Predicting the Highway Costs Index with Machine Learning

Samir Huseynov [†] Luis A. Ribera [‡] Marco A. Palma [††]

Abstract

Big data is increasingly attracting the attention of economists. New machine learning techniques can help to analyze unconventional data structures with a large number of variables relative to the number of observations. This new venue of research offers unique opportunities for analyzing previously untouched fields due to data limitations. Our study introduces a machine learning approach for modeling and forecasting highway construction cost changes. Our Lasso and Random Forest models have high predictive power, and it suggests that the application of machine learning techniques can improve the estimation of actual project costs for optimal allocation of public funds.

Keywords: forecasting, highway cost index, machine-learning.

JEL Codes: C55, D24.

[†]Corresponding author, Research Assistant, Department of Agricultural Economics, Texas A&M University. 2124 TAMU, College Station, Texas, 77843, samirhuseyn@tamu.edu, (979)-777-7173

[‡]Professor, Department of Agricultural Economics, Texas A&M University. 2124 TAMU, College Station, Texas, 77843

[††]Associate Professor, Department of Agricultural Economics, Texas A&M University. 2124 TAMU, College Station, Texas, 77843

1 Introduction

Advancement in technology not only makes our lives more comfortable and also enables to store more and more data about many aspects of daily economic transactions (George et al. 2014, Chen & Zhang 2014). Google, Facebook, and Twitter record every click or a browsing pattern users generate on these platforms. You may be surprised to realize that the pair of shoes you were browsing on eBay at lunchtime, a couple of days ago, has been following you ever since, even when you read the daily news. You may get extra surprised to see that, with the help of various machine learning tools, you have been targeted with the ads of other related products, such as a coat, which matches the pair of shoes well and was frequently shopped by customers similar to you.

Machines can learn with and without the human supervision. It has been shown that Unsupervised Machine Learning (UML) methods can be useful especially when the outcome of interest is unobserved (Mohammed et al. 2016). For instance, a database may contain various information about customers, including zip code, education, gender and so on, but a researcher is interested in clustering the customers into low and high spending groups. Thus, in this case, input variables are available, but the outcome of interest is missing from the dataset. UML algorithms can solve this problem by clustering the customers into different groups based on observed variables. The eventual goal is to find clusters that overlap with spending habits.[1]

However, researchers are still skeptical about the black-box nature of these techniques and prefer using Supervised Machine Learning (SML) methods, which are

[1] See (Hastie et al. 2015) for an in-depth discussion of this topic.

the focus of this article (Athey 2017). Recent years have witnessed a considerable surge in the application of SML to various economic topics. For example, banks in the U.S. use credit scores of customers in making decisions on loan requests. But in other countries, especially in the developing world, financial institutions do not have this tool in their credit appraisal toolkit. With the help of Random Forest SML Bjorkegren & Grissen (2015) develop a model which can successfully predict default probabilities by using pre-paid mobile phone usage data. The authors match customer data of a microfinance institution with the data of a mobile operator in a Caribbean country. This approach enabled the authors to observe customers' SMS, data usage and call activities combined with their financial history. Bjorkegren & Grissen (2015) show that their model can be a good alternative to the credit score system and can reduce defaults by 41%. Furthermore, in developed countries, financial institutions offer almost the entire portfolio of their services via online platforms. Customers can integrate their online debit and credit accounts, which helps banks to record every single individual financial activity. This kind of financial information is increasingly used in the assessment of credit applications and also in the evaluation of potential bankruptcy of business loans. More detailed data also enables to improve predictions on future performance of financial institutions (Chen et al. 2016, Gogas et al. 2017).

Not only private sector and also government institutions have started recording an immense amount of data. "Smart Cities" create a massive opportunity for researchers to analyze a broad range of economic topics from job creation to environmental policy (Kitchin 2014). For instance, the partnership of IBM and the city government

of Rio de Janeiro, Brazil, enabled to combine data streams from thirty agencies, ranging from traffic and public transport to employee information, into one data center (Kitchin 2014). Glaeser et al. (2016) use Google Street View (GSV) data to predict income. Conventionally governments use property taxes as a tool to conduct property value appraisal. But Glaeser et al. (2016) show that by applying SML methods to big urban data, including street view images, they can better predict neighborhood income level and property values. The authors acquired GSV image data of New York City between 2007 and 2014 and linked that data to income data from American Community Survey. After training the v-support vector regression (v-SVR) model in the training set, they made predictions for the hold-out-sample. Glaeser et al. (2016) report that the R^2 of their predictions was above 80%, which means a very successful prediction accuracy. Thus, the fact that industry has already embraced big data and the relevant analytical tools to feed new insights into its daily operations serves as an additional motivation for academia to model this new phenomenon.

