Forecasting Liquidity with Indicators of Financial Stability

Eli Metzner[*]

Econ 491/492

Advisor: Bill English

April 2020

Abstract

This project aims to forecast funding liquidity as a proxy for financial system stability using a set of macroeconomic and financial indicators collected from the previous financial stability literature. Accurate forecasts of liquidity not only inform macroprudential and monetary policy, but also identify the economic variables with the clearest relationship to overall financial stability, providing a clear policy-based motivation for this analysis. Measuring liquidity via bank funding spreads, I utilize lasso regression to fit forecast models of these spreads in the United States and United Kingdom economies based on their one-quarter lagged values and high-dimensional vectors of macro and financial indicators, producing out-of-sample errors up to 15 percent lower than baseline autoregressive models of the spreads. In parallel, I perform natural language processing on the minutes of the Federal Open Market Committee to generate a Financial Stability Sentiment score, meant to gauge policymaker opinions of market conditions and provide an alternative measure of overall financial stability. Though modeling sentiment is difficult given the low signal-to-noise ratio of the information content in FOMC minutes, lasso models of the sentiment score produce a more modest 5 percent forecast improvement over a baseline autoregressive model. I then refit these models with simple feed-forward neural networks and obtain much clearer forecast improvement for the U.K. funding spread and the sentiment score. These results contribute to the financial stability literature by highlighting some of the short-run macroeconomic and financial indicators for financial stability, and by demonstrating the utility of machine learning techniques in forecasting within data-rich policy settings.

[*] Many thanks to Bill English, Sigga Benediktsdottir, and Ray Fair for their patience and insight in turning some very abstract ideas into this concrete project. This work would not have been possible without them. Additional thanks to Mom, Dad, and Ben for their unwavering support and encouragement throughout this year and the past 21 alike. All errors are my own. Code and data are available at github.com/metzner28/liquidity.

1 Introduction

In the wake of the Global Financial Crisis of 2007-2008, financial stability analysis has become a central focus of economic research. Given the importance of financial stability to the economy as a whole, central banks have rapidly developed financial stability reporting frameworks meant to promote transparency and accountability from a regulatory standpoint, as well as to increase public awareness and understanding of the buildup of risks and vulnerabilities in financial systems. According to the Federal Reserve's *Financial Stability Report*, financial systems that are "stable" continue to meet credit and liquidity demands of households and firms even when affected by adverse events or "shocks." In contrast, the effects of shocks to unstable financial systems propagate through the economy, leading to adverse effects on output, employment, and economic activity (for context, the drop in house prices after the subprime mortgage bubble burst could be considered a shock that precipitated the 2008 crisis, due to the vulnerabilities in the financial system that had built up prior to the crisis).

While *shocks* are inherently difficult to predict, *vulnerabilities* build up over time, such that monitoring these vulnerabilities can be informative as to the current stability of the financial system [Board of Governors of the Federal Reserve, 2018]. Much recent literature has been devoted to characterizing these vulnerabilities by identifying macroeconomic and financial indicators that are predictive of stress to the financial system in the United States and internationally. While this work is important in its own right for identifying the data most associated with periods of financial system stress, it is difficult to find literature that has gone beyond basic descriptive statistics and simple regression models predicting the current probability of a financial crisis in attempting to develop a true leading indicator that quantifies systemic risk. This is a two-part problem: first, validating the indicators that capture systemic risk; and second, identifying a suitable outcome variable to serve as a measure of the stability of the financial system as a whole. The choice of such an outcome variable is a nontrivial one given the complexity of the financial system and the many measures of credit and liquidity risk that exist in financial markets.

In this work, I aim to utilize techniques from time series econometrics and machine learning to forecast several variables that might function as proxies for overall financial stability. First, I focus on the TED spread – the difference between LIBOR, a standard interbank lending rate, and the risk-free rate on the 3-month Treasury bill – which quantifies the premium on interbank lending, reflecting liquidity and credit risk in the financial system. I next perform natural language processing on the minutes of Federal Open Market Committee meetings to generate a Financial Stability Sentiment score ("FSS"), similar to the measure described in Correa et al. [Correa et al., 2017]. The FSS should represent a more subjective quantification of financial stability, as its value is based on the information content of observers' opinions about market conditions rather than directly drawn from market conditions themselves. Given each metric's lagged values as well as 26 exogenous indicators collected from the literature, I aim to train forecast models that can successfully predict future values of each of these variables with higher accuracy than a "baseline" autoregressive model. Taken together, the results presented in this work represent an incremental, but nonetheless important, step in the rapidly growing field of financial stability analysis, and could be of use to central banks and other financial institutions to complement their existing risk forecasting tools.

The rest of this paper proceeds as follows. First, I review the existing financial stability literature and describe the datasets of financial stability indicators that will be used in modeling. Next, using data collected for the United States, I perform dimensionality reduction via principal components analysis on the collected indicators to visualize the features that contain the most information; then, using both lasso regression and a variety of simple feed-forward neural networks to model these variables over time, I show that these indicators are reasonably predictive of liquidity risk and financial stability sentiment, improving the out-of-sample prediction error by about 15 percent compared to a baseline AR(1) model for both response variables. For robustness, I then repeat the analysis using a lower-dimensional dataset from the United Kingdom, both extending my findings to an international context as well as testing the models on a sparser dataset. Finally, I discuss my findings and conclude.

2 Literature Review and Data Description

Kiley et al. (2015) develop a powerful framework for monitoring financial system vulnerabilities in the United States economy. They compile a list of 44 financial and macroeconomic indicators – levels and growth of corporate, household, and mortgage debt, bond and lending spreads, leverage, and other macro variables such as house prices and balance of payments series – which have been shown in previous literature to contribute to financial stability, and which are currently monitored in central banks' financial stability reports. They combine the indicators using various statistical techniques to create "heatmaps" that visually display the time series of systemic risk as the distance in probability space of the combined indicators from their mean values. Focusing on several categories of research that emphasize credit booms, funding to meet increased credit demand, inflated asset prices, term mismatch, and wholesale funding as key determinants of financial crises, Kiley et al. group their indicators into three categories: risk appetite in asset markets, nonfinancial imbalances, and financial sector vulnerabilities. The risk appetite category consists of 12 indicators measuring valuations in housing and equity markets, commercial real estate, changes in lending standards, and volatility; nonfinancial sector imbalances include 17 indicators measuring home mortgage and consumer debt, nonfinancial corporate debt, and net savings; and financial sector vulnerabilities include 15 indicators of bank and nonbank leverage, bank funding and run risk, financial system size, and maturity/term mismatch.

Having defined classes of financial stability indicators, they next normalize each indicator series to zero mean and unit variance and parse the normalized indicators into 14 components underlying the overall index, categorizing each indicator and generating the component values from the simple average of the indicators in each of the 14 component categories. They next combine the components into three broader categories in three ways: by taking the arithmetic means, the geometric means, and the root mean squares of the components at each date. The result is three very similar versions of a combined index quantifying financial vulnerabilities. They also create a fourth version of the combined index using principal components analysis, taking the values of the first principal component (technically, the elements of the eigenvector associated with the largest eigenvalue) of the indicators' correlation matrix. Scaling each of these aggregate indices into probability space using the kernel density estimate of the CDF of each, the time series of their indices look very similar, with a steady buildup to a peak around 2008 and a sharp decrease after. This is an interesting result, because it shows a clear trend within the high-dimensional dataset that is almost independent of the dimensionality reduction method used to generate it. Kiley et al. (2015) is undoubtedly a significant foundation in financial stability analysis on its own, but it does not take the next step of using the time series of combined risk indices displayed in the heatmaps to forecast subsequent strain on the financial system such that policymakers can proactively regulate and adjust to mitigate economic impacts [Kiley et al., 2015].

Liquidity in money markets and bank lending markets has been shown to be an indicator not only of efficiency and stability in the financial system, but also of real economic activity, demonstrating its utility as a response variable that, I hypothesize, can be predicted by a set of explanatory financial and macroeconomic indicators. Accurate prediction of funding market liquidity from this set of systemic risk indicators, especially with a one-quarter or one-year lead, could be an important tool in guiding the optimal path of monetary and macroprudential policy to avoid financial system strain. Goldberg (2016) uses gross dealer positions in U.S. government security markets as a broad measure of financial system liquidity, reflecting the role of liquidity in Treasury markets for efficient provision of credit and execution of monetary policy. In his work, Goldberg finds that supply shocks to Treasury market liquidity "capture [financial system] stress episodes well," noting negative liquidity supply shocks during the dot-com boom, 9/11, and the 2008 Crisis. Further, he demonstrates that these liquidity shocks are predictive of unemployment and industrial production, validating Treasury market liquidity as an important signal of stress to the financial system and shocks to economic output [Goldberg, 2016]. Unfortunately, data on gross dealer positions in Treasury markets is only publicly available going back to 1998, resulting in an insufficient sample size to meaningfully train models and test their predictive accuracy.

In the absence of publicly available direct measures of money market liquidity, bank funding spreads have been shown, at least in the case of the 2008 Crisis, to be a reasonable proxy for money market liquidity and for financial system strain. Increasing bank funding spreads represent a larger difference between the rate on interbank lending and the risk-free rate, reflecting increased perceived credit risk and counterparty risk in lending and other markets as banks wary of default of their counterparties require a higher interest rate. For the United States, a common bank funding spread is the TED spread, or the difference between the 3-month LIBOR (London Interbank Offered Rate, a standard interbank lending rate) and the 3-month Treasury yield.[1] Brunnermeier (2009) explains that in times of financial system stress, banks charge higher interest rates for unsecured loans and

[1] The acronym comes from the spread's two components historically, the yield on a Treasury bill and the yield on Eurodollars, US-denominated deposits at European banks, on which the relevant interest rate is LIBOR. Today it is calculated simply as the difference between LIBOR and the 3-month T-bill yield.

the demand for secure collateral – "flight to quality" – increases, pushing Treasury yields down as demand rises. Therefore, the TED spread widens when the financial system is stressed and functions as an indicator of the severity of liquidity shocks and overall financial instability [Brunnermeier, 2009]. Boudt, Paulus, and Rosenthal (2013) also use the TED spread to predict liquidity in funding and equity markets, demonstrating via a two-regime two-stage least squares model that a TED spread of over 48 bp is indicative of a market regime reflecting perceived financial strain, while if the spread is under this value, credit risk is perceived to be low and the supply of liquidity is adequate [Boudt et al., 2017]. I therefore use the TED spread as my measure of funding liquidity in the U.S. financial system.[2]

While the TED spread quantitatively and concretely tracks funding risk in the economy and has clear predictive value in gauging systemic risk, it does not necessarily provide a holistic and qualitative view of financial stability and the buildup of vulnerabilities. To account for this, I turn to the Financial Stability Sentiment score, first introduced by Correa et al. (2017). Correa et al. apply natural language processing and sentiment analysis to central banks' financial stability reports, creating a "financial stability dictionary" that classifies about 300 words related to financial stability as carrying positive or negative sentiment. The score is then constructed by parsing each financial stability report, subtracting the number of "positive" words from the number of "negative" words and dividing the difference by the total number of words in each report. In this way, an elevated FSS corresponds to deteriorating sentiment by the authors of the reports (central banks) as to the health of the financial system in the current period – it parallels the TED spread in that increasing values theoretically correspond to a buildup of financial vulnerabilities, and it can therefore function as an additional response variable to validate the signal carried by the indicators with respect to a subjective quantification of systemic risk. The authors then run a VAR model to determine the predictive power of their FSS scores on several financial cycle variables, and find that decreases in asset valuations both lead and lag associated increases in the FSS – that is, the FSS reflects past financial sector strains as well as leading future strains, and it is not straightforward to decompose the ex-post and ex-ante information content of the FSS. Additionally, they find that an increased FSS is significantly related to contemporaneous risk buildup in financial sector variables

[2] See appendix for an overlay of Treasury market liquidity and the TED spread over the period for which both series are publicly available.

such as the credit-to-GDP ratio and the overall debt service ratio, but when the FSS leads these indicators by one period, the relationship loses significance.

Given this evidence on the FSS's relationship with financial cycle indicators, the authors finally seek to determine its predictive power with respect to banking crises as determined by the previous literature, fitting a probit model where a banking crisis dummy variable (0 or 1 based on whether a banking crisis occurs in a given country at a given period) is regressed on the FSS, the credit-to-GDP gap, and the debt service ratio. Even including these financial cycle indicators to control for omitted variable bias, the FSS maintains predictive power for crises one quarter ahead, with a 1.25 standard-deviation increase in the FSS associated with a 22 percent increase in the probability of a banking crisis in the quarter ahead. However, this relationship is only marginally statistically significant ($0.05 < p < 0.10$). In explaining the weakness of the result, the authors mention that negative central bank communications around crises may not be substantially elevated compared to normal times, as central banks may issue more cautious outlooks or focus on the resiliency of the financial system at such times in order to shore up confidence among consumers and firms, thereby exerting a contradictory effect on the FSS [Correa et al., 2017].

Clearly, there are some issues with the timing and endogeneity of the FSS, as it is inherently responsive to both past variation in financial indicators as well as to expectations of future variation in those indicators – that is, the FSS cannot necessarily differentiate forward-looking from contemporaneous deterioration in financial stability sentiment. However, its demonstrated utility for predicting banking crises at a one-quarter lead is promising enough that I proceed to explore it further. Financial stability reports are only available from the Federal Reserve starting in 2011, so instead of scoring the reports themselves as in Correa et al., I apply their scoring framework to the minutes of Federal Open Market Committee meetings, available continuously since 1967.[3] This should be a reasonable analog that allows us to capture the sentiment of policymakers regarding financial risks, while allowing the use of the full dataset for analysis instead of constraining it to data for 2011 or later. To construct my FSS response variable, minutes for the March, June/July, September, and December FOMC meetings for each year in Kiley's dataset (corresponding to its

[3] Before 1993, the minutes were released as two separate documents, the "Record of Policy Actions" and the "Minutes of Actions," while a single combined document is released for all meetings 1993 or later; I use the Record of Policy Actions before 1993, as it has stronger similarity to the single document released after 1993. Text of the documents used to generate the scores is available online at github.com/metzner28/liquidity.

quarterly observations) were downloaded in html format from the Fed's website and the financial stability dictionary was obtained from Correa et al. I then scrape and score the minutes with a Python script to generate the time series of the FSS, resulting in a series of 118 quarterly observations from Q1 1990 to Q2 2019, to match Kiley's dataset and the TED spread data.

To summarize, I have identified a set of macroeconomic and financial variables hypothesized to contribute to systemic risk, as well as a response variable that has been shown in the literature to directly track funding risk and stress to the financial system; I also construct an alternative response variable to subjectively measure systemic risk through policymaker sentiment. Kiley's original dataset consists of 44 indicators measured from Q1 1990 to Q4 2014, of which 26 are available in every period (i.e. starting in 1990, with no missing data to 2014). I exclude all indicators that are not continuously available in order to generate the largest possible training set – in terms of observations, but not features – for analysis. The resulting 26 indicators are grouped by the specific financial system vulnerability they measure: risk appetite/asset valuations, financial sector vulnerabilities, and nonfinancial sector imbalances. Each of these indicators was updated such that the final U.S. dataset of explanatory variables consists of 118 observations of 26 features, each measured quarterly from Q1 1990 to Q2 2019. Full specifications for the data collection, including full descriptions of each indicator and the sources of updated data for each series, are available in the online appendix.[4] Time series plots of the features in each group, along with each response variable, where each series has been normalized to zero mean and unit variance are included here:

[4] github.com/metzner28/liquidity

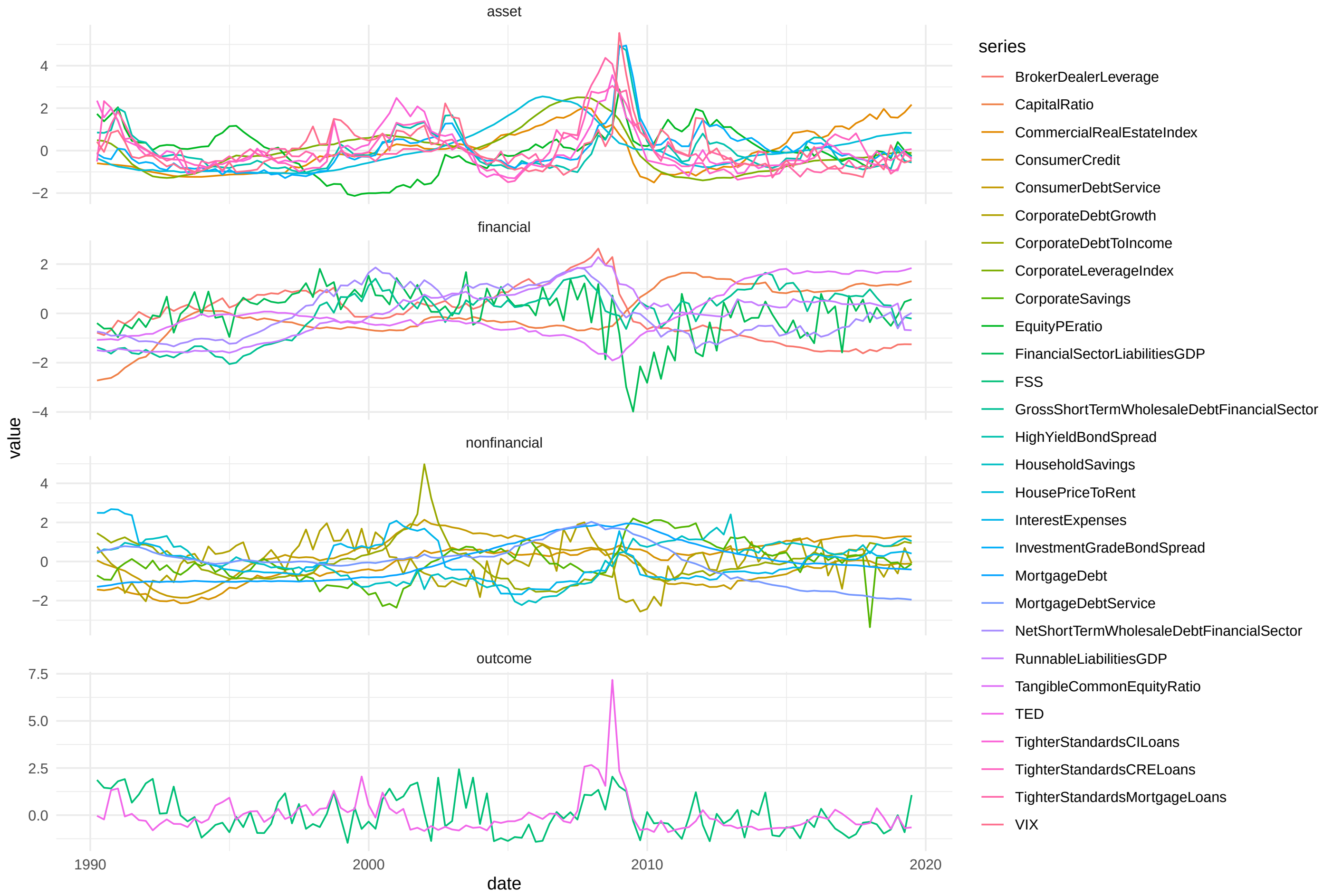

Figure 1: Time Series of U.S. Financial Stability Indicators, TED Spread, and FSS

