Performance Uncertainty and Ranking Significance for Early-Warning Models

Suman Basu (IMF), Roberto Perrelli (IMF), and Weining Xin (IMF)
BoE-FRB-KCL Conference, November 6, 2020

The Promise of ML Needs to be Benchmarked

- Anticipating and preparing for crises are important yet intrinsically difficult
- Early warning system (EWS) is developed to tackle this challenge
  - Kaminsky et al. (1998) for signal extraction approach; Frankel and Rose (1996) for logit regression
- Machine learning (ML) is introduced to the literature and expands the choice set
  - Non-parametric model structure could help prevent overfitting and accommodate more complex relationship
- However, macro data in EWS is **small** in some important aspects, making it different from other ML applications
- Hence models need to be evaluated and ranked carefully based on prediction performance

Performance Uncertainty and Ranking Significance

- Statistical significance in ranking matters, especially when data is small and interpretability is important
  - In case of no significant performance difference, traditional statistical models may be preferred given its interpretability
- To test the significance in performance difference, performance uncertainty arising from sampling needs to be estimated
  - For macro data, which sources of sampling variation matter? Or what are plausible alternative histories?
- This paper touches on performance uncertainty and ranking significance of early-warning models
  - Propose three sources of sampling variation that are important for macro data
  - Construct confidence intervals (CIs) to estimate performance uncertainty
  - Test ranking significance using conditional performance difference

Results: Wide Confidence Intervals, but Significant Performance Difference

- EWS performance varies substantially with histories: CIs are generally wide
  - Interestingly, CIs of signal extraction approach are wider
- Degree of performance uncertainty depends on the source of sampling variation and model algorithm
  - CIs are wider when accounting for some specific sources of variation in SE/RF
- Signal extraction approach performs significantly better than random forests
  - In fixed cutoff testing, for all variations, at 10% significant level
  - But in rolling cutoff testing, greater performance uncertainty and no significance

ML May Win,
but Performance Could Vary with Data

ML Holds Promise, but Not a Panacea

- Suppose that crises follow: $y_{it} = f(X_{it-1}, \epsilon_{it}; \theta_{t-1})$
  - y_{it} is crisis event $\in \{0, 1\}$, f is a non-linear function, X_{it-1} is a vector of explanatory variables, θ_{t-1} is the global regime, and ϵ_{it} is an idiosyncratic shock.
- Machine learning holds promise: non-parametric model structures
  - Accommodate much more flexibility in function f: non-monotonicities, non-linearities, interactions, and etc.
  - Hyperparameters help prevent overfitting
- But macro panel data is small in the context of crisis prediction
  - Global shocks: shifts in global regime θ_{t-1} is infrequent but substantial
  - Idiosyncratic shocks: trajectories of countries or part of their histories are affected
  - Countries are not many and crisis events y_{it} are even fewer and heterogeneous

Three Sources of Sampling Variation

- Assess model performance uncertainty arising from sampling: The extent to which model performance varies if the model is estimated and evaluated on a different dataset, representing a different history

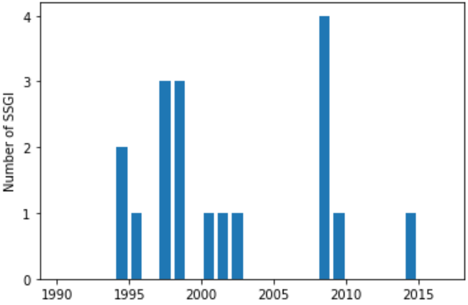
- Three sources of sampling variation, in line with the three aspects in which macro panel data is small
  - Histories of global regimes: Global shocks are infrequent but substantial
    $\Rightarrow$ What if some of the global crisis waves didn't happen?
  - Histories of countries: Idiosyncratic shocks changed countries' trajectories
    $\Rightarrow$ What if some of the countries followed different histories, e.g., by taking different policy actions that prevented or triggered crises?
  - Types of countries: Countries are not many and crises are even fewer
    $\Rightarrow$ What if some of the countries were more like a certain type, e.g., LICs $\rightarrow$ EMs $\rightarrow$ AEs?

