Forecast Quality

23-27 February 2026, Kigali

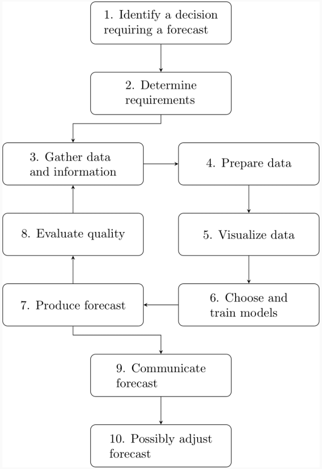
https://workshop.f4sg.org/africast/

Outline

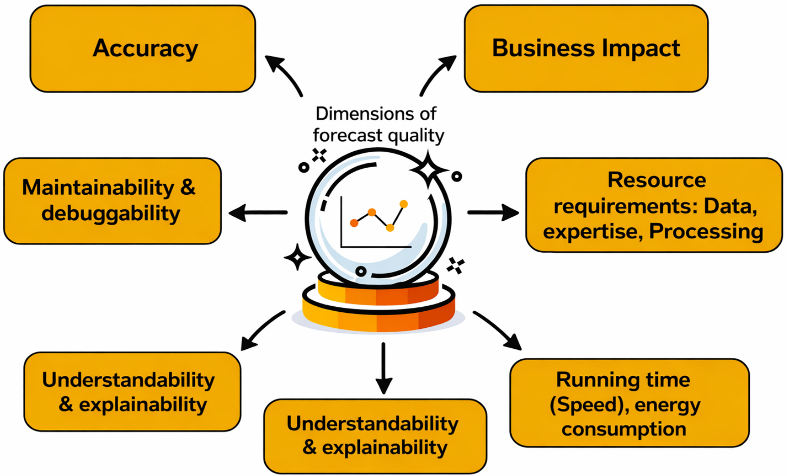
What is a good forecast?

- A good forecast?
- A good forecasting model?
- A good forecasting process?

Forecasting workflow



Quality of forecas(ting)



Translating forecast accuracy into business value

Possibility 1: model the relationship

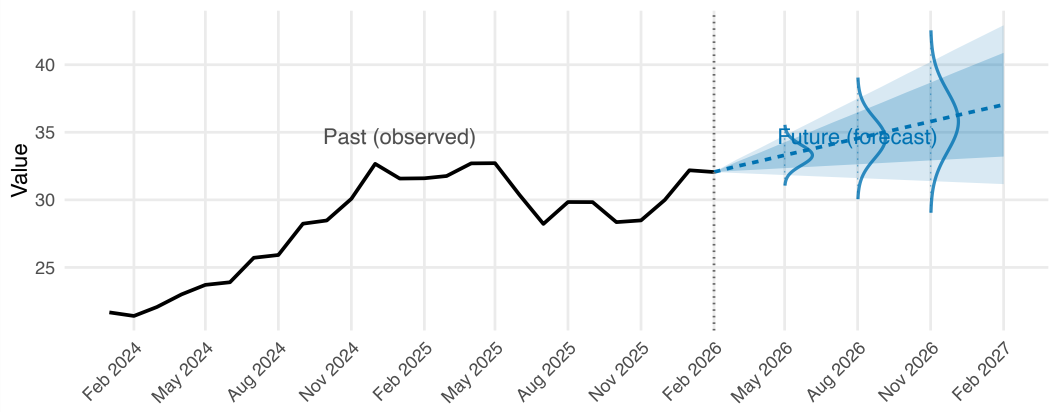
- Collect forecasts with accuracies and associated business values (over time or cross-sectionally)
- Create a statistical/predictive model
  - Regression
  - Machine Learning
  - Predict the business value of improved accuracy
- Simpler than simulation, but requires data on forecast accuracy and business value

Translating forecast accuracy into business value

Possibility 2: simulation

- Model the processes that turn forecasts into decisions (simplify!)
- Simulate the outcome for more accurate forecasts
- More complicated than statistical modeling
  - I consider this a feature, not a bug
  - It forces you to think about the relationship
  - It lays bare the important drivers
  - It allows tweaking other parameters than forecast accuracy
- Therefore: more informative than modeling

Rememberign what a forecast is



Forecast accuracy evaluation

- We mimic the real life situation

–

- We pretend we don’t know some part of data (new data)
- It must not be used for *any* aspect of model training
- Forecast accuracy is computed only based on the test set

Pitfalls for Forecast Evaluation

- **Data Leakage**: Information from the future (validation set) unintentionally influencing the training & evaluation.
- **Inappropriate Benchmarks**: Not comparing complex models against simple, established baselines.
- **Wrong/Ad-hoc Metrics**: Selecting evaluation metrics that do not align with business goals (e.g., using MAPE when zeroes exist in data, or failing to differentiate between over/under forecasting).
- **Small Datasets**: Evaluation on too little data or data that is not representative of the future, leading to unreliable results.
- **Over-reliance on Plots**: Using visual charts for evaluation rather than rigorous, quantitative error measures, particularly in rolling-origin scenarios.

