the effect of churn, each monthly transaction estimate only includes merchants that have been First Data customers within the prior 13 months. These data provide a timely view into purchasing behavior, trading representativeness off with timeliness.

Google Trends is another source of timely, near-real-time source of data that covers a wide range of activity. In many respects, trends gauge public interest in various economically-related issues, as captured through Google's online offerings, including Google Search, Google News, and Froogle. 160 keywords were derived from QSS NAICS definitions and monthly estimates for the

period of 2003 through 2017 were requested via the Google Trends API. The API returned 240 volume indexes that were constructed from a simple random sample of search queries, aggregated into a time series of proportion of total Google search activity, and indexed to the maximum search volume share in the time series.

TABLE 2: Data sources used for this prediction study.

Data	Description	Economic Relevance
Census Bureau Quarterly Services Survey (QSS)	Longitudinal survey of 19,000 US businesses operating in the services sector.	Key input into BEA's Personal Consumption Expenditure (PCE) series.
BLS Current Employment Survey (CES)	Employment estimates released monthly, converted into quarterly average. CES is currently relied on for national accounting estimates. Contains 140 industry series.	Employment trends that coincide and trend with consumption.
BLS Consumer Price Indexes (CPI)	National-level price indices for products and are currently relied on for national accounting estimates. Each CPI is associated to NAICS code based on keyword similarity. Contains 600+ series	Price changes of items that are consumed alongside services.
First Data Credit Card Transactions	Near real-time credit card transaction aggregates, converted from Merchant Class Codes (MCC) to NAICS. Contains 192 industry series.	Contemporaneous measure of consumption.
Google Trends	Monthly activity indices for search queries, Google News topics, and Froogle shopping activity. Converted from search terms to NAICS based on keyword similarity. Contains 240 industry series.	Gauge of interest and prospective buying behavior on the internet.

3.2.4. One-Step Ahead Validation

Of the $n = 31$ observations, $n = 12$ are set aside for validating performance. As our objective is to generalize and apply models, we simulate the PCE estimation process using a one-step-ahead model validation. The model validation technique is an iterative one, producing each $\hat{y}_{it}$ by training on all data $t < T$, then applying the prediction developed on data points $t < T$ to produce a prediction for the observation $t = T$. For each of the 12 validation quarters, we re-train each model by growing the data's time window (see Figure 1), thereby producing predictions that are responsive to evolving economic patterns. In Figure 1 we start with a prediction for *T = 0* and use time periods *T = -1* and less to form this prediction. We next move one step ahead to predict time *T = 1* using information in time periods *T = 0* and lower to form the prediction. The predictions in period *T = 2* and future periods proceed accordingly. While the number of

observations per prediction grows with time, we assume the benefits of greater accuracy and stability among the predictions should affect all models in the same way.

In total, 73,884 model scenarios were trained and produced predictions for 12 consecutive validation quarters, resulting in 886,608 model runs and predictions.

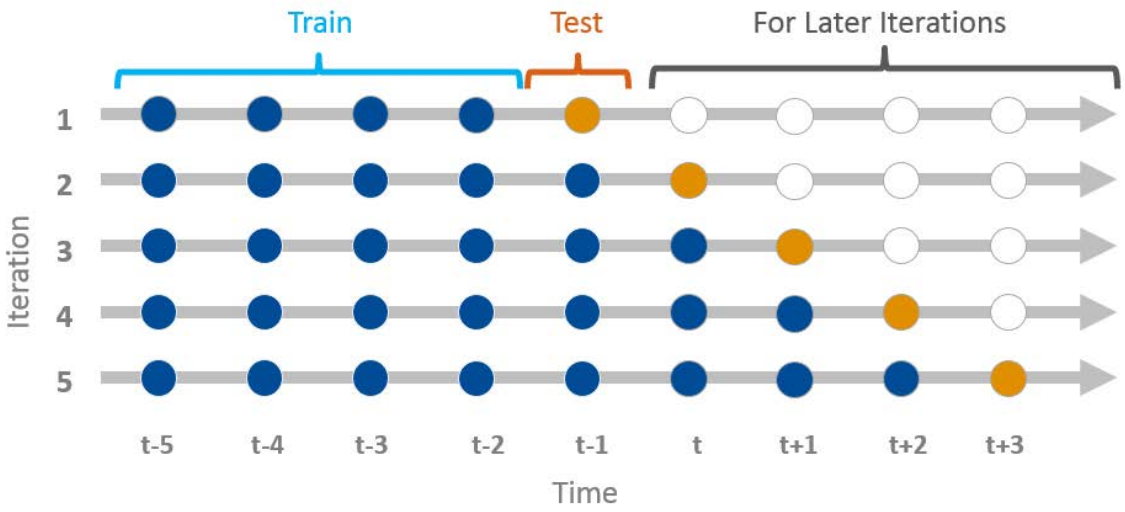

FIGURE 1: One-step ahead model validation design. The X axis represents both the training set (pink) and test set (orange). The Y axis represents the prediction time periods.

4. Evaluating Performance and Revision Reduction

When a large sample is available, a robust model selection framework should include both a model validation step (e.g., one step ahead or k-fold cross validation) to aid in selecting the most generalizable model and a test step to re-validate the chosen model's performance. The sample available for this study, however, is not sufficiently large to afford a test set, thus a model chosen from thousands of candidates may run the risk of overfitting the data. We instead take a conservative approach that evaluates performance by selecting ensembles of models developed under common conditions. For an industry *i*, for example, all models that were trained using a Random Forest would be considered one ensemble, whereas all models that rely on BLS CES would be considered another.

First, we train thousands of models for a cross-section of 188 QSS series covering many industries using one-step-ahead validation. QSS predictions $\hat{y}_{it}$ are converted into PCE component estimates $\hat{C}_m$ for a model *m*:

$$\hat{C}_m = g_c(\hat{y}_{it})$$

where g_c is BEA's PCE estimation process that seasonally adjusts and converts available QSS data into components estimates. Note that some PCE components rely on only one QSS series while others rely on multiple.

