

Forecasting for Data Scientists

Theory Session 3 – Forecast evaluation and Probabilistic Forecasting

Christoph Bergmeir

December, 2025

Universidad de Granada, Spain

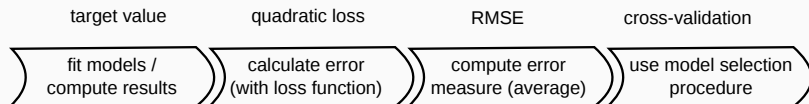
Monash University, Melbourne, Australia

<https://www.cbergmeir.com>

Forecast Evaluation

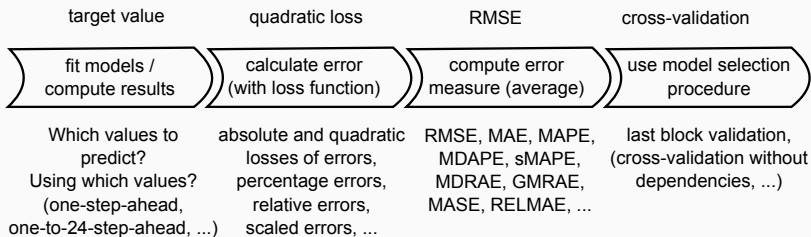
Evaluation

evaluation in general regression



Evaluation

evaluation in general regression



traditional time series forecast evaluation

Data Partitioning and Model Training

Training and test set splitting

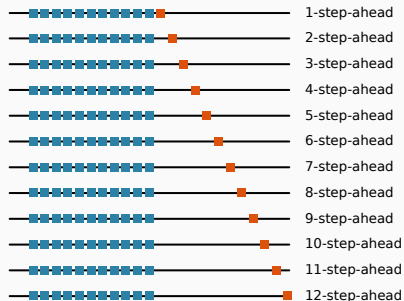
Training and test split:



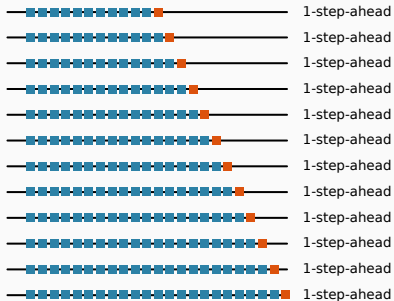
Out of sample evaluation

Fixed and rolling origin evaluation:

Fixed origin

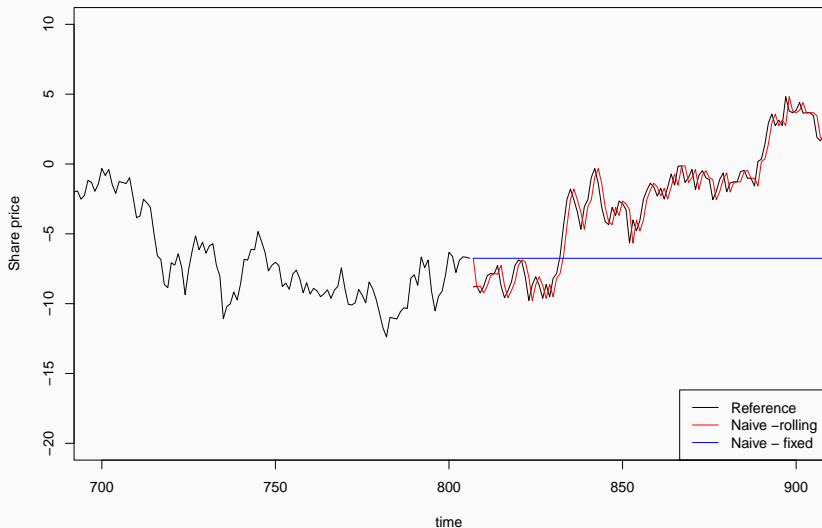


Rolling origin

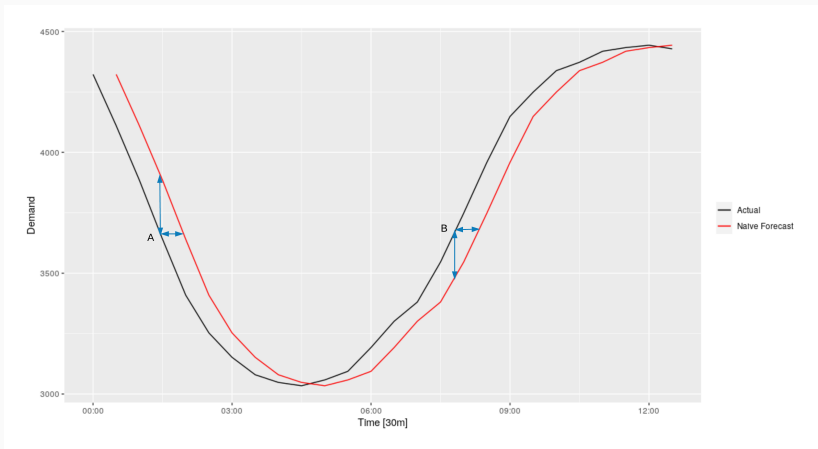


Fixed and rolling origin evaluation are very different

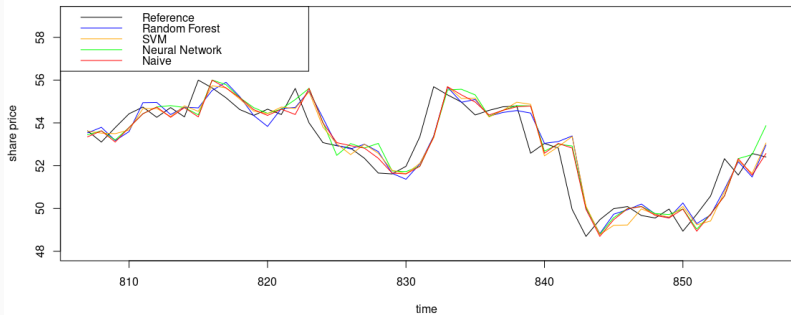
Example: Naive forecast



Plots for rolling origin are deceiving for small horizons



Which one is best?



