



Distinguished Lecture Series 2024/2025

Electricity Price Forecasting I

Rafał Weron*

(with Julia/Python notebooks coded and explained by Arkadiusz Lipiecki)

Department of Operations Research and Business Intelligence, Wrocław Tech, Poland

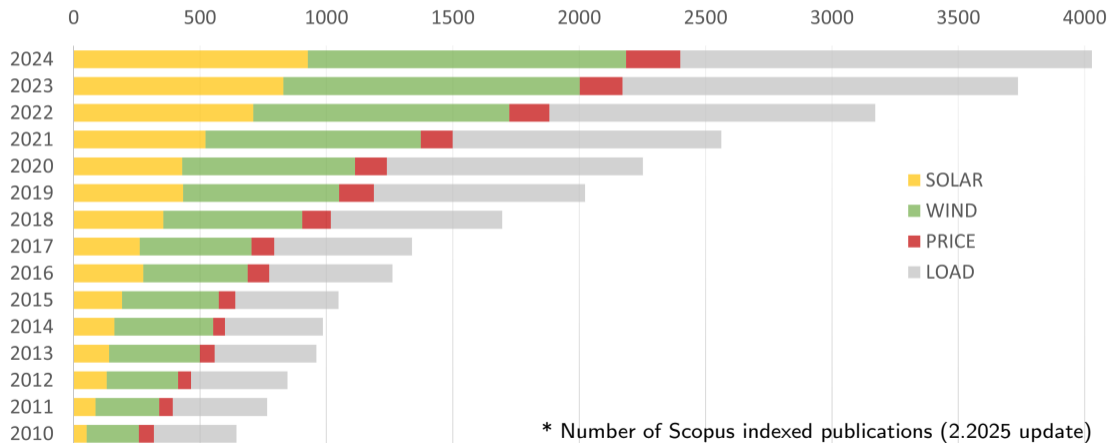
<https://p.wz.pwr.edu.pl/~weron.rafal>



*Based on joint work with K.Bilińska, Y.Chawla, K.Cheć, A.Lipiecki, K.Maciejowska, W.Nitka, T.Serafin, B.Uniejewski, P.Zaleski (Wrocław Tech), C.Challu, K.Olivares (Nixtla), T.Hong (UNCC), K.Hubicka (UBS), A.Jędrzejewski (U.Aveiro), C.Kath (RWE), J.Lago (Amazon), G.Marcjasz (Alpiq), M.Narajewski (Statkraft), J.Nasiadka (Nokia), J.Nowotarski (BNY Mellon), H.Zareipour (U.Calgary), F.Ziel (U.Duisburg-Essen)

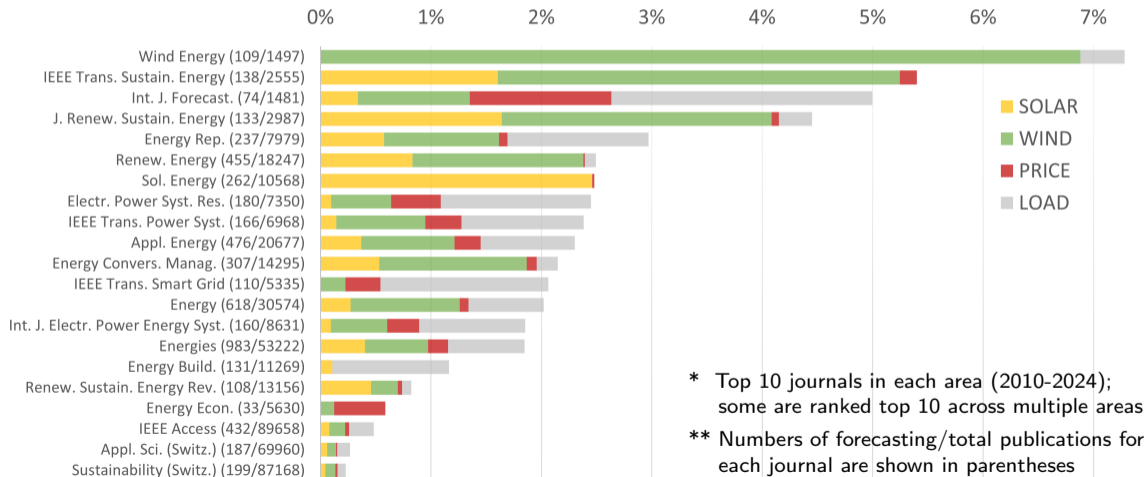
Energy (load, price, wind & solar) forecasting*

(Hong, Pinson, Wang, Weron, Yang & Zareipour, 2020, IEEE OAJPE)



Percentage of energy forecasting publications*

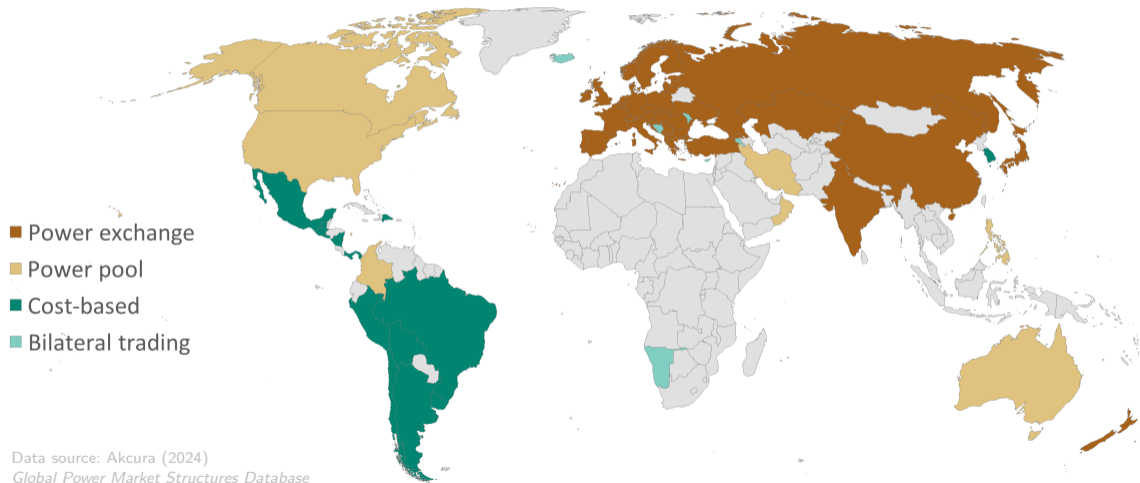
(Hong, Pinson, Wang, Weron, Yang & Zareipour, 2020, IEEE OAJPE)



* Top 10 journals in each area (2010-2024); some are ranked top 10 across multiple areas

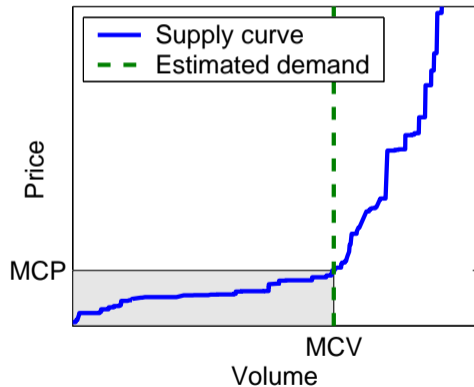
** Numbers of forecasting/total publications for each journal are shown in parentheses

Competitive power market structures across the globe

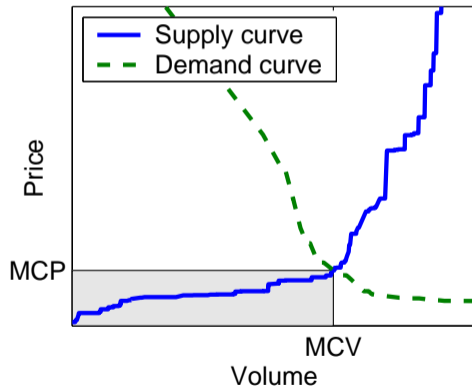


Power pool vs. power exchange

Power pool: one-sided auction



Power exchange: two-sided auction



National vs. zonal vs. nodal pricing

Single National Price

Uniform price clears across entire market.

International examples:



UK



Germany



Poland



Zonal Pricing

System divided into a small number of zones with individual prices.

International examples:



Australia



Denmark



Italy



Norway



Sweden

Key:

Boundaries -
for illustration
purposes only



Nodal Pricing

System divided into many "nodes" with individual prices.

