Measuring forecasting performance
A complex task!

Nikolaos Kourentzes[a]
Juan R. Trapero[b]
Ivan Svetunkov[a]
[a]Lancaster University; [b]Universidad de Castilla-La Mancha

International Symposium on Forecasting 2016
21/06/2106

Lancaster University Management School | Lancaster Centre for Forecasting | Forecasting Society | UCLM

Agenda

1. What is the issue with current metrics?
2. Metrics of bias & accuracy
3. The Mean Root Error (MRE)
4. The Bias Coefficient
5. Examples

LCF

What is good `forecasting performance'?

- Forecasting is important – that's why we are all here!
- Evaluating forecasting performance is necessary, but what constitutes good forecasting performance?
  - **Forecast bias**: on average how much we over/under-forecast
  - Forecast error magnitude (**accuracy**): how big are the errors irrespective of direction
- A `good performing forecast' should be fine at both → these are not always highly correlated!
- How to measure accuracy and bias?

LCF

Metrics

- A lot of research and innovations → mostly motivated by the statistical properties of metrics
- Main focus on accuracy not bias
- What should a good metric do (not all necessary, but nice to have)?
  - Be unbiased and symmetric (unless weighting is desirable), unlike MAPE
  - Scale-independence, unlike MSE & MAE
  - Possible to calculate in a wide range of circumstance, unlike MAPE & GMRAE
  - Easy to interpret (correctly!) to non-statisticians, unlike sMAPE & MASE
  - Report what is supposed to! E.g., for slow moving items most metrics are misleading.

LCF

Metrics of accuracy (1/3)

- Scale dependent: MSE, RMSE, MAE, …
  - Not useful for presenting accuracy across series
  - Consider your loss function
- Percentage errors: MAPE, sMAPE, MAAPE, …
  - Biased (not symmetric) and problematic in calculation
  - MAPE is regarded as easy to interpret, but in fact misleading (not symmetric)
  - sMAPE is just wrong
  - Mean Arctangent Absolute Percentage Error:
    - $\mathrm{MAAPE} = n^{-1} \sum_{j=1}^{n} \left(\arctan \left(\left| \frac{y_j - f_j}{y_j} \right| \right) \right)$
    - Nice idea to avoid scaling issues, but: not-symmetric; undefined when y = f = 0; low sensitivity; interpretation in radians!

LCF

Metrics of accuracy (2/3)

- Scaled errors: MAE/mean, sMAE, sMSE, MASE, …
  - MAE/mean scales on sample used for measurement, not great for slow movers. sMAE & sMSE scale with in-sample mean so less problematic
  - But assumes a lot: why is the mean an appropriate scaling factor?
  - MASE:
    - Similar to MAE/mean, but instead of mean use in-sample Naïve MAE → hard to interpret (different samples/horizons)
    - Also biased, should be using geometric mean

LCF

Metrics of accuracy (3/3)

- Relative errors (relative on individual errors): MRAE, MdRAE, GMRAE, …
  - It is a ratio → use geometric mean
  - GMRAE:
    - Easy to interpret and forces use of benchmark
    - But can be problematic to calculate → Trimming is subjective
- Relative errors (relative on summary errors): RMAE (CumRAE), AvRelMAE, …
  - Retain interpretability while typically easy to calculate
  - Ratio → use geometric mean → AvRelMAE
  - AvRelMAE: almost great! What about slow movers (calculation and loss function)?

LCF

Metrics of bias (1/1 – There are not many!)

- Mean Error (ME)
  - Flagship bias metric, but scale dependent
  - Mean Percentage Error → do not use due to asymmetry!
  - Scaled ME (sME) → similar to sMAE and sMSE, what is your scaling?
  - One more point: ME is not `clean' bias: MSE = Var(f) + ME^2 + σ
    - OK for researchers, but do users understand this?
- Mean Directional Bias (MDB)
  - $$\mathrm{MDB} = n^{-1}\left(\sum_{e_j>0}\mathrm{sgn}(e_j) + \sum_{e_j<0}\mathrm{sgn}(e_j)\right) = n^{-1}\left(n_{pos} - n_{neg}\right)$$
  - Retains only direction, not size of bias → scale independent
  - Bounded between [-1, 1] → so great for benchmarking comparisons
- Special metrics: Periods-In-Stock (PIS), …
  - Developed for particular applications and are not general.