A prediction model applied where revisions are unlikely to reduce revisions will likely add error to official estimates, so it is important to evaluate the reduction in revisions. From the perspective of data quality, an estimate should only be used if revision reductions are consistently expected across a broad distribution of models. We construct two measures to evaluate revision reduction potential: The *Mean Revision Reduction Probability (MRRP)* and the *Proportion of Improved Periods (PIP)*.

Proportion of Improved Periods (PIP). It is easy to imagine that an ensemble can reduce revisions on average, but masks generally poor individual quarter-to-quarter performance. The PIP is the proportion of the test period that would have had a revision reduction had a given model been applied. This measure captures the consistency of revision reductions over time, placing emphasis on cases where there is a net improvement over current BEA methodology.

$$PIP_m = \frac{1}{T}\sum_{i=1}^{T}(|\hat{C}_{m,t} - C_{third,t}| < |\hat{C}_{current,t} - C_{third,t}|)$$

To summarize proportion of improved periods for each component PIP_c, we calculate the proportion of models that yield improvements in the majority of historical quarters:

$$PIP_c = \frac{1}{M}\sum_{m=1}^{M}(PIP_m > 0.5)$$

In small samples, it may be challenging to distinguish models on their performance and to some extent can be viewed as an arbitrary decision. Thus, when PIP_c is high, we would have some surety that a model selected at random could improve component C at least a majority of the time. Conversely, a low PIP_c value indicates that a prediction strategy poses an increased risk of increasing quarterly revisions in component C.

Mean Revision Reduction Probability (MRRP). Whereas PIP captures revision reductions with respect to time, we also consider how often average dollar revision reductions yield improvements to PCE components in the long run. $MRRP$ is based on the Root Mean Square Revision (RMSR) that compares PCE $\hat{C}_m$ to the actual third estimate of PCE resulting in:

$$\mathrm{RMSR}_m = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\mathrm{C}}_m - \mathrm{C}_{third})^2}$$

Similarly, $RMSR$ is calculated for the current projection methodology:

$$\mathrm{RMSR}_{\mathrm{current}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\mathrm{C}}_{\mathrm{current}} - \mathrm{C}_{third})^2}$$

Relative revisions ($\Delta RMSR_m$) are expressed as the dollar difference between $RMSR_m$ and $RMSR_{\mathrm{current}}$, where a negative value indicates a revision reduction:

$$\Delta\mathrm{RMSR}_m = \mathrm{RMSR}_m - \mathrm{RMSR}_{current}$$

Looking across a set of M models, we summarize their collective performance as the Mean Revision Reduction Probability (MRRP) defined as:

$$MRRP_c = \frac{1}{M}\sum_{m=1}^{M}(\Delta\mathrm{RMSR}_m < 0)$$

in which we are interested in the proportion of models that can achieve a net revision reduction. Like PIP, an arbitrary model selected to predict a component with a high $MRRP$ value is more likely to yield revision reductions.

Together, PIP and MRRP can be summarized by taking the harmonic mean:

$$\mu_k = 2 \times \frac{MRRP \times PIP}{MRRP + PIP}$$

where larger values of μ_k indicate more revision reductions. In samples with little power, μ_k could be used as the basis of identifying the number of components that should be included to maximize revision reductions; However, in this study, we use μ_k to examine the revision impacts of applying a prediction strategy at a pre-defined cutoff, namely $\mu_k \geq 0.8$.

5. Results

5.1. QSS Predictions

We sift through the manifold of results to better understand which algorithms, data sets and modeling practices contribute to prediction performance. The process generates 393 sets of predictions for each of the 188 QSS series, representing possible growth paths under a broad set of assumptions.

Taking a closer look at key industries shown in Figure 2, we see that the mass of the out-of-sample predictions tend to follow the variation in the target series. The center mass of the predictions over time also tend to have a central tendency, which suggests that prediction of the QSS growth is generally possible regardless of the modeling scenario.

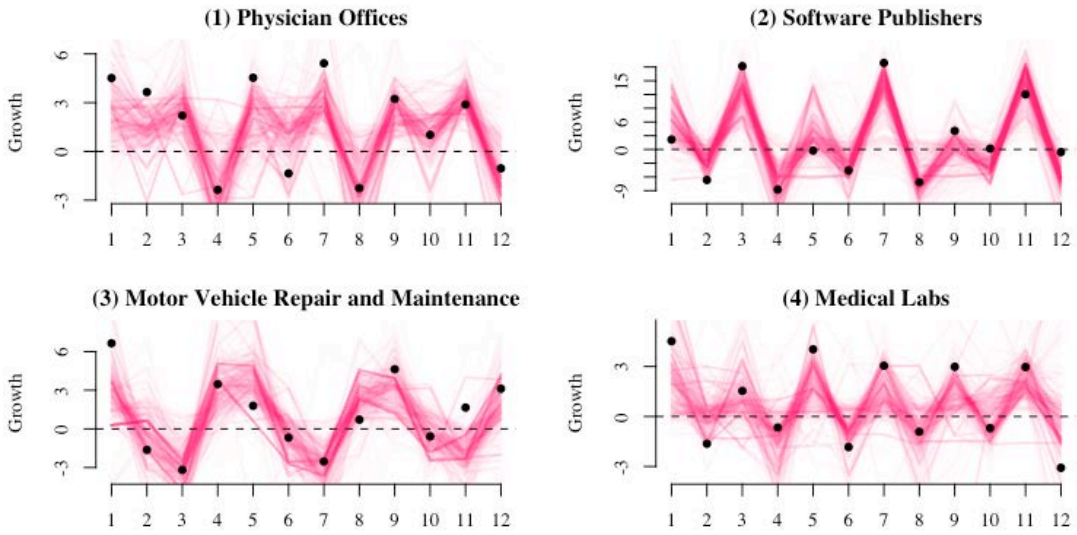

FIGURE 2: Comparison of actual QSS quarterly growths (black dot) with 393 sets of out-of-sample predictions (magenta lines)

Each algorithm reacts to data selection in a different way. The predictions for the physician offices category is a prime example, shown in Figure 3. Stepwise and CART regressions are prone to overfitting the data and are sensitive to high-leverage data points. Rather than producing a diffuse cloud of predictions that have correlated movements, they produce a discrete set of predictions, many of which perform relatively poorly. In contrast, the XG Boost and SVM algorithms produce predictions that are more dispersed. However, none of the poorer prediction paths is particularly prominent; in some of these cases there are algorithms that seem to be flatter, but none that show highly variable fluctuations nowhere near the actual data like the CART algorithm does. Rather, the central tendency in these algorithms is toward the actual data.

