**Introduction:**

* Discuss the implementation of out of sample tests (OOST) of forecasting accuracy
* Summarize the rationale for OOST
* Compare fixed-origin and rolling-origin procedures
* Examine the application of OOST to an individual time series
  + Various problems discussed
* Consider role of OOST in method selection
* Describe the extension of OOST from individual time series to multiple time series and forecasting competitions
* Evaluate the adequacy of OOST in forecasting software

**Section 2: In-sample vs. Out-of-sample Evaluation**

* For a given forecasting method
  + In-sample errors are likely to understate forecasting errors
  + Overfitting and structural changes may further aggravate the divergence between in-sample and post-sample performance
* Updating procedures (smoothing method)
  + One makes each forecast as if they were standing in the immediately prior period
  + General goodness-of-fit is based on *one step ahead errors*
    - Errors made in estimating the next time period from the current time period
* Errors in forecasting into the more distant future will be larger than those that are made one step ahead
* Methods selected by best in-sample fit may not best predict post-sample data
* Tests based on *holdout* samples have become commonplace
* The *fit period* is used to identify and estimate a model
* The *test period* is reserved to assess the model’s forecasting accuracy
* If the forecaster withholds all data about events occurring after the fit period, the forecasting accuracy evaluation is structurally identical to the real world forecasting environment
  + However, peeking at the data while selecting the forecasting method pollutes the environment

**Section 3: Fixed-origin vs. Rolling-origin Procedures**

* The final time in the fit period (T), the point from which the forecasts are generated, is the *forecasting origin*
* The number of time periods between the origin and the time being forecast is the *lead time* or the *forecasting horizon*
* The longest lead time is the N-step ahead forecast
* N denotes the length of the test period
* For OOST, can use single or multiple forecasting origins
* Single forecasting origin
  + See math ex in paper
  + Several shortcomings
    - Because it yields one forecast and one forecasting error, for each lead time, it requires a fairly long test period to produce a forecasting track record
    - Forecasts susceptible to corruption by occurrences unique to that origin
    - Forecasting software is not great
  + Can overcome these problems by updating the forecasting origin
  + Mitigate problems by using multiple time series
  + ONLY way we can assess the post-sample accuracy of forecasts, when we do not know or can not replicate the underlying forecasting methodology
* Rolling origin evaluations
  + Successively update the forecasting origin and produce forecasts from each new origin
  + Provides N(N+1)/2 forecasts, against N from the fixed origin

**Section 4: Issues in Implementing Out-of-sample Evaluations**

* The most fundamental choice is how to split the series between fit and test periods
* Denote the maximal length forecast as H
* N must be at least as large as H
* In rolling origin evaluation, each update of the forecasting origin adds one new observation to the fit period
* For OOST, the principle purpose of rolling window is to level the playing field in multiperiod comparison of forecasting accuracy

**Section 5: Sliding Simulations**

* A sliding simulation is an extension of the rolling origin design to serve as a process for method selection and estimation
* The rolling horizon is used to compare the efficacy of various method-selection rules
* The sliding simulation requires a 3-way division of the time series
* N simulations withheld from the time series serve as a test set
* The remaining period of fit is subdivided between the first T observations, which represent the in-sample fit period, and the remaining P observations, T + 1 and T + P, which constitute the post-sample fit period
* We can use sliding simulation to compare individual-selection and aggregate-selection rules
* For individual selection, we identify the best method for each time series in a batch
* For aggregate selection, we apply to every series in the batch the method that works best in the aggregate

**Section 6: Multiple Time Series: Forecasting Competitions**

* Desirable characteristics for an OOST are *adequacy*, enough forecasts at each lead time, and *diversity*, desensitizing forecast error measures to special events and specific phases of business
* To achieve this, must use rolling origins, multiple test periods, or multiple time series
* There are a lot of external factors that can affect a time series forecast (start date, month, end date, day, season, etc)
* The use of multiple time series, as in forecasting competition, creates a pooled data structure: S time series, s = 1 to S and up to T + N time periods per series
* Sigma s represents the average errors across time series
* When averaging over sigma s:

1. Avoid scale dependent error measures such as root mean squared error (RMSE) or mean absolute deviation (MAD)
2. User percent error measures instead absolute percent error (APE) because they are scale independent -> However, percent errors can be badly skewed if close to 0, so use median absolute percent error (MdAPE) instead of mean absolute percent error (MAPE)
3. Use relative error measures when it is necessary to average over time series that differ in volatility

* There are pros and cons to choosing different types of error measuring for our forecast \*\*see paper 6.2\*\*

**Section 7: Out of Sample Evaluations for Forecasting Software**

* This section just compares and contrasts various forecasting software

**Section 8: Summary**

* For an individual time series, OOST of forecasting accuracy is facilitated by use of rolling origin evaluations
* Rolling origin:
  + Permits more efficient series-splitting rules
  + Allows for distinct error distributions by lead time
  + Desensitizes the error measures to special events at any single origin
  + Applying the procedure across multiple test periods is desirable to mitigate the sensitivity of error measures to single phases of the business cycle
  + Recalibration of parameters of a forecasting equation is essential