For the United Kingdom, Aikman et al. (2018) conduct a similar analysis to that of Kiley et al. (2015), identifying 29 financial stability indicators for the U.K. economy that largely overlap with Kiley's. Drawing from the same categories of indicators previously validated in the early warning literature – leverage, asset prices, levels and growth in debt and credit, and maturity mismatch – the authors again form three broad groups into which they classify the indicators: asset valuations, private nonfinancial sector debt, and terms and conditions on new credit. The private nonfinancial sector category consists of 12 indicators measuring growth and levels of consumer, corporate, and real estate credit and debt service, as well as external leverage via the U.K. current account, external debt, and capital inflows. Asset valuations are divided into financial and real assets, with six financial category indicators measuring term premia on U.K. gilts; the volatility, risk premia, and price-to-earnings ratio of the FTSE index; and investment-grade and high-yield bond spreads. Real asset valuation indicators index house price growth and ratios of house prices to income, as well as commercial real estate price growth. Finally, indicators measuring terms and conditions on new credit include spreads on new mortgage and commercial real estate loans, as well as average loan-to-value ratios on household and commercial loans.

Following Kiley's analysis, the authors again combine the indicators using several "simple and transparent" dimensionality reduction or "weighting" methods and produce indices and accompanying heatmaps of the composite financial stability risk in the U.K. over time. While Kiley et al. focus on characterizing the composite vulnerability indices themselves, Aikman et al. go slightly further, interpreting the relationship between the indices and downside risks to U.K. economic growth. They estimate several regressions to forecast the impact of their risk metrics on GDP growth, projecting the tails of the distribution of U.K. GDP growth to construct a "GDP-at-risk" measure at 1-quarter, 4-quarter, and 12-quarter horizons. Surprisingly, they find that private nonfinancial sector leverage generally has a negative impact on GDP growth in the short run, and an insignificant effect in the long run, while elevated asset valuations have a positive short-run effect and a significantly negative long-run effect. This result casts some doubt on the long-run information content of the selected indicators for the U.K. economy, specifically for nonfinancial sector variables (which make up 12 of the 29 indicators), a fact to keep in mind when testing the short-run relationship between liquidity risk and financial vulnerabilities [Aikman et al., 2018].

To mirror the analysis of the effect of financial vulnerabilities on liquidity for the United Kingdom, a U.K. analog of the TED spread – a response variable that quantifies the cost of interbank lending and therefore reflects general credit and liquidity risk to the economy – is necessary. Benos and Zikes (2016) of the Bank of England investigate liquidity determinants in U.K. gilt markets, noting that "liquidity (or lack thereof) was at the heart of the 2008-09 financial crisis." Using primary dealer transactional data similarly to Goldberg's analysis of liquidity in Treasury markets referenced above, they fit several models determining effects of credit and bank lending spreads on gilt market liquidity, measured via "yield curve noise," a measure of gilt mispricing constructed via the difference between the actual prices of each transaction and the price on an idealized yield curve. Regressing noise on net dealer volume of gilts traded, as well as several credit spreads and volatility indices, they find, in a convenient parallel that should not be surprising, that the spread between the 3-month LIBOR and the 3-month repo rate with gilts as collateral – an analog of the TED spread for the U.K. – provides a measure of liquidity in gilt markets at high significance. Not only does this result validate empirically the more theoretical choice of the TED spread as a response variable for the U.S. analysis (assuming the same effects are at play in U.S. Treasury markets and U.K. gilt markets), but it also points toward a publicly available measure of U.K. funding liquidity that can be utilized to fit a model for the U.K. economy. That said, I use this LIBOR-3M Repo spread as my measure of funding liquidity in the U.K. economy [Benos and Zikes, 2016].

To assemble the U.K. dataset, then, I start with the LIBOR-Repo spread referenced in Benos and Zikes, and note that gilt Repo rates are not publicly available from the Bank of England before 1997 or after 2Q 2018 due to changes in Bank of England policy. Therefore, this dataset will necessarily contain fewer observations than the U.S. analog. Of the 29 original indicators in Aikman et al., 17 are available continuously from Q1 1997 to Q2 2018, and I again make the decision to exclude all indicators that are not continuously available in order to generate the largest possible training set – in terms of observations, but not features – for modeling. The 17 indicators remaining are grouped as in Aikman et al.: asset valuations, private nonfinancial sector debt and leverage, and terms of credit. Each of these indicators was updated such that the final U.K. dataset of explanatory variables consists of 90 observations of 17 features, each measured quarterly from Q1 1997 to Q2 2018. No analog of the FSS is available for the U.K., as the Bank of England did not publicly release minutes of Monetary Policy Committee meetings before 2015. Again, time series plots of the features in each group, along with the funding spread response variable, where each series has been normalized to zero mean and unit variance are included here:

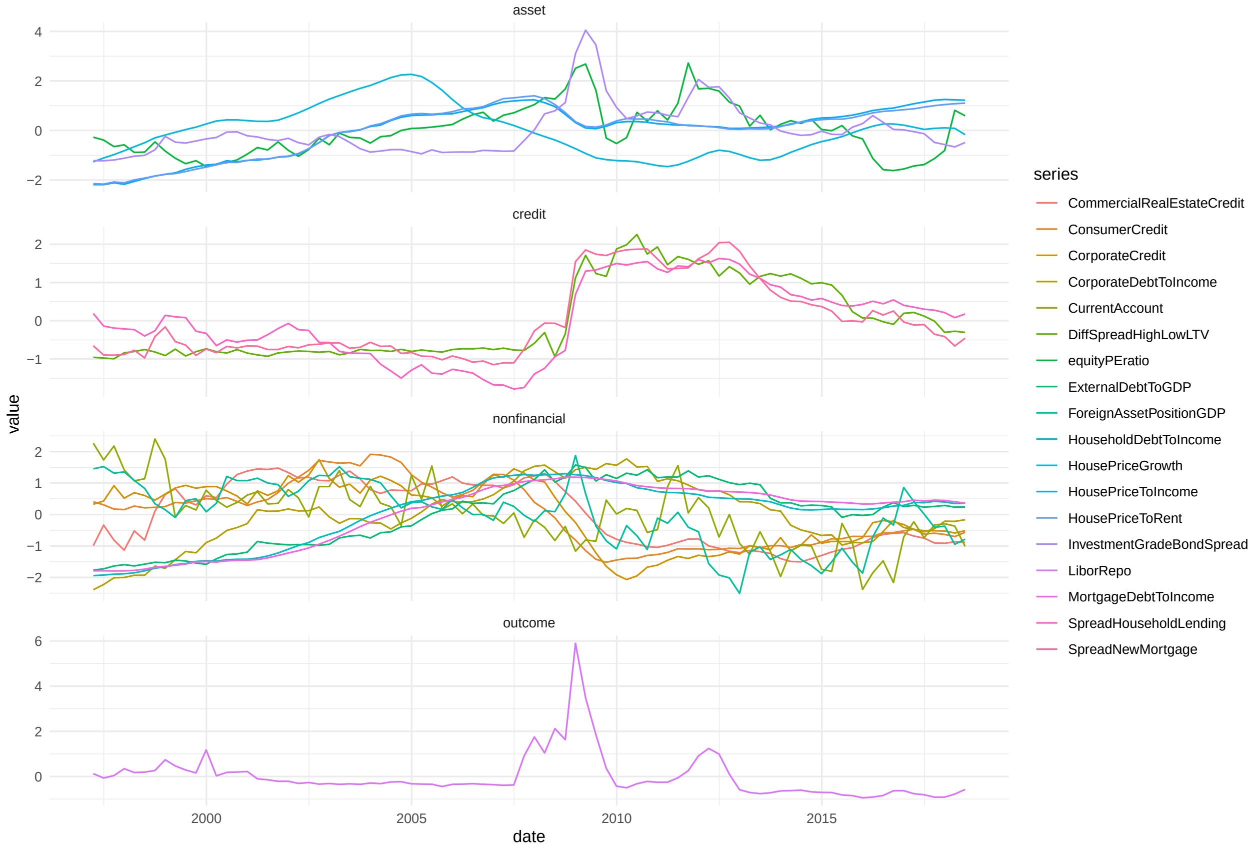

Figure 2: Time Series of U.K. Financial Stability Indicators and 3-month LIBOR-Repo Spread

Despite the analyses of Kiley et al. (2015) and Aikman et al. (2018), it is clearly challenging to glean much predictive and interpretable information from these jumbles of graphs. The data description is meant as a preview of the task at hand, extracting a clean signal from the noisy and high-dimensional sets of indicators collected here that allow accurate prediction of the outcome variables shown in the last panels of each set of graphs.

3 Analysis

3.1 U.S. Data

3.1.1 Exploratory Data Analysis

I begin my analysis with the dataset of U.S. Financial Stability Indicators from Kiley et al. (2015) shown in Figure 1, noting that the time series graphs in the figure reveal that this dataset is high-dimensional and highly correlated. This is an ideal use case for principal components analysis (“PCA”), a linear dimensionality reduction method that generates orthogonal components from linear combinations of the original features. As a first attempt to visualize the indicators that carry the most information, I perform PCA on the matrix of indicators. The indicators were each differenced once to remove concerns of nonstationarity, which can affect the interpretability of PCA loadings.[5] I now plot a heatmap of indicator loadings for the first three principal components, as well as scatterplots of the first two principal components colored by the one-quarter-ahead log TED spread and standardized FSS.

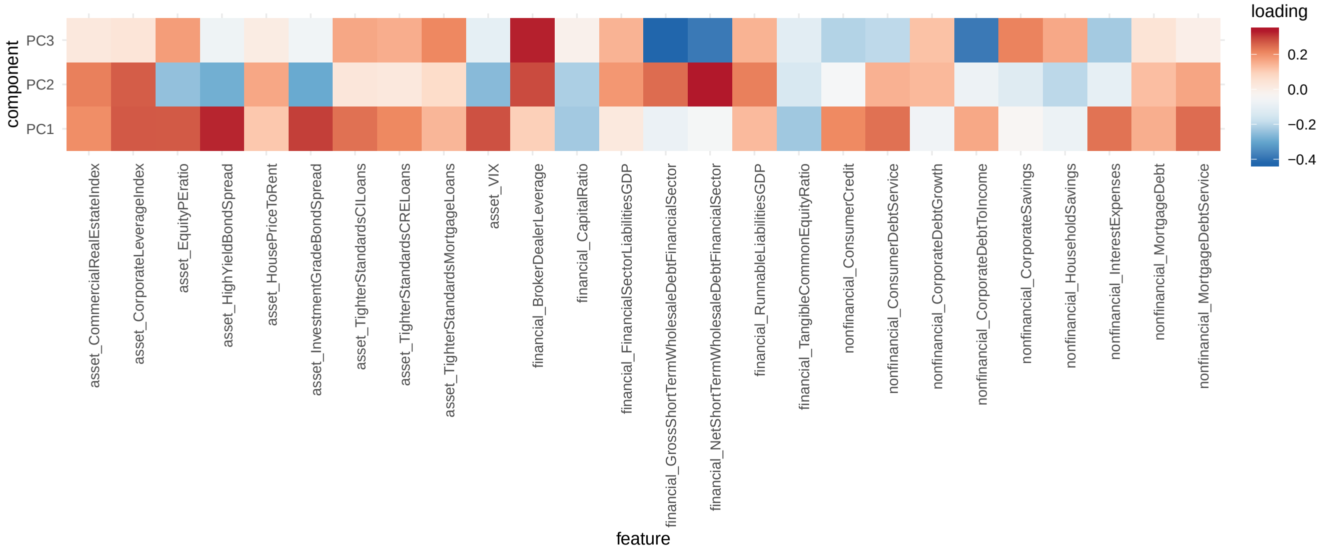

Figure 3: Indicator loadings on first three principal components

[5] Specifics on data stationarity and detrending are available in the online appendix and are explained further during model specification.

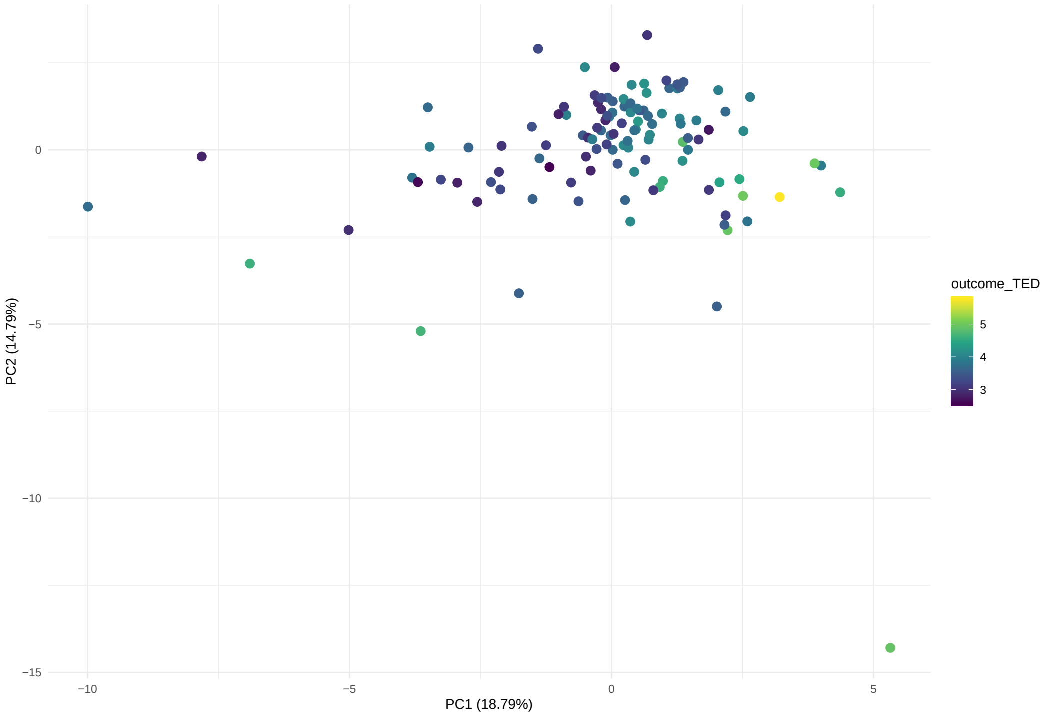

Figure 4: First two PCs, colored by one-quarter-ahead log TED spread

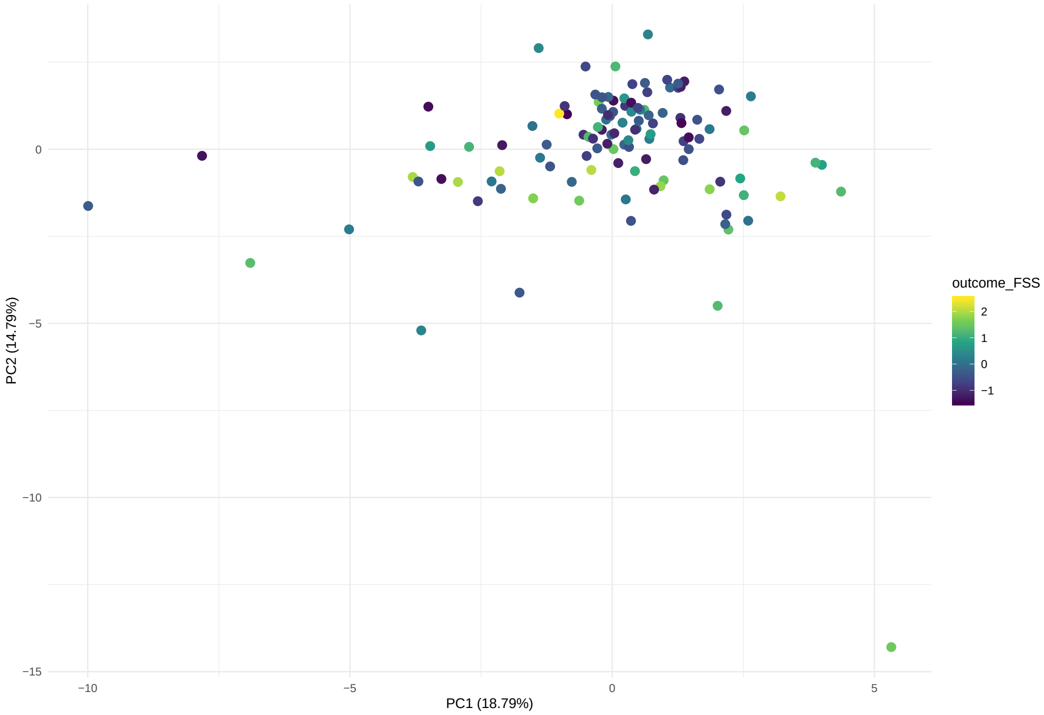

Figure 5: First two PCs, colored by one-quarter-ahead standardized FSS

Together, the first two principal components explain 34 percent of the variance in the indicator dataset, and the first three 42 percent. The indicator loadings in Fig. 3 show that the first principal component seems to pick up variation in asset pricing variables, with positive loadings on all of the asset pricing variables at the left of the heatmap and the highest values on high-yield and investment grade bond spreads, as well as the P/E ratio of the S&P500 and the VIX, a standard volatility index for equity markets. The next two prinicipal components are much noisier, but it appears that the second has correlation with the financial sector indicators shown at the middle of the heatmap, with the largest loadings on broker-dealer leverage and net short-term wholesale debt in the financial sector. Now, even though the variance explained by the first principal component is not as high as would be ideal, Fig. 4 shows that the first principal component is clearly picking up information related to the TED spread – the values of the TED spread increase smoothly (from darker blue to lighter green/yellow points) with the first principal component (on the x-axis). This suggests that the asset pricing variables that have high loadings on the first principal component may be more important than other indicators in predicting funding liquidity. Similarly, while slightly harder to detect a clear pattern, Fig. 5 suggests that variation in the FSS is reflected in the second principal component, with FSS scores increasing down the y-axis (as the values of the second principal component decrease).