International examples:



USA



New Zealand



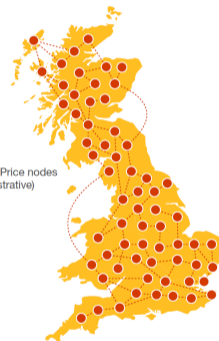
Canada



Singapore

Key:

● GB Price nodes
(illustrative)



Adapted from: National Grid ESO (2022) *Net Zero Market Reform*

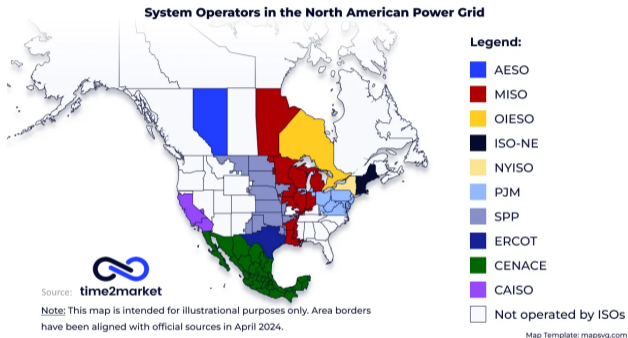
North American landscape

Independent System Operators (ISO)

- Reliable & effective grid operation
- Scheduling of power generation
- Stability of supply (transmission)

ERCOT (Texas) is a **Regional Transmission Organization (RTO)**

- Does not cross state lines



Nodal pricing



Day-ahead (DA) and real-time (RT; ~5% of volume) markets

European landscape

Transmission System Operators (TSO)

- Operate the transmission system
- Ensure the grid is balanced
- Exception: Germany has 4 TSOs

Market coupling

- Price Coupling of Regions – EUPHEMIA (DA)
- Flow-Based Market Coupling (DA)
- XBID mechanism (ID)

Members of the Price Coupling of Regions (PCR) System

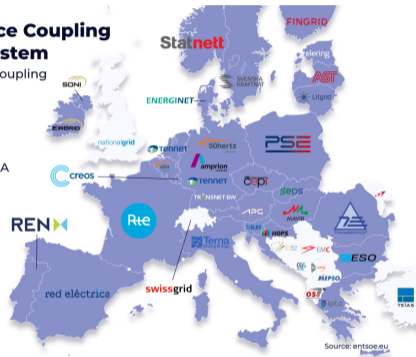
European Day-Ahead Market Coupling

Legend:

■ Members of PCR EUPHEMIA

Not members

Source:  time2market



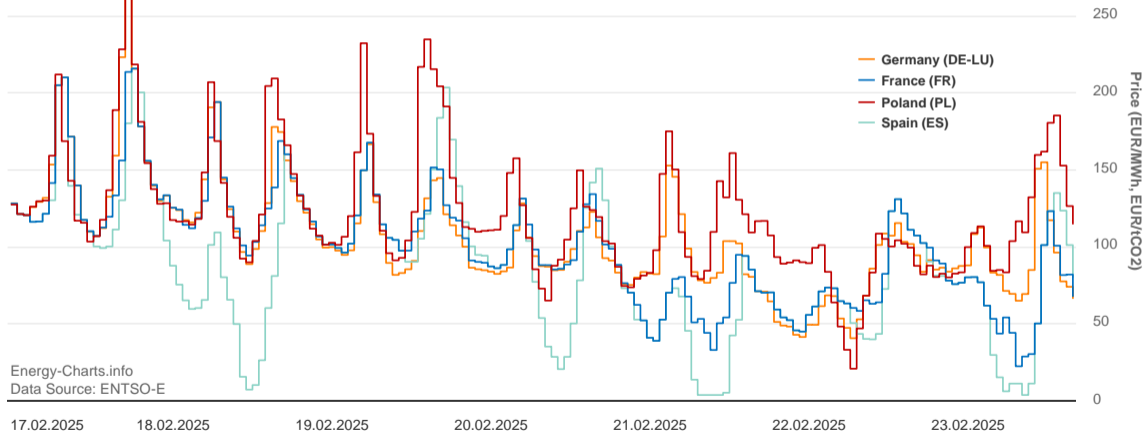
Zonal pricing



Day-ahead (DA) and intraday (ID; 3-20% of volume) markets

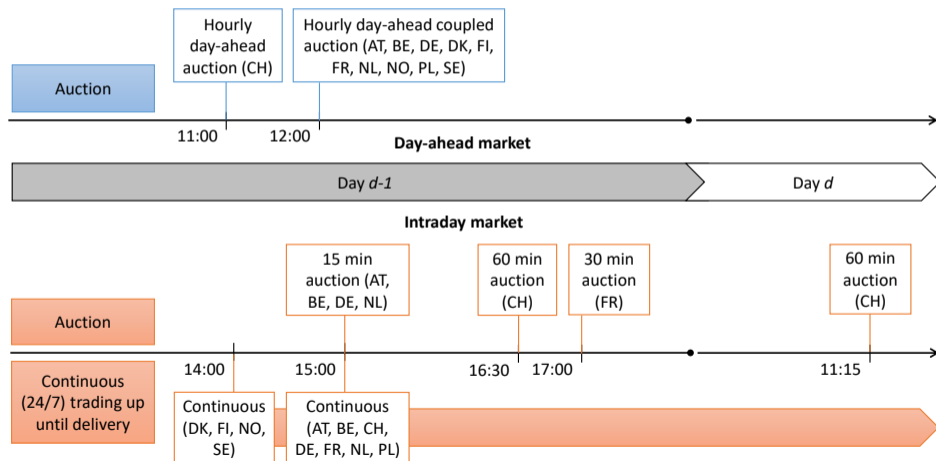
Market coupling \rightarrow (more) similar prices in bidding zones

Day-ahead prices in the European Union in 2025 (week 8)



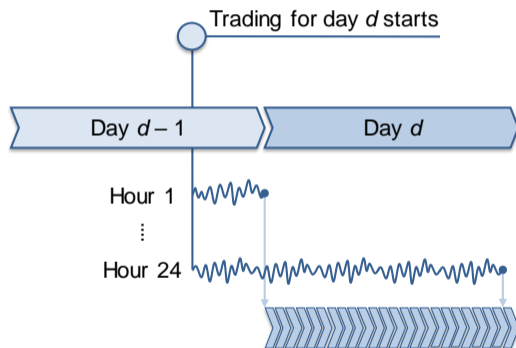
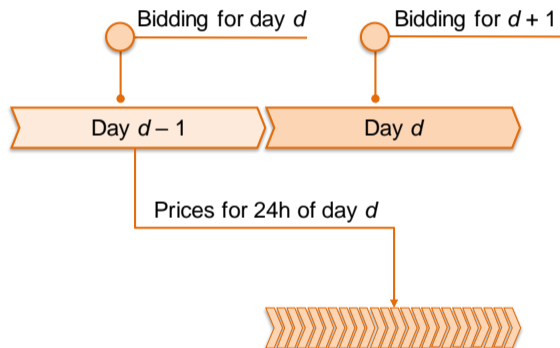
Timeline of DA and ID trading activities in Europe

(Maciejowska, Uniejewski & Weron, 2023, Oxford Res. Enc.)



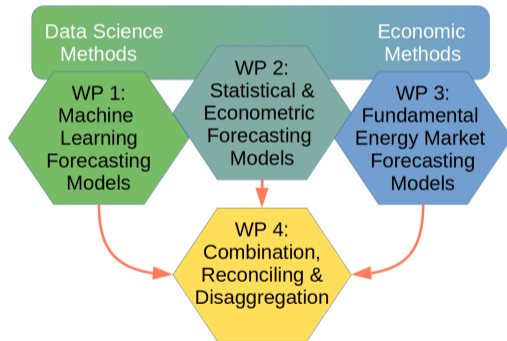
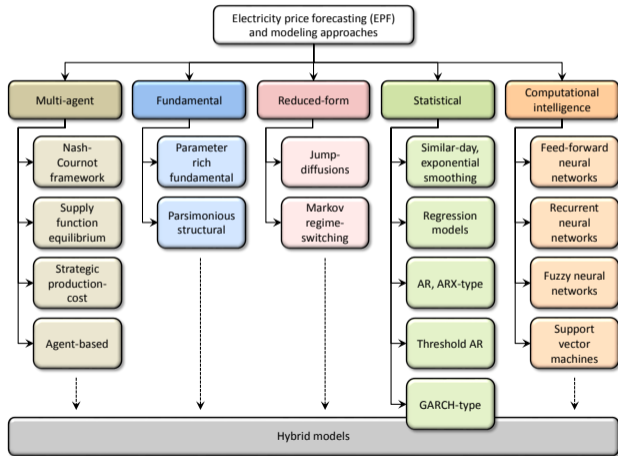
Day-ahead ($> 90\%$ of papers) vs. intraday (real-time) markets

(Maciejowska, Uniejewski & Weron, 2023, Oxford Res. Enc.)



