	Machine Learning Models	

Improving External Validity of Machine Learning, Reduced Form, and Structural Macroeconomic Models using Panel Data

Cameron Fen and Samir Undavia

April 27th, 2021

tructural Models

Machine Learning Models

Conclusion 000

Motivation: How can we augment macroeconomic data to improve the external validity of our models?



Structural Models

Machine Learning Models 000000000 Conclusion 000

Motivation: How can we augment macroeconomic data to improve the external validity of our models?





Answer: Country-by-County Panel Data

- Augment the Size of our Data: We advocate fixing this dearth of data problem by using a panel of 49 other countries + US data
- External Validity: We offer evidence that using a panel dimension when estimating models improves forecasting performance in a wide variety of circumstances
- More Fundamental Conclusions: We demonstrate that this augmented data makes reduced form and structural models more policy invariant
- Flexible Models: We show that this procedure opens up the use of more flexible/non-parametric models like our 18,000 parameter recurrent neural network that outperforms all baselines in this data-rich regime



The Data and Baseline Models

- We evaluate the performance of our models via GDP forecasting
- We use growth rate data from 50 different developed countries (GDP, consumption, unemployment) to forecast GDP growth
- Data Sources: World Bank and Trading Economics via Quandl
- Test set: 2008Q4-2020Q1
- Models:
 - 1 AR(2)
 - 2 VAR(1)
 - 3 Smets Wouters 2007 DSGE
 - 4 Factor Model*
 - 5 Recurrent Neural Network
 - 6 AutoML

Introduction		Machine Learning Models	Appendix
00000			

The Baseline Models

Baseline Models

- AR(2) model: forecasts GDP with two lags in a linear manner
- VAR(1) and VAR(4) models: forecasts GDP using one lag from our three independent variables
- DSGE model: A structural model that uses economic theory to forecast
- Factor model: Uses large cross section of data (248 variables), condenses the information down into lower dimensional factors via PCA, and uses a linear model to regress GDP on the factors and a GDP lag

Appendix: Model Description

	Reduced Form Models	Structural Models	Machine Learning Models	Conclusion	Appendix
00000	000000	000000	00000000	000	000000000000000000000000000000000000000

Reduced Form Models

ntroduction 00000 Reduced Form Models

Structural Models

Machine Learning Models

Conclusi 000

Forecasting Improvement



Horizon-Model

ructural Models

Machine Learning Models

Conclusion 000

Improvement Not Obvious A Priori

- More data does not always lead to better forecasting performance
 Literature on negative transfer, see (Wang 2019) among many
- The new data is not a representative of US GDP performance as it is a different country-biased
- Be However the increase in data will also reduce variance of our models

Structural Models

Machine Learning Models

Conclusion 000

Forecasting Performance Not Constrained to US

Table 1: Average Forecasting Performance

Time (Q's Ahead)	1Q	2Q	3Q	4Q	5Q
AR(2)					
Local Data	5.1	5.1	5.3	5.4	5.4
World Data	4.8	4.8	5.0	5.1	5.1
VAR(1)					
Local Data	5.0	5.0	5.4	5.4	5.5
World Data	4.7	4.8	4.9	5.0	5.0
VAR(4)					
Local Data	8.4	7.2	8.0	7.8	8.8
World Data	4.7	4.8	4.9	5.0	5.0

Introduction 00000 Reduced Form Models

Structural Models

Machine Learning Models

Conclusior 000

Cross Country Interpretation

- The table shows the use of cross country data improves forecasting performance on average across all 50 countries
- Using the cross country data allows the model to learn more fundamental parameter values that generalize across many countries and policy regimes
 - The panel approach is a single model trained on world panel data that makes forecasts for 50 different countries



Conclusion 000

Policy/Country Invariance using Fully Out-of-Sample Predictions



Horizon-Model



Invariance Across Countries

- This table show forecasting performance only using local data (first bar in groups), using world data (second), and using all country data except the in-sample country being forecasted (third)
- Performance is aggregated across all 50 countries
- In all situations, pooled except the in-sample country outperforms using only in-sample country to forecast
- On average having pooled training data except for the in-sample country is responsible for more than 50 percent of the reduction in RMSE gained from moving from only in-sample data to the full panel
- Provides strong evidence that using panel augmentation improves the policy invariance and generalization to unseen countries for reduced-form models