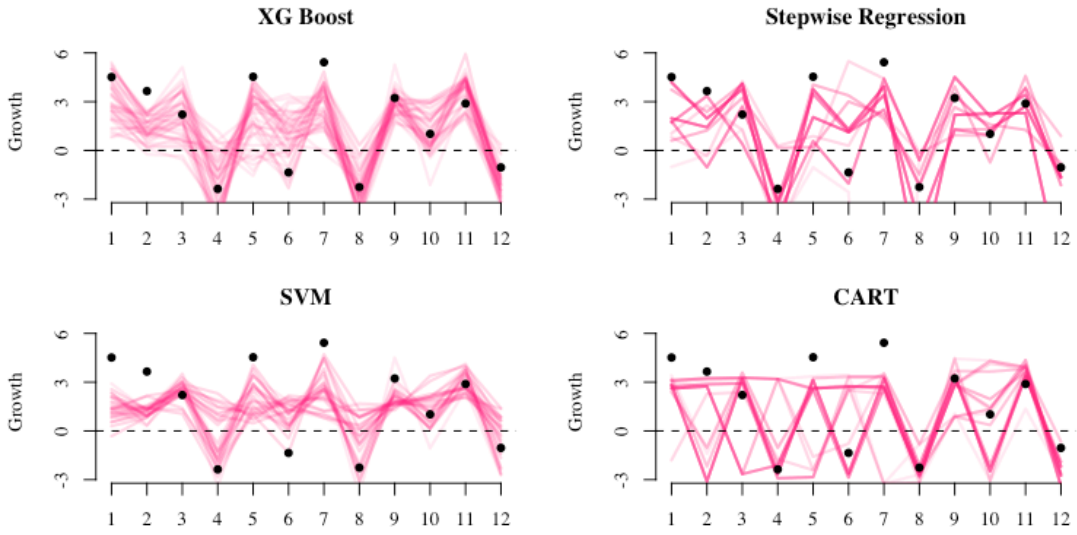

FIGURE 3: Comparison of different modeling assumptions applied to physician services series.

Model selection for prediction use cases is guided by finding the model with the lowest error: Given a series of models, we could choose the model that minimizes a squared loss function. This selection paradigm is effective when the sample size is large; however, as discussed previously, crowning a specific model champion is a foolhardy task with only $n = 12$, as the model selection process may overfit the data.

Instead, we take advantage of the sheer number of out-of-sample predictions to identify conditions that generally maximize predictive performance across a large cross section of 188 series that we study. We estimate a simple fixed effect regression to extract the average contribution of each modeling dimension:

$$RMSE_{i,k,m} = \beta + \alpha_i + \gamma_m + \xi_k + \epsilon_{i,k,m}$$

As we would expect, some industries are more predictable than others due to sampling variability and volatility in the sector; thus, we control for industry fixed effects α_i. γ_m are a matrix of dummy variables for each model type (e.g., Extreme Gradient Boost *xg*, Random Forests *rf*, etc.). ξ_k represents the data and variable selection procedures (e.g. cherry picking, CES, Google, etc.). From the resulting regression, we can determine which modeling strategies tend to perform better in matching the QSS estimates.

TABLE 3: Industry fixed effect regression results with clustered standard errors.

	(1)	(2)	(3)
Constant	5.01 (0.06)***	6 (0.08)***	6.01 (0.09)***
Algorithms (Ref = Stepwise Regression)			
4Q Moving Average	1.97 (0.23)***		2.16 (0.25)***
Ridge Regression	0.04 (0.07)		0.04 (0.07)
LASSO	-0.16 (0.04)***		-0.16 (0.04)***
CART	0.69 (0.11)***		0.69 (0.11)***
Random Forest	-0.55 (0.05)***		-0.56 (0.06)***
Gradient Boosting	-0.42 (0.05)***		-0.43 (0.05)***
SVM Regression	0.25 (0.1)**		0.25 (0.1)**
MARS	1.47 (0.13)***		1.48 (0.13)***
Data (Ref = Google)			
CES		-0.86 (0.1)***	-0.97 (0.11)***
First Data		-0.72 (0.08)***	-0.81 (0.09)***
Consumer Price Indexes		-0.35 (0.06)***	-0.39 (0.07)***
Dependent Lags		-0.83 (0.11)***	-0.87 (0.11)***
Variable Selection (Ref = Kitchen Sink)			
...Cherry Picking		0.22 (0.05)***	0.28 (0.06)***

Number of Data Sets (Ref = 1)			
2 sets		0.36 (0.05)***	0.31 (0.05)***
3 sets		0.81 (0.1)***	0.8 (0.11)***
Fixed Effects	Yes	Yes	Yes
N	73,884	73,884	73,884
R-squared	0.64	0.62	0.65
Adjusted R-squared	0.64	0.62	0.65
Residual Standard Error	2.75	2.83	2.72

Algorithms. Aside from the industry fixed effects, the choice of algorithm appears to have the greatest overall influence on RMSE. Among algorithms, we find that tree-based ensemble techniques offer the greatest improvements: relative to stepwise regression, random forests and gradient boosting reduce RMSEs on average -0.56 and -0.43 percentage points, respectively. LASSO regression offers an improvement over stepwise. In contrast, MARS and moving averages should be avoided due to their overwhelmingly poor performance. It is worth noting that prediction is a game of wins at the margins—if a technique does not perform well across industries, there is still a chance that it can offer consistent accuracy gains for individual industries. Nevertheless, because we don't have the data to assess all of the series individually, we have to assess the performance of these models more generally.