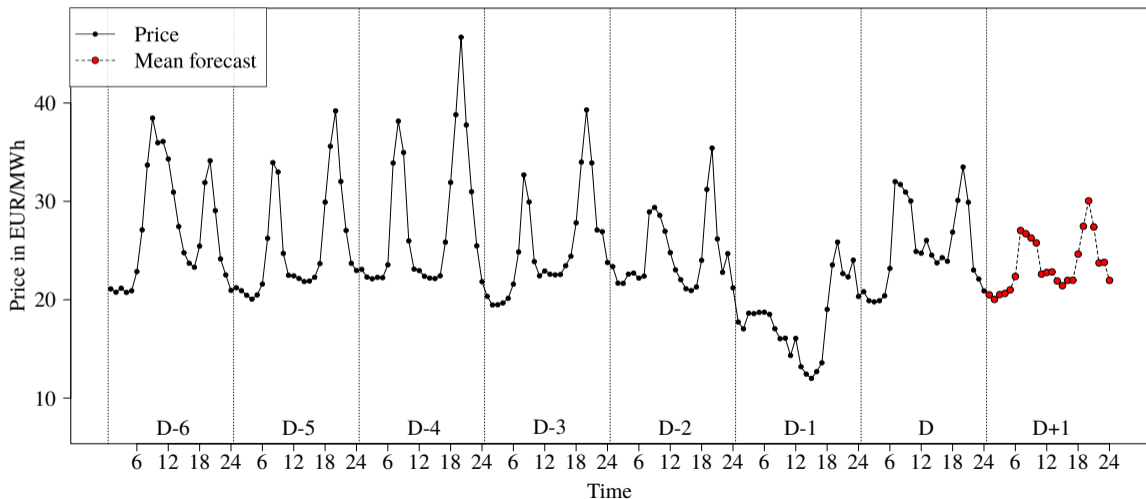
Model taxonomy: 2014 vs. 2022

Weron (2014, IJF) → Weron & Ziel (2022, DFG-NCN project PRIORITY)

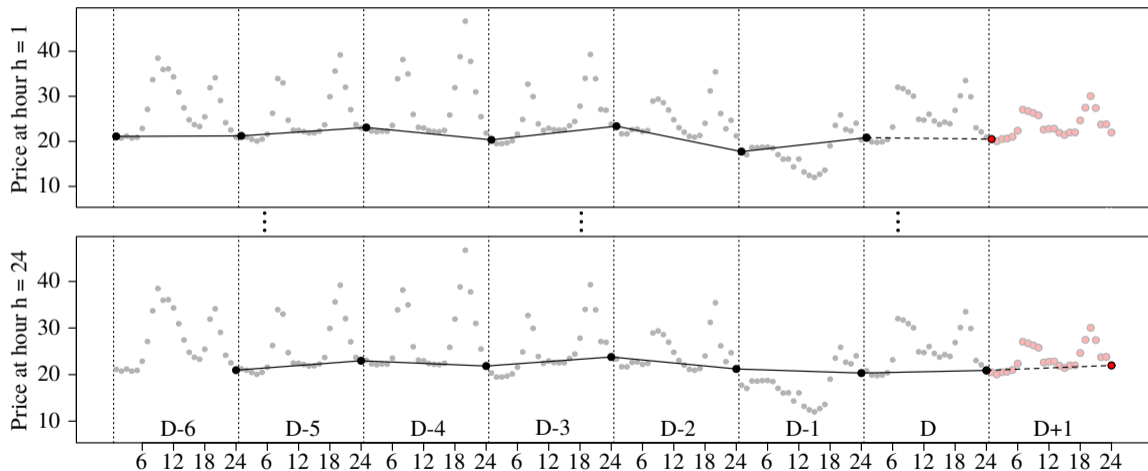


Day-ahead point forecasting: Univariate ...

(Ziel & Weron, 2018, ENEECO)



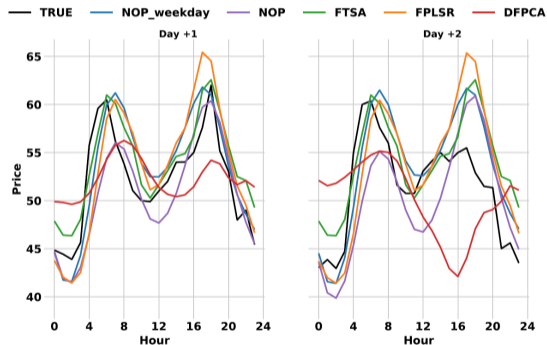
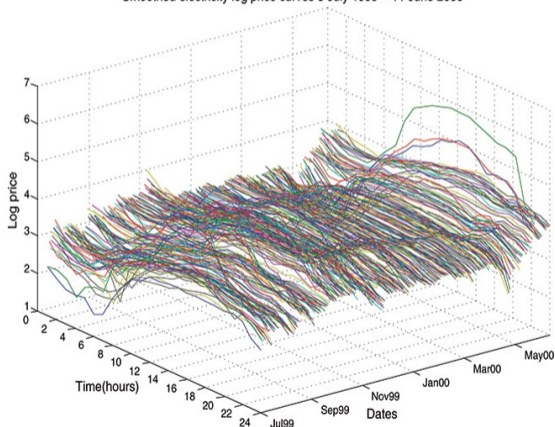
... multivariate ...



... functional (data analysis) ...

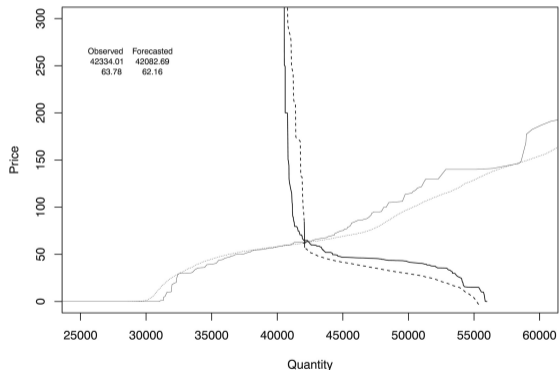
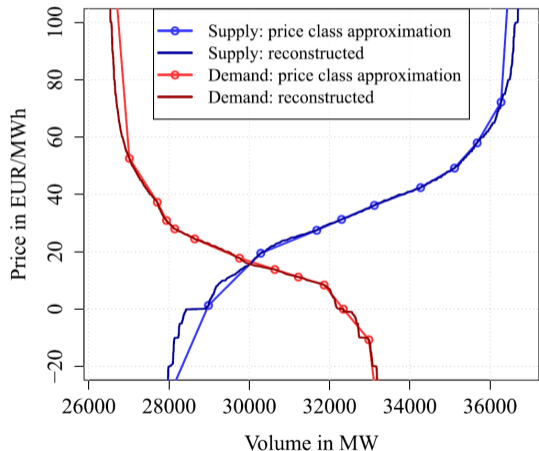
(Chen & Li, 2017, JBES; Chen et al., 2019, Ann.Appl.Stat; Wang & Cao, 2023, Environmetrics)

Smoothed electricity log price curves 5 July 1999—11 June 2000



... or supply & demand curves?

(Ziel & Steinert, 2016, ENEECO → 'X-model'; Shah & Lisi, 2020, J.Forecasting)



1 Introduction

2 'Toy' models

- The forecasting setup
- Naive models
- (Auto)regressive models
- Nonlinear AR models
- Exponential smoothing models
- Supply stack models

3 Beyond point forecasts

4 Forecast accuracy

International Journal of Forecasting 30 (2014) 1030–1081

Contents lists available at ScienceDirect

(2014)

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Review

Electricity price forecasting: A review of the state-of-the-art with a look into the future

Rafał Weron



Renewable and Sustainable Energy Reviews 81 (2018) 1548–1568

Contents lists available at ScienceDirect

(2018)

Renewable and Sustainable Energy Reviews

journal homepage: www.elsevier.com/locate/rsre

Recent advances in electricity price forecasting: A review of probabilistic forecasting

Jakub Nowotarski, Rafał Weron*

Department of Operations Research and Business Intelligence, Wrocław University of Science and Technology, Wrocław, Poland

ing (EPF) over o explain the opportunities he paper also next decade es involving iii) statistical

Applied Energy 293 (2021) 116983

Contents lists available at ScienceDirect

(2021)

Applied Energy

journal homepage: www.elsevier.com/locate/apenergy

Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark

Jesus Lago ^{a,*}, Grzegorz Marcjasz ^b, Bart De Schutter ^a, Rafał Weron ^b

^a Delft Center for Systems and Control, Delft University of Technology, Delft, The Netherlands
^b Department of Operations Research and Business Intelligence, Wrocław University of Science and Technology, Wrocław, Poland

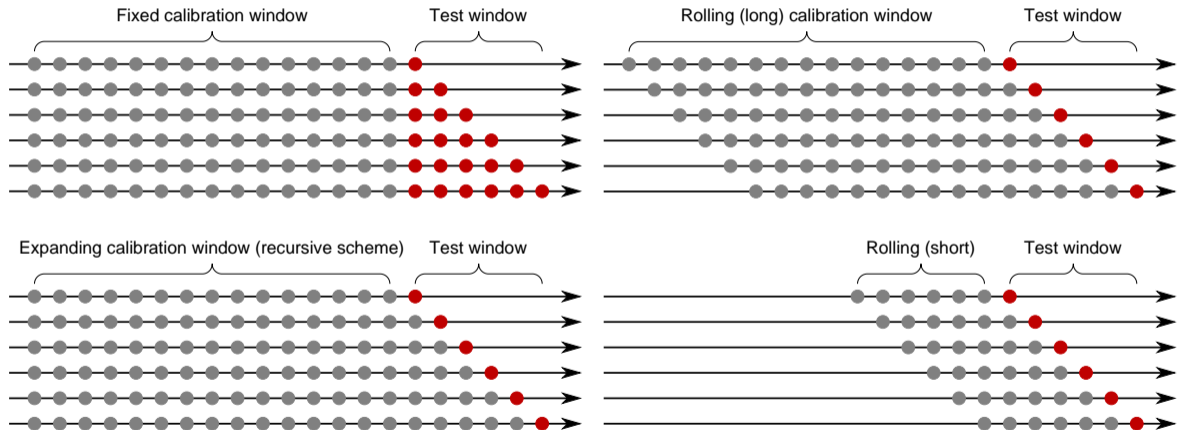
ARTICLE INFO

ABSTRACT

Keywords

While the field of electricity price forecasting has benefited from plenty of contributions in the last two decades,