Fixed origin evaluation

- Easier to implement than rolling origin
- Forecasting competitions mostly operate like this, as test set cannot be disclosed
- Drawback: Not usually how we operate in practice (forecasts every week, every 5 minutes, ...)
- Only one forecast can be evaluated per horizon
- Often difficult to capture the full picture: Summer holidays, Christmas, etc.

Rolling origin evaluation / Time series cross-validation

- Rolling origin eval is also called time series cross-validation (TSCV)
- Can take single or larger steps (leave-one-out versus k -fold TSCV)
- Can be difficult to implement with some of the traditional methods, as they would usually be re-trained for every forecast
- Much more natural with ML methods, where it is normal to have new data and not re-train the model
- With ML models, you would usually take single steps, not retraining for a number of steps (“fold”), and retrain for a number of folds
- Even more important in a global model if all the series are in sync regarding their timestamps.
- In a competition dataset like the M3 or M4, the series are not aligned and already mixed together wildly, so it is less important.
- Beware of leakage from training to test set!

Data leakage

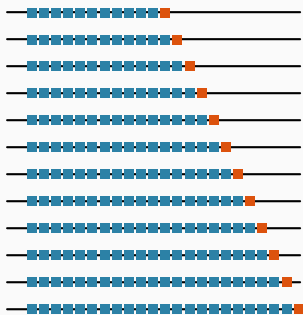
- Not trivial to avoid in forecasting
- Rolling origin: data travels from test to training set
- Hard to completely separate training from evaluation code base

Data leakage (2)

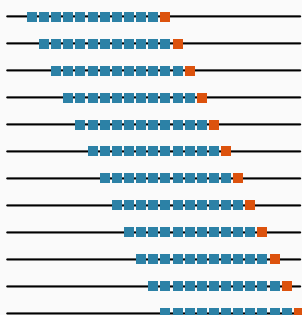
- Usually in ML: Don't normalize (calculate mean, variance) before splitting into training and test set
- In forecasting also problematic:
 - Normalization
 - Any form of smoothing or decomposition
 - Seasonal decomposition: STL, etc.
 - Feature extraction
 - Empirical mode decomposition

Rolling origin evaluation (cont'd)

Expanding window



Fixed window



- Expanding window better if less data available, as training set grows (Bell and Smyl, 2018)
- Also can combine the two: start with expanding window, then move to fixed window

Cross-validation

- We have seen TSCV
- Why not just do a normal cross-validation?



Out of sample evaluation



5-fold cross-validation

Cross-validation (cont'd)

- Problems of serial correlation between the data
- Problems of non-stationarities
- Using data from the future to predict the past doesn't feel right

→ fear that CV will grossly underestimate the generalisation error

- Most traditional methods (ETS, ARIMA) and also RNNs cannot deal well with missing data (that is reserved for testing)

Cross-validation (cont'd)



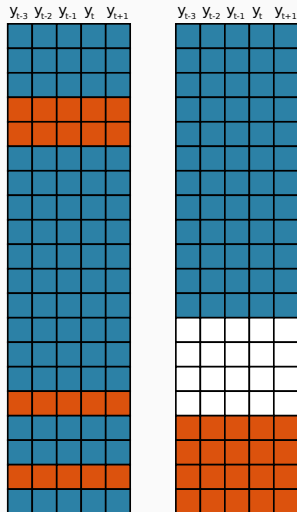
5-fold blocked cross-validation



5-fold non-dep. cross-validation

- Works in the literature to address problems of serial correlation: non-dependent cross-validation or blocked cross-validation (Burman et al., 1994; Racine, 2000; Bergmeir and Benítez, 2012)
- But general problems of non-stationarity remain

Pure AR models and Cross-validation



cross-
validation

OOS
evaluation

Pure AR models and Cross-validation (cont'd)

- Theoretical prove that cross-validation performs well in a purely autoregressive setup, as long as models nest or approximate the true model, as then errors are uncorrelated (Bergmeir et al., 2018)
- As seen before, many Machine Learning methods work like this
- Cross-validation can and should be used without modification to detect overfitting then.
- Underfitting can be detected separately, e.g., by a test for serial correlation.
- Implemented in the `CVar` function in the `forecast` package in R (Hyndman and Khandakar, 2008)

Model diagnostics: Ljung-Box test for serial correlation

- Ljung and Box (1978)
- Detects serial correlation in the residuals, which means the model is underfitting

→ Look at the residuals that you are getting. Is there autocorrelation in there? Are there any seasonal patterns in there?