Outline

1 Forecast quality

2 Data partitioning

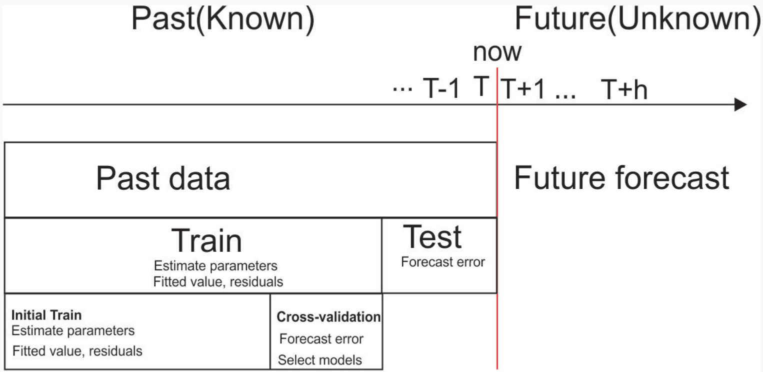
3 Evaluating point forecast accuracy

4 Evaluating distributional forecast accuracy

In-sample (training) vs. out-of-sample (test)

- Fitting and its residual are not a reliable indication of forecast accuracy
- A model which fits the training data well will not necessarily forecast well
- A perfect fit can always be obtained by using a model with enough parameters
- Over-fitting a model to data is just as bad as failing to identify a systematic pattern in the data

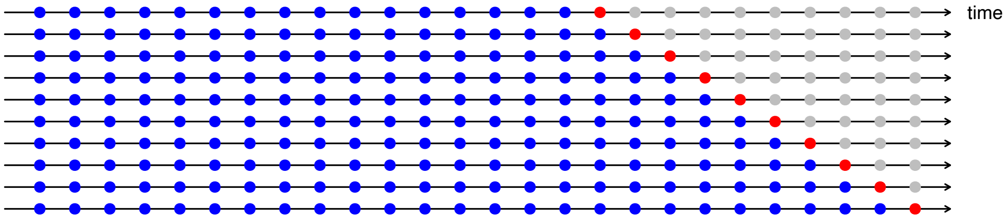
Time series cross-validation (rolling origin)



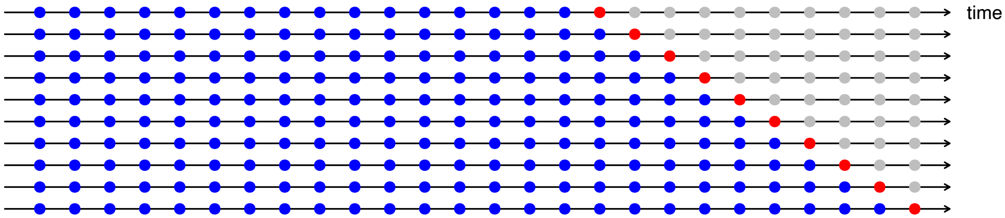
Test size= forecast horizon, h

Cross-validation size=nb of experiment+h-1

Time series cross-validation



Time series cross-validation



Outline

Forecast errors

Forecast "error": the difference between an observed value and its forecast.

$$e_{T+h} = y_{T+h} - \hat{y}_{T+h|T},$$

where the training data is given by $\{y_1, \dots, y_T\}$

Forecast Bias

- Bias and variance have different consequences for a business
  - e.g., always underpredicting leads to persistent stockouts
- Bias can be measured with the **Mean Error**:

$$\text{ME} = \frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)$$

- If ME > 0, model **underpredicts** on average $\Rightarrow$ negatively biased

Measures of forecast accuracy

$$y_{T+h} = (T+h)\text{th observation, } h = 1, \dots, H$$
$$\hat{y}_{T+h|T} = \text{its forecast based on data up to time } T.$$
$$e_{T+h} = y_{T+h} - \hat{y}_{T+h|T}$$

$$\text{MAE} = \text{mean}(|e_{T+h}|)$$
$$\text{MSE} = \text{mean}(e^2_{T+h}) \qquad \text{RMSE} = \sqrt{\text{mean}(e^2_{T+h})}$$
$$\text{MAPE} = 100\text{mean}(|e_{T+h}|/|y_{T+h}|)$$