	Structural Models	Machine Learning Models	
	000000		

Structural Models

Apply the Panel Data to Smets-Wouters 2007



DSGE Model: US Data vs World Data

	Structural Models	Machine Learning Models	Appendix
	000000		

DSGE Model: US Data vs World Data

- The graph shows the RMSE performance of the maximum likelihood Smets-Wouters 2007 models estimated on US only data as is typical (first bar), versus our data set consisting of 27 countries (second bar)
- Stars next to the horizon indicate Diebold-Mariano P-values that the world model forecasts is better than the US model
- Average improvement in RMSE is 25 percent, tempered by the fact that the US model estimated with maximum likelihood has poor performance
- We estimated with maximum likelihood versus Bayesian techniques because it's applicable to many practitioners (GMM, Calibration, and other point estimation techniques) as well as easy comparison to our other forecasts

 Introduction
 Reduced Form Models
 Structural Models
 Machine Learning Models
 Conclusion
 Appendix

 00000
 00000000
 000000000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000

Out-of-Sample Tests





DSGE Model: Out-of-Sample

- The first two bars of each triplet are the same as the previous graph, the third adds a model estimated on all data except US data
 - Thus the forecasts in the third bar are both time step out-of-sample and country out-of-sample
- Despite the policy invariance of DSGE models, the out-of-sample improvements are even better than out-of-sample improvements for reduced form models
 - Suggests to us there is room to make DSGE models better generalizers and more policy invariant
- The out of sample model outperforms the original Bayesian SW DSGE at one and two quarters ahead
- Since the US is the only country that that has data back into the 1960s, part of this could be due to increased parameter stability across space compared to over time...

Introduction 00000 duced Form Models

Structural Models

Machine Learning Models

Conclusio 000

Parameter Stability across Time

Time Similarity vs Country Similarity



	Structural Models	Machine Learning Models	Appendix
	000000		

Parameter Stability: Results Inconclusive

- The first two bars in each triplet are the second and third bar in the previous chart (world data, and out-of-sample data)
- The third bar only uses data post 1995 to test the hypothesis that more recent data leads to better forecasting performance
- Clearly the out-of-sample data is the best model, but the post 1995 data is at least as good as using the entire world panel
- Given the criticism of internal validity of panel models in Pesaran 1995, not really definitive evidence that data across space but over the same time period is more relevant, but not contradictory either

	Machine Learning Models	Appendix
	00000000	

Machine Learning Models: RNN

Introduction 00000	Reduced Form Models	Machine Learning Models ○●○○○○○○○	Appendix 000000000000000000000000000000000000

Our Model

- We use a Recurrent Neural Network (RNN) as our forecast model
- An RNN is a state space model
- Like a linear state space model, the RNN will have latent states similar to the below representation of the linear state transition equation:

$$h_t = Ah_{t-1} + B + \epsilon \tag{1}$$

...and a measurement equation

$$y_t = Ch_t + Dx_t + E + \epsilon \tag{2}$$

- However, the main innovation of an RNN is to use gates, a logistic regression that is element-wise multiplied to the states (next slide)
 - Conditioned on previous information, the model can control the states allowing the eigenvalues of A to be arbitrary

Introduction 00000	Reduced Form Models	Machine Learning Models 00●000000	Appendix 000000000000000000000000000000000000
A G	RU Cell		

State Transition Equations:

$$z_t = \sigma_g(W_z x_t + U_z h_{t-1} + b_z)$$

$$r_t = \sigma_g(W_r x_t + U_r h_{t-1} + b_r)$$

$$\hat{h}_t = \phi_h(W_h x_t + U_h(r_t \odot h_{t-1}) + b_h)$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \hat{h}_t$$

Measurement Equation:

$$y_t = W_y x_t + U_y h_t + b_y$$

		Machine Learning Models	
00000		00000000	

Architecture Structure

Machine Learning Models 000000000

RNN Forecasting Improvement

4 3.5 3 2.5 RMSE US 2 World 1.5 1 0.5 0 -H1*** H2*** H3*** H4** H5

RNN Model: US Data vs World Data

Horizon

	Machine Learning Models	
	00000000	

Machine Learning Models: AutoML etc.