3.1.2 Model Specification and Results

I now specify a set of empirical models of funding liquidity and financial stability sentiment, measuring funding liquidity via the TED spread and sentiment via the FSS. The goal is to produce one-quarter-ahead forecasts of the TED spread and the FSS given their values in the current period and the values of a vector of indicators informing financial system stability, also in the current period. Therefore, the general form of the models I seek to fit is as follows:

$$y_t = \mu + \phi_1 y_{t-1} + \Lambda \boldsymbol{X_{t-1}} + \epsilon_t$$

where y_t is the response variable in period t with mean value μ, Λ is a vector of coefficients on the financial stability indicators, $\boldsymbol{X_{t-1}}$ is a vector of values of financial stability indicators in period $t-1$, and ϵ_t is an error term. While this seems like a good theoretical framework for analysis, there are a few problems with this approach. First, since the response variables and all of the indicators are themselves time series, stationarity is of concern. Similarly to the concerns mentioned above as to the interpretability of PCA loadings calculated from nonstationary data, time series models trained on nonstationary data also suffer from interpretability issues. To solve this problem, I apply Augmented Dickey-Fuller Tests to each series, which showed failure to reject the presence of a unit root in 17 of the 26 indicator series – though not the TED spread or the FSS – with significance at the 5 percent level. Therefore, I difference all 26 series once, resulting in a new dataset of 117 observations where each observation is the difference in each series between period t and period $t-1$.[6] Next, it might also be reasonable to argue that the indicators themselves are endogenous, at least to the TED spread, as a higher cost of interbank lending might influence future lending standards to consumers and firms, for example, or future levels and growth of credit. This issue is avoided based on the timing I have specified – since the models generally attempt to predict the effect of *past* lending standards (and other indicators) on *future* funding liquidity, the right-hand side is pre-determined with respect to the response variables, so there should not be any simultaneity issues that might suggest endogeneity in the model. Given the wide range of the TED spread in the dataset (from about 30 to 350 basis points), I take the log to ensure that linear models provide reasonable fit. Similarly, I standardize the FSS to zero mean and unit variance to increase interpretability. Therefore, I now aim to estimate the following two models:

$$\log(TED_t) = \mu + \phi_1 \log(TED_{t-1}) + \Lambda \boldsymbol{X_{t-1}} + \epsilon_t \tag{1}$$

$$FSS_t = \alpha + \rho_1 FSS_{t-1} + \Omega \boldsymbol{X_{t-1}} + \epsilon_t \tag{2}$$

Focusing now on $\boldsymbol{X}$, the vector of financial stability indicators common to each specification, it is clear that this model still suffers from one serious issue. Inspection of the graphs of the indicators over time reveals that most seem to move together – that is, a substantial number of the indicators are highly correlated. While this is clearly visible in the data, it also makes sense from a theoretical standpoint – asset prices, for example, tend to be correlated over time, and the dataset

[6] While I could have differenced only the 17 series with unit roots, I argue that differencing uniformly makes for a more readily interpretable model. Since there is no real pattern among the series with unit roots (they are distributed about equally between the three indicator categories), I argue that modeling based on period-on-period changes, as opposed to levels, does not materially affect the results. Full results for the ADF tests are available in the online appendix.

of indicators contains 10 series measuring asset prices. Therefore, multicollinearity is of concern, so an accurate and interpretable model cannot be trained directly on the matrix of indicators with traditional least squares methods alone. Building accurate forecasts using high-dimensional datasets has been investigated in past literature, and Stock and Watson (1999, 2002) among others have demonstrated that dimensionality reduction is generally necessary to produce accurate and robust results [Stock and Watson, 1999] [Stock and Watson, 2002].

This issue is complicated by the fact that the final dataset only contains 117 observations – with such a sparse dataset, most nonparametric deep learning and ensemble methods, which usually work very well for difficult regression problems like this one, will invariably overfit the training data and therefore cannot be used to produce meaningful prediction results [James et al., 2017]. Beyond generating an accurate forecast, the forecast models themselves should also be readily interpretable, giving a clear idea of which features in the high-dimensional dataset carry the most information and are most predictive of the outcome variables in the quarter ahead. This is an important criterion in defining an econometric framework, as one of the main challenges set forth in this project involves identifying the macroeconomic indicators that have the largest effect on overall financial stability – clearly interpretable models are necessary to judge the relative effect sizes of each explanatory variable. Therefore, complex nonparametric methods, which tend to be "black box" algorithms that are notoriously difficult to interpret, might not be the best choice for this problem.

These criteria narrow the analysis framework down to a few possibilities: first, we have seen that PCA does a reasonable job of embedding the indicators into a lower-dimensional space with orthogonal features. It would be straightforward to replace the vector of indicators in the model with a vector of values of the principal components of the indicator matrix and perform a principal components regression. This would immediately solve the multicollinearity problem, as the principal components are guaranteed to be orthogonal, so PC regression would be expected to provide decent fit and prediction results. However, principal components regression is not a feature selection method: even if a PC regression model arrived at a suitably minimized out-of-sample error by regressing on the first n principal components of $\boldsymbol{X}$, the model would reveal little to nothing about the indicators that have the most predictive power on the TED spread. Because each principal component is a linear combination of all 26 features, no individual indicators would be "selected" via the model (even though some would have higher weights than others). Keeping in mind that the goal for this analysis is ultimately to specify and fit a set of models that are both readily interpretable and predictive, lasso regression seems a better choice. Lasso models apply a regularization penalty λ to the ℓ_1 norm of the vector of coefficients, shrinking the absolute values of the coefficients and forcing some of the coefficients to zero when λ is sufficiently large. The objective function for a lasso model is as follows, for a coefficient vector $\hat{\beta}$, response variable y, and design matrix $\boldsymbol{x}$, where $||\cdot||_n$ represents the ℓ_n norm of a vector:

$$\hat{\beta} = \arg\min_{\beta} ||y - \boldsymbol{x}\beta||_2^2 + \lambda||\beta||_1$$

The regularization of the coefficient vector – the $\lambda||\beta||_1$ term above – causes lasso models to be *sparse*, involving only a subset of features in the original dataset, and in this way they are much more interpretable than models including all of the original features. When the coefficient estimates are sufficiently regularized, lasso models also guard against overfitting and avoid multicollinearity issues that might be present when fitting a model using all of the highly correlated indicators – solving both of the issues identified when specifying the original model. One drawback to lasso regression is the lack of a closed-form solution to the above minimization problem – this means the models are fit iteratively with a *sequence* of λ to find the value that minimizes the out-of-sample error (at each step, the computer is actually performing a generalization of gradient descent to estimate the coefficients) [James et al., 2017].

I now fit the above specified models using lasso regression, training on observations from Q3 1990 to Q4 2012 (the first 90 usable observations after differencing and taking into account lags) and testing on Q1 2013 to Q2 2019 (the last 26) – this corresponds to about an 80%/20% train/test split. Traditional hyperparameter tuning methods such as cross-validation are not directly applicable here, as the models are trained on time series data. A random k-fold cross-validation split might result in training on the future and testing on the past, which could affect the interpretability of the results by ignoring the time series structure of the data and introducing information from future periods into forecasts of past periods [Zhang, 2012]. Therefore, I choose the optimal regularization penalty λ by evaluating the test error on a sequence of 100 λ ranging from 0.2 to 0, and choosing the value with the minimum test mean squared error to use for the final model. I also fit an autoregressive model without any exogenous indicators as a baseline – comparing the forecast error for the baseline time series model to that of my specification should give an idea of the marginal predictive power gained via the financial stability indicators. Model estimation results, graphs of the regularization paths of the training and test error for each model, and predicted vs. actual time series for the test set are included in Table 1 and Figs. 6-9.

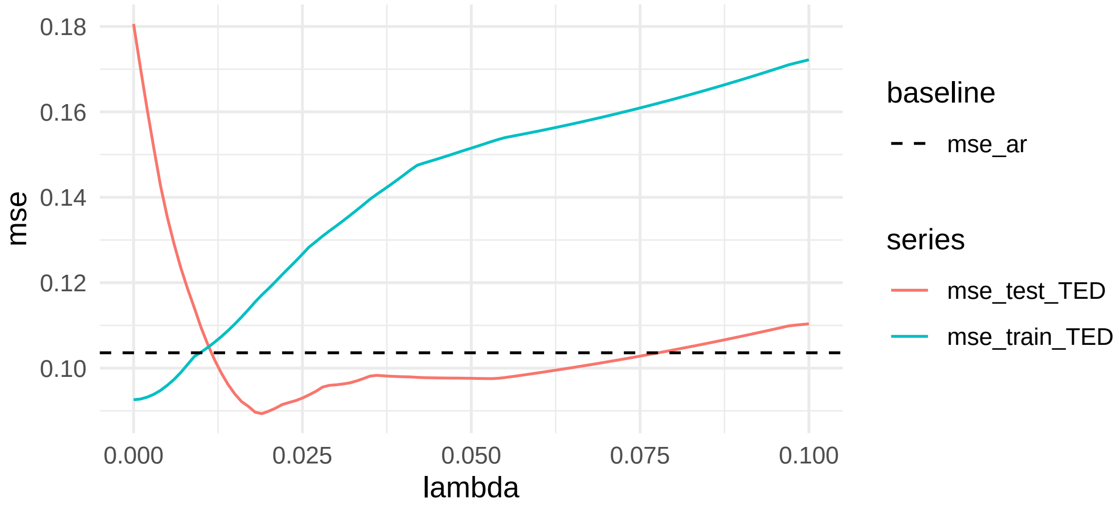

Figure 6: Training and test MSE for TED lasso model, minimum test MSE = 0.089 at $\lambda = 0.019$

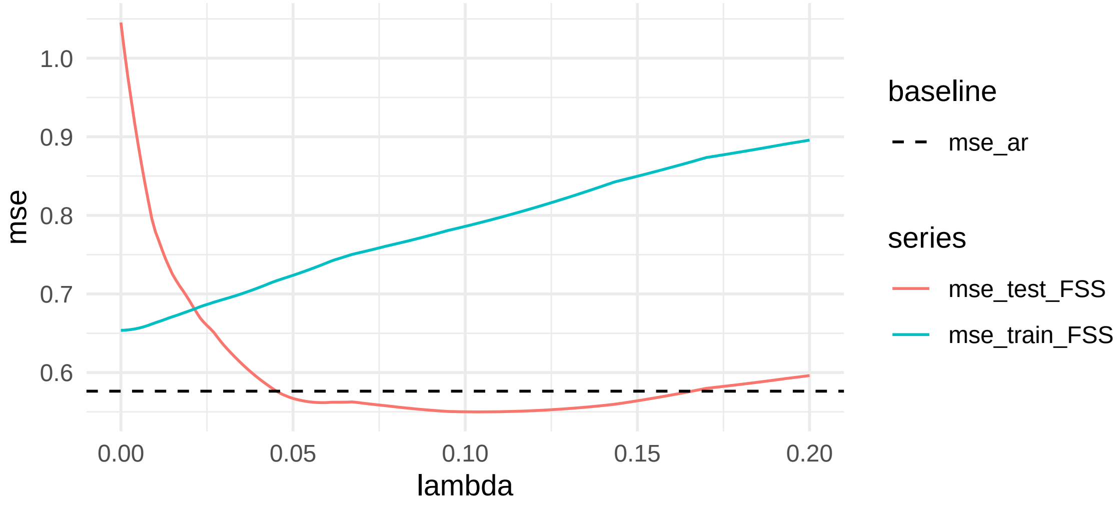

Figure 7: Training and test MSE for FSS lasso model, minimum test MSE = 0.550 at $\lambda = 0.104$

Table 1: Lasso and Baseline AR(1) Model Estimation Results for TED Spread and FSS

Class	Indicator	log(TED)	Baseline (TED)	FSS	Baseline (FSS)
	(Intercept)	1.667	1.022	0.057	0.048
	Lag	0.550	0.725	0.235	0.399
financial	GrossShortTermWholesaleDebtFinancialSector	-0.064			
financial	BrokerDealerLeverage	-0.034			
financial	NetShortTermWholesaleDebtFinancialSector	0.083			
financial	RunnableLiabilitiesGDP				
financial	TangibleCommonEquityRatio	-0.108			
financial	CapitalRatio				
financial	FinancialSectorLiabilitiesGDP	-0.020			
nonfinancial	CorporateSavings				
nonfinancial	ConsumerCredit	-0.069			
nonfinancial	ConsumerDebtService	0.052			
nonfinancial	CorporateDebtToIncome	0.0005			
nonfinancial	CorporateDebtGrowth	0.019			
nonfinancial	HouseholdSavings	-0.024		0.163	
nonfinancial	InterestExpenses				
nonfinancial	MortgageDebt	-0.037		0.052	
nonfinancial	MortgageDebtService				
asset	InvestmentGradeBondSpread				
asset	EquityPEratio				
asset	HousePriceToRent	-0.086			
asset	HighYieldBondSpread	0.023			
asset	CommercialRealEstateIndex	-0.011			
asset	TighterStandardsCILoans				
asset	VIX	0.016			
asset	CorporateLeverageIndex	0.025		-0.092	
asset	TighterStandardsCRELoans	0.081			
asset	TighterStandardsMortgageLoans	0.008		0.091	
Observations		90	90	90	90
R^2		0.683	0.523	0.246	0.162
Test MSE		0.089	0.104	0.550	0.576
Improvement (%)		13.761		4.566	

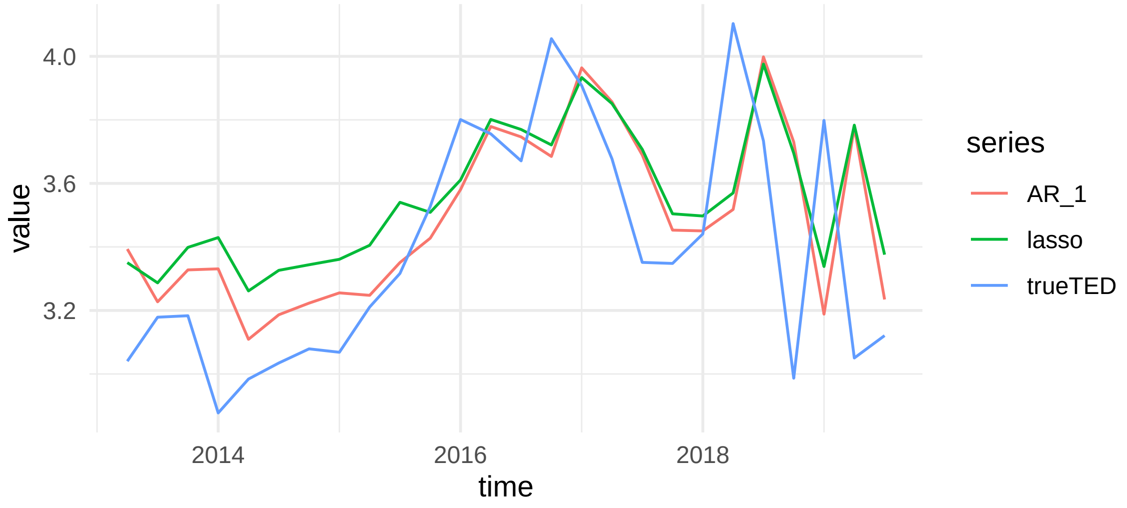

Figure 8: Predicted and actual TED spread on test set for AR(1) baseline and lasso models