Data and variable selection. The data and variable selection dimensions suggest three takeaways.

- There are diminishing returns to adding additional data sources. For example, the coefficients imply that if First Data is added as a data source instead of Google Trends, the reduction in error is -0.81. However, if First Data is added as a second data source, the reduction in RMSE is -0.5 (= -0.8 + 0.3). If First Data is added as a third data source, there is an even smaller reduction in RMSE (-0.3 = -0.8 + (0.8 – 0.3)). Moreover, more data is not necessarily better (e.g., adding google as a second or third data source would increase RMSE).
- Second, models that are constructed on a purely conceptual basis may not necessarily translate into the statistically accurate results. Cherry-picked specifications add an average of 0.28 percentage points to the RMSE, meaning that specifications motivated by conceptual assumptions may omit some useful information from predictions or introduce noise. Thus, relying on the implicit variable selection of the machine learning techniques to surface predictive variables offers some gains.
- Lastly, the Current Employment Survey and dependent lags of QSS, both of which have long been available publicly, on average have the greatest influence on prediction quality.

The CES and CPIs are both currently used for the national economic accounts and if combined with machine learning could likely offer improvements in estimates.

The fixed effects from the above regression also provide estimates of the predictability of each QSS industry series. This is important as some series may be generally harder to predict than others, across all the methods that we consider. The difficulty in predicting a series could be related to a variety of factors such as the volatility of an industry or the sampling error of the series that we are attempting to predict. To investigate the relationship with sampling error, we compare the average prediction error ($\beta_0 + \alpha_i$) and the Census-Bureau-reported average sampling error for the QSS. If there were no prediction error, then all the error would come from sampling and our prediction error would be directly proportional to sampling error (dashed diagonal line), and for a few cases this is nearly the case, such as motor vehicle repair and maintenance, spectator sports, and insurance carriers. However, we find that most prediction error is higher than the sampling error (as expected). Increases in sampling error is problematic for our model predictions, with a one percentage point increase in the target series' sampling error, prediction error increases at a rate of 0.56-percentage points. This serves as a reminder that predictions are only as strong as the targets they mimic.

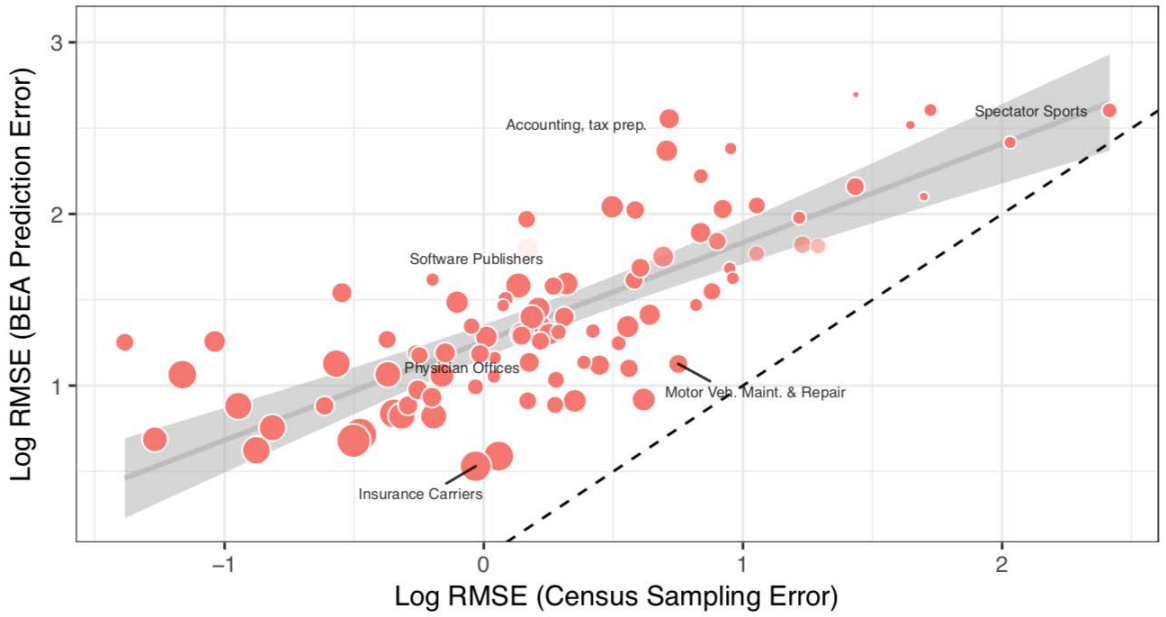

FIGURE 4: Comparison of Survey Sampling Error versus Prediction Error. Each point represents an industry, scaled by its total revenue or expenditure as of 2018-Q1. Transparency denotes statistical significance of fixed effect estimate—solid red indicates highly significant at the 1% level. Dashed diagonal line is the line of equality.

While our goal is to identify the winner of the prediction horse race discussed above, we do not wish to pare the results down too much based only on this regression. We note, however, that algorithm is the single-most important factor in determining RMSE, with a range of about 2.5 percentage points between the best and the worse-performing algorithms. Because along this dimension, the performance is improved by the largest margin, we elect in the following exercise to retain all combinations of all other dimensions (data set, scope), but retain only the most effective algorithm choice. The random forest algorithm generally seems to perform the best, and the second best (grading boosting) is a modification of the random forest algorithm. These two methods are both ensemble techniques, which form estimates based on averaging many nonlinear models. Nonlinear models that are not ensemble methods, such as CART and MARS, perform relatively poorly.