Forecasting setup: Fixed, expanding & rolling windows



'Toy' models for point forecasts

- Let us fix the notation:
 - $P_t = P_{24d+h} = P_{d,h}$ is the **price** for day d and hour h
 - $\hat{P}_t = \hat{P}_{d,h|d-1}$ is the **forecast** of $P_{d,h}$ computed on day $d - 1$ (in the morning)where $t = 1, \dots, T$, $d = 1, \dots, \frac{T}{24}$ and $h = 1, \dots, 24$
- The **prediction error** or **residual** is given by: $\varepsilon_t = P_t - \hat{P}_t$
- Consider five classes of 'toy' models:
 - 1 Naive
 - 2 (Auto)Regressive
 - 3 Nonlinear AR
 - 4 Exponential smoothing
 - 5 Supply stack

Naive models

(Nogales et al., 2002, TPWRS; Weron, 2014, IJF; Lago et al., 2021, APEN)

- Naive (persistent, white noise) forecast:

$$\hat{P}_{d,h}^{\text{naive}} = \begin{cases} P_{d-1,h} & \text{for } d = \text{Tue, Wed, Thu, Fri} \\ P_{d-7,h} & \text{for } d = \text{Mon, Sat, Sun} \end{cases}$$

- Simpler alternatives: $\hat{P}_{d,h}^{(1)} = P_{d-1,h}$ and $\hat{P}_{d,h}^{(7)} = P_{d-7,h}$
- $\hat{P}_{d,h}^{(7)}$ is easier to compute than $\hat{P}_{d,h}^{\text{naive}}$ and, unlike $\hat{P}_{d,h}^{(1)}$, captures weekly effects

Multiple regression

$$y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \cdots + \beta_k x_{k,t} + \varepsilon_t$$

- The predictors can be:
 - different variables, e.g., $x_{1,t}$ – load, $x_{2,t}$ – RES generation
 - lagged values of the same variable, e.g., $x_{1,t} = y_{t-1}$, $x_{2,t} = y_{t-2}$ → **autoregression**
 - or a combination of both
- Coefficients β_1, \dots, β_k measure the **marginal effects**
 - after taking account of the effect of all other predictors
- The **forecast** \hat{y}_t of y_t is obtained by setting $\varepsilon_t = 0$

(Ordinary) Least Squares (OLS) estimation

- The **OLS** chooses β_i 's that minimize the **sum of squared errors (SSE)**:

$$\hat{\beta} = \underset{\beta_i}{\operatorname{argmin}} \sum_{t=1}^T \varepsilon_t^2 = \underset{\beta_i}{\operatorname{argmin}} \sum_{t=1}^T (y_t - \underbrace{(\beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \dots + \beta_k x_{k,t})}_{\hat{y}_t})^2$$

- The estimated coefficients are denoted by $\hat{\beta} = [\hat{\beta}_0, \dots, \hat{\beta}_k]$
- The process of finding $\hat{\beta}_i$'s is called:
 - **estimating** model parameters
 - **fitting** the model to the data
 - **learning (training)** the model

Regression in matrix form

- The multiple regression model for $t = 1, \dots, T$

$$y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \dots + \beta_k x_{k,t} + \varepsilon_t$$

- Can be written in matrix form with **OLS solution**:

$$y = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \implies \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y$$

where $y = (y_1, \dots, y_T)'$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_T)'$, $\boldsymbol{\beta} = (\beta_0, \dots, \beta_k)'$ and

$$\mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & x_{2,1} & \dots & x_{k,1} \\ 1 & x_{1,2} & x_{2,2} & \dots & x_{k,2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1,T} & x_{2,T} & \dots & x_{k,T} \end{bmatrix}$$

Expert ARX-type models

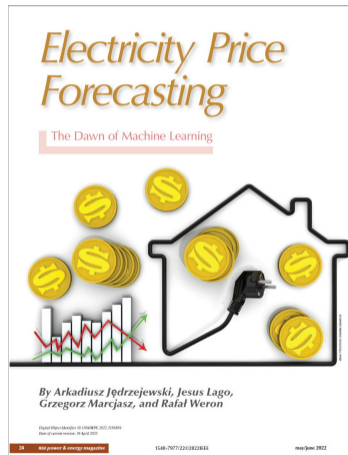
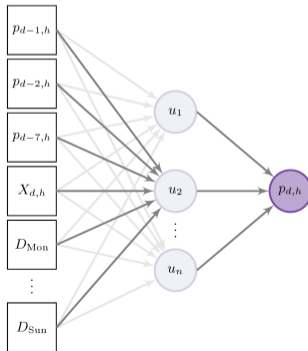
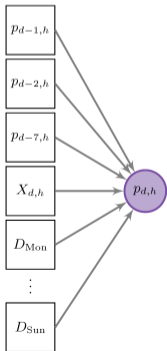
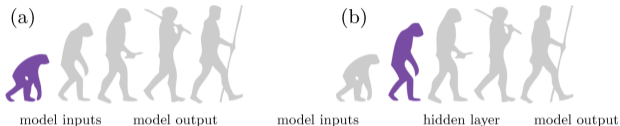
Consider an **autoregressive structure with exogenous variables**:

$$P_{d,h} = \beta_0 + \underbrace{\sum_{i=1}^7 \beta_i P_{d-i,h}}_{\text{AutoRegressive effects}} + \underbrace{\sum_{i=1}^7 \beta_{i+7} D_i}_{D_1 = \text{Mon, ...}} + \underbrace{\sum_{i=1}^K \beta_{i+14} X_{d,h}^{(i)}}_{\text{eXogenous variables}} + \varepsilon_{d,h}$$

- There can be no more dummies than categories \rightarrow set $\beta_0 = 0$ if all D_i 's are used
 - Also set $\beta_0 = 0$ if the mean is removed from $P_{d,h}$ beforehand
- Special cases:
 - Naive model $P_{d,h}^{(1)}$ for $\beta_1 = 1$ and $\beta_{i \neq 1} = 0$
 - **AR(7)**, **sparse AR(7)** with some AR lags missing
 - **ARX(7)** with $K \geq 1$, **sparse ARX(7)** with some AR lags missing

Linear regression vs. single-output (shallow) neural network

(Jędrzejewski, Lago, Marcjasz & Weron, 2022, IEEE-PEM)



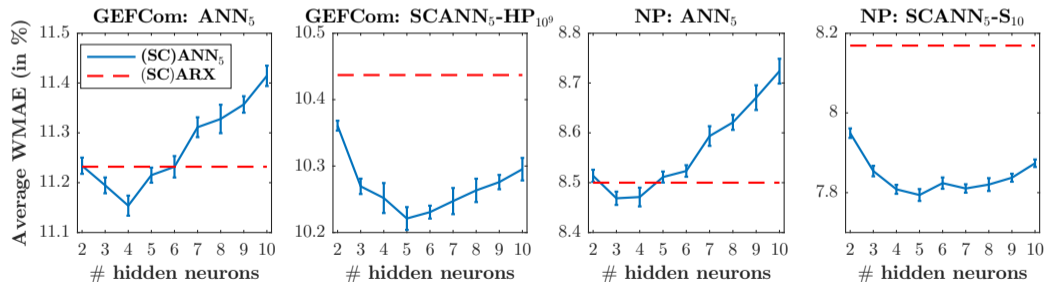
What are the differences?

- Computational complexity
 - Linearity \Leftrightarrow **non-linearity (hidden layers)**
- Optimization
 - **OLS** \Leftrightarrow *back-propagation*, Levenberg-Marquardt algorithm, ...
- Execution time (in MATLAB)
 - **Fast** \Leftrightarrow slow ... ca. $400\times$ slower for one run!
0.061 vs. 24.57 sec. for 7 days on a laptop with i7-1065G7
- Stability
 - Always the same parameters/forecasts \Leftrightarrow different for each run, dependent on starting parameters
 - Solution: **committee machines** \rightarrow **ensemble averaging**

Number of hidden neurons vs. forecast accuracy

(Marcjasz, Uniejewski & Weron, 2019, IJF)

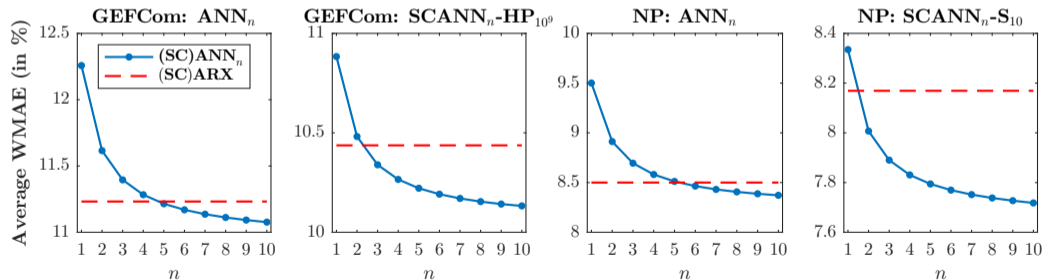
- The higher are the price fluctuations (\rightarrow larger errors) the more neurons are needed
- ... but the dependence is not constant over time (seasonality)