Data, figuratively speaking, is the "input material" of economic research and merely scaling up the input does not necessarily mean that you can increase the output as well. Similar to production, without advancing the techniques and procedures of economic research, it is tough to secure a higher return.

The availability of "Big Data" creates many challenges for economists. From one perspective, big data provides many independent variables that the current economic theory lacks in modeling and connecting those variables to traditional economic topics. Nevertheless, unconventional measures are increasingly becoming a part of

economic models. For instance, the availability of large satellite image data enabled economists to model and explore the relationship between city lights and economic activity (Michalopoulos & Papaioannou 2013, Doll et al. 2006). However, many other similar variables are still "awaiting" their turns to be a part of economic studies.

From another perspective, big data also shifts the focus of economic research. For example, conventional labor economics is interested in the return to schooling, i.e., estimating β or the relationship between one more year education and income. But the availability of huge datasets also helps us to answer other questions, such as deciding which teacher to hire, that is more about predicting $\widehat{y}$ rather than estimating $\widehat{\beta}$ (Mullainathan & Spiess 2017).

Predicting $\widehat{y}$ became a part of economic research long ago. Forecasting of stock returns is the central focus of the financial time series econometrics. Even some widely applied estimation techniques implicitly use predicting, such as the first stage of the instrumental variable method (Mullainathan & Spiess 2017).

The value of machine learning becomes apparent when there is a critical $\widehat{y}$ task, with high dimensional data but without clear guidance of economic theory (Mullainathan & Spiess 2017). For instance, the conventional applied economic research mostly employs datasets where the number of variables is less than the number of observations, and the relationship between input variables and the outcome measure is evident from economic models (Einav & Levin 2014). However, an increasing number of situations in which datasets contain a large number of variables, but a limited number of observations creates new challenges, mainly because of the unusual shape of the datasets ($K \gg N$), lack of initial hypotheses, high correlation among vari-

ables, and the sparsity assumption which implies that there exists a much smaller set of variables that can be used in predicting the dependent variable (Belloni et al. 2012, 2011). In this case, the central question is how many and which covariates are to be selected to predict the outcome of interest. If a subset of variables provides better predictions compared to the complete set variables, then data-driven machine learning approaches can help to identify those variables (Einav & Levin 2014, Mullainathan & Spiess 2017, Belloni et al. 2012).

In this study, we apply several machine learning methods to tackle with a $\widehat{y}$ task that has substantial economic importance and offers a suitable framework to test how this new approach can be helpful. We forecast the Highway Cost Index (HCI) for Texas by using the Least Absolute Selection and Shrinkage Operator (LASSO), Ridge, Elastic Net and Random Forest machine learning methods that recently attracted the attention of economists (Mullainathan & Spiess 2017).

The HCI is a monthly index and is comprised of prices of primary construction materials in highway construction (Wang & Ashuri 2017). The HCI helps to measure the cost of a certain amount of materials to be used in construction compared to prices in 2012 (Wang & Ashuri 2017).[2] Thus the HCI has high importance in public investment budgeting and cost estimation, but it is mostly unexplored by economists, mainly because of the high dimensionality of the data and the lack of relevant techniques to reduce the number of predictors.

[2]ftp://ftp.dot.state.tx.us/pub/txdot-info/cst/hci-binder.pdf

2 The Background of HCI

The U.S. highway construction represents the largest share of civil public spending. In January of 2017, public highway construction spending was 86.7 billion dollars or 32% of the total public construction spending.[3] Nearly half of all the currently active large transportation projects in the United States have exceeded their initial budget (Shane et al. 2009). Thus highway construction cost estimates need to reflect the actual project costs as much as possible for optimal allocation of public funds (Harper et al. 2013).

This study employs a machine learning approach for predicting highway construction costs. Our study is motivated by the fact that accurately forecasting highway construction cost trends will help States (and particularly Texas in our case) to monitor increases in highway construction costs and efficiently match budget appropriations with actual expenditures.[4]

Figure 1 shows that, although the HCI has the upward trend across years, it is prone to short-term fluctuations due to economic conditions (Shahandashti & Ashuri 2015). Thus we model and forecast changes in the HCI during a 12-month-period. Furthermore, our study shows that machine learning methods can be a promising venue for applications with a limited number of observations and a large number of potential regressors.