Crisis Definition and Explanatory Indicators

Sudden Stops in EMs

- 10 sudden stop-prone EMs; 1990-2017
- Sudden stops in net private capital inflows
  - $\frac{\text{Capital inflows}}{\text{GDP}}_t < \frac{\text{Capital inflows}}{\text{GDP}}_{t-1}$ - 2%
  - $\frac{\text{Capital inflows}}{\text{GDP}}_t < \frac{\text{Capital inflows}}{\text{GDP}}_{t-2}$ - 2%
  - Or IMF programs$_t$ $>$ 500% of quota
- With growth impacts
  - $\Delta\%\text{GDP}_t - \frac{1}{5}\sum_{s=1}^{s=5} \Delta\%\text{GDP}_{t-s} < 10^{th}$ percentile
  - Or IMF programs$_{t+1}$ $>$ 500% of quota
- Rare and brutal: 6.4%

Countries	Years
Argentina	1995, 2000, 2008
Brazil	2002
Chile	1998
Indonesia	1997
Malaysia	1997, 2008
Mexico	1994, 2009
Philippines	1998
Russia	2008, 2014
Thailand	1997
Turkey	1994, 1998, 2001, 2008



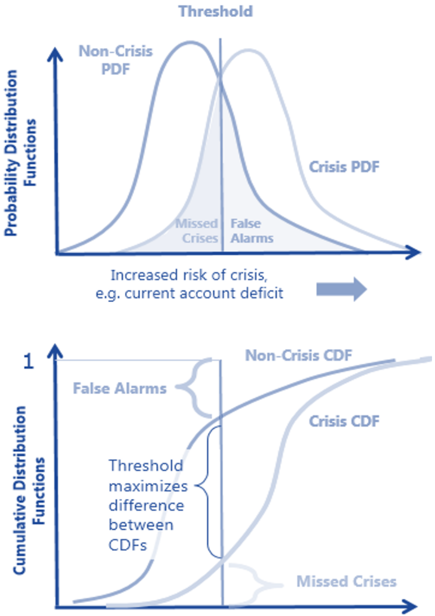
25 Explanatory Indicators Implied by Basu et al. (2019)

Medium-term bubble building:	Shor-term bubble bursting:
5-year inflation	Change in public debt
5-year money growth	Change in reserves
5-year stock price growth	Change in stock price growth
5-year housing price growth	Change in housing price growth
5-year inter-bank liabilities growth	Change in external equity liabilities
5-year REER growth	Change in REER appreciation
5-year private credit growth	Change in private credit
Buffers and mismatch:	**Global factors:**
Current account balance	TED spread
Amortization-to-exports ratio	Percentage of AEs in banking crises
EMBI spread	Inter-bank liabilities to AEs in banking crises
Foreign liabilities-to-domestic credit ratio	Export growth
External debt	
Capital adequacy ratio	
Interest coverage ratio	

Model Choice and Design

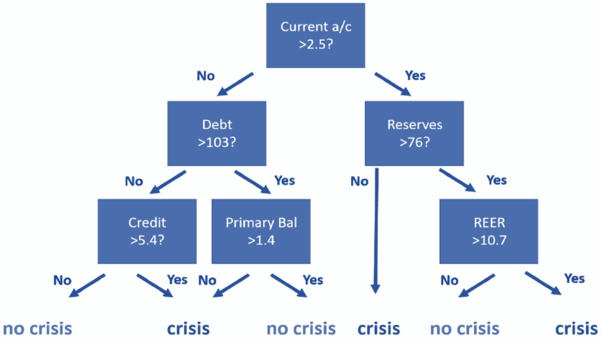
Signal-Extraction Approach (SE)