In subsequent sections, we evaluate the performance of the optimal modeling strategy based on a collection of 47 Random Forest models that were constructed under a variety of conditions.

5.2. Revision Reductions

Upon converting predictions of QSS to PCE estimates, each component of PCE can be evaluated on whether it may lead to to revision reductions relative to current practice based on our measures of improved fit (PIP and MRRP). 71 PCE services sub-components are considered—all of which incorporate one or more QSS series. We find that there is at least one prediction model for each of the 71 components that can improve upon current BEA methods. In large samples, this would be a reasonable finding. However, this is an overly optimistic conclusion for a small sample that lacks statistical power.

Instead, as we mentioned previously, we take a more conservative approach to evaluate models using measures of revision reductions (PIP and MRRP) to identify modeling strategies that on average yield improvements. In principle, one would place greater confidence in predicting a component in which 90% of models can reduce revisions rather than a component in which only 1% of models can meet the task. In low power samples, selecting a specific model from a pool of alternative models is like drawing a model at random. Thus, the chance of overfitting would arguably be less likely in the former case. Comparing across PCE components, the bubble chart shows significant heterogeneity in predictability—higher scores indicate greater surety that a model is not a random improvement. A component in the larger grey area indicates that one in two models can reduce revisions ($p \geq 50$) whereas the smaller box indicates that 8 in 10 models can reduce revisions ($p \geq 80$).

Based on these cutoffs, we find that 20 PCE components have at a least a coin flip's chance or better of seeing revision reductions—three of which have historically averaged at least $1 billion in revisions per quarter. This is not to say that other components are not predictable, but rather there is a far smaller margin of error for selecting a reliable model, especially given the limited

sample size. When reviewing the less predictable components, we find evidence that evaluating components on only one loss function could reduce data quality. MRRP alone would overstate the consistency of revision reductions as improvements could be concentrated in only a minority of time periods. For example, nearly half of the models predicting HHH (For Profit Home Health Care Services) satisfy the condition $\Delta RMSR < 0$, but less than 10% can improve estimates in at least a majority of test periods. Components like HHH have one or two large revision reductions that mask suboptimal performance in all other quarters.

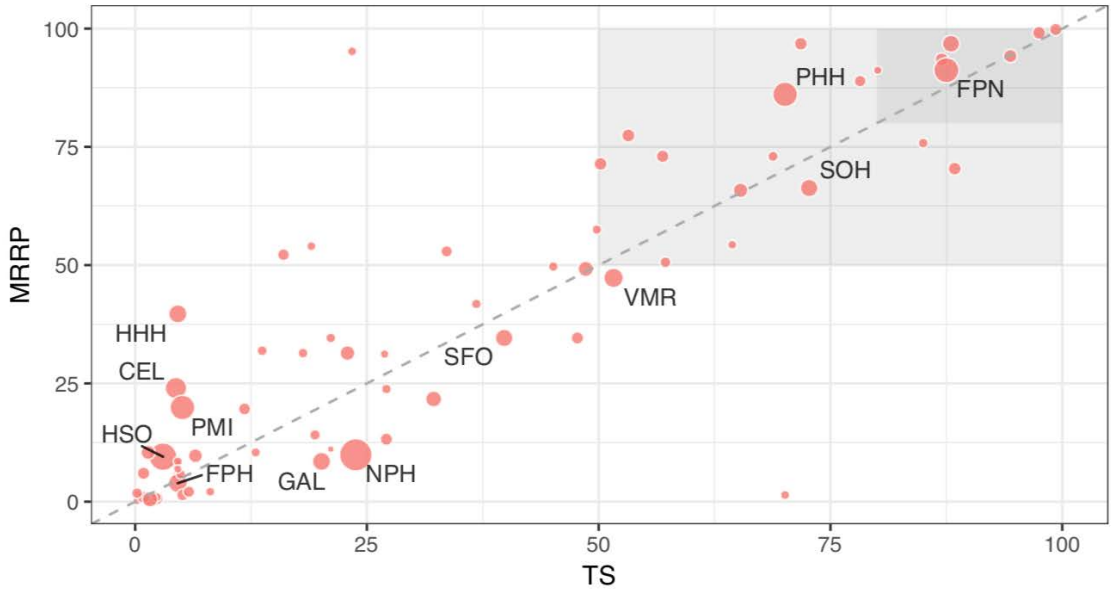

FIGURE 5: Comparison of MRRP and TS for each PCE Services component. Circles are scaled based on average quarterly revisions under current BEA methodology and labeled when revisions exceed $1 billion.

The story becomes more nuanced as we evaluate among alternative modeling strategies for each PCE component. Generally, the consensus, or lack thereof, give clues about what contributes to accuracy. Several components are predictable when applying almost any modeling strategy. Physician Services (PHH) and Specialty Outpatient Care (SOH) fall into this category, which translates as a need for fine-tuning towards optima rather that conducting an exhaustive search. Other components like Non-Profit Hospitals (NPH) have little chance of improvement regardless of the modeling strategy. These two scenarios may be due to a combination of the magnitude in sampling error of the underlying target series and availability of input variables. In contrast, modeling strategies for certain components fail to achieve consensus such as in the case of motor vehicle repair and maintenance (MVR). However, two algorithms stand apart in their ability to reduce revisions. We can infer that accuracy in this case may be more likely a matter of identifying the appropriate functional form.