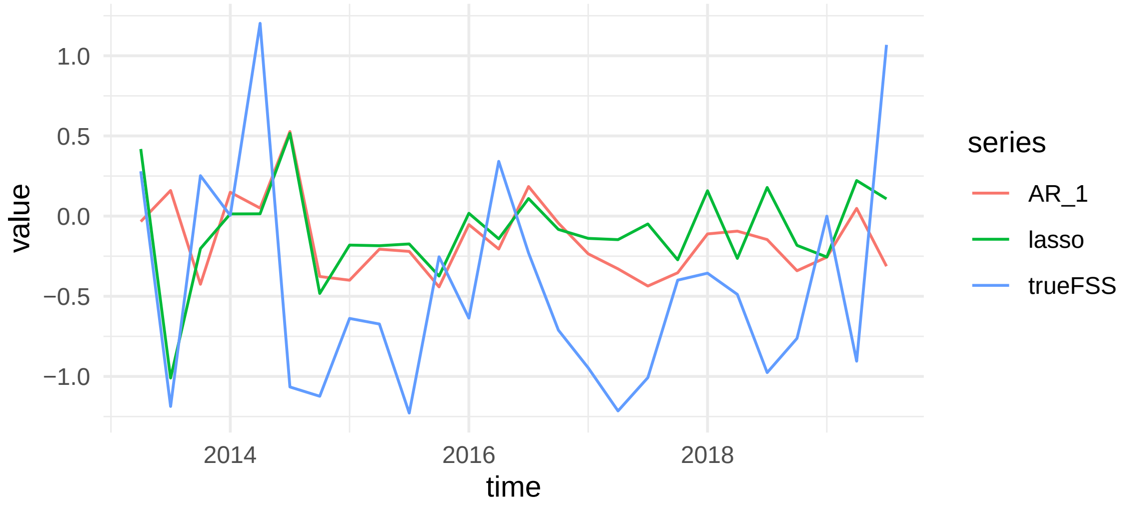

Figure 9: Predicted and actual FSS on test set for AR(1) baseline and lasso models

Before proceeding to interpret the coefficient estimates, I first analyze the model results shown in Table 1. First, both lasso models are sufficiently regularized and perform feature selection as expected: the TED spread model shrinks 8 indicator coefficients to 0, while the FSS model sends all but 4 – 22 total – to 0. As discussed earlier, regularization prevents overfitting, which can be seen in action here as Figs. 6 and 7 show the test error decreasing, and the training error increasing, as a function of λ. This is an ideal visualization of the *bias-variance tradeoff*, a central concept in machine learning: the error can be decomposed into its bias, the systematic difference between the expected value of an estimator and the true value of the parameter being estimated; and its variance, or the sensitivity of the estimator to random noise in the training set. Intuitively, the bias of a model could be minimized by overfitting to the training set – this would reduce the expected difference between the model output and the response variable's true values. However, overfitting necessarily increases the variance of the model, encouraging the estimator to model even random noise in the training data and leading to poor out-of-sample performance. In this case, we can see that the naive models at $\lambda = 0$ have low bias (training error is low) but high variance (test error is very high, reflecting over-sensitivity to the training data and poor out-of-sample performance). Now, increasing the penalization parameter λ regularizes the models and reduces overfitting, decreasing the model's variance while increasing its bias. We can visualize this via the sharp decrease in the test error as λ increases, with a more gradual increase in the training error. Since the ultimate goal here is accurate out-of-sample prediction, I tune the models to maximize out-of-sample performance, even though this necessarily results in a slightly higher training error than a less-regularized model [James et al., 2017].

The model results show that this approach seems to have been successful. Adding in the financial stability indicators results in a clear improvement in out-of-sample performance for both models, with a 13.76 percent decrease in test MSE over the baseline model for the TED spread model and a more modest but nonetheless apparent 4.57 percent decrease for the FSS model. This is promising – regardless of statistical significance or identification, these results demonstrate that there exist combinations of financial stability indicators that improve the forecast accuracy of forward-looking proxies for financial system strain. The TED spread models generally display better fit than the FSS models, as the baseline model has a substantially higher R^2 and lower training and test MSE than the FSS models (the response variables are at similar scales in terms of magnitude and variance, so the MSE values should be roughly comparable even though the units are different for each response). Notably, the R^2 values for the lasso models are both substantially higher than those of the baseline models, suggesting that the indicators account for a fairly substantial proportion of the unexplained variance present in the purely autoregressive forecasts. However, R^2 is strictly nondecreasing in the number of explanatory variables in a given model – that is, adding more variables will never decrease the R^2 – so we cannot conclude anything about the extra predictive power gained by the indicators from the R^2 alone as the lasso models each have many more explanatory variables than the baseline, so some increase in R^2 would be expected even if the indicators were completely unrelated to the response variables. Combined with the improvement in out-of-sample performance over baseline, though, the results strongly suggest that forecasts for funding liquidity and financial stability sentiment can be improved by taking into account previously identified financial stability indicators.

I next seek to determine the statistical significance of the indicators as predictors – to unambiguously determine which features carry the most information about funding liquidity and financial stability sentiment in the quarter ahead. Since lasso regression introduces significant bias into the coefficient estimates, it is not straightforward to perform hypothesis testing within the lasso model itself. The features that remain after regularization can be considered "significant," as they contribute enough to minimizing out-of-sample error that their coefficients are nonzero even with a penalty on their absolute values, though this is inherently ambiguous. Of the features that remain, there is no generally accepted method of generating standard errors to create confidence intervals for the coefficients or test whether they are meaningfully different from zero (though testing significance in sparse regression settings is currently an area of active research in statistics) [Lockhart et al., 2014] [Kyung et al., 2010]. Thus, to perform hypothesis testing, lasso regression is no longer appropriate, and I fit a model by OLS (making the Newey-West correction for autocorrelation of the error terms, considering the time series nature of the data) to test the significance of each coefficient. The best way to do this is to now treat the lasso models as *feature selection* methods only – that is, use the lasso models as a "first stage" to guard against overfitting or including multicollinear features, then with the features that remain in the lasso models, fit another set of models by OLS including these features only, and then perform hypothesis testing. These models will not be equivalent to their lasso counterparts – since they are not regularized, the values of the coefficients will change, and the out-of-sample performance will decrease. That said, this approach should give an idea of the significance of the indicators chosen by the lasso models – if the lasso is performing as expected, the majority of the indicators remaining should be significant, as the non-significant features will have been removed through regularization.

Fitting the models for the log TED spread and the FSS by OLS, including only the indicators with nonzero coefficients as explanatory variables, yields the results in Table 2.

Table 2: OLS Model Results for Selected Features

		Dependent variable:	
		log(TED)	FSS
		(1)	(2)
financial	GrossShortTermWholesaleDebtFinancialSector	−0.251**	
		(0.123)	
financial	BrokerDealerLeverage	−0.053	
		(0.055)	
financial	NetShortTermWholesaleDebtFinancialSector	0.299**	
		(0.118)	
financial	TangibleCommonEquityRatio	−0.130***	
		(0.045)	
financial	FinancialSectorLiabilitiesGDP	−0.020	
		(0.052)	
nonfinancial	ConsumerCredit	−0.116**	
		(0.051)	
nonfinancial	ConsumerDebtService	0.075	
		(0.066)	
nonfinancial	CorporateDebtToIncome	0.004	
		(0.026)	
nonfinancial	CorporateDebtGrowth	0.013	
		(0.047)	
nonfinancial	HouseholdSavings	−0.045	0.268***
		(0.045)	(0.077)
nonfinancial	MortgageDebt	−0.080*	0.190*
		(0.043)	(0.111)
asset	HousePriceToRent	−0.106**	
		(0.051)	
asset	HighYieldBondSpread	0.058	
		(0.047)	
asset	CommercialRealEstateIndex	−0.076	
		(0.069)	
asset	VIX	0.041	
		(0.053)	
asset	CorporateLeverageIndex	0.040	−0.246
		(0.045)	(0.163)
asset	TighterStandardsCRELoans	0.086***	
		(0.020)	
asset	TighterStandardsMortgageLoans	0.015	0.238***
		(0.041)	(0.047)
	lag	0.531***	0.236
		(0.059)	(0.145)
	Constant	1.743***	0.044
		(0.222)	(0.085)
Observations		90	90
R^2		0.721	0.301
Adjusted R^2		0.645	0.259
Test MSE		0.103	0.618
Note:	*p<0.1; **p<0.05; ***p<0.01, Newey-West standard errors		

The OLS results shown in Table 2 allow further analysis of the fit and identification of the TED spread and FSS models, though fitting by OLS without regularization causes a clear decrease in out-of-sample performance, even including only the features selected by the lasso models. The test MSE for these models is not distinguishable from their respective AR(1) baselines, probably because even including only a subset of the original features, the OLS model still overfits the training data as the coefficients' values are not penalized (which we can see immediately, as many of the values are different from their lasso estimates). This is not surprising – again, the goal for this exercise is not prediction – but it confirms the earlier discussion of the necessity of regularization for accurate out-of-sample prediction on high-dimensional datasets.[7]

Regardless, fitting by OLS allows for a more conclusive look at the fit of the models and the significance of the selected features. Starting with the TED spread model, even though the lasso R^2 of 0.683 was treated with suspicion due to the fact that R^2 is nondecreasing in the number of explanatory variables in a given model, the adjusted R^2 for the OLS model – which takes into account the number of explanatory variables and will decrease if predictors are added that do not explain additional variance – is 0.645. This value indicates good fit, and compared to the baseline R^2 of 0.523, means that the financial stability indicators explain about 12 percent of the unexplained variance in the AR(1) baseline. In terms of significance, the results are mixed: while 7 out of the 18 selected indicators (as well as the lagged TED spread) are significant at least at the 10 percent level, it is surprising that the majority of the features selected by the lasso model are not significant at conventional levels. While this might be assumed to indicate an insufficiently regularized lasso model, it is more likely that since λ was specifically chosen to minimize the out-of-sample prediction error, some insignificant features were selected for contributing even very marginally to minimizing the test error.

The FSS lasso model is much more severely regularized than the TED spread model, with only four features selected. Its OLS adjusted R^2 is 0.259, about on line with the lasso R^2. As the baseline R^2 for the FSS was 0.162, the four features selected for the FSS model therefore explain an

[7] Note that the OLS models are fit using only the training set, instead of the full dataset, which would arguably give more robust significance results. This was done to preserve consistency with the lasso models, the idea being that if the full dataset were used to train the lasso models, different features may have been selected. I fit the OLS models to the full dataset in the appendix, but the results do not change meaningfully. Fitting using the training set only also allows comparison of the test MSE of the OLS and lasso models, further illustrating the value of regularization in producing accurate forecasts.

additional 10 percent of the variance in the baseline, a surprisingly strong effect considering that the 18 features selected in the TED spread model only account for 12 percent of the unexplained variance in the TED baseline. Of the four indicators here, three are significant, a result more in line with expectations for the lasso's ability to select meaningful features. Of note, though, the coefficient on the lagged FSS is not significant, which calls into question whether the FSS is even an autoregressive process in its simplest form. It could be that FOMC policymakers base their sentiment more on forward-looking expectations for financial stability and the economy as a whole, rather than an assumed process where policymakers consider and discuss past conditions and then update their outlooks based on changes since the past meeting. It is also a possibility that the financial stability dictionary built by Correa et al. (2017) with financial stability reports in mind is not directly applicable to FOMC minutes without modification.[8] Sentiment analysis is a notoriously tricky problem – even Correa et al. (2017) found unexpected effects and weak significance in their analysis of financial stability reports, which should be more robust to start with than FOMC minutes. Therefore, while it is no doubt encouraging that the financial stability indicators seem to explain almost 10 percent more of the variance in the FSS than the baseline, and the lasso model's out-of-sample performance was better than the baseline, these results are difficult to interpret because hypothesis testing does not confirm that the FSS has a significant autoregressive component.

Having analyzed the predictive accuracy, fit, and significance of each model, I go on to interpret the coefficient estimates. First, for the TED spread model, the coefficient on the lagged value of the spread is large and significant, confirming that there is a significant time series component to the TED spread, as hypothesized. Other indicators that have a significant effect on the spread include the average bank tangible common equity ratio; asset pricing features including the ratio of house prices to rents and the percentage of banks tightening standards for commercial real estate lending; and levels of total consumer credit and mortgage debt. Most of the coefficients on these indicators have the expected sign: the largest effect aside from the lagged TED spread is on the bank tangible common equity ratio, which measures capitalization and leverage in the banking system. According to Kiley et al. (2015), lower TCE ratios contribute to increased vulnerability

[8] To resolve some of these doubts, I explore extensions to the FSS in the appendix.

as more leveraged banks cannot respond as effectively to financial system shocks, such as en masse mortgage loan defaults, etc. Therefore, the coefficient on this indicator is expected to be negative, as it is. Next, the percentage of banks tightening standards for commercial real estate loans also has a large and highly significant effect. As an early warning indicator, Kiley et al. (2015) assert that a decreasing percentage of banks tightening lending standards contributes to a buildup of risk, as higher availability of credit to risky borrowers might increase default risks. However, it is also straightforward to understand how an increasing percentage of banks tightening lending standards might be an indicator of imminent financial system strain, as banks hesitate to extend credit to riskier clients in a more uncertain environment.

Several debt variables – mortgage, consumer, and corporate – also have a significant effect, which is in line with the previous literature suggesting that the credit-to-GDP gap is a reliable indicator of financial instability. Interestingly, though, both gross and net wholesale debt for the financial sector have a nonzero and significant effect on the spread, though they have opposite signs: gross wholesale financial debt, which is simply the ratio of short-term debt issued by the financial sector to GDP, has a negative effect on the TED spread one quarter ahead, while net wholesale debt, which subtracts short-term and government-supplied assets from gross wholesale debt, has a positive effect on the spread (we recall that an increasing TED spread signals increasing financial system strain, so these results might suggest that gross wholesale debt levels tend to decrease vulnerability, while net wholesale debt tends to increase it). This is not the expected result, but it does seem to be a common theme: while most of the significant coefficients are of the expected sign, there are several notable exceptions. The house prices to rent ratio, for example, was highly elevated in the lead-up to the financial crisis, yet the coefficient on it is negative in this case, suggesting that a higher ratio lowers funding spreads one quarter ahead. The same is true with other asset pricing variables, such as the high-yield bond spread, the coefficient of which is expected to be negative as lower spreads signal increased risk appetite leading to financial system strain. It could be that one quarter ahead is not a sufficiently distant lead to pick up on early warning indicators – that is, some of the coefficients are flipped because they are not capturing the buildup of risks to financial stability, but rather the effect once instability and strain on funding liquidity have already set in or are anticipated imminently. It could also be that the coefficient estimates are biased, and their signs are an artifact of regularizing the model. In this case, the models obviously lose interpretability as to the actual effect sizes of the indicators, but the feature selection should still give an idea of which indicators are most predictive of liquidity.

To summarize, the lasso model demonstrates that 18 of Kiley's 26 financial stability indicators are predictive of funding liquidity one quarter ahead, and indicators from all three classes – asset prices, financial sector vulnerabilities, and nonfinancial sector imbalances – are represented almost equally in the final group of selected features. Somewhat surprisingly, though, the indicators that PCA showed to contribute most to the variance in the dataset – most notably the investment-grade bond spread over Treasury yields and the P/E ratio of the S&P 500 – were not among the features selected by the lasso model. The divergence between the results of the unsupervised and supervised modeling techniques here might imply that the dataset of indicators is inherently noisy, and that several of the asset pricing variables identified by Kiley et al. as early warning indicators are not actually predictive of the chosen measures of systemic risk at a one-quarter lag. Instead, they might only add noise, or unexplained variance, to a sparser matrix of indicators that carry real information about financial instability. This interpretation aligns with previous literature, which suggests that equity prices and other asset valuations have low signal-to-noise ratio with higher volatility at much shorter-term frequencies than are useful to meaningfully predict the buildup of systemic risk [Drehmann et al., 2012]. Even with somewhat unexpected features selected, some with coefficients having unexpected sign, the models performed relatively well. Kiley et al.'s dataset of indicators clearly has nontrivial predictive power for funding liquidity, improving the out-of-sample prediction error on the log TED spread by about 14 percent over the AR(1) baseline. Whether these indicators are the only exogenous variables capable of adding predictive accuracy to models of systemic risk remains unclear, but some of them are clearly valuable additions.

Now, an ideal result for the FSS model would involve selection of the same set of features as the TED spread model with similar effect sizes, as this would demonstrate that no matter how financial stability is measured – via funding spreads as a proxy for liquidity or via policymaker sentiment – a consistent set of indicators carry the same predictive power as early warnings for a buildup of systemic risk. Unfortunately, the actual results between the two models differ more than expected. The FSS model only selects four indicators – household savings, mortgage debt, corporate leverage, and tightening standards for mortgage loans. Of these four indicators, three are significant, with household savings and tightening standards on mortgage loans significant at the 1 percent level and mortgage debt significant at the 10 percent level. While mortgage debt levels have the expected positive coefficient (like the TED spread, an increase in the FSS signals buildup of financial system strain), household savings levels, which would be expected to act as a buffer against financial shocks and therefore decrease the FSS in the next period, have an unexpectedly positive effect on the FSS. Perhaps growth in household savings might decrease household consumption to levels where policymakers start becoming wary of muted GDP growth and sentiment deteriorates accordingly, but that seems very implausible and this effect is most likely an optimization artifact. Similar to the effect seen with tightening standards on CRE loans in the TED spread model, the percentage of banks tightening standards on mortgage loans has a significant positive effect on the FSS, when the expected sign would be negative due to tightening standards signaling decreased risk appetite. Again, this is explained by the one-quarter lag time, which likely picks up a co-movement of tightening lending standards and deteriorating sentiment in an already-strained financial environment, rather than a true causal relationship between lending standards and the FSS.

