Statistical and economic evaluation of forecasts in electricity markets: beyond RMSE and MAE

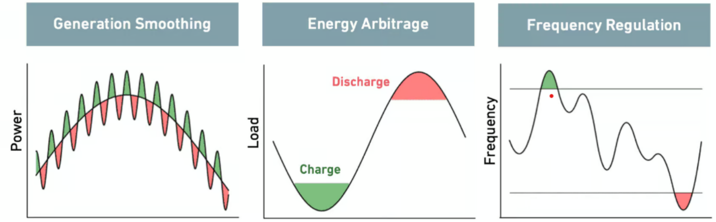
Arkadiusz Lipiecki, Katarzyna Maciejowska, Bartosz Uniejewski

Department of Operations Research and Business Inteligence, Wrocław University of Science and Technology, 50-370 Wrocław Tech. PL

Battery Energy Storage Systems (BESS)

Business models for utility scale BESS:

- Power Quality - keeping frequency and voltage within permissible limits
- Power Reliability - providing electricity in case of supply reduction or interruption
- Increased utilization - optimizing use of existing assets
- Arbitrage - exploiting temporal price differentials



BESS profits from price arbitrage

Battery specification: *Lindberg et al. (2024)*

Consider BESS that earns income from the arbitrage in DA market

$$\pi_t = 0.97 P_{t,h^{dis}} - \frac{1}{0.98} P_{t,h^{ch}} - C,$$

where

- One cycle a day, 1MWh
- h^{ch} and h^{dis} - charging and discharging hours
  - $h^{ch} < h^{dis}$
- Cost of operation, $C = 23.26$ EURO/1MWh
- Decision is made day-ahead → based on forecasts of prices

Forecast evaluation

How to choose the "best" forecast?

- Murphy (1993) identified three distinct aspects of forecasts:
  - consistency - correspondence between forecasters' judgments and their forecasts
  - quality - correspondence between the forecasts and observations
  - value - benefits from using the forecasts
- Forecasts with higher statistical accuracy may not result in a higher economic value:
  - Stratigakos et al. (2022), Lindberg et al. (2024)
- Decision-focused learning (DFL) - adds cost or profit into the estimation process
  - Zhang et al. (2023), Carriere and Kariniotakis (2019)
  - Lack of interpretability, high computation costs
  - Dedicated to particular problem

Accuracy measures

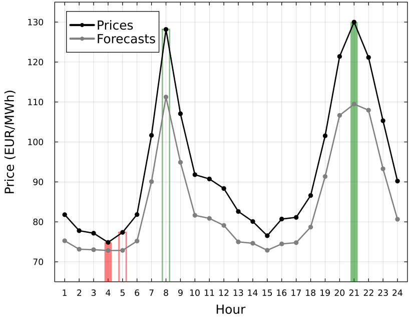
Classical measures of forecast accuracy:

$$RMSE = \sqrt{\frac{1}{24T}\sum_{t=1}^{T} e_t e_t'}$$

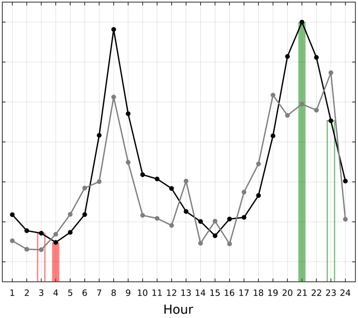
$$MAE = \frac{1}{24T}\sum_{t=1}^{T}\sum_{h=1}^{24} |e_{t,h}|$$

Where $e_{t,h} = P_{t,h} - \hat{P}_{t,h}$ is the forecast error on day t and hour h and e_t is a (24×1) vector of errors

Beyond RMSE and MAE



(a) $RMSE = 10$, $\pi = 24.68$



(b) $RMSE = 10, \pi = 0.16$

Evaluation of forecast quality:

- Dispersion of errors
- Association of forecasts
- Ability to select min/max price

Dispersion and association measures

Dispersion - show how diversified the forecast errors are within a single day

$$Cov_e = \log det\hat{\Sigma},$$

where

$$\hat{\Sigma} = \frac{1}{T}\sum_{t=1}^{T} e_t' e_t.$$

Association - the relationship between forecasts and observations

$$Corr_f = \frac{1}{T}\sum_{t=1}^{T} \rho(P_t, \hat{P}_t),$$

where P_t i $\hat{P}_t$ are (24×1) vectors of daily prices (actual and forecasts)

Selection of hours with min/max price

Min-max hours difference

$$MHD = \frac{1}{T}\sum_{t=1}^{T}\left|h_t^{(min)} - \hat{h}_t^{(min)}\right| + \frac{1}{T}\sum_{t=1}^{T}\left|h_t^{(max)} - \hat{h}_t^{(max)}\right|,$$