LCF

Agenda

1. ~~What is the issue with current metrics?~~
2. ~~Metrics of bias & accuracy~~
3. The Mean Root Error (MRE)
4. The Bias Coefficient κ
5. Examples

LCF

Root Error

- We propose a different loss function that brings some useful properties
  - Retain more information
  - Keep connection between accuracy and bias
  - Geometric interpretation
  - Symmetric & robust
- We calculate the square root of error, **positive errors remain real, negative become imaginary**:

$$z_j = \sqrt{e_j} = a_j + ib_j \qquad\qquad i^2 = -1$$

$$SRE = \sum_{j=1}^{n} \sqrt{e_j} = \sum_{j=1}^{n} a_j + i \sum_{j=1}^{n} b_j$$

$$MRE = n^{-1} SRE.$$

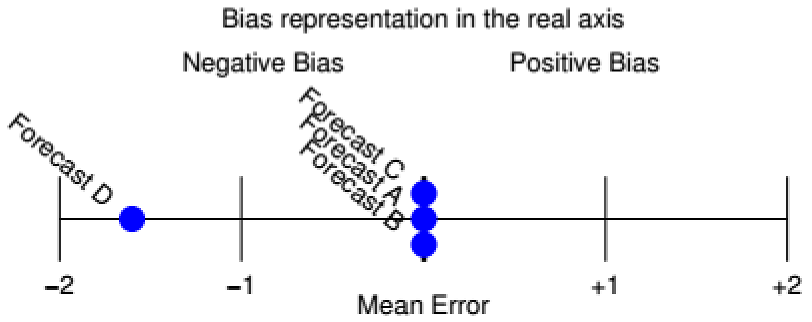
LCF

Root Error - Visualise

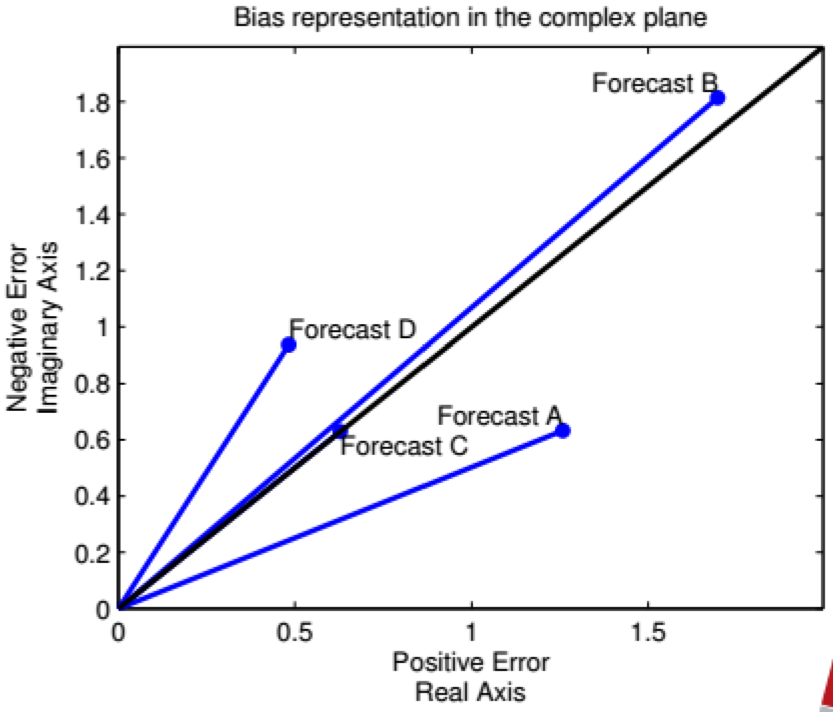
- Consider some forecasts with errors:

  A = (-10, 2, 2, 3, 3); B = (-50, 2, -1, -1, 50); C = (-3, 3, -2, 2,0); D = (-6, -5, 2, 1, 0)