[*Figure* 1 *about here*]

[3] https://www.census.gov/construction/c30/pdf/release.pdf
[4] *https://www.fhwa.dot.gov/policyinformation/nhcci/desc.cfm*

The review of the relevant literature shows that the primary method to analyze small data sets with a large number of variables uses causal and univariate time series models with a small set of variables chosen by researchers (Xu & Moon 2011). However, this approach generates skepticism regarding the selection of input variables. We show that using appropriate Machine Learning methods can introduce a more robust mechanism to select independent variables and it can generate accurate forecasts. An empirical application of our model and forecasting of highway cost estimates illustrates the usefulness of the approach.

3 Studies on HCI

Accurate forecasting of the HCI is essential in avoiding underestimating highway construction costs (Wang & Ashuri 2017). This issue is critical especially for public projects where cost adjustments and the request of additional funds may face bureaucratic barriers and take a long time. Cost indexes facilitate cost estimation, bid preparation, investment planning and also monitoring cost escalations in construction projects (Wang & Ashuri 2017, Zhang et al. 2017).

Shahandashti & Ashuri (2015) develop a forecast model for the National Highway Construction Cost Index (NHCCI) and consider various macroeconomic indicators, spanning from oil price to consumer price index for their model. By using Augmented Dickey-Fuller and Granger Causality test, they determine that crude oil price and average hourly earnings are main indicators in predicting the NHCCI. They employ the Vector Error Correction model to forecast the cost index and report 2.07% the

Mean Absolute Percentage Error (MAPE) for a 12-month-period forecast. But their study employs a one-step-ahead forecast approach, meaning the model always use input data up to $t-1$ period to forecast the changes in the NHCCI for t period.

Joukar & Nahmens (2015) also employ a time series framework and forecast the CCI with its historical values. They report around 20% MAPE for out-of-sample predictions.

Zhang et al. (2017) forecast the Construction Cost Index (CCI) with the visibility graph network approach. The authors use historical values of the CCI to predict its future value both with one-step-ahead and multi-step-ahead approaches. In the 12-month-period multi-step-ahead forecast, they report more than 50% MAPE in their predictions.

Wang & Ashuri (2017) use k nearest neighbor (k-NN) and perfect random tree ensembles (PERT) machine-learning methods to predict the CCI. They employ multi-step-ahead forecast approach and use historical values of the CPI and the producer price index (PPI). For the following 12 months predictions, the authors report 18% and 19% MAPE for the Pert and k-NN methods respectively.

4 Data Collection and Pre-Process

The data was obtained from the Texas A&M Transportation Institute. It consists of 20 individual price items that define the HCI for Texas. Table 1 presents the list of variables.

After dropping observations where one or more variables have missing values,

our monthly dataset spans from January 2002 until March 2016 for a total of 145 observations. Our dependent variable is the HCI.

[*Table* 1 *about here*]

To prepare the dataset for machine learning estimations, we transformed input variables and the outcome measure into relative changes (Wang & Ashuri 2017). The relative change transformation helps to get rid of the potential non-stationarity since our variables have time series nature. We follow Wang & Ashuri (2017) and define our relative change transformations as the following:

$$x_{pt} = \frac{X_{pt} - X_{pt-1}}{X_{pt}},\ y_t = \frac{Y_t - Y_{t-1}}{Y_t} \tag{1}$$

where p denotes the name of transformed input variable X, and $t\epsilon[1, ..., T]$ is the month of the observation. Y indicates the outcome variable, i.e., the HCI.

After the relative change transformation, we generated lags of input variables up to six months to avoid having any looking-ahead bias in our estimations (Wang & Ashuri 2017). Thus, the total number of input variables is 140 in our dataset.

We set aside observations, spanning from September 2002 until March 2015, for our training set (133 observations). We used observations from April 2015 to March 2016 as our test set.

In line with (Wang & Ashuri 2017), after predicting $\widehat{y}_{t+1}$, the prediction of the HCI for April 2015 was calculated as the following:

$$\widehat{Y}_{t+1} = (1 + \widehat{y}_{t+1})Y_t \tag{2}$$

where Y_t represents the actual HCI of March 2015. For the next months we employed the similar strategy:

$$\widehat{Y}_{t+s+1} = (1 + \widehat{y}_{t+s+1})\widehat{Y}_{t+s} \tag{3}$$

where $s\epsilon(1, ..., 11)$.