AutoML Forecasting Improvement on US Test Set

AutoML: US Data vs World Data

			Machine Learning Models		Appendix
00000	000000	000000	000000000	000	000000000000000000000000000000000000000

Forecasting Comparison

Table 2: RMSE of RNN, AutoML, and Baseline Models

Time (Q's Ahead)	1Q	2Q	3Q	4Q	5Q
VAR(4)					
US Data	2.99	3.03	3.10	3.08	3.08
World Data	2.37	2.52	2.56	2.63	2.63
AR(2)					
US Data	2.53	2.88	3.03	3.14	3.13
World Data	2.57	2.62	2.67	2.72	2.72
Smets-Wouters DSGE Bayesian					
US Data	2.79	2.95	2.89	2.80	2.71
Factor					
US Data	2.24	2.48	2.50	2.67	2.86
RNN (Ours)					
US Data	3.46	3.37	3.01	3.23	3.30
World Data	2.35	2.52	2.50	2.62	2.60
AutoML (Ours)					
US Data	2.41	2.58	2.71	2.45	2.92
World Data	1.97	2.32	2.59	2.62	2.61
SPF Median	1.86	2.11	2.36	2.46	2.65

 Introduction
 Reduced Form Models
 Structural Models
 Machine Learning Models
 Conclusion
 Appendix

 00000
 0000000
 00000000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000

Graph of Forecast Performance: 1 Quarter Ahead

	Machine Learning Models	Conclusion	
		000	

Conclusion

Introduction 00000	Reduced Form Models	Machine Learning Models	Conclusion O●O	Appendix 000000000000000000000000000000000000

Conclusion

- I showed how you can improve external validity of a wide variety of models at little cost
- I demonstrated increased policy invariance brought about by using a panel of countries rather than a single country
 - This hints at the model identifying more fundamental parameter values when using the larger dataset
- I also show how this panel data improves the forecasting performance of a wide variety of non-parametric machine learning models–allowing them to outperform all traditional baselines that were augmented with panel data or not

			Machine Learning Models	Conclusion	
00000	000000	000000	00000000	000	000000000000000000000

Thank You

camfen@umich.edu

	Machine Learning Models	Appendix
		•00000000000000000

Appendix

ion Reduced Form Models Structu 0000000 0000

structural Models

Machine Learning Model

Conclusion 000

Graph of Forecast Performance: 1 Quarter Ahead

luction Reduced Form Models Structural Models M 00 000000 000000 00

Machine Learning Models

Conclusion 000

Graph of Forecast Performance: 2 Quarters Ahead

Introduction Reduced Form Models Structural Models Machine Learning Models Conclu 00000 0000000 00000000 000

Conclusion 000

Graph of Forecast Performance: 3 Quarters Ahead

 Introduction
 Reduced Form Models
 Structural Models
 Machine Learning Models
 Conclusion
 Appendix

 00000
 0000000
 00000000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000

Graph of Forecast Performance: 4 Quarters Ahead

 Introduction
 Reduced Form Models
 Structural Models
 Machine Learning Models
 Conclusion
 Appendix

 00000
 0000000
 00000000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000
 000

Graph of Forecast Performance: 5 Quarters Ahead

ntroduction 00000 educed Form Models

Structural Models

Machine Learning Models

Conclusio 000

Dense Layers: A Picture

Introduction 00000	Reduced Form Models 0000000	Machine Learning Models 000000000	Appendix 0000000●00000000000000000000000000000
Dense	Lavers		

• A dense layer in a neural network is simply a vector regression $y = \sigma(\beta x)$ with a link function (called an activation), σ , which makes the model nonlinear

• y is a vector and so
$$\beta$$
 is a matrix

y becomes the multivalued input of the next layer, ie

$$y_3 = \sigma(\beta_3(y_2)) = \sigma(\beta_3(\sigma(\beta_2 x))) \tag{3}$$

The activations σ are essential because a linear combination of linear transformations is still linear so without the activations the model doesn't become more expressive

Dense Layers: An Example

- Our input, x, is a 3 dimensional vector (GDP, consumption, unemployment) with 250 time-steps
- Like in logistic regression, x is input into the first layer: $y_2 = \sigma(\beta_2 x)$
- y_2 is now the input, like x, into the second layer
 - If we want y_2 to be size 128, then by definition β_2 is 128×3

If,

$$y_3 = \sigma(\beta_3 y_2) = \sigma(\beta_3(\sigma(\beta_2 x))) \tag{4}$$

and y_3 is the 1 dimensional output, then since y_2 is 128×250 then β_3 is 1×128

In this case, y₂ is the only hidden layer, but you can imagine having many more hidden layers before producing an output

Back to Model Architecture

- We use a linear model which has two lags of GDP to forecast ahead
- Despite the simplicity, this is one of the workhorse models among forecasting practitioners (Hamilton 1994)
- We find that the AR(2) outperforms the Smets Wouters DSGE at shorter time intervals (1-2 quarters ahead), but is outperformed by Smets Wouters at longer intervals (4-5 quarters ahead)
- A factor model augmented with one GDP lag seems to dominate the AR(2)