- Simple algorithm
  - Identify variable-specific threshold
  - Aggregate variable-specific flags
- Pros:
  - Simple to implement and easy to interpret
  - Able to impose priors and not data hungry
  - Exhaustively tested: won horse race in Berg et al., 2005
- Cons: Cannot address
  - Non-monotonicities
  - Non-linearities
  - Interactions



Machine Learning (ML): Random Forests

- Ensemble models based on decision trees
    - Split samples sequentially and recursively
    - Ensemble trees into forests
    - Impute using surrogates
- Pros: Capture
    - Non-monotonicities
    - Non-linearities
    - Interactions
- Cons:
    - Difficult to interpret
    - Easy to manipulate
    - Overfitting



Cutoff-Based Testing Procedure

- Fixed cutoff
  - Estimate up to year 2007, and test on years afterwards
  - Stable performance with large test set

- Rolling cutoff
  - Estimate up to year t, and test on year $t+1$ and $t+2$
  - Average performance over five test sets with cutoff year $t = 2007, 2009, 2011, 2013$, and 2015
  - Difficult to manipulate and assess performance updating over the GFC

- Evaluation metrics:
  - Sum of errors $= \frac{\# false\ \ alarms}{\# noncrises} + \frac{\# missed\ \ crises}{\# crises}$.
  - AUC for reference

Performance Uncertainty Estimation

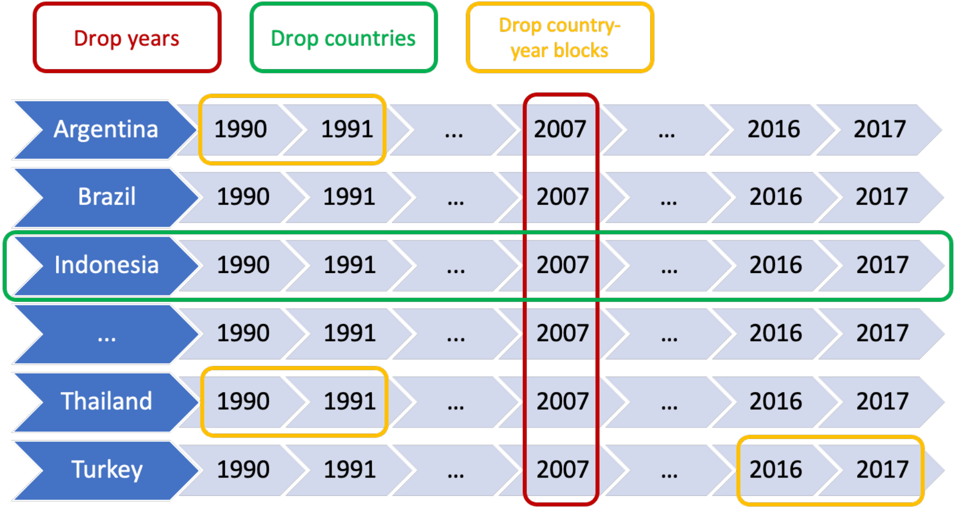
Resampling on the Entire Sample

- Three steps: (1) generating new samples; (2) estimating and testing models; (3) constructing confidence intervals
- Procedure as follows:
    1. Perform resampling on the original entire sample S to obtain a new sample S_j.
    2. Split the new sample S_j into training set and test set based on cutoff rules.
    3. Estimate different models (signal extraction model and random forests) on the same training set and tested on the same test set. Model performance on the test set are then calculated.
    4. Repeat 1.-2. for 200 times, and construct confidence intervals using the model performance calculated.