5 Methodology

The Least Absolute Shrinkage and Selection Operator (LASSO), Ridge and Elastic Net Estimations

We follow Friedman et al. (2010) in setting up our model for Lasso, Ridge and Elastic Net estimations. Our response variable is $Y\epsilon R$ and the matrix of input variables is $X\epsilon R^p$. Furthermore, we approximate our outcome variable as the following:

$$E(Y|X = x) = \beta_0 + x^\intercal\beta \tag{4}$$

We normalize our variables and use $P_\alpha(\beta) = (1-\alpha)\frac{1}{2}\|\beta\|^2_{\ell_2} + \alpha\|\beta\|_{\ell_1}$ penalty to obtain a sparse solution in the presence of a large number of covariates (Belloni et al. 2012, 2011, Tibshirani 1996, Friedman et al. 2010). Generally, we estimates the following model (Hastie et al. 2015, Friedman et al. 2010):

$$\underset{(\beta_0,\beta)\epsilon R^{p+1}}{min}\left[\frac{1}{T}\sum_{t=1}^{T}\left(y_t-\beta_0-x_t^{\top}\beta\right)^2-\lambda P_{\alpha}(\beta)\right] \tag{5}$$

where for each outcome variable there are p independent variables and $p > T$ (T is the total number of observed time periods). Note that in our case $p = 140$ and $T = 133$.

This problem can be solved with *coordinate descent*. Let's denote the objective function in equation 5 with $R(\beta_0, \beta)$. Suppose we estimate $\widetilde{\beta}_0$ and $\widetilde{\beta}_\ell$ for $\ell \neq j$ and the goal is the partially optimize with respect to β_j. The gradient at $\beta_j = \widetilde{\beta}_j$ exists if $\widetilde{\beta}_j \neq 0$ and $\widetilde{\beta}_j > 0$:

$$\frac{\partial R}{\partial \beta_j}|_{\beta=\widetilde{\beta}} = -\frac{1}{T}\sum_{t=1}^{T} x_{tj}(y_t-\widetilde{\beta}_0-x_t^{\top}\widetilde{\beta})+\lambda(1-\alpha)\beta_j+\lambda\alpha \tag{6}$$

When, $\widetilde{\beta}_j \neq 0$ and $\widetilde{\beta}_j < 0$ we have a similar expression. Then the solution progressed through the coordinate-wise updates and has the following form:

$$\widetilde{\beta}_j \leftarrow \frac{S\left(\frac{1}{T}\sum_{t=1}^{T} x_{tj}(y_t-\tilde{y}_i^{(j)}),\lambda\alpha\right)}{1+\lambda(1-\alpha)} \tag{7}$$

where, $\tilde{y}_i^{(j)} = \widetilde{\beta}_0 + \sum_{\ell\neq j} x_{t\ell}\widetilde{\beta}_\ell$ is the fitted value when x_{tj} is excluded and $y_t - \tilde{y}_i^{(j)}$ is the partial residual for β_j. S is the soft-thresholding operator.[5]

Depending on the value of α we can estimate three different models. When $\alpha = 1$ we estimate Lasso. We obtain Ridge and Elastic Net regressions for $\alpha = 0$ and $\alpha = 0.5$ respectively.[6]

[5] Please, refer to Friedman et al. (2007) for details.

[6] Friedman et al. (2010) present a detailed discussion about the steps of estimations of models.

If there are m highly correlated predictors, Lasso picks only one and ignores the rest. But contrary to Lasso, Ridge method keeps all highly correlated variables, nevertheless shrinks their coefficients towards each other (Friedman et al. 2010). So, in Ridge estimation, every predictor will get $1/m$th size of its coefficient that it would get in the case of fitting alone. Elastic Net estimation is a compromise between Ridge and Lasso, as it discards only extremely correlated predictors and then shrinks coefficient like Ridge.

Lasso behaves well when there are many highly correlated predictors (Friedman et al. 2010). We assume that most of our predictors are highly correlated. For instance, it is reasonable to assume that the prices of cement and asphalt move together. We also follow Friedman et al. (2010) and check whether the ridge or the elastic net regressions can outperform the Lasso concerning predictions of the HCI. Friedman et al. (2010) particularly focused on the elastic net estimation and showed that in the case of high multicollinearity the elastic net may perform better regarding the selection of the most important independent variables.