Back to Baselines

Baseline Models: DSGE (Smets and Wouters 2007)

- Smets Wouters is New Keynesian DSGE model that is essentially an extension of the Christiano et. al. 2005 model estimated in a Bayesian framework
- The model is geared towards forecasting rather than macroeconomic analysis
- Despite limited attention paid by practitioners, we think this model deserves more attention, especially at longer time horizons
- The model outperforms AR(2) and factor models at longer horizons, but is unable to detect the great recession at shorter horizons which leads to under-performance

Back to Baselines

Baseline Models: Factor Models

- These models were introduced by Stock and Watson 2002 and Forni et. al. 2000
- These models take a large cross section of data (in our case 248 data series) in order to forecast and PCA regression, in our case, reduce the large cross section to 8 factors
- We extend the factor model highlighted in Fred-QD (McCraken and Ng 2020) by combining factors estimated in a pseudo-out-of-sample manner along with a lag of GDP for forecasting
- This model is the most formidable competitor to the neural network, however, it under-performs over long horizons likely because of variance issues and the limited predictive power of the cross section of variables

Out of Sample Forecast Comparison: Expansions

Time (Q's Ahead)	1Q	2Q	3Q	4Q	5Q
VAR(1)					
US data	2.3	2.6	2.9	3.0	3.0
World data	2.1	2.2	2.2	2.2	2.2
AR(2)					
US data	1.7	1.7	1.8	1.9	1.9
World data	1.6	1.6	1.6	1.5	1.5*
Smets Wouters DSGE					
US data	1.8	1.8	1.7	1.6	1.5*
Factor					
US data	1.6	1.6	1.6	1.9	2.1
GRU*					
Best	1.8	2.3	2.0	2.0	1.9
Mean Forecast	1.7	1.7	1.7	1.7	1.7
Median Forecast	1.7	1.7	1.7	1.7	1.7
SPF Median	1.4	1.5	1.5	1.5	1.5

*All GRU models use entire world data panel

tructural Models

Machine Learning Models

Conclusion 000

Out of Sample Forecast Comparison: LSTM Model

Table 3: LSTM Forecast Performance

Time (Q's Ahead)	1Q	2Q	3Q	4Q	5Q
Best RMSE	2.4	2.6	2.6	2.6	2.6*
RMSE of Mean	2.4	2.6	2.5	2.6	2.6*
RMSE of Median	2.4	2.5	2.5	2.6	2.6*
Mean RMSE	2.4	2.6	2.5	2.6	2.6*
Std Dev RMSE	0.05	0.07	0.05	0.06	0.06

Appendix: Additional Results

Introduction

educed Form Models

Structural Models

Machine Learning Model: 000000000 Conclusion 000

Model Bias and Variance

	Forecast Bias							
	(1-Qtr)	(2-Qtrs)	(3-Qtrs)	(4-Qtrs)	(5-Qtrs)			
GRU	0.459*	0.480*	0.506*	0.620*	0.644**			
	(0.343)	(0.369)	(0.365)	(0.379)	(0.375)			
SPF	0.331	0.600**	0.723**	0.804**	0.901***			
	(0.274)	(0.302)	(0.335)	(0.347)	(0.372)			
DSGE	1.459***	1.513***	1.300***	1.058***	0.827***			
	(0.354)	(0.378)	(0.544)	(0.386)	(0.385)			
AR2	0.937***	1.317***	1.636***	1.795***	1.780***			
	(0.351)	(0.381)	(0.380)	(0.383)	(0.384)			
Factor	0.432*	0.163	0.459	0.533*	0.699**			
	(0.328)	(0.449)	(0.367)	(0.390)	(0.414)			

Table 4:

Notes:

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

Reliability of Data Trained on US data vs World Data

Table 5: GRU Monte Carlo Simulations

Time (Q's Ahead)	1Q	2Q	3Q	4Q	5Q
GRU World Data					
Mean RMSE	2.4	2.6	2.5	2.6	2.6*
Std Dev RMSE	0.06	0.06	0.05	0.06	0.06
GRU US Data					
Mean RMSE	2.6	2.8	2.7	2.7	2.8
Std Dev RMSE	0.34	0.34	0.33	0.24	0.44

Back to World vs US Data

Forecasting Performance Table

Back to Results

Graph of Forecast Performance: Tile