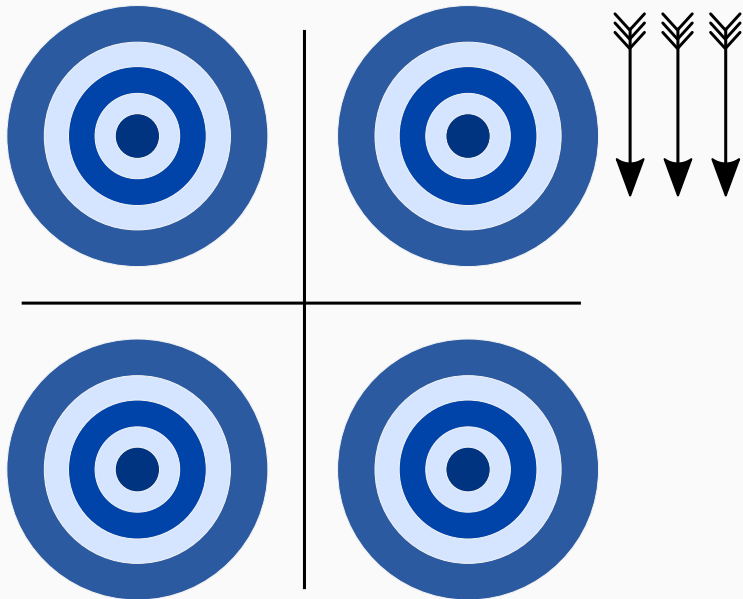
→ If yes, model is still not using all the information available in the data

Errors and error measures

Errors and error measures

- Overviews by Hyndman and Koehler (2006) and Hewamalage et al. (2023)
- Still controversial, no solution that always works
- Still papers that come up with their own ad-hoc measure
- Measures often misused

Back to the basics: Measuring variance, bias



Measuring variance, bias (2)

- Bias and variance have different consequences for a business (to always underpredict may lead to having always stockouts)
- Bias can be measured with the mean error:

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

- If $ME > 0$, model underpredicts on average, it is negatively biased
- If $ME < 0$, model overpredicts on average, it is positively biased
- Thus, bias is $-ME$

Measuring variance, bias (3)

- We can measure the variance as:

$$\text{Sample Variance} = \frac{1}{n-1} \sum_{t=1}^n ((y_t - \hat{y}_t) - \text{mean}(y_t - \hat{y}_t))^2$$

- If we assume the true population mean is zero (there is no bias):

$$\text{Error Std Dev} = \sqrt{\frac{1}{n} \sum_{t=1}^n (y_t - \hat{y}_t)^2}$$

Scale-dependent errors and error measures

Scale-dependent errors: squared error (SE), absolute error (AE):

$$SE_t = (y_t - \hat{y}_t)^2$$

$$AE_t = |y_t - \hat{y}_t|$$

Corresponding error measures, e.g., RMSE and MAE:

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^n (y_t - \hat{y}_t)^2}$$

$$MAE = \frac{1}{n} \sum_{t=1}^n |y_t - \hat{y}_t|$$

- MSE is equal to $\text{Bias}^2 + \text{Var}$, so if model is unbiased, RMSE and standard dev of the error are the same
- is minimised by predicting the mean of the forecast distribution
- penalises large errors more heavily
- minimising RMSE leads to (mean-)unbiased forecasts

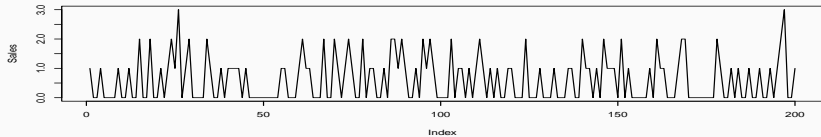
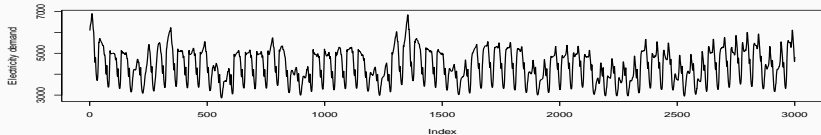
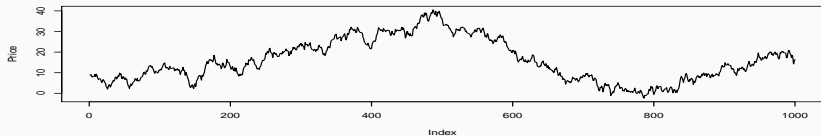
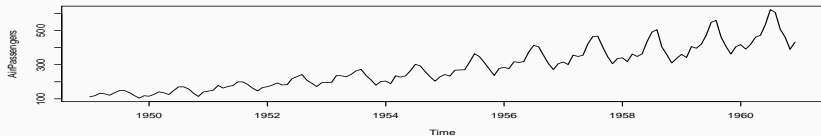
MAE

- sometimes also called mean absolute deviation (MAD)
- is minimised by predicting the median of the forecast distribution
- more robust to outliers and large errors
- minimising MAE can lead to (mean-)biased forecasts if forecast distribution is skewed
- for series with small integer values, minimising MAE will lead to predictions that are small integers (as median of forecast distribution is an integer)
- for intermittent series, minimising MAE can lead to predicting only zeros, and heavily biased forecasts towards underprediction, as median of the forecast distribution is zero.

Scale-dependent error measures: RMSE and MAE

- Advantages: Are on the same scale as the data, are interpretable
- Problem: If series are on very different scales, some series can dominate the evaluation
- We need a scale-free measure.
- We need to divide by “something”
- After 30 years of research in forecasting, we have still not found this “something” in a way that it works under any possible non-stationarity and series characteristics.
- There are over 40 error measures proposed in the literature that we are aware of (sMAPE, MASE, etc.)