We start our analysis in the training set. In the training set, we estimate Lasso, Ridge and Elastic Net models to find the optimal λ^* that minimizes the mean squared errors (MSE) in predictions of models. We randomly divide the training data into ten groups of equal size and define nine groups as the "pre-validation training" set and the $10th$ as a "pre-validation test" set.

We estimate a model with a specific value of λ and mark the MSEs in predictions of a fitted model for each λ value. We repeat this procedure 100 times, and λ^* is found when the MSE of a model is minimized. Figure 2 showcases one example from

our estimations.

[*Figure 2 about here*]

Random Forest

Regression trees have been developed to bridge linear models with non-parametric estimations (Faraway 2016, Breiman 2001, Liaw et al. 2002). This approach does not presume any specific model before starting analyses. Instead, the focus is to develop a model by learning patterns in the dataset via certain algorithms. While growing a tree, at each node $\sqrt{p}$ number of input variables are randomly drawn with replacement. A point, along with the range of each randomly drawn variable, is chosen to do a split to minimize the Residual Sum of Squares (RSS) of the outcome variable. For instance, at each partition of an input variable, we get two parts, and the RSS of the outcome variable is calculated as the following (Faraway 2016):

$$RSS(Partition) = RSS(part_1) + RSS(part_2) \tag{8}$$

The node is assigned to the input variable which helps to get the minimum RSS for the outcome variable. We continue to grow nodes and consequently the tree until we cannot improve (or minimize) the RSS.

Particularly, we use bootstrap aggregating (or bagging) in Random Forest estimation. For $b = 1, ..., B$: 1) We draw a sample with replacement from (X, Y). 2) We fit a regression tree to (X_b, Y_b). 3) We predict the outcome variable with the tree and compare the prediction with the true value from hold-out-sample in the training

set (Faraway 2016). The eventual goal is to minimize the MSE of the prediction.

With this pace, Random Forest grows hundreds of trees. Each tree of the Random Forest can be thought as a different model, so the Random Forest employs not one but hundred models in its final decision. Figure 3 depicts how increasing the number of trees in the forest affects the prediction error rates in the training set.

[*Figure 3 about here*]

We grow our forest with the training set and then we predict the HCI in the test set. In the test set, we ask each tree of the forest to predict the HCI. Since each tree is a different model, we get different predictions, and the final forecast is the average of forecasts made by the forest (Faraway 2016).

For the illustration purposes, we present a simple tree in Figure 4. For instance, this particular tree starts with the fifth lag of the relative change of the price of Class C Concrete. If the relative change is more than 0.02 the analysis continues along the right branch and then checks the first lag of the relative change in the price of Class A Concrete and makes a prediction about the change in the HCI for the next period.

[*Figure 4 about here*]

After estimating models and predicting the next 12 months values of HCI, we calculate the MAPE of predictions. The MAPE is calculated as the following:

$$MAPE = |(\widehat{Y}_t - Y_t)|/Y_t| \tag{9}$$

6 Prediction Results and Conclusion

Figure 5 presents predictions of the models. It is evident that the models predict the HCI very well until the eighth month. We observe a slight deviation of the predictions from the actual values between eighth and eleventh months. But our predictions revert to the actual HCI in the 12th month.

[*Figure* 5 *about here*]

Table 2 reports the MAPE of predictions. It can be observed that the MAPE doubles in the following two months after the seventh month. But it starts to decrease after the tenth month. On average, our models have 12% MAPE for the next 12-month multi-step-ahead forecast.

[*Table* 2 *about here*]

Our results prove that employing Machine-learning methods can yield significant improvements in cost index forecasts compared to time series models. The MAPE of our findings is very close to the outcomes of Wang & Ashuri (2017), it further hints that project planners and engineers can save a considerable amount of costs by employing these new techniques.

The good prediction performance of the employed SML methods shows that these techniques can be good alternatives to time series models, especially in the multi-step-ahead forecasting (Wang & Ashuri 2017). Furthermore, it also suggests that SML methods can be successfully employed in predicting other cost indexes, such as

the Building Construction Index (BIC) (Wang & Ashuri 2017).