Overall, the FSS model results cast some doubt on the overall analysis strategy presented here, as, even in the most generous interpretation, it seems that the TED spread and the FSS are not equivalent response variables. It could be that the two responses measure different components of systemic risk and are therefore influenced by non-intersecting subsets of financial stability indicators – there is some evidence for this within the PCA plots, which show that the TED spread varies with the first principal component of Kiley's dataset, while the FSS varies (albeit less smoothly) with the second principal component. Correa et al. test several variants of the FSS, both counting only words classified as negative in the dictionary in the score's numerator, as well as adding, instead of subtracting, the positive and negative words in the numerator. I explore these variants in the appendix, but the results are not meaningfully different than for the FSS generated here. Again, sentiment analysis is already a technique that lacks robustness, and applying a dictionary created for financial stability reports to FOMC minutes could have added even more noise to any signal that might have been present within the minutes themselves.

3.1.3 Nonparametric Methods

Next, I turn to more speculative techniques to improve the out-of-sample performance of the empirical models. As discussed, deep learning methods generally require more data than we have here, and they are usually prone to overfitting when trained on sparse datasets. However, simple feed-forward neural networks with one or two hidden layers and some form of regularization to prevent overfitting have been used with some success to predict economic time-series data, so I attempt here to forecast the TED spread and the FSS using neural networks [Terasvirta et al., 2005] [Nakamura, 2005]. The number of nodes and layers in a neural network generally determines its bias and variance (at a high level – this is in no way a strict relationship and trial and error is often necessary to determine the best architecture for a given dataset), so to mitigate overfitting issues I constrain the neural network architectures tested to no more than two hidden layers with no more than 16 nodes each [Zhang, 2012]. I also utilize dropout between layers, a regularization technique in which some percentage of the units in each layer, as well as their connections to the next feed-forward layer or the output layer, are randomly "dropped" during each training epoch, essentially creating a sparser neural network for testing with lower average weights at each unit (as the weights are updated after each epoch) [Srivastava et al., 2014]. Each neural network in the table below was fit with a stochastic gradient descent optimizer with learning rate $\eta = 0.025$ over 100 training epochs (these hyperparameters were also chosen by trial and error), and the MSE on the test set was used for determining the best architecture for the TED spread and the FSS.[9]

[9] It is important to note here that the test MSE varies widely between architectures, and in fact – due to the nondeterministic nature of the stochastic gradient descent optimization algorithm – the test MSE will vary even when fitting the same architecture repeatedly. Because replicability is a necessary component of sound research, this is a disclaimer that running the code used to fit the neural networks (provided in the online appendix) will not necessarily produce this exact array of test MSE values every time, though the relative values and relationships between architectures and test MSE should be consistent.

Table 3: Description of Feed-Forward Neural Network Architectures

Architecture	Units (1)	Units (2)	Dropout (1)	Dropout (2)	Activation (1)	Activation (2)
1	8				relu	
2	8		0.4		relu	
3	12		0.4		relu	
4	16		0.4		relu	
5	8		0.25		relu	
6	12		0.25		relu	
7	16		0.25		relu	
8	8		0.25		sigmoid	
9	12		0.25		sigmoid	
10	16		0.25		sigmoid	
11	8	8	0.4	0.4	relu	relu
12	12	12	0.4	0.4	relu	relu
13	16	16	0.4	0.4	relu	relu
14	8	8	0.25	0.25	relu	relu
15	12	8	0.4	0.25	relu	relu
16	16	8	0.4	0.25	relu	relu
17	8	8	0.25	0.25	relu	sigmoid
18	12	8	0.25	0.25	relu	sigmoid
19	16	8	0.25	0.25	relu	sigmoid

Table 4: Neural Network Results for log(TED) and FSS Forecasts

Architecture	Test MSE log(TED)	Improvement (%)	Test MSE FSS	Improvement (%)
1	0.334		1.106	
2	0.108		0.817	
3	0.134		0.638	
4	0.15		0.509	11.7
5	0.142		0.797	
6	0.202		0.679	
7	0.086	16.9	0.771	
8	0.173		0.509	11.7
9	0.167		0.61	
10	0.175		0.563	2.3
11	0.175		0.732	
12	0.142		0.648	
13	0.14		0.854	
14	0.126		0.485	15.8
15	0.137		0.702	
16	0.129		0.742	
17	0.211		0.685	
18	0.215		0.654	
19	0.133		0.685	

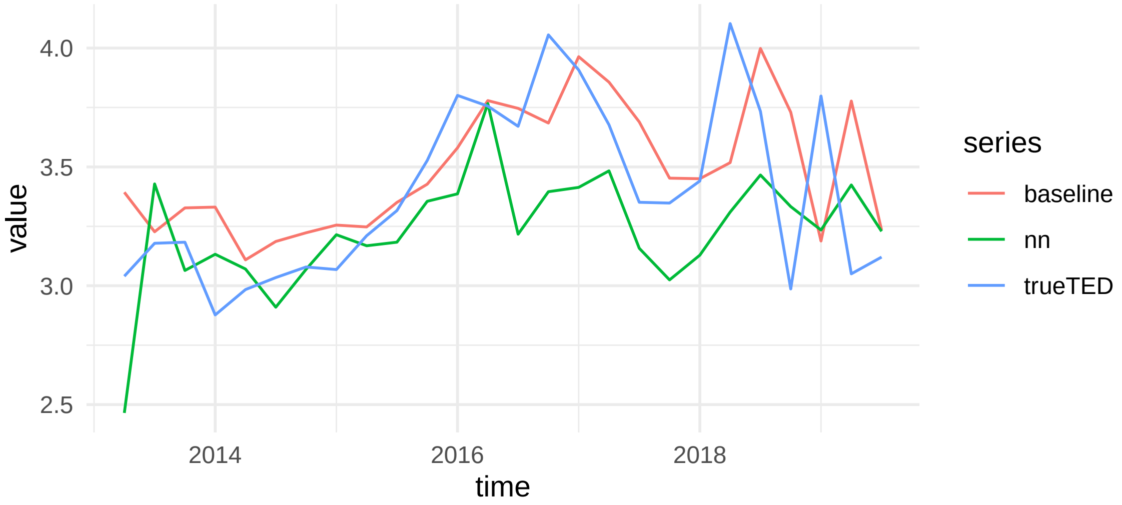

Figure 10: Predicted and actual TED spread on test set for AR(1) baseline and neural network models

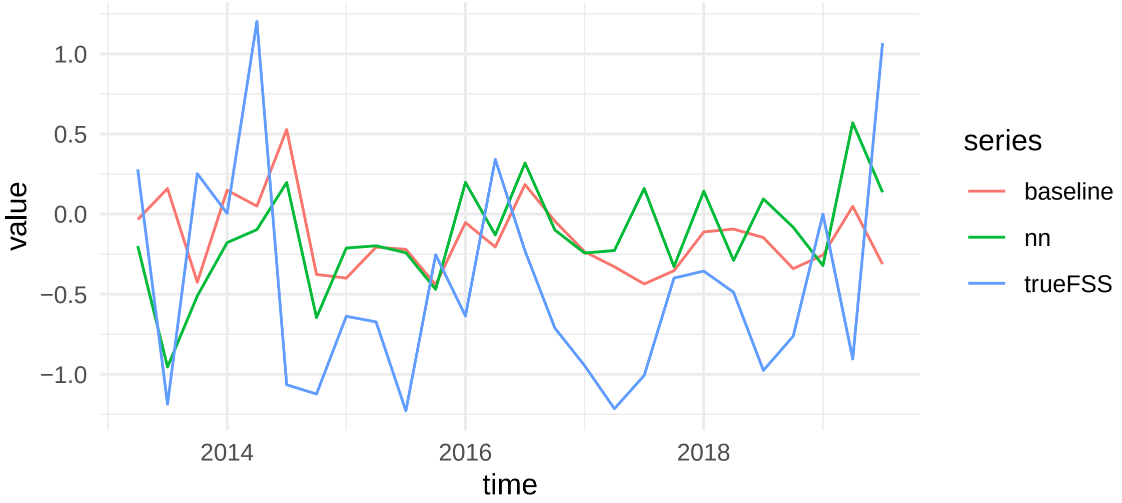

Figure 11: Predicted and actual FSS on test set for AR(1) baseline and neural network models

First, a key limitation here is the lack of interpretability inherent to neural networks. As "black box" machine learning methods, neural networks do not produce a clear correspondence between trained node weights and biases and features in the dataset. Therefore, there is no way to generate a list of effect sizes or selected features from a trained neural network, so the only way to evaluate its effectiveness is via its out-of-sample performance. This makes neural networks useful for forecasting, but not for feature selection or hypothesis testing of significant features.

With these limitations in mind, though, neural networks seem to marginally improve forecasting performance for the TED spread and significantly improve performance for the FSS. As shown in the neural network model results in Tables 3 and 4, Architecture 7, which has one hidden layer with 16 nodes and ReLU activation and a dropout rate of 25 percent, forecasts the TED spread with a 16.9 percent improvement in test MSE over the baseline AR(1) model specified earlier. This performance is in line with the lasso model, which improved the forecast error by 14 percent. The 3 percentage-point increase in out-of-sample performance of the neural network over the lasso model is marginal, and the exact value might be inconsistent, but we can conclude that for forecasting the TED spread, a neural network seems to perform about as well as the lasso model estimated earlier, if not slightly better.

The FSS forecast shows more robust improvement, with the best neural network architecture improving performance over the AR(1) baseline by 16 percent, 11 percentage points better than the 5 percent improvement gained by the FSS lasso model. Unlike for the TED spread, this architecture had two hidden layers, each with 8 nodes, 25 percent dropout, and ReLU activation. It is not surprising to see ReLU activation generally outperforming sigmoid activation (the *activation* applied to each layer is the nonlinearity applied to the output of each node in the layer, and the ReLU activation, which outputs the maximum of 0 and the node's value, is generally agreed to be preferable in most cases), though it is slightly unexpected to see a model with two hidden layers outperform sparser models with such a small dataset [Zhang, 2012]. These results are encouraging, as they confirm that the financial stability indicators carry information about the TED spread – fitting the liquidity model in two separate ways, with two different techniques, gives very similar improvement in test MSE. They also provide a significant improvement in predictions of the FSS, bringing the improvement over baseline in out-of-sample performance for the FSS model up to about the same level as for the TED spread model.

3.2 U.K. Data

3.2.1 Exploratory Data Analysis

With an analysis framework built to test the information carried by the Kiley et al. (2015) set of financial stability indicators for the U.S. economy, I now turn to the data of Aikman et al. (2018) to test whether these results are reproducible in another developed economy. Time series for the Aikman et al. (2018) dataset, as well as for the 3-month LIBOR-Repo spread (which I call the LR spread), are shown in Fig. 2 above. It is immediately clear that the main challenge to this robustness check is data availability. There are fewer series available in Aikman's dataset – 17 as opposed to 26 – as well as fewer observations, as gilt repo rates are not available before 1997 or after 2Q 2018. Therefore, the U.K. analysis I perform here is not only a check on the robustness of the U.S. results, but also a test as to whether my analysis can generalize to even more data-sparse environments. I begin by performing PCA on the full dataset of indicators, again differenced once as for the Kiley et al. (2015) dataset, and visualizing the results via a heatmap of indicator loadings on the first three principal components and a scatterplot of the first two principal components colored by the log LR spread:

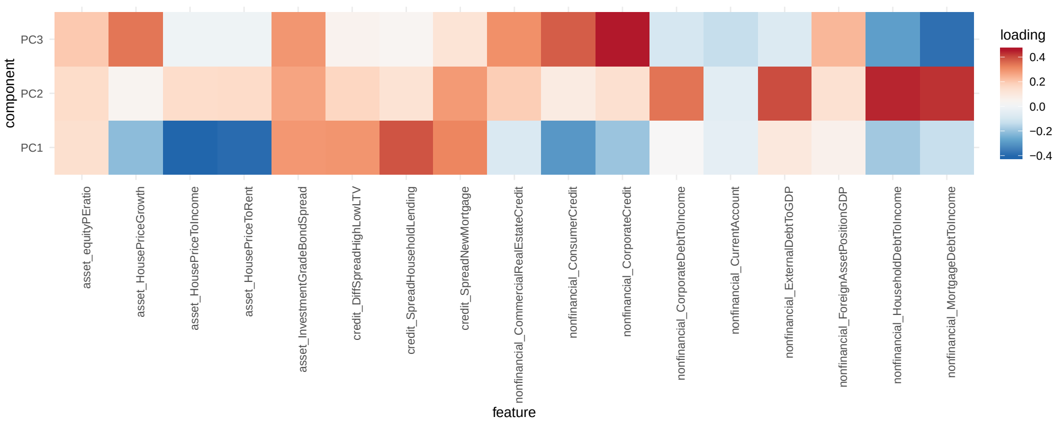

Figure 12: Indicator loadings on first three principal components

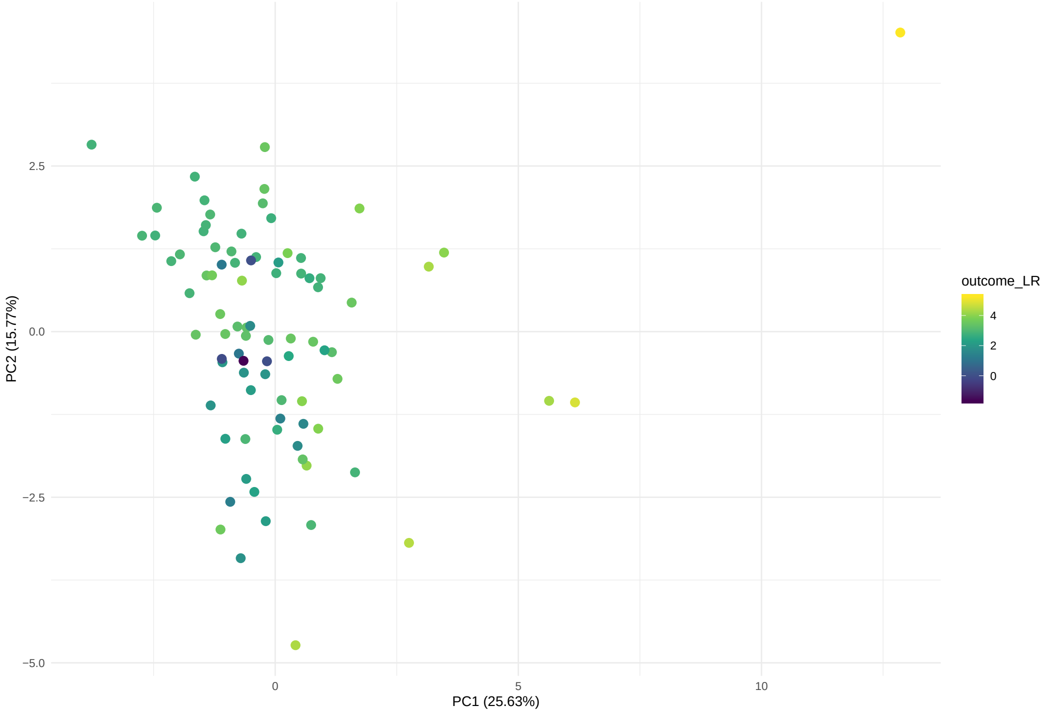

Figure 13: First two PCs, colored by one-quarter-ahead log LR spread

Promisingly, the first two PCs explain 41 percent of the variance in Aikman's dataset, and the first three 50 percent. This is slightly higher than the variance explained by the first PCs of the Kiley et al. dataset, and suggests that the PCA results here will be readily interpretable. Focusing now on the indicator loadings on the first three principal components, the first PC broadly picks up variation in asset and credit variables, with strongly negative loadings on household lending spreads and and positive loadings on housing price-to-rent and price-to-occupant-income ratios. The second and third PCs capture variation in nonfinancial sector indicators, with the highest-magnitude loadings on corporate credit and household and mortgage debt-to-income ratios. Now, similar to the results using the U.S. dataset, Fig. 13 shows a relatively smooth variation of the log LR spread with the first principal component (i.e., along the x-axis of the plot). The first robustness check of the previous results has therefore succeeded – in two different economies, we now see that the bank funding spread varies with the principal component of financial stability indicators that explains the most variance within the indicator matrix, an important finding that confirms the relationship between funding liquidity and the information contained within the matrix of financial stability indicators.

There is one very clear outlier in the upper right-hand corner of the plot at the LR spread's highest value, which occurs in Q4 2008, the heart of the financial crisis. Unlike the Kiley et al. dataset, in which the TED spread also hits its peak during this period, the outlier is clearly visible on the scatterplot of principal components of the Aikman et al. dataset, while for the Kiley et al. dataset the maximum value of the TED spread occurs at average values for both PCs and is therefore close to many other points in the plot. This is an interesting difference that might suggest that the first two PCs here explain more of the variance in the funding spread over time than in the U.S. dataset – we know that they explain more of the variance in the indicator matrix, but it appears that this translates well to variance in the funding spread in this case as well. Overall, the PCA results for the U.K. confirm the robustness of this analysis for the U.S. We conclude that in two independent economies, the funding spread tends to vary with the first principal component, or the principal component explaining the most variance, in the matrix of financial stability indicators.