Problems: Non-stationarity, Non-normality

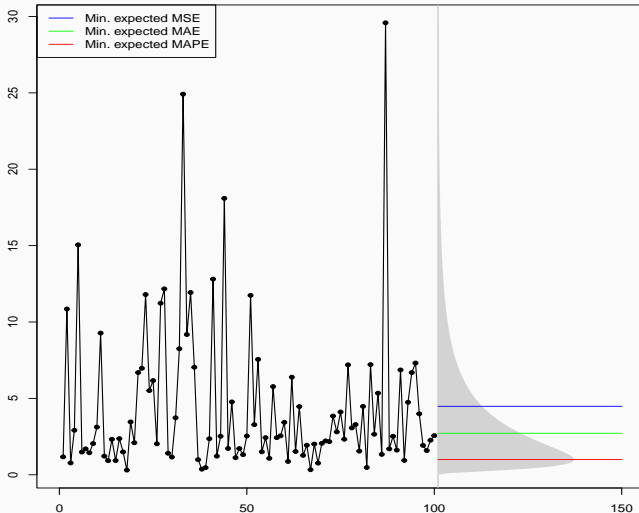


Percentage Errors (PE)

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

- Problem: Cannot be used if y_t is zero and is distorted when y_t is small. Was originally used for inventory count data.
- Is also not symmetric: exchanging the prediction and the true value changes the result
- Is minimised by the “(-1)-median” of the forecast distribution

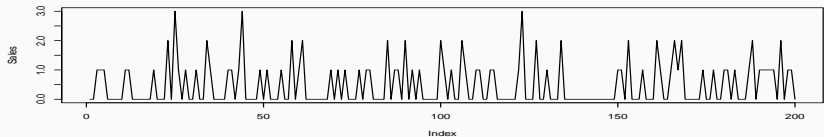
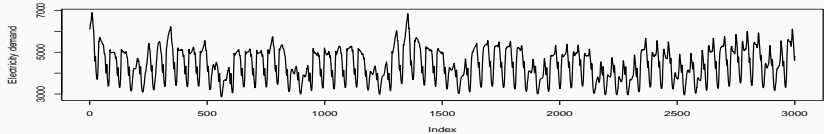
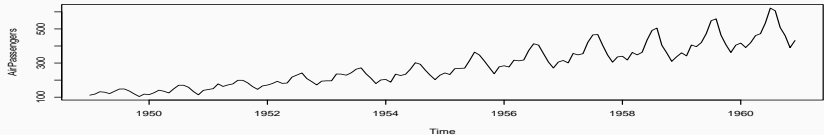
Different measures are minimal under different summary statistics of the forecast distribution



(also see Kolassa (2020))

50% of error can have very different consequences

MAPE tends to focus more on the series with small values as they tend to have large errors



Solution: symmetric MAPE (sMAPE)

$$\text{sMAPE} = \frac{1}{n} \sum_{t=1}^n \left| 100 \frac{y_t - \hat{y}_t}{\frac{|y_t| + |\hat{y}_t|}{2}} \right| = 200 \frac{1}{n} \sum_{t=1}^n \left| \frac{y_t - \hat{y}_t}{|y_t| + |\hat{y}_t|} \right|$$

- Achieves symmetry from before, but breaks another symmetry: overprediction is penalised less than underprediction (as dividing by the prediction)
- Still has problems with zeros; if both actual and prediction are zero, it is not defined
- If there is a zero in the data and we don't predict an exact zero, the sMAPE is maximal.
- This has large implications for intermittent data: If you use sMAPE to evaluate intermittent forecasts, your main concern will be to predict zeros as exact zeros.
- Big advantage: sMAPE is bounded by 200, will never go higher.

modified sMAPE (Suilin, 2017)

$$\text{msMAPE} = 200 \frac{1}{n} \sum_{t=1}^n \left| \frac{y_t - \hat{y}_t}{\max(|y_t| + |\hat{y}_t| + \epsilon, 0.5 + \epsilon)} \right|$$

With a default of $\epsilon = 0.1$

- ad-hoc solution for problems with zeros and small values
- will make the evaluation skewed and is not minimised by any meaningful summary statistic of the forecast distribution

Mean Arctangent Abs. Percentage Error (MAAPE)

$$\text{MAAPE} = \frac{1}{n} \sum_{t=1}^n \arctan \left(\left| \frac{y_t - \hat{y}_t}{y_t} \right| \right)$$

- Proposed by Kim and Kim (2016)
- Authors argue it is as interpretable as MAPE
- Advantage: It is $\pi/2$ instead of ∞ , when y_t is zero
- Has the same problem as sMAPE that if y_t is zero, it always is maximal, no matter what \hat{y}_t is
- Therefore, not good for intermittent data

Relative errors and error measures

Use a benchmark method B (usually the naive forecast)

Relative Errors:

$$RE_t = \frac{y_t - \hat{y}_t}{y_t - \hat{y}_{tB}}, \quad MRAE = \frac{1}{n} \sum_{t=1}^n \left| \frac{y_t - \hat{y}_t}{y_t - \hat{y}_{tB}} \right|.$$

→ Same problems as before, if true values are zero and predictions from the benchmarks are zero, etc.

Relative Error Measures:

$$RelMAE = \frac{MAE}{MAE_B}.$$

Only has benefits if we are evaluating many predictions on the same scale. If series on different scales and only one forecast per series, it has the same problems as before.