Measures of forecast accuracy

$$
\begin{aligned}
y_{T+h} &= (T+h)\text{th observation, } h = 1, \dots, H \\
\hat{y}_{T+h|T} &= \text{its forecast based on data up to time } T. \\
e_{T+h} &= y_{T+h} - \hat{y}_{T+h|T}
\end{aligned}
$$

$$
\begin{aligned}
\text{MAE} &= \text{mean}(|e_{T+h}|) \\
\text{MSE} &= \text{mean}(e_{T+h}^2) \qquad \text{RMSE} = \sqrt{\text{mean}(e_{T+h}^2)} \\
\text{MAPE} &= 100\text{mean}(|e_{T+h}|/|y_{T+h}|)
\end{aligned}
$$

- MAE, MSE, RMSE are all scale dependent.
- MAPE is scale independent but is only sensible if $y_t \gg 0$ for all t, and y has a natural zero.

RMSE

$$\text{MSE} = \text{Bias}^2 + \text{Var}$$

- If the model is unbiased, RMSE and the standard deviation of the errors are equal
- MSE is minimised by predicting the **mean** of the forecast distribution
- Penalises large errors more heavily than small ones
- Minimising RMSE leads to mean-unbiased forecasts

MAE

- Also called **Mean Absolute Deviation (MAD)**
- Minimised by predicting the **median** of the forecast distribution
- More robust to outliers and large errors than MSE
- Minimising MAE can lead to mean-biased forecasts if the forecast distribution is skewed
- For series with small integer values, minimising MAE leads to integer predictions (since the median of a discrete distribution is an integer)
- For **intermittent series**, minimising MAE can lead to predicting only zeros — heavily biased towards underprediction, since the median of the forecast distribution is zero

Mean Absolute Percentage Error (MAPE)

$$\text{PE}_t = 100\frac{y_t - \hat{y}_t}{y_t}, \qquad \text{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|100\frac{y_t - \hat{y}_t}{y_t}\right|$$

- **Cannot be used** if $y_t = 0$, and is distorted when y_t is small (was originally designed for inventory count data)
- **Not symmetric**: exchanging the prediction and the true value gives a different result
  - e.g., predicting 50 when truth is 100 gives 50%, but predicting 100 when truth is 50 gives 100%
- Minimised by the (-1)-median of the forecast distribution

MASE

Mean Absolute Scaled Error

$$\text{MASE} = \text{mean}(|e_{T+h}|/Q)$$

MASE

Mean Absolute Scaled Error

$$\text{MASE} = \text{mean}(|e_{T+h}|/Q)$$

- For non-seasonal series, scale uses naïve forecasts:

$$Q = \frac{1}{T-1}\sum_{t=2}^{T}|y_t - y_{t-1}|$$

- For seasonal series, scale uses seasonal naïve forecasts:

$$Q = \frac{1}{T-m}\sum_{t=m+1}^{T}|y_t - y_{t-m}|$$

where m is the seasonal frequency

MASE

Mean Absolute Scaled Error

$$\text{MASE} = \text{mean}(|e_{T+h}|/Q)$$

- For non-seasonal series, scale uses naïve forecasts:
$$Q = \frac{1}{T-1}\sum_{t=2}^{T}|y_t - y_{t-1}|$$
- For seasonal series, scale uses seasonal naïve forecasts:
$$Q = \frac{1}{T-m}\sum_{t=m+1}^{T}|y_t - y_{t-m}|$$
where m is the seasonal frequency

Proposed by Hyndman and Koehler (IJF, 2006).

Measures of forecast accuracy

Root Mean Squared Scaled Error

$$\text{RMSSE} = \sqrt{\text{mean}(e_{T+h}^2/Q)}$$

- For non-seasonal series, scale uses naïve forecasts:

$$Q = \frac{1}{T-1}\sum_{t=2}^{T}(y_t - y_{t-1})^2$$

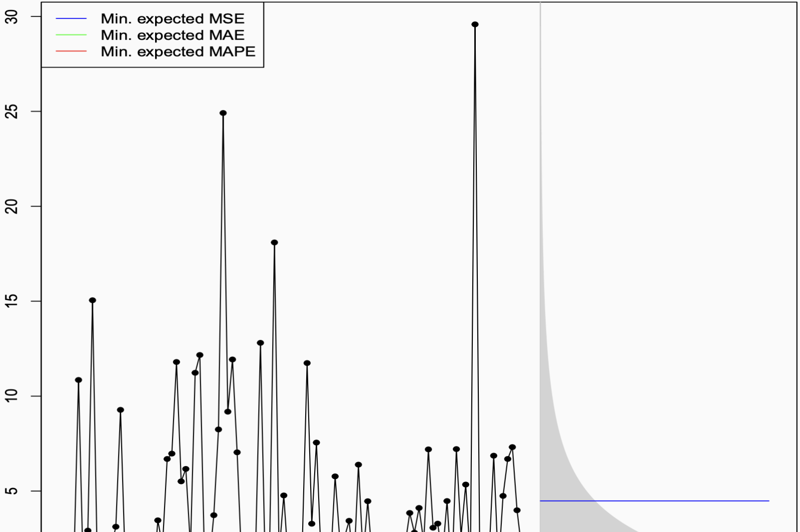
- For seasonal series, scale uses seasonal naïve forecasts:

$$Q = \frac{1}{T-m}\sum_{t=m+1}^{T}(y_t - y_{t-m})^2$$

where m is the seasonal frequencyq

Proposed by Hyndman and Koehler (IJF, 2006).

Different measures are minimal under different summary statistics of the forecast distribution



Relative Error Measures

Use a benchmark method B (usually the naïve forecast)

Relative Errors:

$$\text{RE}_t = \frac{y_t - \hat{y}_t}{y_t - \hat{y}_t^B}, \qquad \text{MRAE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t - \hat{y}_t^B}\right|$$

- Same problems as before: undefined when both $y_t = 0$ and $\hat{y}_t^B = 0$

Relative MAE:

$$\text{RelMAE} = \frac{\text{MAE}}{\text{MAE}_B}$$

- Only beneficial when evaluating many forecasts on the **same scale**
- If series are on **different scales** and there is only one forecast per series, it inherits the same problems as relative errors

Aggregating Error Measures

Three dimensions along which we can average:

Dimension	Example
Horizons	Average over $h = 1, 2, \dots, H$ steps ahead
Rolling origins	Average over multiple train/test splits
Time series	Average over many different series

Recommendations

- If you currently use MAPE or sMAPE, **switch to something else**
- Do not invent your own measure — it is harder to get right than it appears
- Some argue against using multiple measures simultaneously, as different metrics are minimised by different summaries of the forecast distribution
- If you don’t need a scale-free measure, better stick to MAE, RMSE
- **Suggested approach:**
  - Choose a **primary metric** (e.g. RMSSE) that aligns with your loss function
  - Use other measures for **sanity-checking**

Recommendations

If you need a scale-free measure:

- For broadly benchmarking methods without requiring interpretability (the standard scenario for forecasting methodology papers) → use **RMSSE**
- Use **MASE** only if:
  - there are reasons to elicit the **median** of the forecast distribution, **and**
  - all compared methods use L_1 loss
- If evaluating a **mix of methods** trained with different losses (L_1, L_2, or similar) → report **both MASE and RMSSE**

Outline

Prediction interval accuracy using winkler score

Winkler proposed a scoring method to enable comparisons between prediction intervals:

- it takes account of both coverage and width of the intervals.

Winkler score

$$W(l_t, u_t, y_t) = \begin{cases} u_t - l_t & \text{if } l_t < y_t < u_t \\ (u_t - l_t) + \frac{2}{\alpha}(l_t - y_t) & \text{if } y_t < l_t \\ (u_t - l_t) + \frac{2}{\alpha}(y_t - u_t) & \text{if } y_t > u_t \end{cases}$$

Mean Scaled Interval Score (MSIS)

$$\text{MSIS} = \frac{\frac{1}{h}\sum_{t=n+1}^{n+h}\left(q_t^{[u]} - q_t^{[l]} + \frac{2}{\alpha}(q_t^{[l]} - y_t)\mathbf{1}_{y_t < q_t^{[l]}} + \frac{2}{\alpha}(y_t - q_t^{[u]})\mathbf{1}_{y_t > q_t^{[u]}}\right)}{\frac{1}{n-m}\sum_{t=m+1}^{n}|y_t - y_{t-m}|}$$

- Evaluates a **prediction interval** by combining:
  - the **width** of the interval
  - the **magnitude of violations** for points falling outside the interval

Quantile score

Quantile score

$$Q_{p,t} = \begin{cases} 2(1-p)(f_{p,t} - y_t), & \text{if } y_t < f_{p,t} \\ 2p(y_t - f_{p,t}), & \text{if } y_t \geq f_{p,t} \end{cases}$$

Continuous Ranked Probability Score (CRPS)