Number of runs (ensemble size) vs. forecast accuracy

(Marcjasz, Uniejewski & Weron, 2019, IJF)

- The more runs (\rightarrow longer computational time) the better
- The prediction error decays as a power law



Additive Holt-Winters method: Component form

Forecast eq.: $\hat{y}_{t+h|t} = \ell_t + hb_t + s_{t+h-m(k+1)}$

$$\text{Smoothing equations: } \begin{cases} \ell_t = \alpha(y_t - s_{t-m}) + (1 - \alpha)(\ell_{t-1} + b_{t-1}) & \text{Level} \\ b_t = \beta(\ell_t - \ell_{t-1}) + (1 - \beta)b_{t-1} & \text{Trend} \\ s_t = \gamma(y_t - \ell_t) + (1 - \gamma)s_{t-m} & \text{Seasonality} \end{cases}$$

where

- h is the forecast horizon (steps ahead), m is the period
- k is the integer part of $\frac{h-1}{m} \Rightarrow$ estimates come from the final period of the sample

$0 \leq \alpha, \beta, \gamma \leq 1$ are estimated numerically by minimizing the **sum of squared errors**:

$$\sum_{t=1}^T \varepsilon_t^2 = \sum_{t=1}^T (y_t - \hat{y}_t)^2$$

The family of exponential smoothing methods

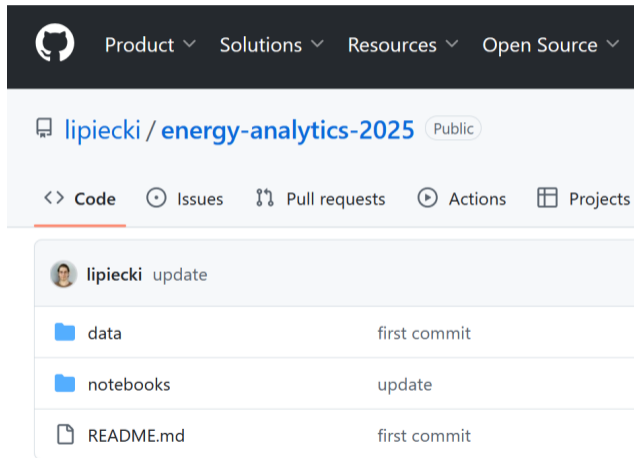
(Hyndman & Athanasopoulos, 2021, OTexts; Hyndman, Koehler, Ord & Snyder, 2008, Springer)

Trend Component	Seasonal Component		
	N	A	M
	(None)	(Additive)	(Multiplicative)
N (None)	(N,N)	(N,A)	(N,M)
A (Additive)	(A,N)	(A,A)	(A,M)
A_d (Additive damped)	(A_d ,N)	(A_d ,A)	(A_d ,M)

Some of these methods we have already seen using other names:

Short hand	Method
(N,N)	Simple exponential smoothing
(A,N)	Holt's linear method
(A_d ,N)	Additive damped trend method
(A,A)	Additive Holt-Winters' method
(A,M)	Multiplicative Holt-Winters' method
(A_d ,M)	Holt-Winters' damped method

Python snippet: ToyModels.ipynb



The screenshot shows the GitHub interface for the repository 'lipiecki / energy-analytics-2025'. The repository is public. The navigation bar includes 'Code', 'Issues', 'Pull requests', 'Actions', and 'Projects'. The 'Code' tab is selected. Below the navigation bar, a commit history table is visible, showing updates by user 'lipiecki'.

File	Commit Message
data	first commit
notebooks	update
README.md	first commit

Supply stack model

(Weron & Ziel, 20**)

Fundamental approach from the subclass of **parsimonious structural** models

Assumptions:

- **Island grid**, i.e., no imports or exports
- The power plant park is composed of J units
- Every unit $j = 1, \dots, J$ is characterized by its
 - **installed capacity** AC_j (in MW)
 - **marginal cost** MC_j (e.g., in EUR, USD) of producing an additional MWh by generator j



Supply stack model cont.

(Weron & Ziel, 20**)

- Consider a park composed of $J = 15$ units
 - Roughly corresponds to Germany in 2018
 - 10 different types of generators
- The **merit order curve** is given by

$$MO(x) = MC_{j(x)}$$

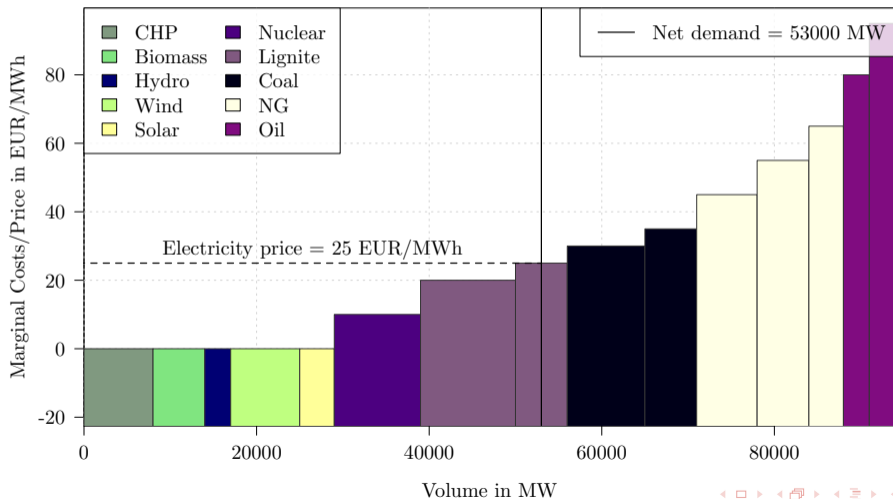
where x is the volume in MW and

- $j(x) = \max_j \{CC_j \leq x\}$ is the **marginal unit**
- $CC_j = \sum_{i=1}^j AC_i$ is the **cumulative capacity**

j	AC_j	MC_j	Type	CC_j
1	8,000	0	CHP	8,000
2	6,000	0	Biomass	14,000
3	3,000	0	Hydro	17,000
4	8,000	0	Wind	25,000
5	4,000	0	Solar	29,000
6	10,000	10	Nuclear	39,000
7	11,000	20	Lignite	50,000
8	6,000	25	Lignite	56,000
9	9,000	30	Coal	65,000
10	6,000	35	Coal	71,000
11	7,000	45	NG	78,000
12	6,000	55	NG	84,000
13	4,000	65	NG	88,000
14	3,000	80	Oil	91,000
15	3,000	95	Oil	94,000

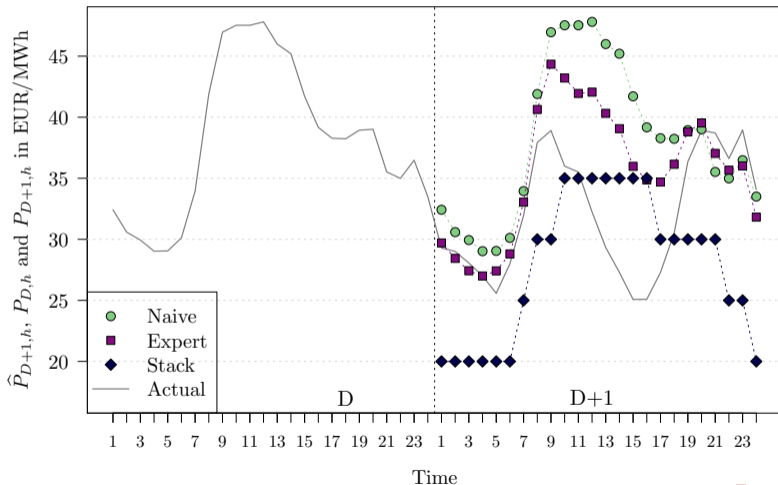
Supply stack model: Price setting

Net demand of 53,000 MW yields a spot price of 25 EUR/MWh



Point forecasts of 3 simple models: Germany, Sun 30.07.2017


Naive $\hat{P}_{d,h}^{\text{naive}}$, sparse AR(7; lags = 1,2,7, dummies = Mon, Sat, Sun), and the supply stack model



1 Introduction

2 'Toy' models

3 Beyond point forecasts

- Probabilistic forecasts
- Reliability & sharpness
- Postprocessing point forecasts
- Historical simulation
- Conformal prediction 

4 Forecast accuracy



Review

Electricity price forecasting: A review of the state-of-the-art with a look into the future

Rafał Weron



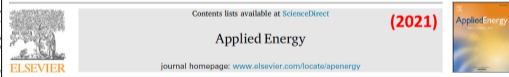
Recent advances in electricity price forecasting: A review of probabilistic forecasting

Jakub Nowotarski, Rafał Weron*

Department of Operations Research and Business Intelligence, Wrocław University of Science and Technology, Wrocław, Poland

ARTICLE INFO

Keywords:
Electricity price forecasting
Probabilistic forecasts
Reliability
Sharpness
Day-ahead markets
Autoregression
Neural networks



Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark

Jesus Lago ^{a,*}, Grzegorz Marcjasz ^b, Bart De Schutter ^a, Rafał Weron ^b

^a Delft Center for Systems and Control, Delft University of Technology, Delft, The Netherlands

^b Department of Operations Research and Business Intelligence, Wrocław University of Science and Technology, Wrocław, Poland