Our results show that by applying machine learning techniques and by scrutinizing hard-to-analyze data structures previously, economists can achieve better predictions, more satisfactory answers to old questions and eventually can develop new researches methods (Einav & Levin 2014). Moreover, a more comprehensive multidisciplinary approach, which utilizes recent advancements in computing algorithms and statistics to solve practical economic problems, can boost applied economic research (Athey 2017).

References

Athey, S. (2017), 'Beyond prediction: Using big data for policy problems', *Science* **355**(6324), 483–485.

Belloni, A., Chen, D., Chernozhukov, V. & Hansen, C. (2012), 'Sparse models and methods for optimal instruments with an application to eminent domain', *Econometrica* **80**(6), 2369–2429.

Belloni, A., Chernozhukov, V. & Hansen, C. (2011), 'Inference for high-dimensional sparse econometric models', *arXiv preprint arXiv:1201.0220* .

Bjorkegren, D. & Grissen, D. (2015), 'Behavior revealed in mobile phone usage predicts loan repayment'.

Breiman, L. (2001), 'Random forests', *Machine learning* **45**(1), 5–32.

Chen, C. P. & Zhang, C.-Y. (2014), 'Data-intensive applications, challenges, techniques and technologies: A survey on big data', *Information Sciences* **275**, 314–347.

Chen, N., Ribeiro, B. & Chen, A. (2016), 'Financial credit risk assessment: a recent review', *Artificial Intelligence Review* **45**(1), 1–23.

Doll, C. N., Muller, J.-P. & Morley, J. G. (2006), 'Mapping regional economic activity from night-time light satellite imagery', *Ecological Economics* **57**(1), 75–92.

Einav, L. & Levin, J. (2014), 'Economics in the age of big data', *Science* **346**(6210), 1243089.

Faraway, J. J. (2016), *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*, Vol. 124, CRC press.

Friedman, J., Hastie, T., Höfling, H., Tibshirani, R. et al. (2007), 'Pathwise coordinate optimization', *The Annals of Applied Statistics* **1**(2), 302–332.

Friedman, J., Hastie, T. & Tibshirani, R. (2010), 'Regularization paths for generalized linear models via coordinate descent', *Journal of statistical software* **33**(1), 1.

George, G., Haas, M. R. & Pentland, A. (2014), 'Big data and management', *Academy of management Journal* **57**(2), 321–326.

Glaeser, E. L., Kominers, S. D., Luca, M. & Naik, N. (2016), 'Big data and big cities: The promises and limitations of improved measures of urban life', *Economic Inquiry* .

Gogas, P., Papadimitriou, T. & Agrapetidou, A. (2017), 'Forecasting bank failures and stress testing: A machine learning approach'.

Harper, C. M., Molenaar, K. R., Anderson, S. & Schexnayder, C. (2013), 'Synthesis of performance measures for highway cost estimating', *Journal of Management in Engineering* **30**(3), 04014005.

Hastie, T., Tibshirani, R. & Wainwright, M. (2015), *Statistical learning with sparsity*, CRC press.

Joukar, A. & Nahmens, I. (2015), 'Volatility forecast of construction cost index using general autoregressive conditional heteroskedastic method', *Journal of Construction Engineering and Management* **142**(1), 04015051.

Kitchin, R. (2014), 'The real-time city? big data and smart urbanism', *GeoJournal* **79**(1), 1–14.

Liaw, A., Wiener, M. et al. (2002), 'Classification and regression by randomforest', *R news* **2**(3), 18–22.

Michalopoulos, S. & Papaioannou, E. (2013), 'Pre-colonial ethnic institutions and contemporary african development', *Econometrica* **81**(1), 113–152.

Mohammed, M., Khan, M. B. & Bashier, E. B. M. (2016), *Machine Learning: Algorithms and Applications*, CRC Press.

Mullainathan, S. & Spiess, J. (2017), 'Machine learning: an applied econometric approach', *Journal of Economic Perspectives* **31**(2), 87–106.

Shahandashti, S. & Ashuri, B. (2015), 'Highway construction cost forecasting using vector error correction models', *Journal of Management in Engineering* **32**(2), 04015040.

Shane, J. S., Molenaar, K. R., Anderson, S. & Schexnayder, C. (2009), 'Construction project cost escalation factors', *Journal of Management in Engineering* **25**(4), 221–229.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 267–288.

Wang, J. & Ashuri, B. (2017), 'Predicting enr construction cost index using machine-learning algorithms', *International Journal of Construction Education and Research* **13**(1), 47–63.