Mean Error

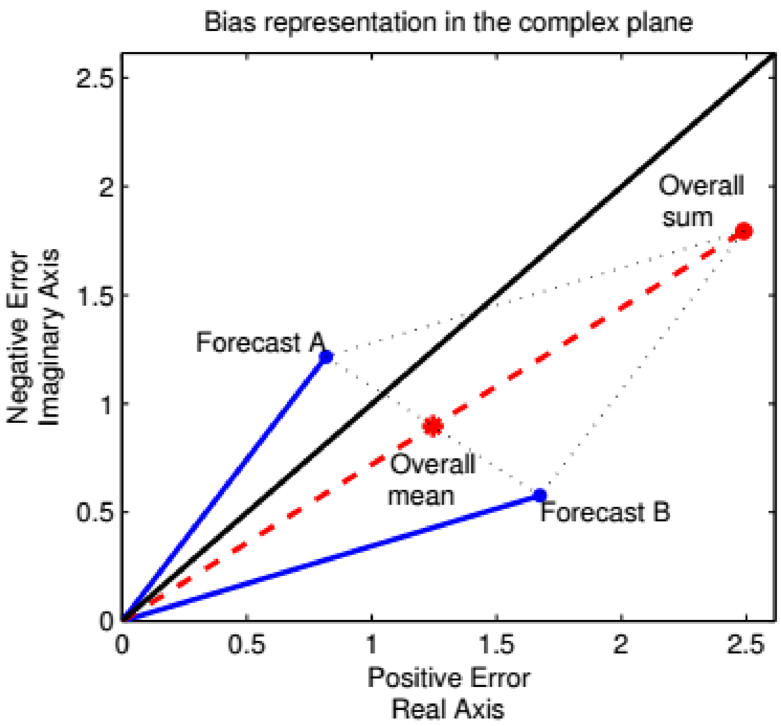


Mean Root Error

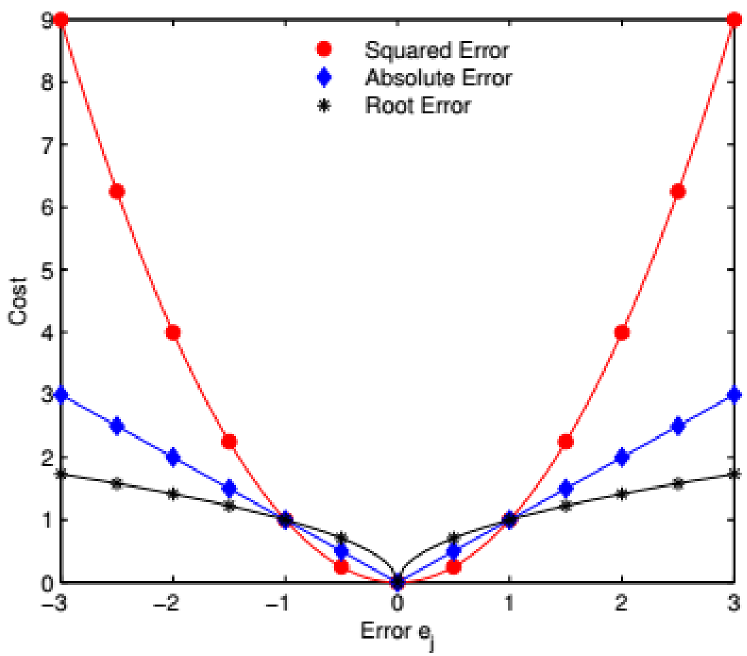


Root Error – Properties

Geometric interpretation of contribution of each forecast error



Robust & symmetric loss



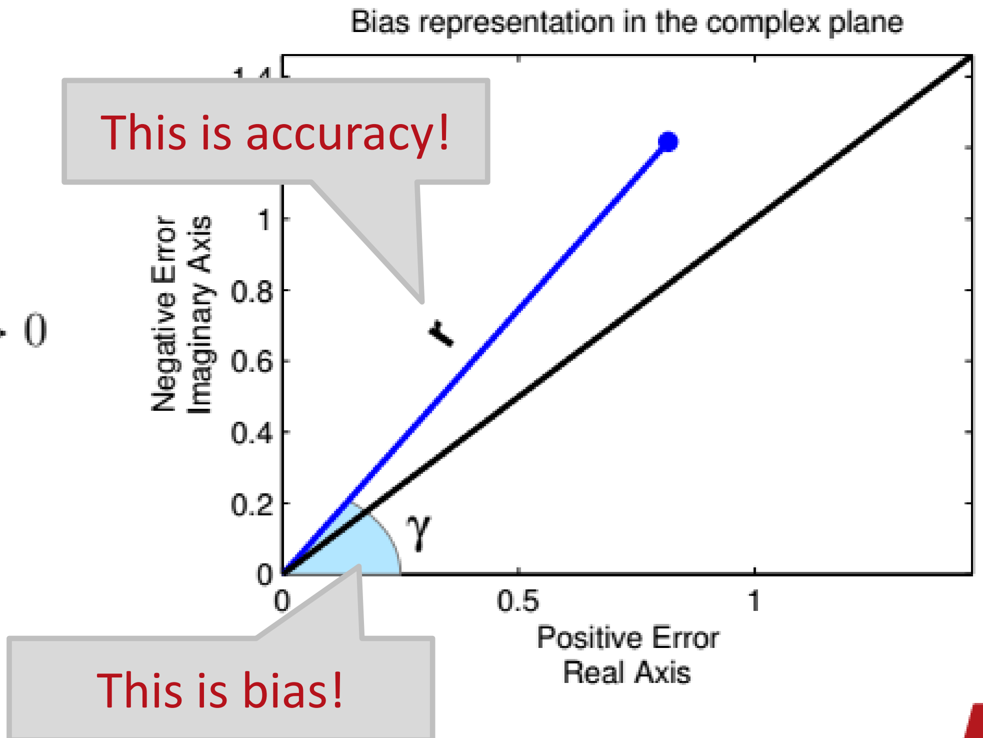
LCF

Root Error – Representations

- Any complex number has a polar coordinates view
- Using the polar we can get the magnitude r and angle γ

$$r = \sqrt{a^2 + b^2} = |z|,$$

$$\gamma = \begin{cases} \arctan(\frac{b}{a}), & \text{if } a > 0 \\ \pi/2, & \text{if } a = 0 \text{ and } b > 0 \\ \pi/4, & \text{if } a = b = 0 \end{cases}$$



- So the root error always contains both bias and accuracy and shows how they are connected!

The bias coefficient κ

- π/4 is the unbiased behaviour. We can normalise γ to a scale and unit free bias metric, the bias coefficient κ:

$$\kappa = 1 - \frac{\gamma}{\pi/4}$$

- Bounded between [-1, 1]. -1 is always negatively biased, and 1 is the opposite. 0 is unbiased.
- No units or scale: can be used to benchmark across forecasts, forecasters, companies, sectors, …
- Can be calculated always
- Has an intuitive interpretation: you are biased 100 κ %

LCF

Comments on Root Error

- Can be scaled to become scale independent (important for the accuracy side)

$$\eta_j = \sqrt{\frac{e_j}{s}} = \frac{z_j}{\sqrt{s}}$$