Jackknifing along Three Dimensions

- Jackknife resampling: imposing priors while preserving panel data structure
- Three aspects of small data nature $\Rightarrow$ three sources of sampling variation $\Rightarrow$ three dimensions of jackknifing
  - Global shocks are infrequent but substantial
    $\Rightarrow$ Histories of global regimes: what if some of the global crisis waves didn't happen?
    $\Rightarrow$ Drop years
  - Idiosyncratic shocks changed countries' trajectories
    $\Rightarrow$ Histories of countries: what if some of the countries followed different histories?
    $\Rightarrow$ Drop country-year blocks
  - Countries are not many and crises are even fewer
    $\Rightarrow$ Types of countries: what if some of the countries were more like a certain type?
    $\Rightarrow$ Drop countries
- Also consider an i.i.d. jackknifing to compare

Jackknifing along Three Dimensions



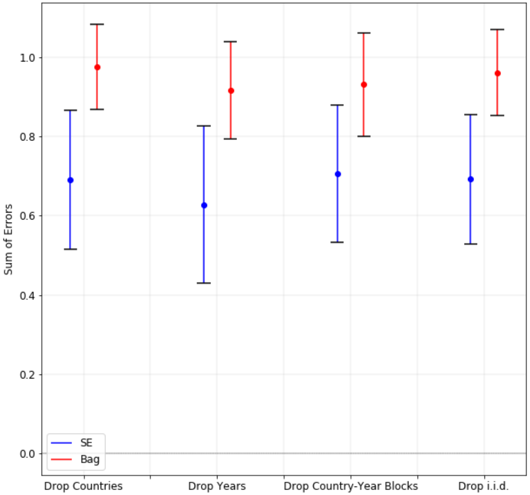
Construct Confidence Intervals

- Unlike standard jackknifing that drops one single observation, 5% of data is dropped for sufficient variation while preserving enough data
  - Macro panel data is cross-sectionally and temporally correlated
  - Dropping one year is to drop $1/38 \approx 2.6\%$ of data

- $\hat{\theta}$ the estimator of model performance obtained from the original sample;
  $\hat{\theta}^*_j$ the estimator of model performance obtained from the jackknifing sample $j = 1, 2, \ldots, 200$
  1. Order the estimators obtained from the jackknifing $\hat{\theta}^*$ such that $\hat{\theta}^*_1 \leq \ldots \leq \hat{\theta}^*_J$, with subscript denoting the jth element in the ordered list.
  2. For a significance level α, select the $[J \cdot \alpha/2]$th and $[J \cdot (1-\alpha)/2]$th elements from the above ordered list of estimators, i.e., $\hat{\theta}^*_{J \cdot \alpha/2}$ and $\hat{\theta}^*_{J \cdot (1-\alpha/2)}$.
  3. Construct the two-tailed confidence interval of $\hat{\theta}$ with the significance level α as $\left[2\hat{\theta} - \hat{\theta}^*_{J \cdot (1-\alpha/2)}, 2\hat{\theta} - \hat{\theta}^*_{J \cdot \alpha/2}\right]$.

Individual Confidence Intervals are Wide

- Signal extraction approach
  - Largest CI generated by dropping years
    $\Rightarrow$ Global regimes matter
    $\Rightarrow$ Likely because it distinguishes only the set of crises and non-crises
- Random forests
  - Largest CI generated by dropping country-year blocks
    $\Rightarrow$ Countries' histories matter
    $\Rightarrow$ Likely because it learns aggressively from individual crisis events
- Overlapping CIs
  - Not necessarily indicate insignificant difference.



Ranking Significance Assessment

CIs for Conditional Performance Difference

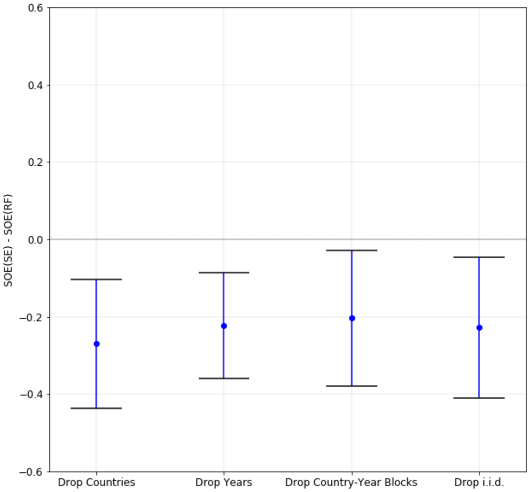
- To rank models, they should be estimated and tested on the same training and test set respectively, i.e., compared within each history
  - Not fair to compare models estimated/tested on a dataset with and without the GFC