Xu, J.-w. & Moon, S. (2011), 'Stochastic forecast of construction cost index using a cointegrated vector autoregression model', *Journal of Management in Engineering* **29**(1), 10–18.

Zhang, R., Ashuri, B., Shyr, Y. & Deng, Y. (2017), 'Forecasting construction cost index based on visibility graph: A network approach', *Physica A: Statistical Mechanics and its Applications* .

Table 1

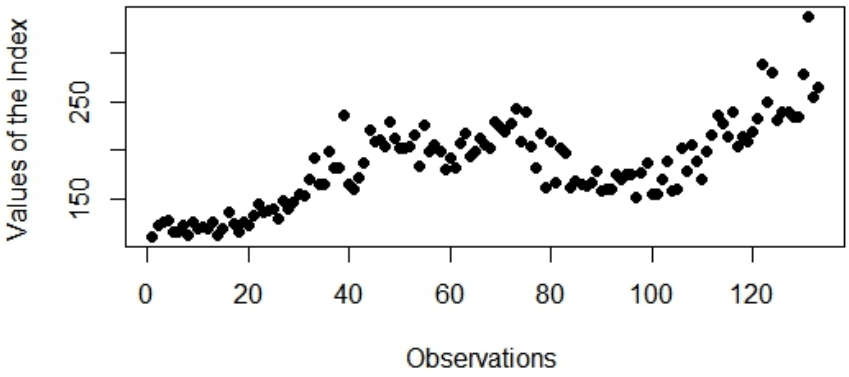
List of Independent Variables

Excavation	Embankment	Lime
Lime Treatment	Cement	Flexible Base
Surface Asphalt	Surface Treatment Aggregate	Class A Concrete
Hot Mix Asphaltic Concrete	Class C Concrete	Class S Concrete
Bridge Rail	Bridge Slab	Regular Beams
Drilled Shafts	Corrugated Metal Pipe	Concrete Box Culverts
	Concrete Riprap	*RetainingWalls*

Table 2

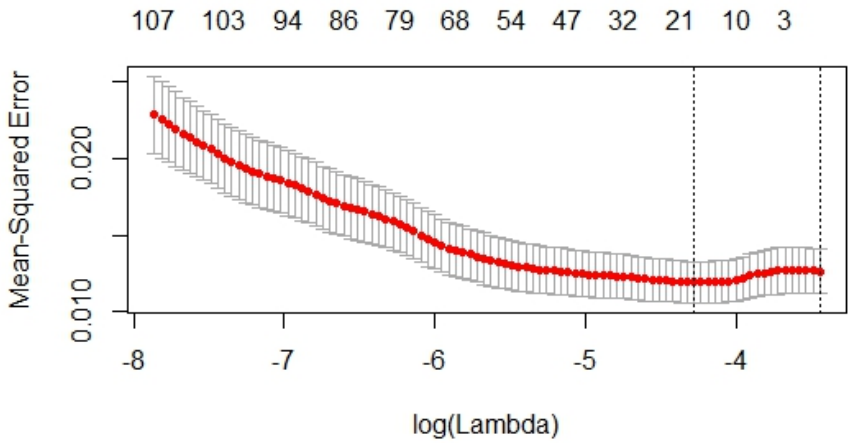
Prediction Performance (MAPE)

	(Rdige)	(Elastic Net)	(Lasso)	(Random Forest)
Lead 1	0.04	0.04	0.05	0.03
Lead 2	0.04	0.06	0.06	0.04
Lead 3	0.07	0.01	0.01	0.09
Lead 4	0.06	0.05	0.05	0.07
Lead 5	0.10	0.11	0.11	0.08
Lead 6	0.01	0.02	0.01	0.02
Lead 7	0.08	0.05	0.08	0.09
Lead 8	0.16	0.17	0.17	0.17
Lead 9	0.32	0.32	0.33	0.33
Lead 10	0.20	0.16	0.16	0.21
Lead 11	0.23	0.16	0.16	0.26
Lead 12	0.11	0.18	0.18	0.09
Average	0.12	0.11	0.12	0.12