- Scaling factor can be anything (mean, standard deviation, MAE of in-sample Naïve, …). Scaling does not affect the bias side of the metric.
- It can be shown that accuracy part of RE can be translated into GMRAE (or equivalently GMRSE).
- It can be shown that MDB is RE without the size of errors.
- The `bias' of RE is not the bias of ME! As the accuracy of MAE is not the accuracy of MSE…

LCF

Agenda

1. ~~What is the issue with current metrics?~~
2. ~~Metrics of bias & accuracy~~
3. ~~The Mean Root Error (MRE)~~
4. ~~The Bias Coefficient κ~~
5. Examples

Fast Moving Goods

- Experiment on 229 FMCG

Metric	Naïve	ETS	MAPA	Naïve	ETS	MAPA
		Mean			Median	
sME	0.029	-0.020	**-0.014**	**0.020**	-0.036	-0.038
MPE %	$-\infty$	$-\infty$	$-\infty$	**-19.84%**	-20.91%	-22.42%
sMSE	1.961	1.205	**1.165**	1.641	0.943	**0.921**
sMAE	0.955	0.756	**0.744**	0.954	0.746	**0.728**
MAPE %	∞	∞	∞	49.43%	43.39%	**42.11%**
MAAPE	**0.437**	0.498	0.495	**0.436**	0.472	0.468
sGRMSE	0.595	0.562	**0.553**	0.612	0.553	**0.541**
κ %	**6.31%**	-10.56%	-11.60%	**5.66%**	-10.50%	-14.46%
|sMRE|	0.752	0.633	**0.625**	0.764	0.631	**0.622**

- The table tells us: **it can be calculated always**, **robust to extremes** (small difference mean vs. median) and therefore retains ranking of methods.