Calculating error measures per series and across series

- Three dimensions along which we can potentially average:
 - horizons
 - rolling origins
 - different time series
- Each dimension can collapse: only one horizon, only one origin, only one time series
 - this can cause problems with divisions by zero, small numbers, and others

Calculating error measures per series and across series (2)

- We want to aggregate and divide by a normalising factor
- If we divide first, we have problems with division by zero and numerical instability
- If we aggregate first, we need to make sure that what we aggregate is on the same scale
- That usually means we don't want to aggregate from different series before normalising.
- Already within a single series there can be problems (see the WAPE)

Calculating error measures per series and across series (3)

Weighted (Mean) Absolute Percentage Error (WMAPE)

$$\text{WMAPE} = \frac{\sum_{t=1}^n |y_t - \hat{y}_t|}{\sum_{t=1}^n |y_t|}$$

- Belt (2017) and Kolassa and Schütz (2007) argue for this measure to be used in demand forecasting
- Called “weighted” as MAE can be seen as a weighted (scaled) version of MAPE
- WMAPE is also called MAD/mean ratio (if all actuals are non-negative)

Squared and bias error measures

Weighted Bias Percentage Error (WBPE):

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

Weighted Root Mean Squared Percentage Error (WRMSPE):

$$\text{WRMSPE} = \frac{\sqrt{\sum_{t=1}^n (y_t - \hat{y}_t)^2}}{\sum_{t=1}^n |y_t|}$$

As discussed before, WRMSPE measures for an unbiased model the standard deviation

Squared and bias error measures (2)

- In retail/e-commerce, it is beneficial to monitor both bias and overall error.
- If one overall number needs to be measured, practitioners have proposed to use $WAPE+WBPE$

Calculating WAPE/WRMSPE per series and across series

- In the WAPE, we divide by $\sum_{t=1}^n |y_t|$
- This sum can run over each series individually or over all series
- Running over all series:
 - It is now an MAE (i.e., a scaled measure) globally scaled to be a percentage
 - Essentially same as MAE, but slightly more interpretable
 - As in MAE, series with higher values get more importance in the error
- Running per series:
 - Scale-free
 - Only works if we have more than one forecast, otherwise same as MAPE
 - If the whole test set is zero, we divide by zero

Advantages and disadvantages of WAPE/WRMSPE

Advantages:

- Only divides by zero if whole test set is zero
- Minimised by the median of the forecast distribution
- Interpretable as a percentage

Disadvantages:

- Only consistent estimator when series is stationary (Hyndman, 2025)
- Can be seen as RelMAE where benchmark method is a constant zero: works well if a constant zero is a reasonable benchmark

Mean Absolute Scaled Error (MASE)

- Proposed by Hyndman and Koehler (2006)
- Defined as the MAE divided by the MAE of the (seasonal) naive forecast over the training part of the time series.

$$\text{MASE} = \frac{\sum_{t=1}^n |\hat{y}_t - y_t|}{\frac{n}{m-s} \sum_{k=s+1}^m |y_k - y_{k-s}|}$$

- RMSSE: Equivalent with squared errors
- Used in the M5 competition

$$\text{RMSSE} = \sqrt{\frac{\sum_{t=1}^n (\hat{y}_t - y_t)^2}{\frac{n}{m-s} \sum_{k=s+1}^m (y_k - y_{k-s})^2}}$$

MASE and RMSSE: Advantages

Resolve many problems that other measures have:

- are minimised by the mean/median of the forecast distribution
- are scale-free
- hardly any problems with zeros or intermittency
- cope well with many types of non-stationarity
 - remainder of the benchmark needs to be stationary for estimator to be consistent
 - when using naive: series needs to be difference stationary
- Interpretation: If $MASE < 1$, forecast is on average better than naive

MASE and RMSSE: Problems

- As with other measures, MASE/RMSSE depend on how meaningful the chosen benchmark is
- The naive forecast works a lot better on certain (smooth) series than on others.
- Different benchmarks in use (naive, seasonal naive, could also use more complex methods like ARIMA, SES, etc.)
- Difficult to choose benchmark with fixed origin forecasting: many different horizons are in use
- Series can have significant changes between training and test set

→ In practice often difficult to interpret

→ High or low MASE is not necessarily equivalent with good or bad (absolute) performance.

Examples for problems with MASE/RMSSE

Should you use MASE or RMSSE?

- Why did the original authors propose MASE and not RMSSE?
 - Original paper proposed MASE to stay close to MAPE so that people would switch
 - Idea was that MASE was easier to understand and more likely to be adopted
 - No big difference between MASE and RMSSE for classical models that assume Gaussian errors
- RMSSE adequate in many more situations than MASE (equivalent to a discussion RMSE vs MAE)