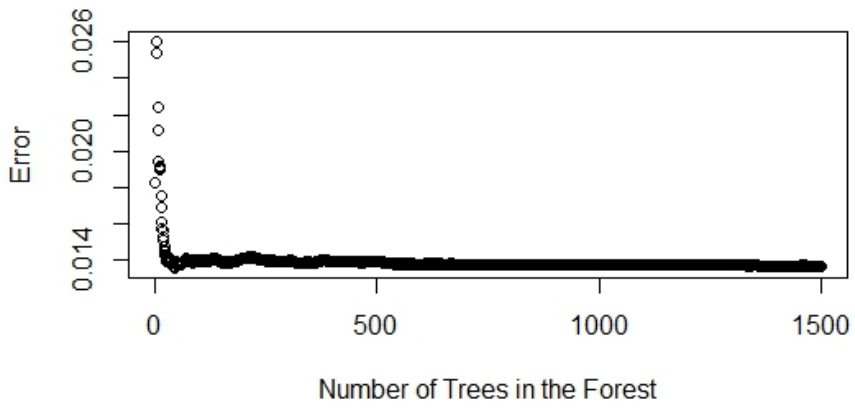
This graph depicts the values of the HCI for Texas from January 2002 until March 2016. It is evident that the HCI closely follows the economic activity as we observe a decline around the 2007-08 financial crises (which starts around the 72nd observation in our dataset). The subsequent economic recovery also increases the HCI.

Figure 1



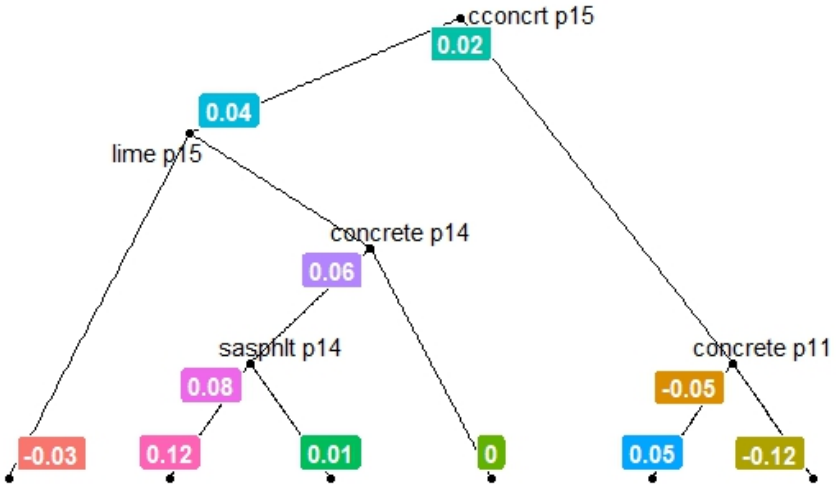
In this particular example, we find λ^* by picking different values for the λ and predicting the NHCCI in Nine Months. The top axis shows the number of chosen input variables by Lasso and the bottom axis shows the corresponding penalty. The Left axis presents the MSE in predictions. The dashed lines indicate the one standard error region of the λ^* that minimizes the MSE. Consider that this region also coincides with a very few number of input variables (which is evident from the top axis).

Figure 2



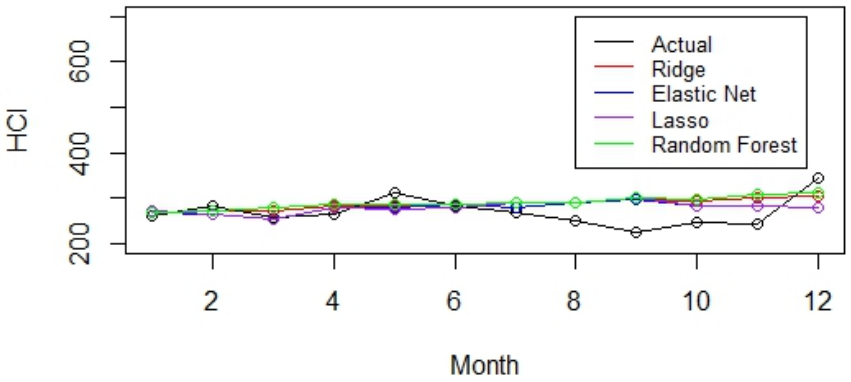
We choose to grow 1500 trees in each forest. This graph shows the performance of Random Forest in developing a model to predict the NHCCI in four months. Prediction error rate doesn't decrease after a certain number of trees.

Figure 3



This tree is constructed to describe how Random Forest classifies observations.

Figure 4



Performance of Models in Predicting the HCI.

Figure 5