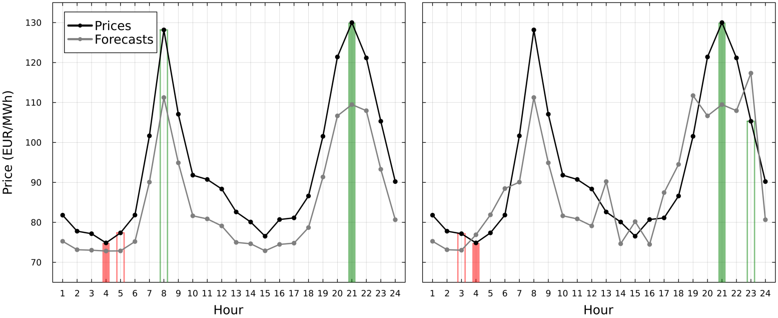
Min-max price difference

$$MPD = \frac{1}{T}\sum_{t=1}^{T}\left|P_{t,h_t^{(min)}} - P_{t,\hat{h}_t^{(min)}}\right| + \frac{1}{T}\sum_{t=1}^{T}\left|P_{t,h_t^{(max)}} - P_{t,\hat{h}_t^{(max)}}\right|,$$

where:

- $h_t^{(min)}$ and $h_t^{(max)}$ - hours of the min/max price on the day t
- $\hat{h}_t^{(min)}$ and $\hat{h}_t^{(max)}$ - hours of the min/max forecasted price

Forecast properties: $RMSE = 10$ and $MAE = 9$



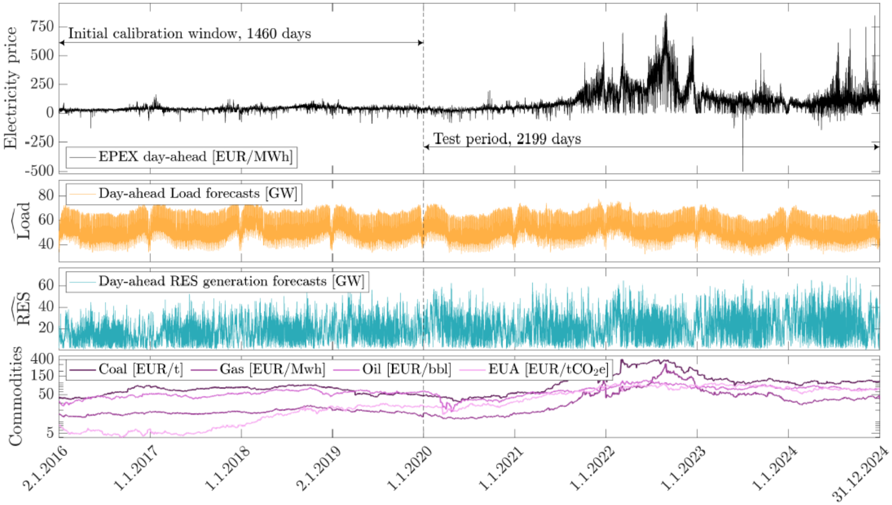
(a) $\pi = 24.68$ (b) $\pi = 0.16$

Additionally:

(a) Low Cov-e, high Corr-f, MPD $= 0.91$ and MHD $= 6.5$

(b) High Cov-e, low Corr-f, MPD $= 13.50$ and MHD $= 1.5$

Data, Germany EPEX



Pool of forecasts

Research is based on a pool of forecasts:

- three model types:
  - ARX - autoregressive models with exogenous variables: *Ziel and Weron (2018), Maciejowska et al. (2023)*
  - NARX - a non-liear, NN counterpart of ARX (single layer, 5 neurons, average over 5 runs): *Jedrzejewski et al. (2022)*
  - LEAR - a reach model that allows for cross-hour relationships, L_1 normalization (LASSO): *Uniejewski (2024)*
- eight model specifications
- seven calibration window sizes: 56, 84, 112, 182, 365, 730, 1460
- forecast averages: across calibration windows
- For each day: 192 predictions of all 24 hourly prices

Average daily profit per 1 MWh of traded electricity

Year	Oracle	Max	Min	Δ	$\%\Delta$
2020	6.23	4.32	0.45	3.87	89.5%
2021	47.43	43.57	33.63	9.94	22.8%
2022	143.18	134.05	115.52	18.53	13.4%
2023	65.44	60.22	52.32	7.9	13.1%
2024	109.05	102.57	83.38	19.19	18.7%

Correlation of profits and forecast quality measures

Pool of forecasts

For each day in the evaluation period, there is a pool of forecasts based on different models → calculate measures and average profits