ARTICLE INFO

Keywords:

ABSTRACT

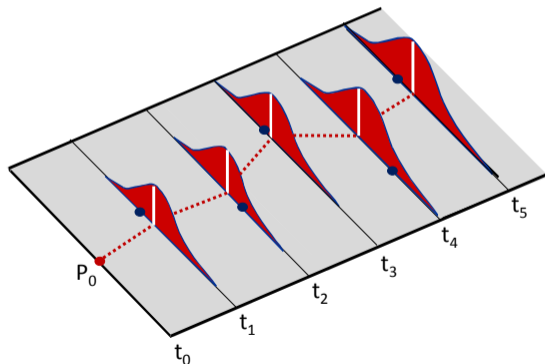
While the field of electricity price forecasting has benefited from plenty of contributions in the last two decades,

Probabilistic (interval, density) forecasting

(Gneiting & Katzfuss, 2014, Annu Rev)

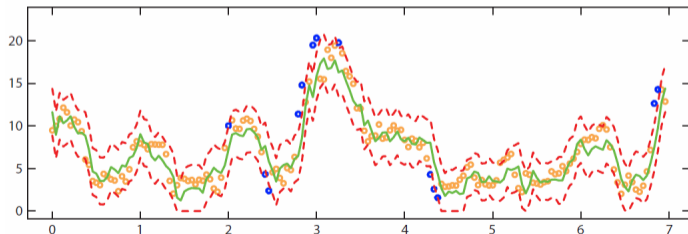
We cannot observe the true underlying distribution \Rightarrow we cannot compare the *predictive distribution* \hat{F} with the actual one F ... only with past observations

Gneiting et al. (2007a, 2007b, 2014) argue that probabilistic forecasting aims to '*maximize the **sharpness** of the predictive distributions, subject to **reliability***'



Reliability (calibration, unbiasedness)

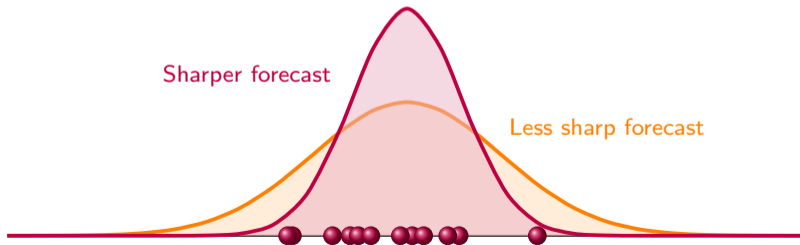
- Refers to the **statistical consistency** between \hat{F} and the observations
- If a 90% PI covers 90% of the observed prices, then this PI is said to be:
 - reliable (Pinson et al., 2007; Pinson & Kariniotakis, 2010)
 - well calibrated (Gneiting et al., 2007a, 2007b, 2014)
 - unbiased (Taylor, 1999)
- Example: 13 \circ or 'misses' and 155 \circ or 'hits' \rightarrow the coverage is $\frac{155}{168} \approx 92\%$



Sharpness

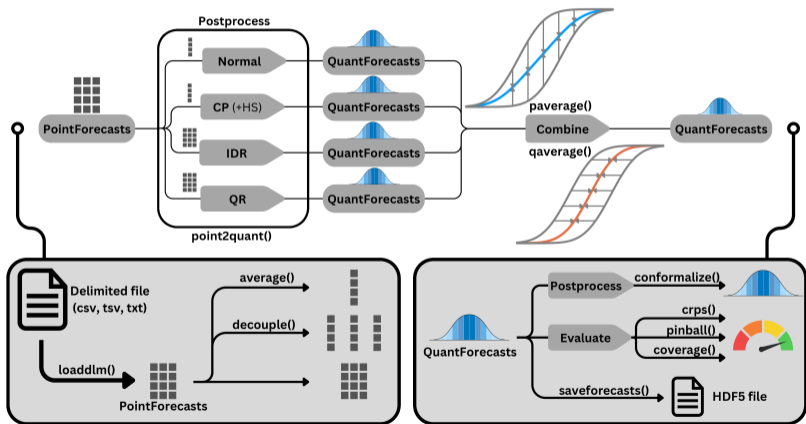
(Pinson et al., 2007, Wind En; Gneiting & Raftery, 2007, JASA; Gneiting & Katzfuss, 2014, Annu Rev)

- Refers to the **concentration** or **tightness** of the predictive distributions
 - Derives from the idea that reliable predictive distributions of null width correspond to perfect point predictions
 - Reliability is a joint property of the predictions and the observations
 - **Sharpness** is a property of the forecasts only



Postprocessing point forecasts

(Vannitsem et al., 2018, Elsevier; Chen et al., 2024, Ann Appl Stat; Lipiecki et al., 2024, ENEECO)



Experiment yourself:



The 'normal' benchmark

- Assume that the prediction errors are $N(\mu, \sigma^2)$ -distributed
- Training corresponds to estimating $\hat{\mu}$ and $\hat{\sigma}$ of $\varepsilon_t = y_t - \hat{y}_t$ for $t \in \mathcal{S}$
 - \mathcal{S} is the training set (or calibration window)
- The τ -th quantile conditional on \hat{y}_t is obtained via:

$$\hat{q}_{\tau|\hat{y}_t} = \hat{y}_t + \hat{\mu} + \hat{\sigma} F_N^{-1}(\tau)$$

where $F_N^{-1}(\tau)$ is the inverse of the standard normal CDF, i.e., with $\mu = 0, \sigma = 1$

Historical simulation

(Hendricks, 1996, EPR; Alexander, 2008, Wiley; Nowotarski & Weron, 2018, RSER)

- A model-independent approach that computes

$$\hat{q}_{\tau|\hat{y}_t} = \hat{y}_t + Q_{\tau}(\varepsilon_t)$$

where $Q_{\tau}(\varepsilon_t)$ is the sample τ -quantile of $\varepsilon_t = y_t - \hat{y}_t$ for $t \in \mathcal{S}$

- The term **historical simulation (HS)** can be traced back to the early 1990s and the beginnings of Value-at-Risk (VaR)
- Similar to **bootstrapped residuals** (see, e.g., Hyndman & Athanasopoulos, 2021, FPP3), but each ε_t is sampled exactly once

FEDERAL RESERVE BANK of NEW YORK

ECONOMIC POLICY REVIEW

Evaluation of Value-at-Risk Models Using Historical Data

April 1996 Volume 2, Number 1

JEL classification: G11, G15, G28



Author: Darryll Hendricks

Recent studies have underscored the need for market participants to develop reliable methods of measuring risk. One increasingly popular technique is the use of "value-at-risk" models, which convey estimates of market risk for an entire portfolio in one number. The author explores how well these models actually perform by applying twelve value-at-risk approaches to 1,000 randomly chosen foreign exchange portfolios. Using nine criteria to evaluate model performance, he finds that the approaches generally capture the risk that they set out to assess and tend to produce risk estimates that are similar in average size. No approach, however, appears to be superior by every measure.

Conformal prediction

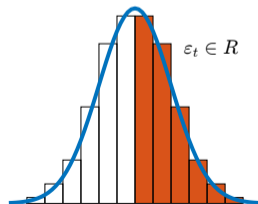
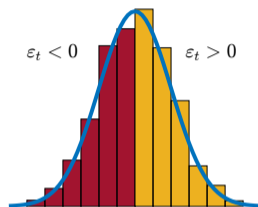
(Vovk et al., 2005, Springer; Kath & Ziel, 2021, IJF; Lipiecki et al., 2024, ENEECO)

- For $t \in \mathcal{S}$ calculate the so-called **non-conformity scores** $\lambda_t = |\varepsilon_t| = |y_t - \hat{y}_t|$, then compute

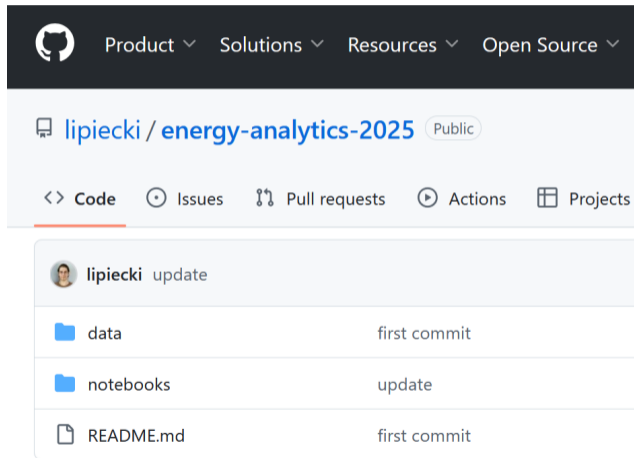
$$\hat{q}_{\tau|\text{hat}y_t} = \hat{y}_t - \mathbb{1}_{\tau \leq 0.5} Q_{2\tau}(\lambda) + \mathbb{1}_{\tau \geq 0.5} Q_{2(1-\tau)}(\lambda)$$

where $Q_{\tau}(\lambda)$ is the τ -th sample quantile of λ_t

- This version is called **inductive** or **split** CP, however, a 'split' is not needed if \hat{y}_t 's are already available
- HS works with ε_t 's, CP with $|\varepsilon_t|$'s \rightarrow **symmetric** \hat{F}
- Using $\varepsilon_1, \dots, \varepsilon_T$, HS approximates the whole distribution, CP only the positive half \rightarrow **smoother** \hat{F}



Python snippet: UncertaintyQuantification.ipynb



The screenshot shows the GitHub interface for the repository 'lipiecki / energy-analytics-2025'. The repository is public. The navigation bar includes 'Code', 'Issues', 'Pull requests', 'Actions', and 'Projects'. The 'Code' tab is selected. Below the navigation bar, there is a commit history table showing updates by the user 'lipiecki'.