3.2.2 Model Specification and Results

Now I specify an empirical model of the bank funding spread in the U.K. economy, which I again fit by lasso regression, OLS, and neural networks. From earlier, recall the general form of the model: I seek to forecast the log LR spread one quarter ahead given its value in the current period as well as a vector of differenced financial stability indicators, representing the changes in their values from the previous period to the current period. That is:

$$\log(LR_t) = \nu + \delta_1 \log(LR_{t-1}) + \Theta \boldsymbol{A_{t-1}} + \epsilon_t$$

where, similar to the models specified earlier, the log LR spread has mean value ν, $\boldsymbol{A_{t-1}}$ is a vector of the Aikman et al. financial stability indicators, reflecting the difference in each indicator from period $t-2$ to $t-1$, with coefficient vector Θ, and ϵ_t is the error term. Following the earlier analysis framework, I now fit this model with lasso regression, as well as an AR(1) baseline for comparison, and report the results in Table 5. Because of the difference in observations between the Kiley et al. dataset and this one, I alter the train-test split: instead of training on 80 percent of the data and testing on the most recent 20 percent, I split closer to 85 percent training and 15 percent test to give the model enough observations to work with. Attempts to train the model on less data than this reveal that with smaller training sets, it does not converge – the test error does not decrease as a function of λ as in the previous models, showing that for small enough training sets, the model immediately overfits to the training data and produces results that do not generalize. Results for the LR spread model as specified above are included in Figs. 14 and 15, as well as Table 5.

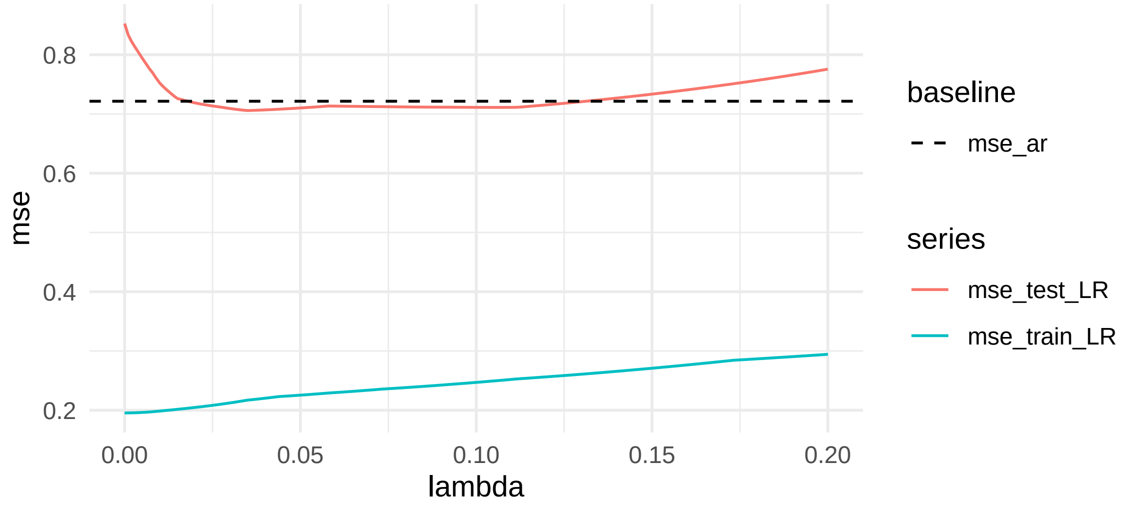

Figure 14: Training and test MSE for LR lasso model, minimum test MSE = 0.706 at $\lambda = 0.035$

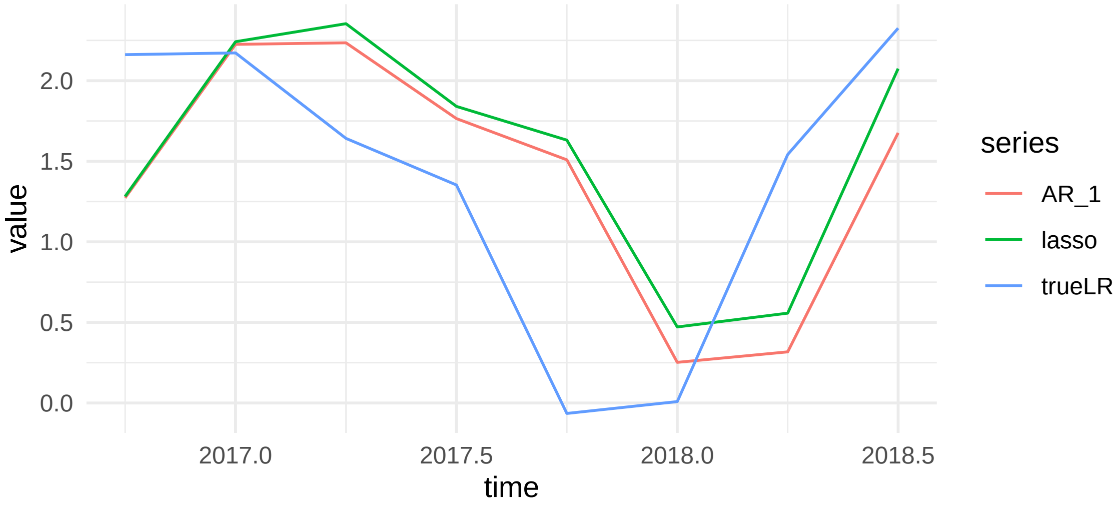

Figure 15: Predicted and actual LR spread on test set for AR(1) baseline and lasso models

Table 5: Lasso and Baseline AR(1) Model Estimation Results for LR Spread

Class	Indicator	log(LR)	Baseline (LR)
	(Intercept)	0.528	0.310
	Lag	0.813	0.886
nonfinancial	MortgageDebtToIncome	0.036	
nonfinancial	ConsumerCredit	-0.067	
nonfinancial	HouseholdDebtToIncome		
nonfinancial	CommercialRealEstateCredit		
nonfinancial	CorporateCredit	0.023	
nonfinancial	CorporateDebtToIncome		
nonfinancial	CurrentAccount		
nonfinancial	ForeignAssetPositionGDP	0.022	
nonfinancial	ExternalDebtToGDP		
asset	HousePriceToIncome		
asset	HousePriceToRent		
asset	HousePriceGrowth		
asset	InvestmentGradeBondSpread	0.007	
asset	EquityPEratio	0.102	
credit	SpreadNewMortgage		
credit	DiffSpreadHighLowLTV	0.031	
credit	SpreadHouseholdLending		
Observations		76	76
R^2		0.787	0.751
Test MSE		0.706	0.722
Improvement (%)		2.178	

First, the lasso results in Table 5 show that the model is sufficiently regularized, with 7 of 17 features selected and a test MSE initially decreasing in λ. The one major issue with this model, though, is that Fig. 14 shows that the test error is significantly higher than the training error at all λ, even though I compensate for the sparser dataset by expanding the size of the training set. As discussed previously, training and test errors can be interpreted as visualizations of the bias-variance tradeoff, and a training error consistently lower than the test error for all λ means that while the coefficient estimates are not expected to be biased, there is high variance in the estimates. Essentially, no matter how much regularization is applied, the models still overfit the training set and basically model random noise within that data. This is one of the risks of modeling in data-sparse environments – generalization is always more difficult with fewer observations on which to train and test. The sparsity of the U.K. dataset explains the significantly elevated test MSE for both the baseline and lasso models. The log LR spread has a range similar to the log TED spread, but the baseline test MSE for the log LR spread is about seven times that of the baseline model for the log TED spread. Another clue pointing toward overfitting issues is the high R^2 value for both models – at 0.751 for the baseline and 0.787 for the lasso, both R^2 values are by far the highest of any of the models fit so far. However, this only means the models explain most of the variance in the training set – it says nothing about their ability to generalize to the test set, which in this case is not impressive. To that end, the U.K. indicators are only responsible for an improvement in out-of-sample performance of 2 percent over baseline, along with an increase of only 4 percent in R^2. Keeping in mind that R^2 is strictly nondecreasing in the number of explanatory variables, a 4 percent increase when adding 8 features is trivial and probably does not demonstrate that the indicators explain any additional variance in the log LR spread compared to the baseline, nor do the added variables provide a nontrivial increase in out-of-sample forecast accuracy compared to the baseline.

Continuing with the analysis framework set up in the U.S. context, I now interpret the above lasso results as a feature selection method, and fit the model with all selected features via OLS to perform hypothesis testing and determine which of the selected features have a statistically significant effect on the log LR spread forecast. These results are reported in Table 6.

Table 6: OLS Model Results for Selected Features, LR Spread Model

		log(LR)
nonfinancial	MortgageDebtToIncome	0.070*
		(0.041)
nonfinancial	ConsumerCredit	−0.147*
		(0.081)
nonfinancial	CorporateCredit	0.116*
		(0.070)
nonfinancial	ForeignAssetPositionGDP	0.051
		(0.033)
asset	InvestmentGradeBondSpread	−0.010
		(0.076)
asset	EquityPEratio	0.119***
		(0.045)
credit	DiffSpreadHighLowLTV	0.053
		(0.071)
	.lagLR	0.848***
		(0.083)
	Constant	0.421
		(0.265)
Observations		76
R^2		0.798
Adjusted R^2		0.773
Test MSE		0.708
Note:	*p<0.1; **p<0.05; ***p<0.01, Newey-West standard errors.	

The model results show that 4 of 7 selected features are significant at the 10 percent level or greater. As in the earlier analysis, we might have expected the lasso model to select almost exclusively significant features, but this is again not the case (though 4 of 7 is at least a majority). The test MSE for the OLS model is indistinguishable from that of the lasso model, which itself is basically the same as the AR(1) model. This is a disappointing result that shows that the U.K. indicators do not necessarily add predictive power to the log LR spread forecast, and it is confirmed by the adjusted R^2 for the OLS model, which at 0.773 is only 2 percent higher than the AR(1) baseline – as expected given the lasso model, the additional variance explained by the U.K. financial stability indicators is marginal at best.

The indicators with statistically significant effects on the log LR spread include several expected features: mortgage and total consumer debt-to-income ratios and consumer credit levels, each of which have the expected positive sign in the lasso estimation, though the sign on the consumer credit feature flips when the model is estimated via OLS (recall again that an increasing LR spread corresponds to increasing funding risk, so growing levels of debt, which should be stressors for the financial system, should elevate the spread). Mortgage debt and consumer credit levels both had significant effects on the log TED spread as well, but the last significant feature in the Aikman et al. dataset, the P/E ratio of the main stock market index (the FTSE All-Shares Index for the U.K., and the S&P 500 for the U.S.) has a significant positive effect on the LR spread, while the equivalent P/E ratio for the U.S. dataset was not selected by the lasso model. For the U.K. model, the coefficient on this feature has the expected sign, as a higher P/E ratio signals elevated risk appetite and is considered in both Kiley et al. (2015) and Aikman et al. (2018) to correspond to buildup of financial vulnerabilities. Interestingly, one of the main results seen in the U.S. analysis is that most asset pricing variables do not tend to exert a significant (or really any) effect on funding spreads – both the P/E ratio of the S&P 500 and the Baa bond spread were not selected by the lasso model. However, both equivalent variables for the U.K. were selected, and the P/E ratio is significant at 1 percent here, with the second-largest effect size of the selected features.

This result can be interpreted as an indication that the determinants of funding liquidity in the U.K. economy do not fully parallel those of the U.S. economy. However, it should also be considered that this robustness check does not give a true comparison between the two economies, as the U.K. dataset leaves out some key indicators that had significant effects for the U.S. Notably, financial sector vulnerability measures – three of which had highly significant effects on the log TED spread – are not available continuously for the U.K. and are therefore not included in the dataset. This could cause omitted variable bias, such that some of the effects that would have been attributed to the financial sector variables if they were present in the final dataset were instead attributed to asset pricing variables. Given the data availability constraint, there is not necessarily a clear solution to this issue. To conclude analysis of the lasso and OLS models for funding spreads in the U.K. economy, the results have shown that the sparser dataset of U.K. financial stability indicators does not carry significant information about funding spreads in the U.K. economy. Forecast accuracy when adding the vector of indicators to the model does not appreciably increase over the AR(1) baseline, which itself has a test error much higher than expected.

3.2.3 Nonparametric Methods

Finally, I test the same array of neural network architectures as implemented earlier for the U.S. dataset, to test whether nonparametric methods can outperform regression models in the U.K. case. Again, each architecture was trained with a stochastic gradient descent optimizer at learning rate $\eta = 0.025$ over 100 epochs, and the architecture descriptions and results are reported in Table 7.

Table 7: Neural Network Architectures and Results for log(LR) Forecast

Units (1)	Units (2)	Dropout (1)	Dropout (2)	Activation (1)	Activation (2)	Test MSE	Improvement (%)
8				relu		0.544	24.6
8		0.4		relu		0.861	
12		0.4		relu		0.954	
16		0.4		relu		0.795	
8		0.25		relu		0.698	3.2
12		0.25		relu		0.784	
16		0.25		relu		0.724	
8		0.25		sigmoid		2.239	
12		0.25		sigmoid		1.476	
16		0.25		sigmoid		1.523	
8	8	0.4	0.4	relu	relu	1.153	
12	12	0.4	0.4	relu	relu	1.103	
16	16	0.4	0.4	relu	relu	1.237	
8	8	0.25	0.25	relu	relu	1.436	
12	8	0.4	0.25	relu	relu	1.696	
16	8	0.4	0.25	relu	relu	1.435	
8	8	0.25	0.25	relu	sigmoid	2.171	
12	8	0.25	0.25	relu	sigmoid	2.092	
16	8	0.25	0.25	relu	sigmoid	1.512	

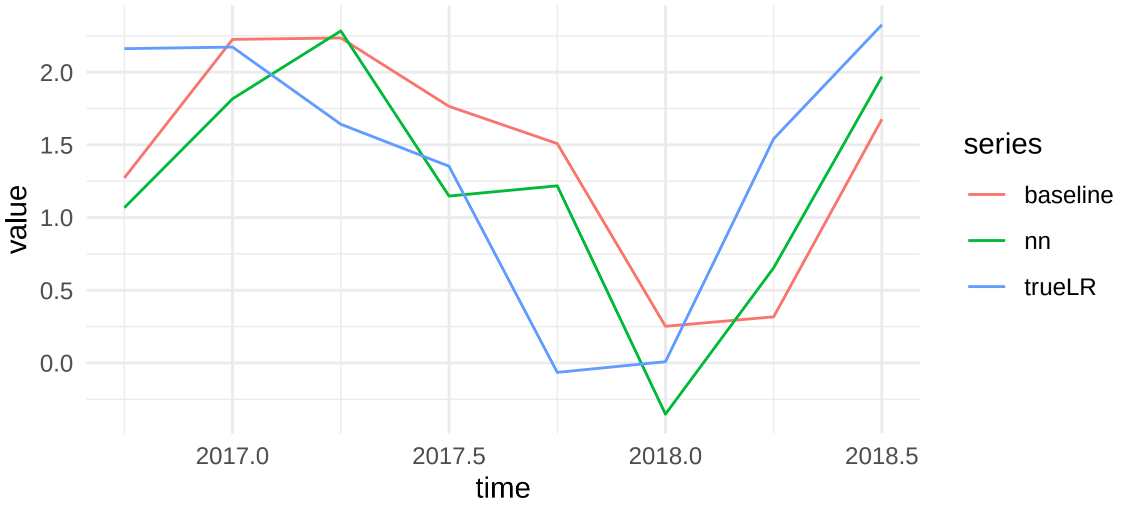

Figure 16: Predicted and actual LR spread for AR(1) baseline and neural network models

These results illustrate the earlier discussion about the influence of neural network architectures on the bias-variance tradeoff of a given model: we can see that the more complicated architectures have consistently elevated test MSE compared to the simpler ones with fewer nodes and layers. While this effect was somewhat visible in the analysis of the U.S. dataset, it is much clearer here. That said, the simplest model – one layer with 8 nodes, ReLU activation, and no dropout, performs by far the best on the U.K. data, resulting in a 24.6 percent improvement in out-of-sample performance over the AR(1) baseline. This is a surprising effect size given that the lasso model was by all measures unsuccessful in improving forecast accuracy over the baseline, and it confirms that, at least for the FSS and the U.K. data, fitting the model with a simple neural network results in a nontrivial increase in forecast accuracy over the baseline.

Unfortunately, again, neural networks are a "black box" nonparametric machine learning method – therefore, even though these results demonstrate that neural networks increase the forecast accuracy for liquidity in the U.K., it is not apparent exactly how they do so, and there is no equivalent to the lists of coefficients and hypothesis tests provided by more traditional regression methods to determine the most predictive financial stability indicators in this context. While it would be ideal to construct an interpretable model that also performs better than the baseline, at least for the U.K. economy this does not seem to be possible, in all likelihood because of data sparsity and constraints on availability over the sampling period.

4 Discussion

The main goals of this project were to contribute to the financial stability literature by determining the signal carried by the indicators chosen by Kiley et al. (2015) and Aikman et al. (2018); to develop models that forecast bank funding spreads, and therefore the overall risk to financial stability, more accurately than simple time series methods alone; and to investigate the utility of machine learning techniques beyond more traditional econometrics in building these forecasting tools. Overall, results were mixed – when considered as a whole, there are some encouraging components to this project, but more work undoubtedly needs to be done to develop them further.