Slow Moving Goods

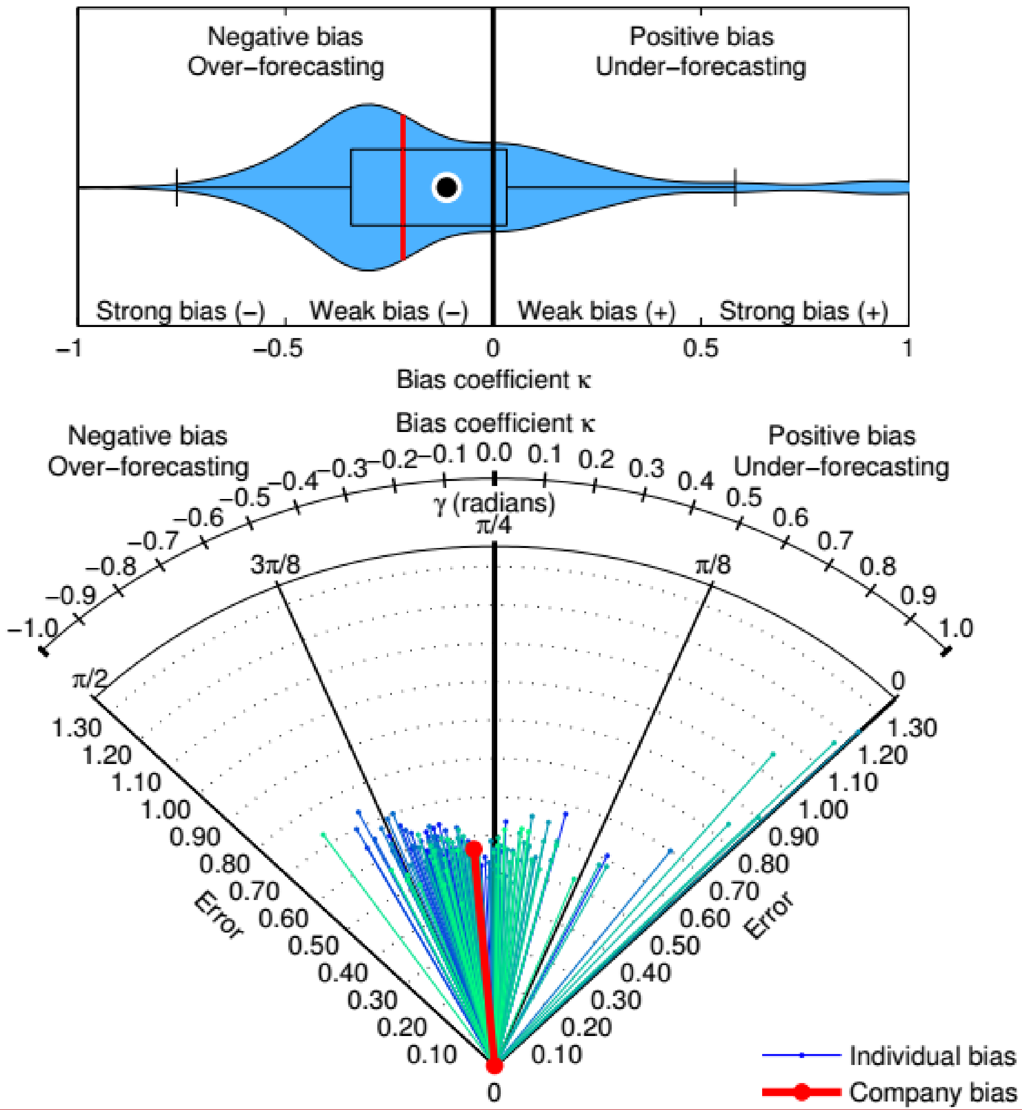
- 5,000 slow moving series. Compare against the meaningless zero-forecast.

Metric	SBA	MAPA	Zero Forecast		SBA	MAPA	Zero Forecast	
	Mean				Median			
sME	-0.013	**-0.006**	0.265	✓	-0.153	-0.146	**0.117**	–
sAPIS	22.474	22.281	**20.172**	–	17.488	17.154	**6.327**	–
sMSE	**5.348**	5.350	5.419	✓	0.167	0.163	**0.131**	–
sMAE	0.497	0.491	**0.265**	–	0.363	0.359	**0.117**	–
MAAPE*	1.498	1.498	-	–	1.501	1.501	-	–
κ %	-70.7%	**-70.2%**	100.0%	✓	-77.1%	**-76.4%**	100.0%	✓
|sMRE|	0.513	0.507	**0.127**	–	0.513	0.504	**0.104**	–

MAAPE could not be calculated for the Zero Forecast as in many cases AAPE was indeterminate. Therefore no best method is identified.