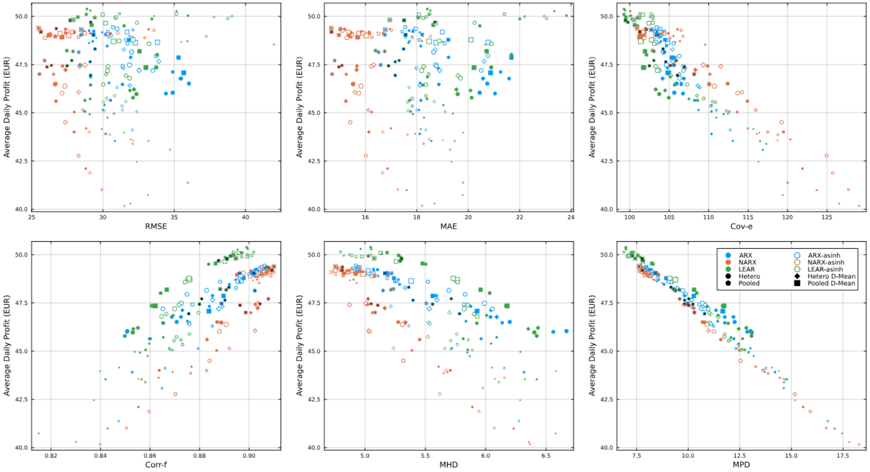
Model 1	Model 2	Model 3	Model 4	Model 5	…	Model N
$RMSE_1$	$RMSE_2$	$RMSE_3$	$RMSE_4$	$RMSE_5$	…	$RMSE_N$
MAE_1	MAE_2	MAE_3	MAE_4	MAE_5	…	MAE_N
⋮	⋮	⋮	⋮	⋮		⋮
π_1	π_2	π_3	π_4	π_5	…	π_N

Correlation of profits and forecast quality measures

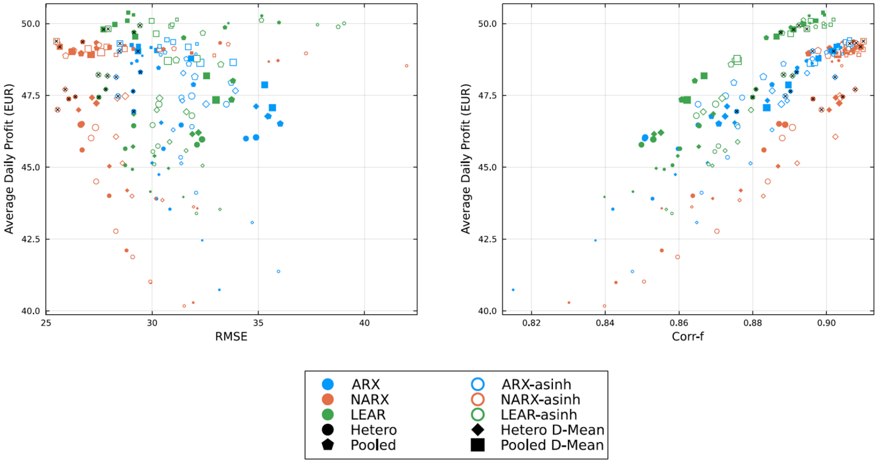
Spearman coefficient

Measure	ARX	NARX	LEAR	All
RMSE	-0.401	-0.277	-0.060	-0.169
MAE	-0.378	-0.377	-0.002	-0.172
Cov-e	-0.728	-0.861	-0.739	-0.780
Corr-f	0.896	0.843	0.865	0.802
MHD	-0.820	-0.791	0.794	-0.755
MPD	-0.969	-0.965	-0.962	-0.964

Correlation of profits and forecast quality measures

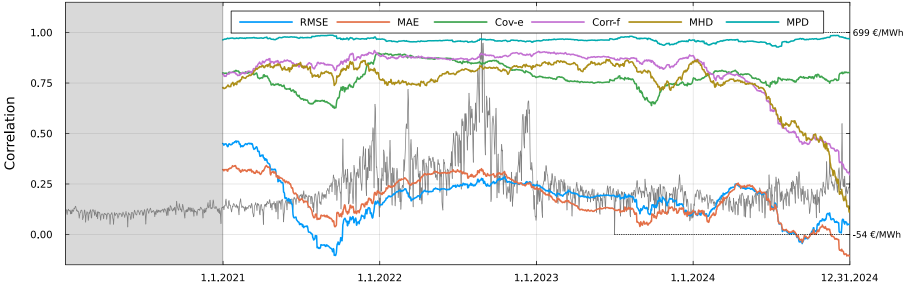


Correlation of profits and RMSE/Corr-f



Correlation of profits and quality measures

Time evolution, 365 moving window



Conclusions

- Forecast quality can be measures with various metrics: accuracy (RMSE, MAE), dispersion, association etc.
- Forecast accuracy measures are weakly associated with profits:
  - $\rho_{RMSE} = -0.169$ and $\rho_{RMSE} = -0.172$
  - confirm previous observations
- Relation between dispersion and association measures with profits is stronger:
  - dispersion: $\rho_{Cov_e} = -0.780$
  - association: $\rho_{Corr_f} = 0.802$
- The strongest correlation is between *MPD* and profits:
  - *MPD*: $\rho_{MPD} = -0.964$
  - *MHD* correlation is lower than for $Corr_f$: $\rho_{MHD} = -0.755$

Conclusions

When we take a more detailed look

- There are substantial accuracy differences between model types
  - NARX - highest accuracy
  - ARX - lowest accuracy

  However there are almost no differences between models in terms of best profits
- Pooled regression (parameters of all hours are estimated jointly)
  - has higher RMSE/MAE for ARX and LEAR than heterogeneous estimation process
  - brings the highest profits for all models (in particular NARX)
- Looking beyond RMSE/MAE shows that simple models are competitive in terms of capturing properties needed to bring high profits