With regard to the first goal, validation of the information content of financial stability indicators identified in previous work, I claim moderate success. The heatmaps and composite indices of financial stability generated in Kiley et al. (2015) reperesent a valuable starting point and provide a strong theoretical framework for data-driven financial stability analysis, but the empirical validation of each of the indicators provided by Kiley et al. does not go beyond demonstration of a causal relationship between the *composite index*, not the individual components, and the credit-to-GDP gap, a generally agreed-upon leading indicator of financial instability. Further, the aggregation methods used are agnostic to the importance of each indicator – by combining indicators via simple arithmetic and geometric averages, the weight on each indicator is uniformly the reciprocal of the number of indicators. This weighting scheme effectively values indicators that carry more information equally to those that might be repetitive or only add noise to the indicator matrix. While Kiley et al. try using the first principal component of the normalized indicator matrix as weights, arguably a more effective strategy as it guarantees that the indicators contributing most to the variance within the dataset are weighted highest, the final index is a simple arithmetic average of indicator values, effectively discarding information as to the relative importance of each indicator in quantifying systemic risk.

To build on this result, I sought to directly measure the predictive power and significance of each indicator, ideally building on the broad demonstration of a correlation between a composite index of the indicators and endogenous measures of the financial cycle. The first step to solving this problem was choosing a response variable exogenous to the indicators themselves – unlike the credit-to-GDP gap mentioned in Kiley et al. – as a proxy for systemic risk, the idea being that the indicators that have the largest predictive power for this response variable are the ones that are the most important in measuring systemic risk. One positive and robust result from this work is the choice of bank funding spreads in the period ahead as a suitable response variable. Besides the theoretical backing that the literature gives for the relationship between liquidity and financial stability, the analysis in this work confirms this connection in both the U.S. and U.K. economies: scatterplots of the first two principal components of each indicator matrix colored by the log funding spread for each economy demonstrate that the log funding spread one period ahead varies with a clear gradient along the first principal component, or the component explaining the most variance, in each indicator matrix. While not conclusively demonstrative of a causal relationship between liquidity and the financial stability indicators, it does show robustly and generally that the level of the log funding spread is related to the levels of the indicators, validating the choice of funding spreads as a response in modeling financial stability over time.

Notwithstanding this promising unsupervised analysis, the actual effects of each indicator on the funding spread in each economy, as determined by lasso and OLS regressions, were somewhat ambiguous. The features selected by the lasso model – which are guaranteed to contribute to forecast accuracy over baseline – were not all significant when re-fit by OLS, and several indicators had coefficients with signs opposite to what would be expected given the past financial stability literature. Most notably, the housing price-to-rent ratio for the TED spread model and consumer credit levels for both the TED spread and LR spread models had highly significant negative effects on their respective funding spreads, despite being almost canonical early-warning indicators for the buildup of financial vulnerabilities. One of the key assumptions I make in modeling the effect of the indicators on funding spreads is that the time horizon for each indicator to affect the financial system in a quantifiable way is constant, and that this constant time horizon is one quarter. Admittedly, this choice was based on data availability more than anything else, as larger lags would have further reduced the effective size of the datasets and therefore made the models even harder to train accurately. It could be, however, that the time horizon is not constant across indicators, or that a one-quarter lead is not enough time for the more long-term indicators with less quarter-to-quarter variation to exert a meaningful effect on bank funding spreads. In this case, a VAR-driven model with multiple lags of each of the selected indicators might be a more productive approach to identifying the dynamic effects of the indicators on overall financial stability. Future work should either focus on a strategy along these lines to validate the results presented here, or add more lags to the existing models to pick up longer-term indicator effects on systemic risk buildup.

All that said, though, if we interpret the selected features as the ones that have the strongest effect on funding liquidity *in the short run*, the results make more sense. For the TED spread, I find that the most significant features are the average bank TCE ratio (which measures whether banks are adequately capitalized and not over-leveraged), which has the expected negative effect on the TED spread and is significant at 1 percent; and the percentage of banks tightening standards for commercial real estate loans, which is also significant at the 1 percent level. While it makes sense that the TCE ratio would have a negative effect on the TED spread at any time horizon (as a better-capitalized banking system is likely more robust to shocks and better able to absorb counterparty risk), a higher percentage of banks tightening standards for commercial real estate or mortgage loans is expected to decrease vulnerabilities over the medium term, due to the elevated risk appetite signaled by banks effectively loosening lending standards as in the lead-up to the 2008 financial crisis. However, in the short run, we might expect tightening lending standards to be associated with an increase in the funding spread, as tightening lending standards might indicate that banks are already wary of a riskier lending environment and therefore signal increased credit risk, reflected in the funding spread one quarter ahead. Therefore, the results presented here should be interpreted carefully – the models successfully isolate the variables associated with bank funding spreads in the short run, but the indicators with the most significance when examined via this approach are by no means the only ones that inform financial stability at all time horizons.

To test whether the results generalize to different measurements of the financial cycle, I utilize the financial stability sentiment score (FSS) introduced by Correa et al. (2017) to quantify risk via the information content of FOMC minutes. I hypothesized that the FSS would function in parallel to the TED spread in measuring systemic risk, as Correa et al. showed that the FSS (when generated using central bank financial stability reports) is significantly correlated with contemporaneous risk buildup in financial sector variables. I further assumed that the same set of indicators would significantly affect both response variables, confirming their importance to the buildup of systemic risk via two independent measures.

Unfortunately, this hypothesis was incorrect: the lasso model for the FSS only selected four features, which (although all overlapped with the features selected by the TED spread model) were not the ones with the highest significance or the largest effect size in the FSS model, and some had the opposite sign or drastically different effect size between models. The lack of robustness in the combined results can be attributed partially to the noise in the FSS itself, which probably arises because the financial stability dictionary was created with financial stability reports ("FSRs"), not FOMC minutes, in mind. All else equal, FSRs have much more information and sentiment related to financial stability than FOMC minutes, and therefore a higher signal-to-noise ratio in the underlying index, creating a more robust and clear FSS than the one I generate here to measure systemic risk in FOMC minutes. In fact, the autoregressive component of the FSS is not significant in the OLS model, suggesting a small ρ if the FSS even is an autoregressive process in the first place.[10] Overall,

[10] Correa et al. also introduce some modified versions of the FSS algorithm in their paper, which I explore in the appendix, though I find that the results are not markedly different.

the analysis presented here did identify some short-run determinants of bank funding liquidity in the United States, partially validating Kiley et al.'s approach to financial stability monitoring in the United States, even though the FSS approach was less successful.

The next two goals for the project go hand-in-hand, as I sought to develop more accurate forecasts using the high-dimensional indicator datasets via machine learning methods, thereby demonstrating the value of these less-traditional econometric techniques in financial stability analysis and macroprudential policy more generally. There was more consistent success with respect to these goals. The problem of generating a meaningful signal from an inherently noisy and multicollinear high-dimensional dataset of financial stability indicators is nontrivial, and traditional regression methods alone would not have been capable of fitting robust models that generalize well out-of-sample in such a data-rich environment (especially with the comparative lack of observations in the datasets) [Stock and Watson, 1999] [Stock and Watson, 2002]. The approach using lasso regression to select features from the datasets and guard against overfitting was successful for the TED spread, producing a forecast that performed almost 14 percent better out-of-sample than a baseline autoregressive model. Not only does this result demonstrate the predictive value of the financial stability indicators in forecasting liquidity, but it also displays the value of lasso regression in generating forecasts in a data-rich environment.

Further, though lasso regression was not as successful in forecasting the FSS or the LR spread, simple neural networks performed well on these noisier time series, bringing the forecast improvement over baseline up to and over the levels seen for the TED spread model. While modeling the FSS with a neural network resulted in over 15 percent forecast improvement (about on line with the lasso model's improvement on the TED spread), the neural network model for the LR spread produced over 24 percent improvement over baseline, the largest improvement of any of the models in this work (though given the small size of the test set for the U.K. data and the nondeterministic nature of neural network outputs, the "confidence interval" for the 24 percent improvement statistic is necessarily large).

This work has therefore succeeded in developing more accurate forecasts of liquidity and sentiment using financial stability indicators, improving forecast accuracy of each response variable in this work by at least 15 percent over the autoregressive baseline. Accordingly, my results suggest that not only nontraditional regression methods, but also nonparametric methods traditionally reserved for applications with orders of magnitude more data than we have here, can in fact contribute to financial stability analysis. Constructing optimal monetary and macroprudential policy in economies with ever-increasing data collection is an opportunity and a challenge considered and faced by many central banks, and this project contributes an analysis framework that could complement central banks' forecasting tools in the future [Bernanke and Boivin, 2001].

5 Conclusion

Broadly, this work has shown that data monitored by central bank financial stability authorities and highlighted in the literature for their ability to quantify risk are in fact nontrivially predictive of systemic risk and financial system strain as measured by funding spreads and financial stability sentiment. Including the indicators in forecasts of proxies for systemic risk improves accuracy by at least 15 percent compared to forecasts based on only previous values of the proxies themselves. This result shows the utility of the additional data and the forecasting methods used in gaining a complete understanding of the state and stability of the financial system. Since liquidity has been shown to be a strong proxy for real economic activity [Goldberg, 2016], accurate forecasting of liquidity via bank funding spreads could form an integral part of central banks' and financial stability authorities' outlooks and inform the optimal path of monetary and macroprudential policy, both in terms of the policy interest rate and the countercyclical capital buffer ("CCyB")[11] [Basel Committee on Banking Supervision, 2018]. The CCyB seems to be the most logical policy endpoint for the forecasts developed here, given that central banks and financial stability authorities could react to concerning liquidity forecasts by directly adjusting the CCyB, thereby ensuring that banks are better-capitalized and more prepared to absorb possible adverse events stemming from the buildup of vulnerabilities in the near future.

Future work along these lines might focus on the (possible) non-constant time horizon within which each indicator affects the funding spread. A VAR-driven approach including selected indicators at multiple lags would be valuable in identifying the lags at which each indicator exerts the most significant effect on future liquidity. To that end, though, more data is necessary. Especially

[11] The CCyB ensures that banks are adequately capitalized given the current macroeconomic environment pursuant to the Basel III banking resiliency framework.

with the approaches outlined here, more data is always better, and the results presented here would undoubtedly be more robust if training models on datasets with more than 90 observations were possible. A more thorough search of the primary literature might identify additional indicators – possibly even indicators available at higher frequency than the quarterly data I have worked with here – that could add further value to liquidity forecasts and enable more granular estimates of future systemic risk. The value of data-driven monetary and macroprudential policy cannot be overestimated, and this work will hopefully inspire future efforts to explore the utility of machine learning and "big data" in policymaking and forecasting environments.[12]

References

[Aikman et al., 2018] Aikman, D., O'Neill, C., Levina, I., Galletly, R., Burgess, S., Bridges, J., and Varadi, A. (2018). Measuring risks to UK financial stability. *Bank of England Staff Working Papers*, (738).

[Awazu Pereira da Silva and von Peter, 2018] Awazu Pereira da Silva, L. and von Peter, G. (November 2018). Financial instability: can big data help connect the dots? *Bank for International Settlements.*

[Basel Committee on Banking Supervision, 2018] Basel Committee on Banking Supervision (March 2018). Towards a sectoral application of the countercyclical capital buffer: A literature review. *Bank for International Settlements.*

[Benos and Zikes, 2016] Benos, E. and Zikes, F. (May 2016). Liquidity determinants in the uk gilt market. *Bank of England Staff Working Papers.*

[Bernanke and Boivin, 2001] Bernanke, B. S. and Boivin, J. (July 2001). Monetary policy in a data-rich environment. *National Bureau of Economic Research.*

[Board of Governors of the Federal Reserve, 2018] Board of Governors of the Federal Reserve (November 2018). Financial stability report.

[12] All code and data used here are available at github.com/metzner28/liquidity for further exploration – if you've read this far, please go check it out!

[Boudt et al., 2017] Boudt, K., Paulus, E. C., and Rosenthal, D. W. (2017). Funding liquidity, market liquidity and TED spread: A two-regime model. *Journal of Empirical Finance*, 43:143–158.

[Brunnermeier, 2009] Brunnermeier, M. K. (2009). Deciphering the liquidity and credit crunch 2007-2008. *Journal of Economic Perspectives*, 23(1):77–100.

[Correa et al., 2017] Correa, R., Garud, K., Londono, J. M., and Mislang, N. (2017). Sentiment in central banks' financial stability reports. *International Finance Discussion Papers*, 1203.

[Drehmann et al., 2012] Drehmann, M., Borio, C., and Tsatsaronis, K. (June 2012). Characterising the financial cycle: don't lose sight of the medium term! *Bank for International Settlements Working Papers*.

[Goldberg, 2016] Goldberg, J. E. (2016). The supply of liquidity and real economic activity. *Federal Reserve Board*.

[James et al., 2017] James, G., Witten, D., Hastie, T., and Tibshirani, R. (2017). *An introduction to statistical learning: with applications in R*. Springer.

[Kiley et al., 2015] Kiley, M. T., Aikman, D., Lee, S. J., Palumbo, M. G., and Warusawitharana, M. (2015). Mapping heat in the US financial system. *FEDS Notes*, 2015(1574).

[Kyung et al., 2010] Kyung, M., Gill, J., Ghosh, M., and Casella, G. (2010). Penalized regression, standard errors, and bayesian lassos. *Bayesian Analysis*, 5(2):369–412.

[Lockhart et al., 2014] Lockhart, R., Taylor, J., Tibshirani, R. J., and Tibshirani, R. (2014). A significance test for the lasso. *The Annals of Statistics*, 42(2):413–468.

[McLaughlin et al., 2018] McLaughlin, J., Palmer, N., Minson, A., and Parolin, E. (March 2018). The OFR financial system vulnerabilities monitor. *Office of Financial Research Working Papers*.

[Nakamura, 2005] Nakamura, E. (2005). Inflation forecasting using neural networks. *Economics Letters*, 86:373–378.

[Srivastava et al., 2014] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958.

[Stock and Watson, 2002] Stock, J. H. and Watson, M. W. (April 2002). Macroeconomic forecasting using diffusion indices. *Journal of Business & Economic Statistics*, 20(2):147–162.

[Stock and Watson, 1999] Stock, J. H. and Watson, M. W. (July 1999). Forecasting inflation. *Journal of Monetary Economics*, (44):293–335.

[Terasvirta et al., 2005] Terasvirta, T., van Dijk, D., and Medeiros, M. (2005). Linear models, smooth transition autoregressions, and neural networks for forecasting macroeconomic time series: a re-examination. *International Journal of Forecasting*, 21:755–774.

[Zhang, 2012] Zhang, G. P. (2012). Neural networks for time series forecasting. *Handbook of Natural Computing*, pages 462–474.

6 Appendix

6.1 FSS Extensions

The analysis of the U.S. economy above was complicated by the fact that the FSS score response variable was highly noisy – by inspection, it is clear that the time series of the FSS as specified by Correa et al. and applied to FOMC minutes does not resemble that of the TED spread, calling into question my choice of the FSS as a suitable proxy for systemic risk. Even when applied to financial stability reports (which, as discussed above, should have a much higher signal-to-noise ratio), Correa et al. still find several noisy and surprising results, so they introduce two extensions of the FSS scoring algorithm, which I investigate here. Recall that the FSS is specified by Correa et al. as the difference between the number of words in a given text classified as "negative" and "positive" in their financial stability dictionary, divided by the total number of words in the text. Now, the "negative FSS" or "FSS-" omits the positive words from the numerator, so it is simply the number of negative words over the total; and the "excitement FSS" or "FSS*" takes the sum, rather than the difference, of negative and positive words as the numerator. There are reasonable arguments to be made for each of these scoring algorithms as opposed to the original. Words classified as positive by the dictionary might not actually signal mitigation of central banks' outlooks and might only add noise to the negative sentiment signal carried by the negative-classified words. Alternatively, both negative and positive words might signal increased focus of central bankers on financial stability, and they therefore might carry equivalent information content – adding the two together might boost the overall signal in the scores at each period. All three of the above scoring algorithms were applied to quarterly FOMC minutes from March 1990 to June 2019, and time series of the scores, each normalized to zero mean and unit variance, are included here.

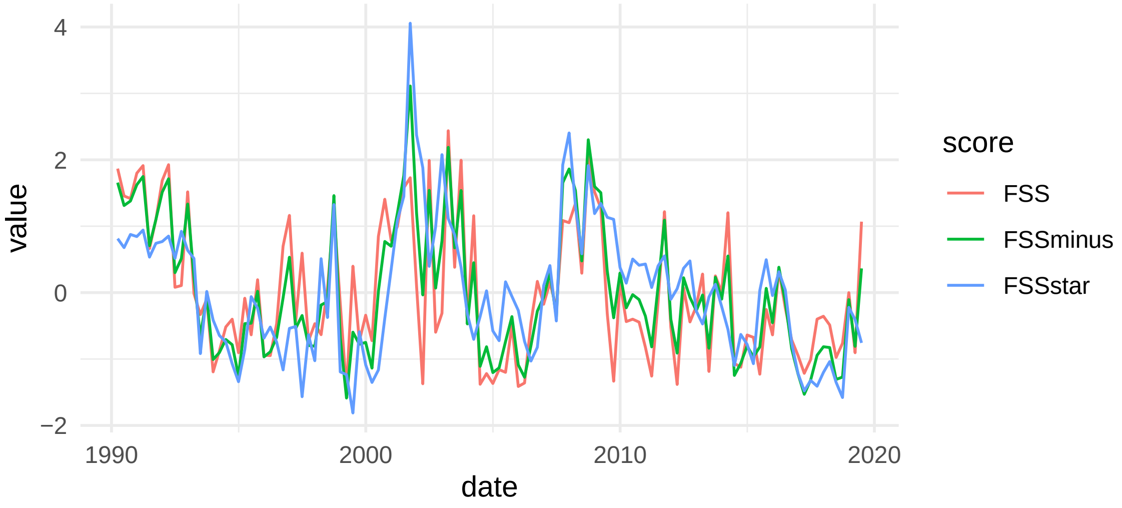

Figure A1: Time series of FSS and Extensions

The scaled scores look relatively similar – there are not many clear differences evident in the above figure besides a clear peak for both the negative and excitement scores around 2001 which is not mirrored by the original FSS. Based on the time series graphs alone, it seems unlikely that the FSS extensions will change the previous modeling results significantly, but I rerun the models anyway for robustness. Using the same analysis framework as above, I run lasso models for the FSS (simply reproducing the results from the analysis above), FSS-, and FSS*, and then test the significance of the selected features in each model by OLS. Lasso and OLS model results are included in Tables A1 and A2, respectively.