Error measures: Summary

Scaling	Error Measures	Count Data (>>0)	Seasonality	Trend (Linear/Exp.)	Unit Roots	Heteroscedasticity	Structural Breaks (With Scale Differences)			Intermittence	Outliers	
							Forecast Horizon	Training Region	Forecast Origin			
None	RMSE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	
	MAE	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	
Actual Values	OOS Per Step	MAPE	✗	✓	✓ ¹	✓ ¹	✓	✓	✓	✗	✗	
		RMSPE	✓	✗	✓	✓ ¹	✓	✓	✓	✗	✗	
		sMAPE	✓	✓	✓	✓	✓	✓	✓	✓	✓	
		msMAPE	✓	✓	✓	✓	✓	✓	✓	✓	✓	
		WAPE	✓	✓	✗	✗	✓	✗	✓	✓	✗	✗
	Per Series	WRMSPE	✓	✓	✗	✗	✓	✗	✓	✓	✓ ¹	✗
		In-Sample Per Series	✓	✓	✗	✗	✓	✗	✗	✗	✓	✗
	OOS All Series	sMSE	✓	✗	✗	✗	✓	✗	✗	✗	✓ ¹	✗
		ND	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓
		NRMSE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗
Benchmark Errors	OOS Per Step	MRAE	✓ ¹	✓ ¹	✗	✗	✓	✓	✓ ¹	✓	✗	✓ ¹
		MdRAE	✓ ¹	✓ ¹	✗	✗	✓	✓	✓	✓	✗	✓ ¹
		GMRAE	✓ ¹	✓ ¹	✗	✗	✓	✓	✓ ¹	✓	✗	✓ ¹
		RMRSE	✓ ¹	✓ ¹	✗	✗	✓	✓	✓ ¹	✓	✓ ¹	✗
	Per Series	Relative Measures	✓ ¹	✓ ¹	✗	✗	✓	✓	✓ ¹	✓	✓ ¹	✓ ¹
		In-Sample Per Series	MASE	✓ ¹	✓ ¹	✓	✓	✓	✓	✓ ¹	✗	✗
	In-Sample All Series	RMSSE	✓ ¹	✓ ¹	✓	✓	✓	✓	✓ ¹	✗	✓ ¹	✗
		✓ ¹	✓ ¹	✓	✓	✓	✓	✓	✓	✓	✓	
None	Transformations	✓	✓	✓ ¹	✓	✓ ¹	✓	✓	✓	✓	✓	

Table 1: Checklist for Selecting Error Measures

Error measures: Summary (2)

- To get a scale-free measure, we need to divide by a normalising factor
- Due to the potential non-stationarity and non-normality (e.g. intermittency) of the series, it has turned out to be extremely difficult to do this in a way that always works
- Today, the RMSSE and MASE are standard error measures, e.g., in the recent M5 competition
 - but have problems with structural breaks between training and test set
 - may not be interpretable, especially with long horizons / fixed origins
- It depends on the characteristics of your data which measure will be adequate

Error measures: Recommendations

- If you currently use the MAPE or sMAPE, switch to something else
- Do not invent your own measure, it will be more difficult than you think
- Some people (S. Kolassa) argue against using multiple measures
- Different measures are minimised by different summary statistics over the forecast distribution
- My take: choose a primary metric (RMSSE) that coincides with your loss function (L2)
- And then using the others for sanity-checking seems reasonable

Error measures: Recommendations (2)

- When building a global model, often the series have meaningful scales (SKUs, dollars, ...)
- If you don't need a scale-free measure, better stick to MAE, RMSE

Error measures: Recommendations (3)

If you need a scale-free measure:

- if you will benchmark different methods broadly to conclude which one is best and you don't need interpretability
 - this is the standard scenario for scientific papers about forecasting methodology
 - use RMSSE
 - use MASE only if there are reasons to elicit the median of the forecasting distribution, and if all your compared methods use this loss
 - if you evaluate over a mix of methods, some trained with L1 loss, some with L2 (or similar): report both MASE and RMSSE

Error measures: Recommendations (4)

If you need a scale-free measure and interpretability:

- use WAPE, WRMSPE, WBPE; proceed with caution
- WAPE, WRMSPE, WBPE could also be normalised with a sum over the training set instead of the test set (??)
- use RMSSE/MASE but with the denominator running over the test set if horizons are long, and with the same (fixed-origin) setup for your benchmark method (??)

Probabilistic forecasting

Probabilistic forecasting

Main ways for probabilistic forecasting:

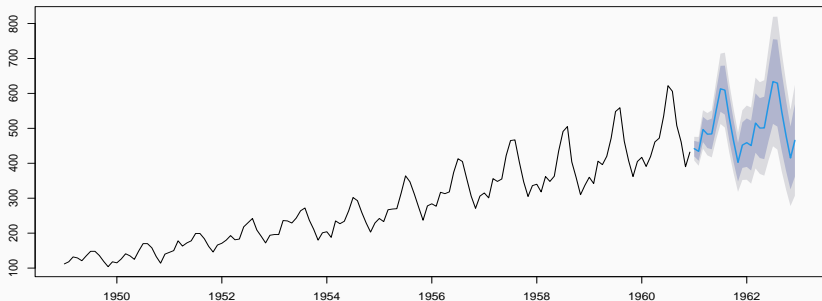
- use analytical prediction intervals
- bootstrapping
- using a Bayesian model (MCMC sampling)
- forecast the parameters of a distribution
- use quantile regression (pinball loss)
- determine uncertainty empirically through backtesting (conformal prediction)
- specialised approaches, e.g., levelset approach (Hasson et al., 2021)

(see also the review by Kneib et al. (2023))

Use analytical prediction intervals

- Possible for some (well-understood) models
- Usually assume normally-distributed errors
- ETS, ARIMA do this (see Hyndman and Athanasopoulos (2018))
- intervals tend to be too narrow (Bermudez et al., 2010)

Forecasts from ETS(M,Ad,M)



Simulation and bootstrapping

- slow
- intervals can also be too narrow if, e.g., they only consider parameter uncertainty
- either need to bootstrap residuals (from an additional validation set) or assume a distribution
- then simulate forecasting paths by feeding the generated/bootstrapped values back into the model
- see Hyndman and Athanasopoulos (2018), Section “Neural Networks”

MCMC sampling

- needs to be a Bayesian model
- Examples: LGT (Smyl et al., 2025, Long et al. (2025)), Orbit (Ng et al., 2020), Bayesian ETS (Bermudez et al., 2010)
- slow