File	Commit Message
data	first commit
notebooks	update
README.md	first commit

1 Introduction

2 'Toy' models

3 Beyond point forecasts

4 Forecast accuracy

- Absolute and square errors
- Percentage errors
- Scaled and relative errors
- Testing for coverage
- CRPS and the pinball score
- DM-type tests 



Review

Electricity price forecasting: A review of the state-of-the-art with a look into the future



Rafał Weron



Recent advances in electricity price forecasting: A review of probabilistic forecasting

Jakub Nowotarski, Rafał Weron*

Department of O

ARTICLE INFO

Keywords:
Electricity price forecasting
Probabilistic forecasting
Reliability
Sharpness
Day-ahead market
Autoregression
Neural network



Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark

Jesus Lago ^{a,*}, Grzegorz Marcjasz ^b, Bart De Schutter ^a, Rafał Weron ^b

^a Delft Center for Systems and Control, Delft University of Technology, Delft, The Netherlands

^b Department of Operations Research and Business Intelligence, Wrocław University of Science and Technology, Wrocław, Poland

ARTICLE INFO

Keywords:

ABSTRACT

While the field of electricity price forecasting has benefited from plenty of contributions in the last two decades,

Measures of (point) forecast accuracy

(Hyndman & Koehler, 2006, IJF; Weron, 2014, IJF; Kolassa, 2020, IJF; Lago et al., 2021, APEN)

- Mean Absolute Error

$$\text{MAE} = \frac{1}{T} \sum_{t=1}^T |\varepsilon_t|$$

- (Root) Mean Square(d) Error

$$\text{MSE} = \frac{1}{T} \sum_{t=1}^T \varepsilon_t^2 \quad \text{and} \quad \text{RMSE} = \sqrt{\frac{1}{T} \sum_{t=1}^T \varepsilon_t^2}$$

where $|\varepsilon_t| = |P_t - \hat{P}_t|$ is the **absolute** and $\varepsilon_t^2 = (P_t - \hat{P}_t)^2$ is the **square(d) error**

- Note: MSE is minimized by the mean, but MAE by the median
⇒ when using OLS measure forecast accuracy with MSE, not MAE

Percentage errors: MAPE and DMAE

- Mean Absolute Percentage Error

$$\text{MAPE} = \frac{1}{T} \sum_{t=1}^T \frac{|\varepsilon_t|}{P_t}$$

where $\frac{|\varepsilon_t|}{P_t}$ is the **absolute percentage error** can be used to compare across datasets

- MAPE works well when $P_t \gg 0$, e.g., in load forecasting
- Is unreliable for electricity prices or temperatures (can be ≤ 0)
- Instead of dividing by P_t we can divide by the daily mean $\overline{P_{24}} = \frac{1}{24} \sum_{t=1}^{24} P_t$ to obtain the **Daily-weighted MAE** for day d :

$$\text{DMAE}_d = \frac{1}{24} \frac{1}{\overline{P_{24}}} \sum_{t=1}^{24} |\varepsilon_t| = \frac{1}{24} \frac{1}{\overline{P_{24}}} \sum_{h=1}^{24} |\varepsilon_{d,h}|$$

Percentage errors: Symmetric MAPE (sMAPE)

See also <https://robjhyndman.com/hyndsight/smape>

- MAPE puts a heavier penalty on negative than on positive ε_t
- Makridakis (1993) proposed the 'symmetric MAPE':

$$\text{sMAPE}_M = \frac{200}{T} \sum_{t=1}^T \frac{|\varepsilon_t|}{|P_t + \hat{P}_t|} = \frac{200}{T} \sum_{t=1}^T \frac{|P_t - \hat{P}_t|}{|P_t + \hat{P}_t|}$$

- Armstrong's (1985) version had no $|\cdot|$ in the denominator
- Both have a problem when $|P_t + \hat{P}_t| \approx 0$
- Chen & Yang (2004) defined it as (also dropped the '100'):

$$\text{sMAPE}_{CY} = \frac{2}{T} \sum_{t=1}^T \frac{|\varepsilon_t|}{|P_t| + |\hat{P}_t|} = \frac{2}{T} \sum_{t=1}^T \frac{|P_t - \hat{P}_t|}{|P_t| + |\hat{P}_t|}$$

- Still, it is undefined when $P_t = \hat{P}_t = 0$

Scaled errors

(Hyndman & Koehler, 2006, IJF)

- The **Mean Absolute Scaled Error** is defined by:

$$\text{MASE} = \frac{1}{T} \sum_{t=\tau+1}^T \frac{|\varepsilon_t|}{e_m} = \frac{1}{T \cdot e_m} \sum_{t=\tau+1}^T |\varepsilon_t|$$

where

- $e_m = \frac{1}{\tau-m} \sum_{t=m+1}^{\tau} |P_t - P_{t-m}|$ is the MAE of a naive prediction **on the training set**
- m is the period for seasonal data (e.g., $m = 4$ for quarterly)
- **Interpretation:** if $\text{MASE} < 1$ then our prediction is better than naive (on the training set), if $\text{MASE} > 1$ then it is worse

Relative errors

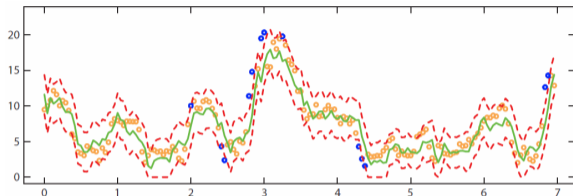
(Hyndman & Koehler, 2006, IJF; Lago et al., 2021, APEN)

- MASE is problematic when:
 - forecasting methods use different calibration windows / training sets
 - P_t exhibits 'long' periods of higher/lower values
- Lago et al. (2021, APEN) argue the a better metric is the **relative MAE**:

$$\text{rMAE} = \text{relMAE} = \frac{\text{MAE}_{\text{method}}}{\text{MAE}_{\text{benchmark}}}$$

- The benchmark can be a naive model (as in MASE)
- Can easily be applied to other metrics, e.g., the RMSE
- **Interpretation:** if $\text{rMAE}, \text{rRMSE} < 1$ then our prediction is better than the benchmark, if $\text{rMAE}, \text{rRMSE} > 1$ then it is worse

Unconditional coverage (UC)



- (Empirical) coverage is measured by

$$I_t = \begin{cases} 1 & \text{if } P_t \in \text{PI} \rightarrow \text{o 'hit'} \\ 0 & \text{if } P_t \notin \text{PI} \rightarrow \text{o 'miss'} \end{cases}$$

and should match the nominal rate

$$\mathbb{P}(P_t \in \text{PI}) = \mathbb{P}(I_t = 1) = (1 - \alpha)$$

- Some studies report only the so-called **PI Coverage Probability**

$$\text{PICP} = \frac{1}{T} \sum_{t=1}^T I_t \cdot 100\%$$

- Other subtract it from the nominal coverage to obtain the so-called **Average Coverage Error**

$$\text{ACE} = \text{PICP} - \text{PINC}$$

where **PINC** – **PI Nominal Coverage**

UC and the Kupiec test

(Kupiec, 1995, J Derivatives)

- Checks whether $ACE = 0$ or $\mathbb{P}(I_t = 1) = (1 - \alpha)$, given that \circ are independent
 - Equivalent to testing that I_t is i.i.d. Bernoulli with mean $(1 - \alpha)$
 - Rejects the null ('good PI') if the percent of misses is statistically different from α
- The **likelihood ratio** statistics for unconditional coverage:

$$LR_{UC} = -2 \log \left\{ \frac{(1 - c)^{n_0} c^{n_1}}{(1 - \pi)^{n_0} \pi^{n_1}} \right\} \sim \chi^2(1)$$

- $c = (1 - \alpha)$ is the nominal coverage rate
- $\pi = \frac{n_1}{n_0 + n_1}$ is the percentage of \circ 'hits'
- n_0 and n_1 are respectively the number of 0's and 1's in I_t

Independence, conditional coverage and the Christoffersen test

(Christoffersen, 1998, IER)

- In the Kupiec (1995) test the **clustering** of **'misses'** does not matter, only the total number of violations plays a role
- Christoffersen (1998) introduced the **Independence** and **Conditional Coverage** tests
- **Ind** is tested against an explicit first-order Markov alternative
 - Like LR_{UC} , also $LR_{Ind} \sim \chi^2(1)$
- **CC** is simply a joint test for **Ind** and **UC**
 - If we condition on the first observation, then

$$LR_{CC} = LR_{UC} + LR_{Ind} \sim \chi^2(2)$$

Continuous Ranked Probability Score (CRPS)

(Gneiting & Raftery, 2007, JASA; Gneiting & Katzfuss, 2014, Annu Rev; Nitka & Weron, 2023, ORD)