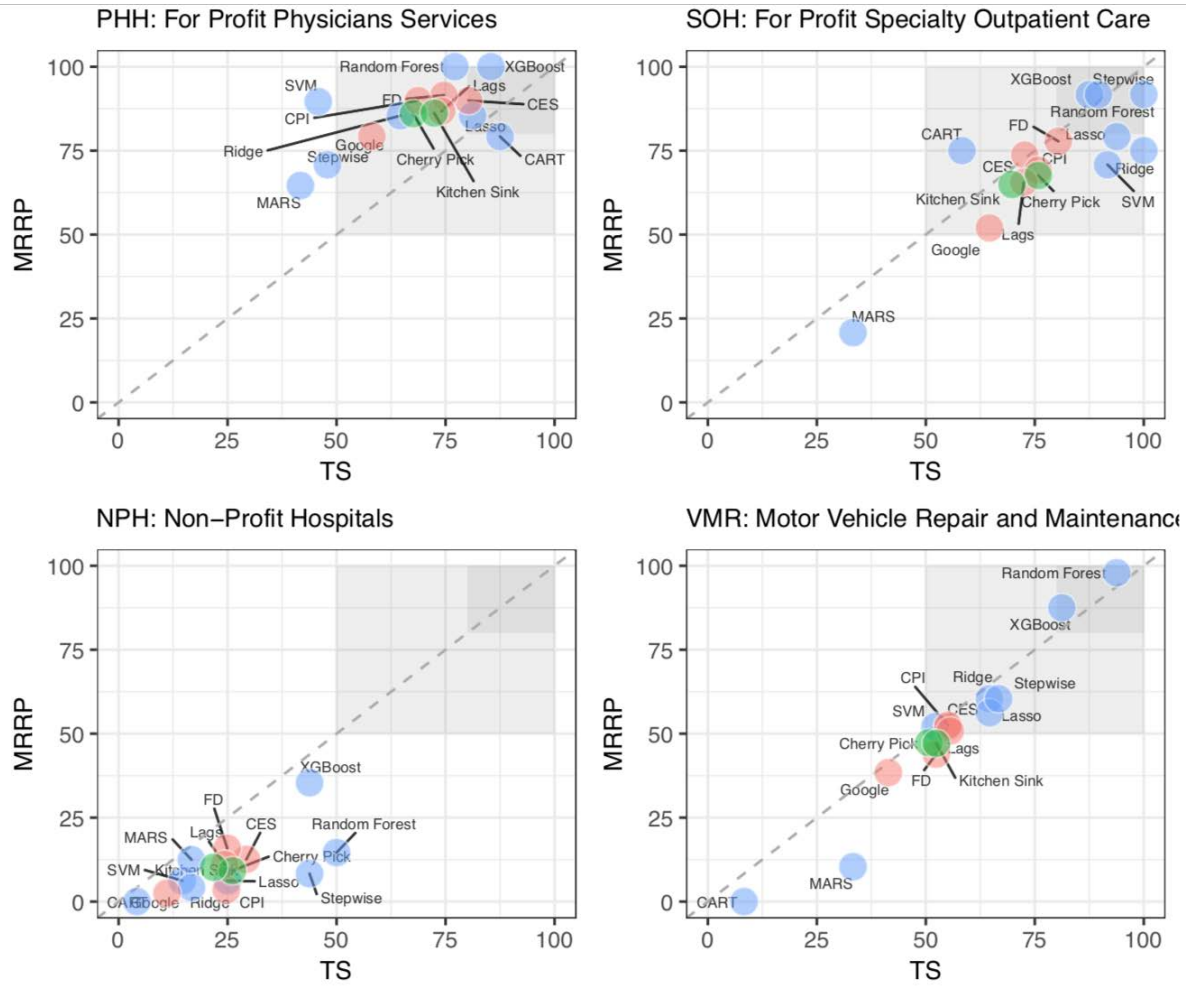

FIGURE 6: Comparison of MRRP and TS for four PCE Services components by modeling component.

While revision reductions for individual sub-components can be easily evaluated, the ability to achieve revision reductions among top line measures (e.g., overall PCE and PCE services) is more challenging due to *offsetting*. Given two sub-components that are added together to estimate a more aggregate PCE component, one may have upward revision reductions and the other may have downward revision reductions. When added together, the revision reductions may partially offset one another, muting the magnitude of improvement to top-line measures. We estimate net revision reductions for the most versatile modeling strategy, Random Forest. Only PCE components where $\mu_k \geq 80$ are included in the calculations. The following impact analysis reflects the contributions of 20 PCE components, each of which has an ensemble of 50 models reflecting a broad range of assumptions.

TABLE 4: Estimated Revision Reductions in historical test sample when only applying Random Forest to components that have at least an 80-percent chance of improvement. Percent correct direction indicates if the ensemble mean's growth accurately anticipates the actual series' direction of growth (positive or negative). GO NP denotes Gross Output for Nonprofit.

	Percent				Levels ($Mil)		Direction	
Component	10th	Mean	Median	90th	Mean	Median	ML	Current
PCE	5.59	12.17	13.11	18.33	2054.75	2213.61	100	100
..PCE Services	0.2	10.3	11.78	19.72	1552.69	1775.76	100	100
....Health Care	2.23	11.27	12.64	18.99	1442.62	1618	100	100
....Transportation	2.91	25.57	26.7	43.86	1100.38	1149.29	75	67
....Recreation	4.28	8.47	8.28	12.75	349.73	341.88	92	83
....Education	1.74	3.25	3.11	5.16	17.6	16.83	100	100
....Professional and Other	1.38	4.2	3.72	7.02	77.84	68.89	75	67
....Personal Care and Clothing	21.8	27.37	28.24	31.03	513.85	530.18	92	83
....Social Services and Religious	10.29	14.21	14.7	17.82	155.06	160.42	83	83
....Household Maintenance	-24.25	10.94	16.71	34.38	45.49	69.49	100	92
....GO NP Social Services	0.07	0.43	0.47	0.74	9.37	10.2	33	33
....GO NP Prof Advocacy	26.24	36.99	41.03	47.8	235.12	260.79	100	100

Starting from the topline, we find that overall PCE revisions would have been reduced on average 12% with an ensemble median of 13%, translating to approximately $2 billion in net revision reductions. The ensemble's upper shoulder suggests that some of the better performing models within the ensemble could achieve as much as a 21.3% revision reduction ($3.6 billion); However, individual model selection would only be possible when statistical power is sufficiently large in the validation sample. Within PCE services, several components attain even larger revision reductions, with health care and transportation services leading (in absolute terms) with average 11.3% and 25.6% improvements, respectively.