Determine uncertainty empirically through backtesting

- Also called conformal prediction, where it has some theory behind it
- often used by companies in practice
- leads to more realistic prediction intervals
- potentially needs a lot of past data and rolling origin forecasts (large validation sets)

→ combine with cross-validation and bootstrapping schemes

Forecast the parameters of a distribution

- e.g., assuming a normal distribution: μ, σ
- DeepAR (Salinas et al., 2019): Normal distribution and negative binomial distribution
- NGBoost (Duan et al., 2020)
- GAMLSS (R. A. Rigby and D. M. Stasinopoulos, 2005, Ziel (2022)), LightGBMLSS (Maerz, 2025)
- Have to assume a certain distribution
- Good if we have limited amounts of data, or knowledge of the distribution

Quantile regression (pinball loss)

- implemented in, e.g., MQ-RNN (Wen et al., 2017), LightGBM
- no distribution assumptions need to be made; therewith better if a lot of data are available
- fast to compute and easy to implement
- only certain quantiles can be obtained, not the full distribution
- in practice, often 5 or 7 quantiles are enough anyway
- with, e.g., a Neural Network, we can fit different quantiles at the same time, by having multiple outputs
- can interpolate between quantiles to get full distribution (Gasthaus et al., 2019)

Pinball loss function

$$\begin{aligned}L_u(y, q_t^{[u]}) &= (y - q_t^{[u]})u && \text{if } y \geq q_t^{[u]} \\ &= (q_t^{[u]} - y)(1 - u) && \text{if } q_t^{[u]} > y\end{aligned}$$

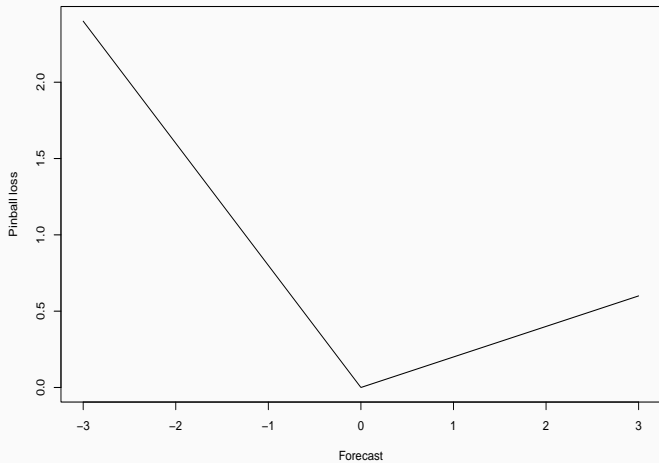
y is the true value

u is the target quantile, for example: $u = 0.9$

$q_t^{[u]}$ is the quantile forecast for target quantile u

It can be proved that minimizing the pinball loss results in the most accurate quantile

Pinball loss (2)



Evaluation of probabilistic forecasts

Simple hit/miss rule

- can be used to evaluate forecasting interval
- directly interpretable (e.g., “80% of forecasts are within the interval”)
- doesn't consider the magnitude of the error
- Problem: trivial solutions are possible by grossly over- and underpredicting certain amounts of times

Mean Scaled Interval Score (MSIS)

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

- used in the M4 competition (Makridakis et al., 2018)
- evaluates a prediction interval
- sum over the size of the intervals and the magnitude of error for points that lie outside of the interval
- If model is optimised for pinball loss, it will not necessarily perform well under MSIS (Smyl, 2020)

Mean Scaled Interval Score (MSIS) (2)

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

Scaled Pinball Loss (SPL)

$$SPL[u] = \frac{1}{h} \frac{\sum_{t=n+1}^{n+h} (u(y_t - q_t^{[u]}) \mathbb{1}_{q_t^{[u]} \leq y_t} + (1 - u)(q_t^{[u]} - y_t) \mathbb{1}_{q_t^{[u]} > y_t})}{\frac{1}{n-m} \sum_{t=m+1}^n |y_t - y_{t-m}|}$$

- 90th quantile forecast: $u = 0.9$, $q_t^{[0.9]}$
- 10th quantile forecast: $u = 0.1$, $q_t^{[0.1]}$
- $\mathbb{1}$ is the indicator function
- n is the time index of the last observation in the training set
- h is the forecast horizon

Weighted Scaled Pinball Loss (WSPL)

- used in the M5 uncertainty track
- Combines, e.g., each series' $SPL[0.1]$ and $SPL[0.9]$ together by taking the average
- Then (optionally) weighs SPL by a series-specific weight.

$$WSPL = \sum_{i=1}^n w_i \times \frac{1}{k} \sum_{j=1}^k SPL[u_k]$$

- k represents the number of quantiles
- We can set w_i to $\frac{1}{n}$ for all n time series to weight them equally
- Lower WSPL scores indicate more precise forecasts

Evaluation of full forecast distributions: CRPS

- Overviews by Gneiting and Raftery (2007); Jordan et al. (2017)
- Continuous Ranked Probability Score (CRPS)

$$\text{CRPS}(F, y) = \int_{-\infty}^{\infty} (F(x) - \mathbf{1}_{x \geq y})^2 dx$$

- y is the observed value, $F(x)$ is the predicted CDF
- integrates over all quantiles
- generalises the MAE to the probabilistic case

And there are more measures...