- The estimator now is a conditional performance difference that is calculated within each jackknifing sample, and confidence intervals are constructed for it

- $\hat{\theta}_1$ for SE and $\hat{\theta}_2$ for RF
  1. Calculate the difference in performance estimators obtained from the original sample, $\Delta\hat{\theta} = \hat{\theta}_1 - \hat{\theta}_2$
  2. Calculate the differences in performance estimators obtained from the jackknifing, $\Delta\hat{\theta}_j^* = \hat{\theta}_{1,j}^* - \hat{\theta}_{2,j}^*$ for $j = 1, 2, \dots, J$
  3. Same as previous procedure but for $\Delta\hat{\theta}$ and $\Delta\hat{\theta}_j^*$

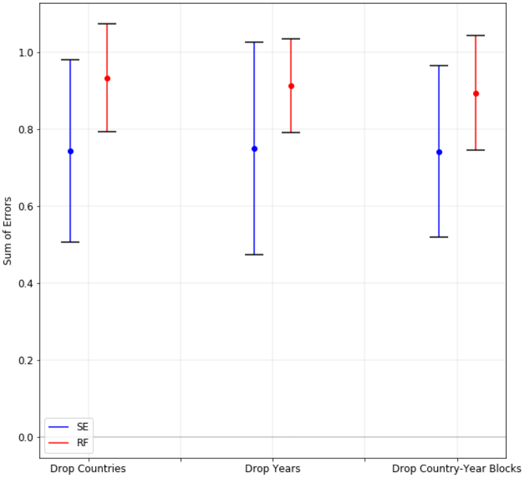
Performance Difference is Significant

- $H_0 : \Delta\theta = 0$
  - θ denote sum of errors
  - Difference between SE and RF
  - Whether zero is inside the CI
- SE performs significantly better than ML
  - When accounting for all variations
  - At 10% confidence level
  - Despite of overlapping individual CIs

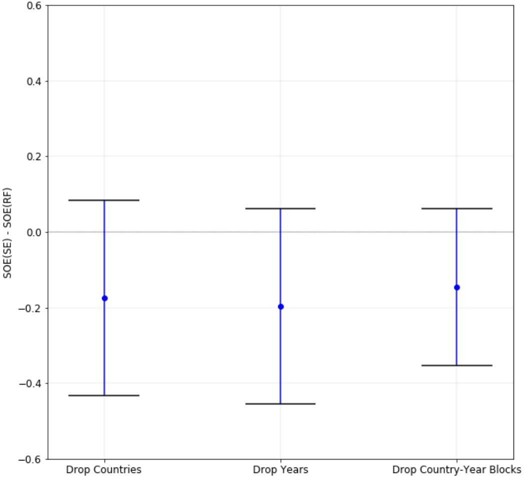


Greater Uncertainty and No Significance in Rolling Cutoff Testing

Individual Performance



Conditional Performance Difference



Conclusions & Next Steps

- EWS performance varies substantially with histories: CIs are generally wide
  - Interestingly, CIs of signal extraction approach are wider
- Degree of performance uncertainty depends on the source of sampling variation and model algorithm
  - SE: CIs are wider when accounting for variations in global regimes
  - RF: CIs are wider when accounting for variations in country histories
- Signal extraction approach performs significantly better than random forests
  - In fixed cutoff testing, for all variations, at 10% significant level
  - But in rolling cutoff testing, greater performance uncertainty and no significance
- Next steps: how CIs depend on (i) number of variables; (ii) percentage of data dropped; (iii) random seed variation alone ...

Thank you!