While the shape of growth is matched by the models, the ability to correctly anticipate the direction of growth—whether it is positive or negative—has apparent effects on the levels. Anticipating a deceleration when growth is accelerating reduces estimate quality and magnifies revisions. We evaluate the performance of the ensemble average relative to current performance using the validation period. As would be expected, current BEA methods are able to anticipate direction of growth in most periods. While we do not find improvements among higher aggregate components of PCE, the prediction ensemble marginally improves subcomponents with one quarter improvement.

6. Conclusion

In this paper, we illustrate a suitable use of machine learning techniques for macroeconomic estimation. We focus on improving data quality by reducing revisions to PCE service components. Our proposed approach provides predictions of advanced estimates using machine learning techniques and identifies PCE components for which prediction-based improvements are likely.

In general, non-parametric techniques such as Random Forest and Gradient Boosting offer marked gains in prediction accuracy and are well-adapted to conducting implicit variable selection at scale. Furthermore, these techniques can accommodate the typical ill-posed problem, sifting through quantities of data without significant loss in prediction quality.

One key evaluation point for macroeconomic prediction is its ability to detect economic downturns. As the current incarnation of the QSS does not span the 2008-2009 recession, it is not possible to test for downturns although it may be applied to anticipating other indicator series. Prior studies such Chauvet and Potter (2013) found that commonly macroeconomic techniques for forecasting output, such as autoregressive models of a variety of builds, generally perform well during expansions, but poorly in recessions. While we are unable to test the machine learning models in this context, we can foresee the likely performance of these non-parametric techniques during recessionary periods by taking note of the core assumptions. Like linear models, non-parametric algorithms are designed for stationary processes. Unlike linear models, the predicted value $\widehat{y_t}$ are bounded by the range of *y* in the training sample. In small samples that do not span recessions, we can assume that the shape of economic growth can be predicted, but the depth of a contraction will likely be understated. A model switching mechanism such as a Markov Switching Model should be incorporated to provide greater flexibility to use both non-parametric and parametric extrapolators.

There are opportunities to improve the stability of predictions while increasing revision reductions. One extension is to train an additional model to marshal predictions and cut through the noise of less reliable models. Model averaging as in the case of Hansen (2007) can improve predictions subject to a linear constraint. More generally, model stacking techniques offer a more flexible solution in which a supervised machine learning algorithm trains on values of $\hat{y}_i$ from the validation set to produce predictions. In either case, additional training observations would be required for the averaging and stacking model to learn which underlying models are in fact predictive. As the sample size is a constraint, we may adopt the leave-one-out model validation strategy as described in Cornwall et al. (forthcoming) to expand the training sample while meeting Granger Causality criteria.

This study also finds that prediction error will only grow with sampling error, as expected; therefore, industries with large sampling error limit the ability for the current strategy to predict highly variable PCE components. One approach to overcome sampling error is to consider a top-

down hierarchical forecasting model (Hyndman et *al*., 2016), predicting the top-line estimates of PCE, then sharing growth by component by modeling conditional probabilities. A benefit is that each component is logically consistent with parent series and have a decent degree of accuracy among low error series, but sampling error and noise may still pose a challenge. An alternative but more costly solution involves improving the underlying survey's sample design by oversampling strata with large sampling error. We recognize this would incur greater cost relative to a modeling strategy but may be a necessity for estimate quality.

This paper shows that using both traditional and alternative data sources can contribute to improved predictions. However, there are issues outside of the prediction methodology that should also be considered. For instance, while private data sources may lead to better predictions, the cost, quality and availability of these data sources may change for external reasons (e.g., a company failing or a change in management). Users of alternative data sources should be mindful of the long-term availability and stability of these sources. Nevertheless, these concerns will be relevant irrespective of the methods that are applied, and it is worth noting that a benefit of the ML approach is that it reduces the reliance on a single data source.

While the macroeconomic literature incorporating machine learning is in its nascent stages, we show that computationally intensive algorithms do in fact offer measurable improvements for estimates of the PCE Services component of GDP. There is considerable scope for future research to apply these techniques to other components of GDP, as well as other national statistics.

7. References

Aladangady, Aditya, Shifrah Aron-Dine, Wendy Dunn, Laura Feiveson, Paul Lengermann, and Claudia Sahm. Forthcoming. *From Transactions Data to Economic Statistics: Constructing Real-Time, High-Frequency, Geographic Measures of Consumer Spending.* National Bureau of Economic Research.

Bańbura, Marta, Domenico Giannone, and Lucrezia Reichlin. 2011. *Nowcasting*. Edited by M.P. Clements and D.F. Hendry. Oxford: Oxford University Press.

Bańbura, Marta, Domenico Giannone and Michele Modugno, and Lucrezia Reichlin. 2013. "Now-Casting and the Real-Time Data Flow." Edited by G. Elliott and A. Timmermann. *Handbook of Economic Forecasting, Volume 2a*. Amsterdam: Elsevier.

Biau, Olivier, and Angela D'Elia. 2013. "Euro Area GDP Forecasting Using Large Survey Datasets: A Random Forest Approach." *Working Paper*.

Bok, Brandyn, Daniele Caratelli, Domenico Giannone, Argia Sbordone, and Andrea Tambalotti. 2017. “Macroeconomic Nowcasting and Forecasting with Big Data.” *Federal Reserve Bank of New York Staff Reports*. https://www.newyorkfed.org/medialibrary/media/research/staff_reports/sr830.pdf.

Breiman, Leo. 2001. “Random Forests.” *Machine Learning* 45 (1): 5, 32.

Breiman, Leo, Jerome H. Friedman, R.A. Olshen, and Charles J. Stone. 1984. *Classification and Regression Trees*. Boca Raton: Chapman & Hall/CRC.