- logarithmic score
- energy score
- variogram score
- for multivariate forecasting: Ziel and Berk (2019)

Thank You

<https://www.cbergmeir.com>

bergmeir@ugr.es

References i

- F. Bell and S. Smyl. Forecasting at Uber: An introduction, 2018. URL <https://eng.uber.com/forecasting-introduction/>. Accessed 2 September 2020.
- T. Belt. When is forecast accuracy important in the retail industry? effect of key product parameters. *Master's thesis, Aalto University.*, 2017.
- C. Bergmeir and J. M. Benítez. On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191:192–213, 2012.

- C. Bergmeir, R. J. Hyndman, and B. Koo. A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, 120:70–83, 2018.
- J. Bermudez, J. Segura, and E. Vercher. Bayesian forecasting with the holt-winters model. *Journal of the Operational Research Society*, 61(1):164–171, 2010.
- P. Burman, E. Chow, and D. Nolan. A cross-validators method for dependent data. *Biometrika*, 81(2):351–358, 1994. ISSN 00063444.

- T. Duan, A. Anand, D. Y. Ding, K. K. Thai, S. Basu, A. Ng, and A. Schuler. Ngboost: Natural gradient boosting for probabilistic prediction. In *International Conference on Machine Learning*, pages 2690–2700. PMLR, 2020.
- J. Gasthaus, K. Benidis, Y. Wang, S. S. Rangapuram, D. Salinas, V. Flunkert, and T. Januschowski. Probabilistic forecasting with spline quantile function rnns. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1901–1910, 2019.

- T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.
- H. Hasson, B. Wang, T. Januschowski, and J. Gasthaus. Probabilistic forecasting: A level-set approach. *Advances in Neural Information Processing Systems*, 34, 2021.
- H. Hewamalage, K. Ackermann, and C. Bergmeir. Forecast evaluation for data scientists: common pitfalls and best practices. *Data Mining and Knowledge Discovery*, 37(2): 788–832, 2023. doi: 10.1007/s10618-022-00894-5.

References v

- R. Hyndman and Y. Khandakar. Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, 27(3):1–22, 2008.
- R. Hyndman and A. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688, 2006.
- R. J. Hyndman. Wape: Weighted absolute percentage error, 2025. URL <https://robjhyndman.com/hyndsight/wape.html>.
- R. J. Hyndman and G. Athanasopoulos. *Forecasting: principles and practice, 2nd edition*. OTexts, otexts.com/fpp2, 2018.

- A. Jordan, F. Krüger, and S. Lerch. Evaluating probabilistic forecasts with scoringrules. *arXiv preprint arXiv:1709.04743*, 2017.
- S. Kim and H. Kim. A new metric of absolute percentage error for intermittent demand forecasts. *International Journal of Forecasting*, 32(3):669–679, 2016. ISSN 0169-2070.
- T. Kneib, A. Silbersdorff, and B. Säfken. Rage against the mean—a review of distributional regression approaches. *Econometrics and Statistics*, 26:99–123, 2023.

- S. Kolassa. Why the "best" point forecast depends on the error or accuracy measure. *International Journal of Forecasting*, 36(1):208–211, 2020.
- S. Kolassa and W. Schütz. Advantages of the mad/mean ratio over the mape. *Foresight: The International Journal of Applied Forecasting*, 6:40–43, 01 2007.
- G. M. Ljung and G. E. Box. On a measure of lack of fit in time series models. *Biometrika*, 65(2):297–303, 1978.

- X. Long, D. F. Schmidt, C. Bergmeir, and S. Smyl. Fast gibbs sampling for the local-seasonal-global trend bayesian exponential smoothing model. *Statistics and Computing*, 35(3):77, 2025.
- A. Maerz. Lightgbmlss, 2025. URL <https://statmixedml.github.io/LightGBMLSS/>. Accessed 2 October 2025.
- S. Makridakis, E. Spiliotis, and V. Assimakopoulos. The m4 competition: Results, findings, conclusion and way forward. *International Journal of Forecasting*, 34(4):802–808, 2018.

- E. Ng, Z. Wang, H. Chen, S. Yang, and S. Smyl. Orbit: Probabilistic forecast with exponential smoothing. *arXiv preprint arXiv:2004.08492*, 2020.
- R. A. Rigby and D. M. Stasinopoulos. Generalized additive models for location, scale and shape,(with discussion). *Applied Statistics*, 54:507–554, 2005.
- J. Racine. Consistent cross-validatory model-selection for dependent data: hv-block cross-validation. *Journal of Econometrics*, 99(1):39–61, 2000.

- D. Salinas, V. Flunkert, J. Gasthaus, and T. Januschowski. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 2019. ISSN 0169-2070.
- S. Smyl. A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. *International Journal of Forecasting*, 36(1):75–85, 2020.
- S. Smyl, C. Bergmeir, A. Dokumentov, X. Long, E. Wibowo, and D. Schmidt. Local and global trend bayesian exponential smoothing models. *International Journal of Forecasting*, 41(1):111–127, 2025.

- A. Suilin. kaggle-web-traffic.
https://github.com/Arturus/kaggle-web-traffic/blob/master/how_it_works.md, 2017.
Accessed: 2018-11-19.
- R. Wen, K. Torkkola, B. Narayanaswamy, and D. Madeka. A Multi-Horizon quantile recurrent forecaster. In *Neural Information Processing Systems*, Nov. 2017.
- F. Ziel. M5 competition uncertainty: Overdispersion, distributional forecasting, gamlss, and beyond. *International Journal of Forecasting*, 38(4):1546–1554, 2022.

F. Ziel and K. Berk. Multivariate forecasting evaluation: On sensitive and strictly proper scoring rules. *arXiv preprint arXiv:1910.07325*, 2019.