LCF

Visualisations



Conclusions

- A lot of work on accuracy, limited work on bias metrics → both are important
- A new metric: Root Error that contains both accuracy and bias
- The metric itself is complex, but the calculation of its components is trivial:
  - Accuracy: symmetric & robust and can be scaled
  - Bias: Robust & scale independent
- Bias coefficient: great for benchmarking
- Powerful visualisations → geometric interpretation of metric.
- Works as intended for several types of application.
- Connection between bias & accuracy permits modelling highly nonlinear behaviour easily.

LCF

Thank you for your attention!
Questions?

Working paper available on request!

Nikolaos Kourentzes

email: n.kourentzes@lancaster.ac.uk

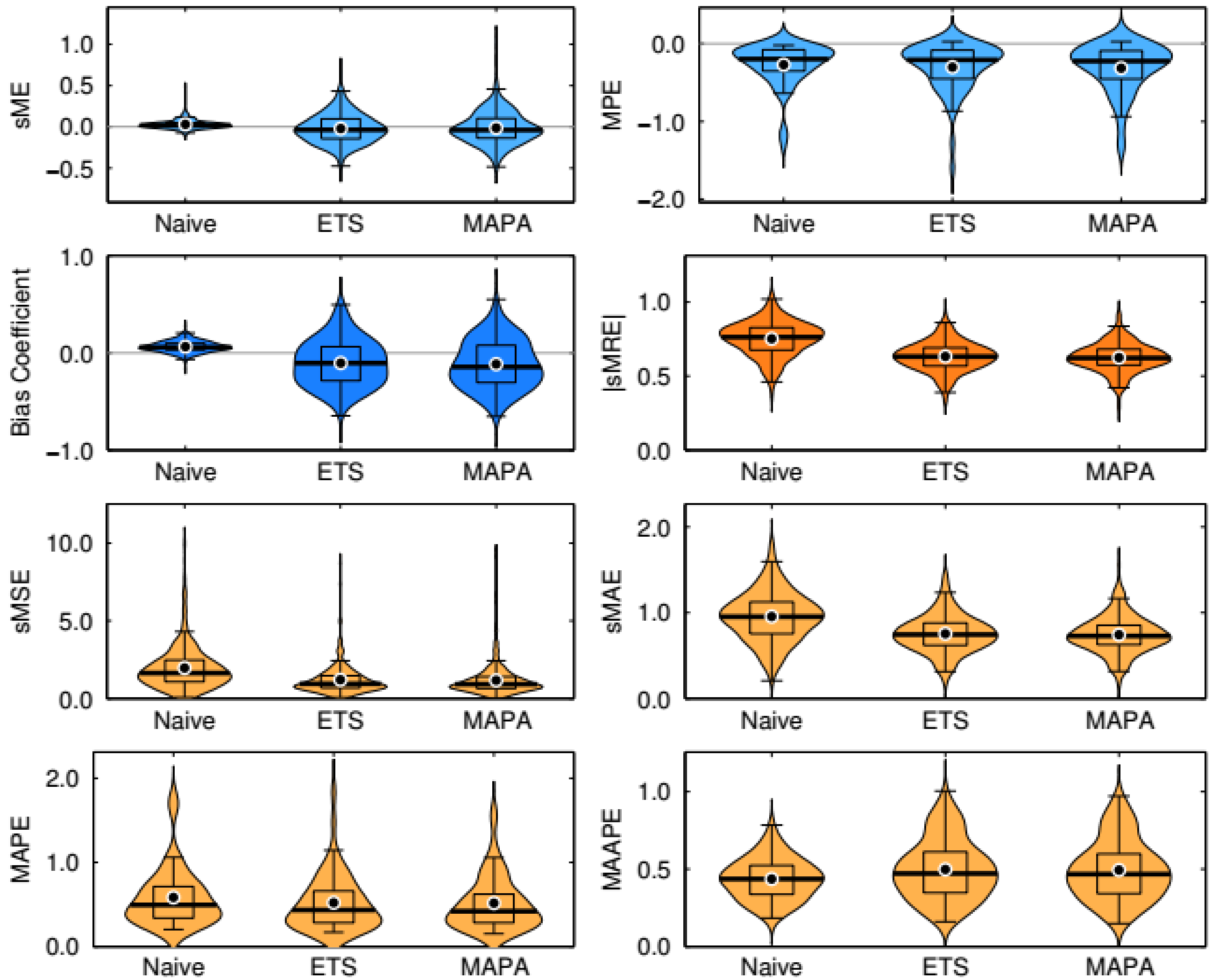
blog: http://nikolaos.kourentzes.com

Lancaster University Management School | Lancaster Centre for Forecasting

Appendix

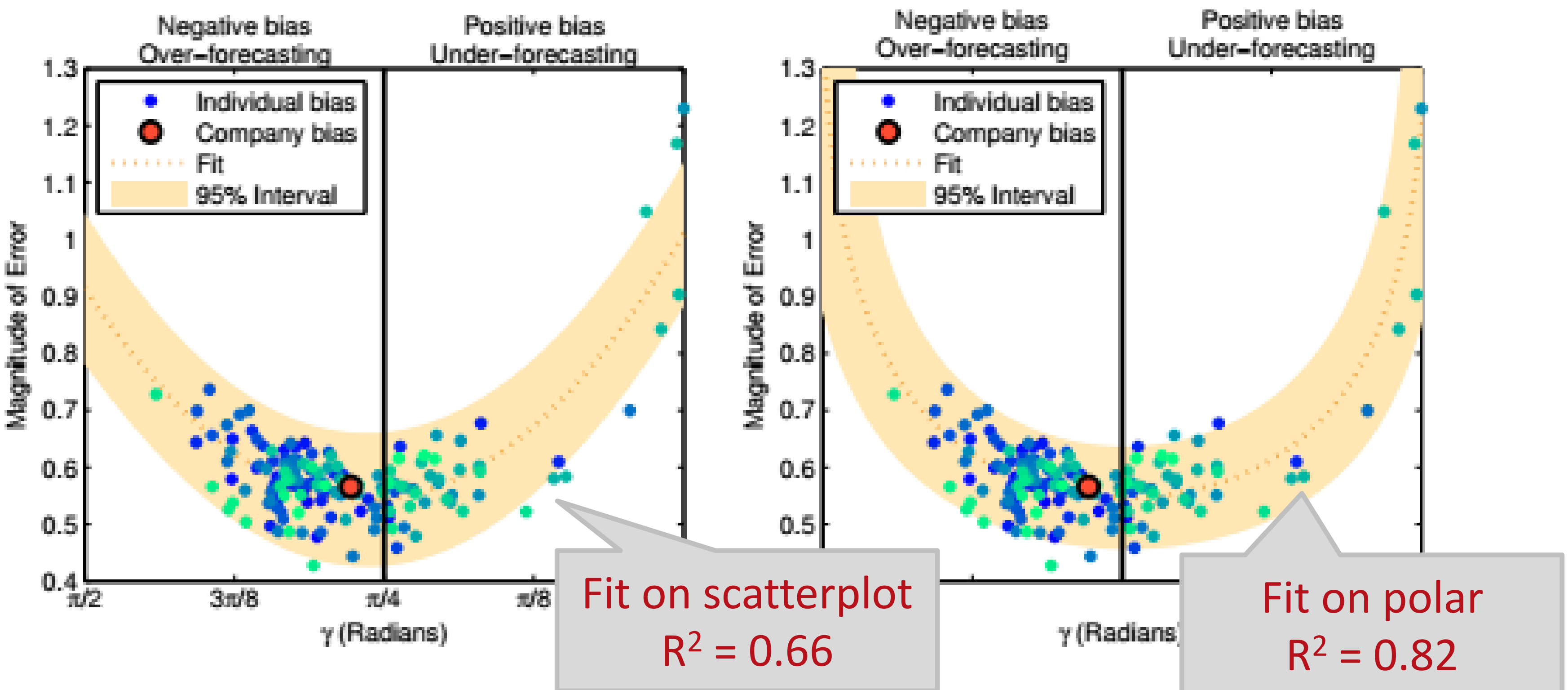
Lancaster University Management School | Lancaster Centre for Forecasting

Fast Moving Goods



Judgemental adjustments: RE trick!

- Fit a polynomial to explain the connection between forecast bias and forecast error of **final adjusted forecasts**.



- Retains the connection between bias & accuracy, allows capturing highly nonlinear behaviours easily.