- The CRPS is the standard metric for evaluating probabilistic forecasts:

$$\text{CRPS}(\hat{F}, x) = \int_{-\infty}^{\infty} (\hat{F}(y) - \mathbb{1}_{\{x \leq y\}})^2 dy$$

where \hat{F} is the predictive distribution and x is the observation, e.g., electricity price

- It is a **proper scoring rule**, i.e., quoting the true distribution as the forecast is an optimal strategy in expectation
- Problem: in practice we often work with a finite set of quantile forecasts

CRPS and the pinball score

(Gneiting & Raftery, 2007, JASA; Nowotarski & Weron, 2018, RSER; Nitka & Weron, 2023, ORD)

- The CRPS can be approximated by:

$$\text{CRPS}(\hat{F}, x) \approx \frac{2}{M} \sum_{i=1}^M \text{PS}(\hat{q}, x, q_i)$$

where

- $q_1 < \dots < q_M$ is an equidistant dense grid of probabilities, e.g., 99 percentiles
- $\hat{q} \equiv \hat{F}^{-1}(q)$ is the quantile forecast for quantile level $q \in (0, 1)$

and the **pinball score** is defined as:

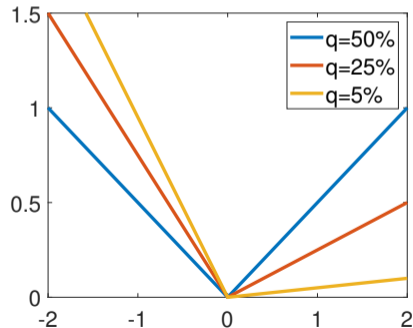
$$\text{PS}(\hat{q}, x, q) = \left(\mathbb{1}_{\{x < \hat{q}\}} - q \right) (\hat{q} - x)$$

- Note: The scaling factor of **2** is usually omitted in practice

Pinball score (or loss) in more detail

$$\text{PS}(\hat{q}, x, q) = (\mathbb{1}_{\{x < \hat{q}\}} - q) (\hat{q} - x) = \begin{cases} (1 - q) (\hat{q} - x) & \text{for } x < \hat{q} \\ q (x - \hat{q}) & \text{for } x \geq \hat{q} \end{cases}$$

- Also known as the **quantile score**, **check function** or the **linlin/bilinear/newsboy loss**
- For an **Aggregate PS** (or **APS**) average:
 - across all t in the test period
 - across all quantiles \rightarrow **CRPS**



Testing for equal predictive performance

(Diebold & Mariano, 1995, JBES; Diebold, 2015, JBES)

- When faced with forecasts from two (or more) models we can rank them based on some **score function** (the lower the better):

$$\hat{S} = \frac{1}{T} \sum_{t=1}^T S(\hat{F}, x)$$

- But if we want to know whether the forecasts of model 1 are **significantly** better (more accurate) than those of model 2, then we need to use a test
- The most popular is the **DM test** for **unconditional** predictive ability

Testing for equal predictive performance cont.

- The test of **Giacomini & White (2006, Econometrica)** accounts for parameter estimation uncertainty and tests **conditional predictive ability (CPA)**
- DM and GW tests can be used for nested and non-nested models if the calibration window does not grow with sample size (**Giacomini & Rossi, 2013**)
- This:
 - 🙄 rules out expanding windows
 - 😎 admits fixed and rolling windows

Testing for equal predictive performance cont.

- **Model confidence set** of Hansen et al. (2011, *Econometrica*) is similar to DM
- But uses bootstrap to approximate the distribution of the test statistics
- **Forecast encompassing** of Harvey et al. (1998, *JBES*)
- The null says that the forecasts of model 1 do not include more information than those of model 2

Diebold-Mariano (DM) test

- It is an asymptotic z-test with null that the mean of the **loss differential series**:

$$d_t = S_1(\hat{F}, P_t) - S_2(\hat{F}, P_t)$$

is zero, where $S_i(\cdot, \cdot)$ is the score function for model i , e.g., $|\varepsilon_t|$, ε_t^2 , Pinball score

- How to use it? Compute the Diebold-Mariano statistic for $t = 1, \dots, T$:

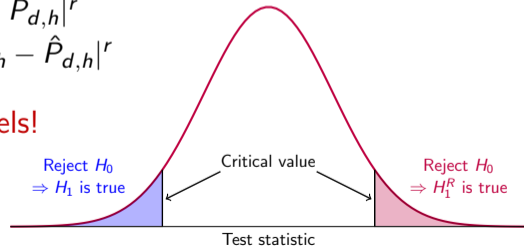
$$DM = \sqrt{T} \frac{\hat{\mu}_{d_t}}{\hat{\sigma}_{d_t}}$$

where $\hat{\mu}_{d_t}$ and $\hat{\sigma}_{d_t}$ are the mean and standard deviation of d_t

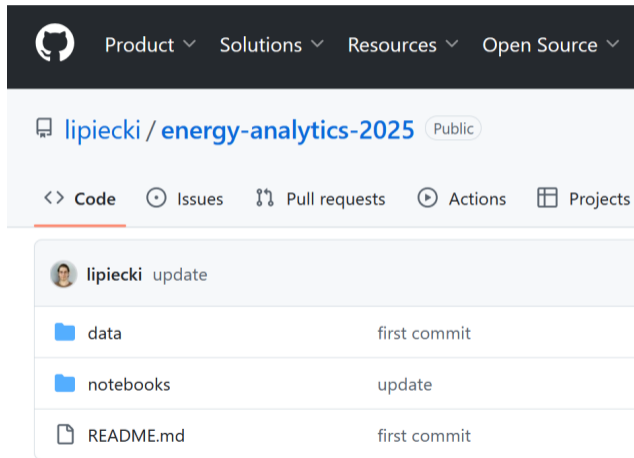
- The **null hypothesis** of no differences is equivalent to $H_0 : \mathbb{E}(d_t) = 0$

Diebold-Mariano (DM) test cont.

- If d_t is **covariance stationary**, the DM test statistics is asymptotically normal
- In practice we test twice, using one-sided tests with alternatives
 - $H_1 : \mathbb{E}(d_t) < 0$, i.e., forecasts of model 1 are **better** than those of model 2
 - $H_1^R : \mathbb{E}(d_t) > 0$, i.e., forecasts of model 1 are **worse** than those of model 2
 e.g., at the $\alpha = 5\%$ significance level
- Due to intraday correlation of electricity prices we test:
 - For each hour: $S_{i,h}^r(\hat{P}_{d,h}, P_{d,h}) = |P_{d,h} - \hat{P}_{d,h}|^r$
 - Jointly for 24h: $S_i^r(\hat{P}_d, P_d) = \sum_{h=1}^{24} |P_{d,h} - \hat{P}_{d,h}|^r$
- The DM test compares forecasts, not models!



Python snippet: DieboldMariano.ipynb



The screenshot shows the GitHub interface for the repository 'lipiecki / energy-analytics-2025'. The repository is public. The navigation bar includes 'Code', 'Issues', 'Pull requests', 'Actions', and 'Projects'. The 'Code' tab is selected. Below the navigation bar, there is a commit history table showing updates by the user 'lipiecki'.


File	Commit Message
data	first commit
notebooks	update
README.md	first commit

Articles & working papers on <https://p.wz.pwr.edu.pl/~weron.rafal/Publ>


Rafał Weron

Professor of Management Science (Energy Forecasting)


Home
Publications
Projects
S3 Seminar
Conferences
Students





K. Maciejowska, B. Uniejewski, R. Weron (2023) *Forecasting electricity prices*, in "Oxford Research Encyclopedia of Economics and Finance", Oxford University Press, DOI: [10.1093/acrefore/9780190625979.013.667](https://doi.org/10.1093/acrefore/9780190625979.013.667). Working paper version available from arXiv: <https://doi.org/10.48550/arXiv.2204.11735>



A. Jędrzejewski, J. Lago, G. Marcjasz, R. Weron (2022) *Electricity price forecasting: The dawn of machine learning*, IEEE Power & Energy Magazine 20(3), 24-31 (doi: [10.1109/MPE.2022.3150809](https://doi.org/10.1109/MPE.2022.3150809)). Working paper version available from arXiv: <https://arxiv.org/abs/2204.00883>



 **Highly Cited Paper** J. Lago, G. Marcjasz, B. De Schutter, R. Weron (2021) *Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark*, Applied Energy 293, 116983 ([doi: 10.1016/j.apenergy.2021.116983](https://doi.org/10.1016/j.apenergy.2021.116983))

-  The **epftoolbox** including Python codes for the two benchmark models (LEAR, DNN) and datasets is available from [GitHub](#)

Contents

- › Statistics
- › Monographs, reviews and edited volumes
- › Peer-reviewed articles in JCR-listed journals
- › Peer-reviewed articles in non JCR-listed journals
- › Book chapters
- › Conference papers
- › Popular science and other papers
- › Forthcoming publications, submitted papers and work in progress
- › Theses

Contact

- › [Department of Operations Research and](#)

See yourself:

Rafał Weron (Wrocław Tech, PL)

IIF Lecture: Electricity Price Forecasting, part I

UNC Charlotte, ISEA2025, 3-4.03.2025

63 / 63