Table A1: Lasso model results for FSS and Extensions

Class	Indicator	FSS	FSS-	FSS*
	(Intercept)	0.057	0.060	0.040
	.lagFSS	0.235	0.465	0.562
financial	GrossShortTermWholesaleDebtFinancialSector			
financial	BrokerDealerLeverage			0.0002
financial	NetShortTermWholesaleDebtFinancialSector			
financial	RunnableLiabilitiesGDP			
financial	TangibleCommonEquityRatio			
financial	CapitalRatio			
financial	FinancialSectorLiabilitiesGDP			-0.089
nonfinancial	CorporateSavings			
nonfinancial	ConsumerCredit			0.005
nonfinancial	ConsumerDebtService			0.065
nonfinancial	CorporateDebtToIncome			
nonfinancial	CorporateDebtGrowth			
nonfinancial	HouseholdSavings	0.163	0.091	
nonfinancial	InterestExpenses			-0.035
nonfinancial	MortgageDebt	0.052		0.021
nonfinancial	MortgageDebtService			0.045
asset	InvestmentGradeBondSpread			
asset	EquityPEratio			0.059
asset	HousePriceToRent			-0.050
asset	HighYieldBondSpread			
asset	CommercialRealEstateIndex			
asset	TighterStandardsCILoans			
asset	VIX			
asset	CorporateLeverageIndex	-0.092	-0.077	-0.188
asset	TighterStandardsCRELoans			
asset	TighterStandardsMortgageLoans	0.091	0.072	0.054
Observations		90	90	90
R2		0.246	0.434	0.568
Test MSE		0.550	0.455	0.297
Improvement (%)		4.566	-9.331	-15.336

Table A2: OLS Model Results for FSS Extensions

		FSS	FSS-	FSS*
nonfinancial	HouseholdSavings	0.268*** (0.077)	0.198*** (0.044)	
financial	BrokerDealerLeverage			0.068 (0.076)
financial	FinancialSectorLiabilitiesGDP			−0.150* (0.086)
nonfinancial	ConsumerCredit			−0.008 (0.091)
nonfinancial	ConsumerDebtService			0.129 (0.142)
nonfinancial	InterestExpenses			−0.064 (0.074)
nonfinancial	MortgageDebt	0.190* (0.111)		0.083 (0.060)
nonfinancial	MortgageDebtService			0.041 (0.066)
asset	EquityPEratio			0.120* (0.068)
asset	HousePriceToRent			−0.111 (0.086)
asset	CorporateLeverageIndex	−0.246 (0.163)	−0.174 (0.107)	−0.269*** (0.080)
asset	TighterStandardsMortgageLoans	0.238*** (0.047)	0.208*** (0.049)	0.057 (0.065)
	lagFSS	0.236 (0.145)		
	lagFSSminus		0.487*** (0.081)	
	lagFSSstar			0.530*** (0.092)
Constant		0.044 (0.085)	0.045 (0.075)	0.029 (0.073)
Test MSE		0.618	0.501	0.314
Observations		90	90	90
R^2		0.301	0.469	0.582
Adjusted R^2		0.259	0.444	0.517
Note:	*p<0.1; **p<0.05; ***p<0.01, Newey-West standard errors.			

First, the results reveal a clear similarity between the FSS and the FSS-: the lasso model selects four features for the original FSS, of which three – household savings, corporate leverage, and the percentage of banks tightening standards for mortgage loans, are selected as the only features in the FSS- model. These three features have the same sign and similar effect sizes in both models, and hypothesis testing via OLS reveals that each also has the same significance level in each model. Household savings and tighter standards for mortgage loans, both nonfinancial sector indicators, have a positive effect on the FSS and FSS- that is significant at the 1 percent level, while the effect of the corporate leverage index, an asset pricing variable, is not significant for either model. As in the original FSS model, the results have unexpected sign, as increases in household savings levels should provide a buffer against financial shocks and therefore a decrease in overall systemic risk – they should then have a negative effect on both the FSS and FSS-, which both increase in response to buildup of vulnerabilities. Similarly, tightening standards on mortgage loans signals decreased risk appetite by lenders and should also decrease overall vulnerabilities, but it has a positive effect on both the FSS and FSS-. As explained in the original analysis of the FSS model, it is straightforward to see how a one-quarter lag would not be sufficient to pick up the vulnerabilities induced by loosening mortgage loan standards, and instead how the model would associate tightening lending standards in an already-deteriorating financial environment with elevated sentiment indices. The FSS* does not seem to mirror the results for the other two indices, as the FSS* lasso model selects 11 features, of which only two, the ratio of financial sector liabilities to GDP and the corporate leverage index, are significant when tested by OLS. Notably, household savings levels, which were significant at the 1 percent level in the FSS and FSS- models, is not even a selected feature for the FSS*, and the only highly significant feature in the FSS* model is the corporate leverage index, which while selected for the other two indices, was not significant.

Unlike the FSS model, the autoregressive terms for both the FSS- and FSS* models were significant at the 1 percent level, confirming that these two indices are in fact autoregressive processes. Because the autoregressive term is highly significant in both of these cases, though, the lasso model does not improve forecast performance for either index – in fact, the lasso model for the FSS- has a test MSE 9.3 percent higher than its AR(1) baseline, while that of the FSS* model was 15.3 percent higher than baseline. Since the models of the FSS extensions failed in improving forecast performance over the autoregressive baseline – and in fact decrease performance, even in a regularized setting – we conclude that the financial stability indicators carry no meaningful information about the FSS- or FSS*. For both extensions, the autoregressive terms were highly significant, while this was not the case for the original FSS, for which adding the financial stability indicators did result in a marginal increase in forecast accuracy over baseline. This makes sense in light of the lack of predictive power added by the indicators to the extension forecasts – if both extensions are purely autoregressive processes, exogenous explanatory variables would only add noise to the time series signal and result in decreased forecast performance, which is clearly visible in the case of the FSS- and FSS*. It remains confusing, then, that some of the indicators have statistically significant effects on the extensions if they do not improve out-of-sample performance for the respective models. This apparent contradiction is a good reminder of the limitations of my analysis framework: it is not guaranteed that features selected by the lasso model to improve out-of-sample performance will be significant when tested by OLS, nor is it guaranteed that features with a significant effect in an OLS model will meaningfully improve out-of-sample performance – besides selected features, there is no correspondence between the fitted lasso and OLS models for each response variable, and since the models are fit in different ways subject to different objective functions, the results may not always line up.

The results here favor an interpretation of financial stability sentiment in which words classified as "negative" by the financial stability dictionary contribute the most information content to the sentiment score, as the model results for the FSS- are in general very similar to those for the original FSS. That is, removing the subtraction of "positive" words from the FSS algorithm does not meaningfully alter the features that affect the score, nor the statistical significance or effect sizes of those features. Removing the subtraction of positive words from the algorithm does, however, restore statistical significance to the autoregressive term in the model and cause the AR(1) baseline forecast to perform better than the lasso model, which supports the conclusion that the time series of the proportion of *negative words only* in the FOMC minutes is an autoregressive process, while the time series of the *net negative words* – where words classified as positive by the dictionary effectively cancel out negative words – is not necessarily a purely autoregressive process, and its forecast is improved by taking into account the financial stability indicators.

Adding the positive and negative words together, as in the FSS* model, results in a meaningfully different set of features affecting the forecast compared to the FSS and FSS-, but most of these features are not statistically significant when tested via OLS. The FSS*, therefore, seems to be a noisier measure of financial stability sentiment than the FSS and FSS-. Neither of the extensions of the FSS seem a much clearer choice as a response variable than the original FSS, and the model results for the FSS have shown some robustness given the FSS- results as well. Therefore, while the FSS may be a noisy measure of financial stability, it does show some internal consistency and robustness to modifications in calculation in terms of the indicators that exert the largest effect on its forecast. However, the forecast improvement due to the financial stability indicators, which was marginal but present for the original FSS, is not similarly robust and does not appear for the FSS extensions, which instead are forecast most accurately by purely autoregressive models. Overall, the results suggest that the "core" of the FSS – the proportion of negative words – is a purely autoregressive process for which the indicators do not meaningfully improve forecast performance, but the *net* proportion of negative words is marginally informed by the financial stability indicators. The proportion of "total" financial stability words – the sum of negative and positive words – is a much noisier index for which several exogenous features are selected without a statistically significant effect on the scores or an improvement in forecast performance over baseline.

6.2 OLS Model Extensions

Here I fit the OLS models of the log TED spread and FSS to the full dataset, as opposed to only the training data, to test whether the significance of the selected features changes when including more observations. These results, as well as the previous results using only the training set from section 3.1, are shown in Table A4.

Table A3: OLS Model Results for log(TED) and FSS, Training Set and Full Dataset

		Dependent variable:			
		log(TED)	log(TED)	FSS	FSS
		(1)	(2)	(3)	(4)
financial	GrossShortTermWholesaleDebtFinancialSector	-0.251^{**}	-0.165^{**}		
		(0.123)	(0.065)		
financial	BrokerDealerLeverage	-0.053	-0.034		
		(0.055)	(0.053)		
financial	NetShortTermWholesaleDebtFinancialSector	0.299^{**}	0.226^{***}		
		(0.118)	(0.061)		
financial	TangibleCommonEquityRatio	-0.130^{***}	-0.123^{***}		
		(0.045)	(0.032)		
financial	FinancialSectorLiabilitiesGDP	-0.020	-0.026		
		(0.052)	(0.038)		
nonfinancial	ConsumerCredit	-0.116^{**}	-0.109^{***}		
		(0.051)	(0.036)		
nonfinancial	ConsumerDebtService	0.075	0.062		
		(0.066)	(0.050)		
nonfinancial	CorporateDebtToIncome	0.004	0.004		
		(0.026)	(0.023)		
nonfinancial	CorporateDebtGrowth	0.013	0.011		
		(0.047)	(0.035)		
nonfinancial	HouseholdSavings	-0.045	-0.038	0.268^{***}	0.176^{***}
		(0.045)	(0.033)	(0.077)	(0.058)
nonfinancial	MortgageDebt	-0.080^{*}	-0.062^{*}	0.190^{*}	0.146
		(0.043)	(0.034)	(0.111)	(0.095)
asset	HousePriceToRent	-0.106^{**}	-0.117^{***}		
		(0.051)	(0.034)		
asset	HighYieldBondSpread	0.058	0.024		
		(0.047)	(0.049)		
asset	CommercialRealEstateIndex	-0.076	-0.103^{**}		
		(0.069)	(0.044)		
asset	VIX	0.041	0.072^{*}		
		(0.053)	(0.043)		
asset	CorporateLeverageIndex	0.040	0.059	-0.246	-0.253^{*}
		(0.045)	(0.036)	(0.163)	(0.146)
asset	TighterStandardsCRELoans	0.086^{***}	0.086^{***}		
		(0.020)	(0.022)		
asset	TighterStandardsMortgageLoans	0.015	0.015	0.238^{***}	0.219^{***}
		(0.041)	(0.035)	(0.047)	(0.049)
	lagTED	0.531^{***}	0.569^{***}		
		(0.059)	(0.049)		
	lagFSS			0.236	0.260^{**}
				(0.145)	(0.120)
	Constant	1.743^{***}	1.573^{***}	0.044	-0.023
		(0.222)	(0.181)	(0.085)	(0.073)
Observations		90	116	90	116
R^2		0.721	0.708	0.301	0.280
Adjusted R^2		0.645	0.650	0.259	0.247
Note:	*p<0.1; **p<0.05; ***p<0.01, Newey-West standard errors.				

The results show relatively minimal differences between fitting the OLS models on the 90 observations of the training set vs. the 116 observations of the full dataset. For the TED spread model, all of the indicators that were significant on the training set retain that significance on the full dataset – three of the seven significant indicators are significant at the 1 percent level on the full dataset, as opposed to the 5 percent level on the training set, demonstrating that increasing the number of observations generally gives higher statistical significance. Additionally, two asset pricing indicators that were not significant on the training set, the VIX and the commercial real estate price index, gain significance at the 10 percent and 5 percent levels, respectively, on the full dataset. While this might mean that these variables only have a significant association with the TED spread over the past few years (the observation period of the test set), a more likely explanation is that their significance was borderline on the training set, and adding the additional observations in the full dataset decreased their standard errors to the appropriate confidence levels. Overall, this does not change the analysis of the TED spread model, though it is encouraging that our results are largely retained by adding more observations into the data from which the model is estimated.

For the FSS model, the significance levels of the largest coefficients, on household savings levels and the percentage of banks tightening standards on commercial real estate loans, do not change when more observations are added, though the difference between the results on the full dataset vs. the training set is larger than for the TED spread model. Mortgage debt, which was significant at the 10 percent level on the training set, loses significance on the full dataset. The corporate leverage index, which was not significant on the training set, gains significance at the 10 percent level on the full dataset. Both of these changes seem marginal, though, as the significance level is only borderline in the first place. More interestingly, though, while the autoregressive component of the FSS is not significant on the training set, it gains significance at the 5 percent level on the full dataset, providing some evidence that the FSS is in fact an autoregressive process given a large enough sample size. This was in doubt given the results on the training set, but we can be more confident in specification of the model considering these results. Notably, the value of ρ, the coefficient on the lag, does not change meaningfully, but the standard error decreases such that significance is gained at the 5 percent level. This result suggests that the underlying autoregressive component of the FSS is present and constant throughout the full sampling period, but since ρ is relatively small, a larger number of observations than only the training data were necessary to estimate ρ at high confidence.

6.3 TED spread and Treasury market liquidity

To support the evidence from the previous literature that the TED spread is a reasonable proxy for funding liquidity in the United States, I include here overlaid time series of dealer positions in Treasury markets and the TED spread, both normalized to zero mean and unit variance, weekly from June 1998 to February 2020, the period over which both series are publicly available.

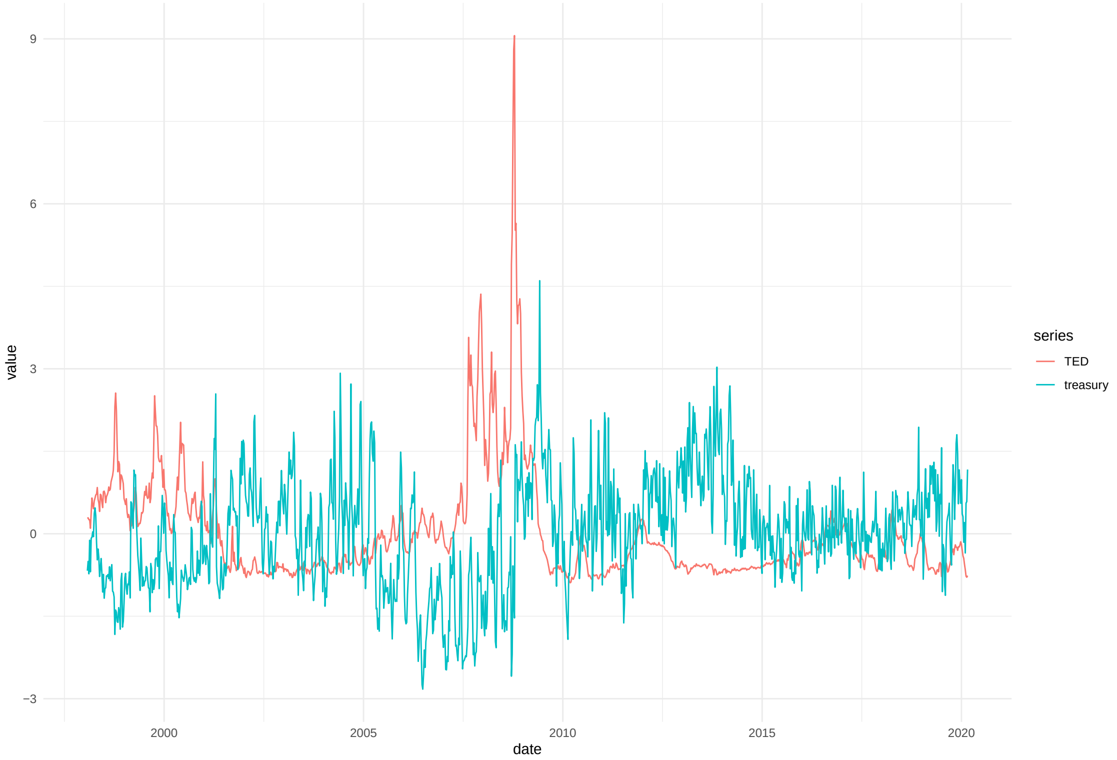

Figure A2: Time series of weekly dealer positions in Treasury markets and TED spread, 1998-2020

The visualization shows a clear inverse relationship between liquidity and the TED spread, as explained in Brunnermeier (2009) and Boudt et al. (2013). As liquidity is withdrawn from the market in the lead-up to the 2008 Crisis, the TED spread increases sharply, then returns to mean levels as QE and other expansionary monetary policy takes hold and injects funding liquidity into the financial system. This relationship lends further confidence to the choice of the TED spread as an appropriate bank funding spread to model and forecast liquidity.