Cavallo, Alberto, and Roberto Rigobon. 2016. “The Billion Prices Project: Using Online Prices for Measurement and Research.” *Journal of Economic Perspectives* 30 (2): 151, 178.

Cavallo, Alberto, W. Erwin Diewert, Robert C. Feenstra, Robert Inklaar, and Marcel P. Timmer. 2018. “Using Online Prices for Measuring Real Consumption Across Countries.” *AEA Papers and Proceedings* 108: 483, 487.

Chauvet, Marcelle, and Simon Potter. 2013. *Forecasting Output*. Vol. 2. Amsterdam: Elsevier.

Chen, Baoline, and Kyle Hood. 2018. “Nowcasting Private Consumption of Services in the U.S. National Accounts with a Bridging with Factor Framework.” *5th Annual Conference of the Society for Economic Measurement (Xiamen, China).*

Chen, Tianqi, and Carlos Guestrin. 2016. “XGBoost: A Scalable Tree Boosting System,” KDD ’16,. New York, NY, USA: ACM, 785, 794. http://doi.acm.org/10.1145/2939672.2939785.

Choi, Hyunyoung, and Hal Varian. 2012. “Predicting the Present with Google Trends.” *The Economic Record* 88 (s1): 2, 9.

Copas, J.B. 1983. “Regression, Prediction and Shrinkage.” *Journal of the Royal Statistical Society, Series B* 45: 311–54.

Cornwall, Gary, Jeffrey A. Mills, Beau A. Sauley, and Huibin Weng. forthcoming. “Predictive Testing for Granger Causality via Posterior Simulation and Cross Validation.” *Advances in Econometrics*.

Diebold, Francis X., and Roberto S. Mariano. 1995. “Comparing Predictive Accuracy.” *Journal of Business & Economic Statistics* 13 (3): 253, 263.

Drucker, Harris, Christopher J.C. Burges, Linda Kaufman, Alexander J. Smola, and Vladimir N. Vapnik. 1996. “Support Vector Regression Machines.” *Advances in Neural Information Processing Systems*.

Efroymson, M. A. 1960. “Multiple Regression Analysis.” *Mathematical Methods for Digital Computers*. Edited by A. Ralston and H. S. Wilf. New York: John Wiley.

Einav, Liran, and Jonathan Levin. 2013. “The Data Revolution and Economic Analysis.” *Innovation Policy and the Economy* 14. National Bureau of Economic Research: 1, 24.

Friedman, Jerome H. 1991. “Multivariate Adaptive Regression Splines.” *The Annals of Statistics* 19 (1): 1, 141.

Friedman, Jerome H. 2001. “Greedy function approximation: A gradient boosting machine.” *The Annals of Statistics* 29 (5), 1189-1232.

Ghysels, Eric, Pedro Santa-Clara, and Rossen Valkanov. 2004. “The MIDAS Touch: Mixed Data Sampling Regression Models.” *CIRANO Working Papers*. https://EconPapers.repec.org/RePEc:cir:cirwor:2004s-20.

Giannone, Domenico, Lucrezia Reichlin, and David Small. 2008. “Nowcasting: The Real-Time Informational Content of Macroeconomic Data.” *Journal of Monetary Economics* 55 (4): 665, 676.

Fragoso, Tiago M., Wesley Bertoli and Francisco Louzada. 2018. “Bayesian Model Averaging: A Systematic Review and Conceptual Classification,” International Statistical Review, 86 (1), 1–28.

Hansen, Bruce E. 2007. “Least Squares Model Averaging.” *Econometrica* 75 (4): 1175, 1189.

Hansen, Bruce E. and Jeffrey S. Racine. 2012. “Jackknife Model Averaging,” Journal of Econometrics, 167, 38–46.

Hoerl, Arthur E., and Robert W. Kennard. 1970. “Ridge Regression: Biased Estimation for Nonorthogonal Problems.” *Technometrics* 32 (1): 80, 86.

Hyndman, Rob J., Roman A. Ahmed, George Athanasopoulos, and Han L. Shang. 2016. “Optimal Combination Forecasts for Hierarchical Time Series.” *Computational Statistics & Data Analysis* 55 (9): 2579, 2589.

Kadane, Joseph B. and Nicole A. Lazar. 2004. “Methods and Criteria for Model Selection,” Journal of the American Statistical Association, 99(465), 279–290.

Kapetanios, George, and Fotis Papailias. 2018. “Big Data & Macroeconomic Nowcasting: Methodological Review.” *Discussion Papers from Economic Statistics Centre of Excellence, ESCoE DP-2018-12*.

Milborrow, Stephen. 2018. *Earth: Multivariate Adaptive Regression Splines*. https://CRAN.R-project.org/package=earth.

Mullainathan, Sendhil, and Jann Spiess. 2017. “Machine Learning: An Applied Econometric Approach.” *Journal of Economic Perspectives* 31 (2): 87–106. doi:10.1257/jep.31.2.87.

Nyman, Rickard, and Paul Ormerod. 2017. “Predicting Economic Recessions Using Machine Learning Algorithms.” https://arxiv.org/abs/1701.01428v1.

Rajkumar, Veg. 2017. “Predicting Surprises to GDP: A Comparison of Econometric and Machine Learning Techniques.” *Thesis, MIT Sloan School of Management*.

Rossi, Barbara. 2013. *Advances in Forecasting Under Instability*. Edited by G. Elliott and A. Timmermann. Amsterdam: Elsevier.

Stock, James H., and Mark W. Watson. 2006. *Forecasting with Many Predictors*. Vol. 1. Amsterdam: Elsevier.

Tibshirani, Robert. 1996. “Regression Shrinkage and Selection via the Lasso.” *Journal of the Royal Statistical Society, Series B* 58 (1): 267, 288.

Varian, Hal. 2014. “Big Data: New Tricks for Econometrics.” *Journal of Economic Perspectives* 28 (2): 3, 28.