

Neural Networks and Deep Learning

Lecture 13: Deep learning for energy and financial time series; time-series foundation models

Prof. dr. Daniel Traian PELE

Bucharest University of Economic Studies (ASE), Department of Statistics and Econometrics
Romanian Academy, Institute for Economic Forecasting (IPE)
IDA Institute Digital Assets ▪ AI4EFin ▪ MSCA Digital Finance

Learning objectives

- Set up a day-ahead forecasting problem without leakage: information set at gate closure, rolling origin, benchmarks.
- Write and estimate the Lasso-estimated autoregressive model (LEAR), a deep neural network (DNN) with 24 outputs and the neural basis expansion analysis (N-BEATS) on Romanian prices.
- Explain how time-series foundation models, Chronos and the Time-series Foundation Model (TimesFM), forecast without training on our data, and test them.
- Compare forecasts with the Diebold–Mariano test and calibrate probabilistic forecasts with conformal prediction.
- Use a foundation model for Value-at-Risk and backtest it on the S&P 500 and Bitcoin.

Abbreviations (1/2)

AIC	Akaike information criterion
ASE	Bucharest University of Economic Studies
CET	Central European Time
CQR	conformalised quantile regression
DE-LU	Germany–Luxembourg bidding zone
DeepAR	deep autoregressive recurrent network (Salinas et al., 2020)
DM	Diebold–Mariano test
DNN	deep neural network (here: a multilayer perceptron with 24 outputs)
ENTSO-E	European Network of Transmission System Operators for Electricity
ES	expected shortfall
EUR	euro
GARCH	generalised autoregressive conditional heteroskedasticity
HS	historical simulation
HU	Hungary (bidding zone)
IPE	Institute for Economic Forecasting, Romanian Academy
ISE	Fraunhofer Institute for Solar Energy Systems
LEAR	Lasso-estimated autoregressive model
LLM	large language model
LR	likelihood ratio
LSTM	long short-term memory network
MAD	median absolute deviation
MAE	mean absolute error
MAPE	mean absolute percentage error
MLP	multilayer perceptron

Abbreviations (2/2)

MW	megawatt
MWh	megawatt-hour
N-BEATS	neural basis expansion analysis for time series
OPCOM	Romanian electricity and gas market operator
PatchTST	patch time-series transformer
ReLU	rectified linear unit
RMSE	root mean squared error
RO	Romania (bidding zone)
TimesFM	Time-series Foundation Model (Google Research)
VaR	Value-at-Risk

Outline

- The forecasting problem and the data
- A strong linear benchmark and a deep network
- Deep forecasting architectures
- Time-series foundation models
- Probabilistic forecasts and conformal calibration
- Foundation models for financial risk

Two business decisions based on day-ahead forecasts

- **Energy: bidding on the day-ahead market**
  - A supplier, a hydro producer or a battery operator bids for the 24 hours of tomorrow before 12:00 Central European Time (CET) today
  - The price of every hour is set by one auction coupled across Europe; the Romanian electricity and gas market operator (OPCOM) runs it in Romania
  - An error of 1 EUR/MWh on 100 MW for 24 hours costs 2,400 euro per day
- **Finance: the Value-at-Risk (VaR) of a portfolio**
  - A bank reports every day the loss that is exceeded with probability 1 %; the Basel rules count the exceptions
  - A quantile forecast, not a point forecast
- **The question of this lecture**
  - Do networks trained on our data, or foundation models pre-trained on millions of other series, beat simple benchmarks?

(Weron, 2014; Lago et al., 2021)

The data: Romanian power prices and load, the S&P 500 and Bitcoin

Series	Frequency	Source	Used for
Day-ahead prices: Romania (RO), Germany–Luxembourg (DE-LU), Hungary (HU)	hourly	energy-charts.info, Fraunhofer Institute for Solar Energy Systems (ISE), from exchange and European Network of Transmission System Operators for Electricity (ENTSO-E) data	price forecasts
Romanian load	hourly	energy-charts.info, ENTSO-E	load forecasts
S&P 500, Bitcoin (BTC/USD)	daily	Yahoo Finance	Value-at-Risk

- Prices from 1 Jan 2015 to 31 Oct 2026: 103 595 hourly prices; the test year is 1 Nov 2025 to 31 Oct 2026: 364 complete days.
- Missing hours stay missing; the two clock-change days per year are kept only when all 24 hours exist.

Daily mean day-ahead prices in Romania

- **In one sentence:** a model trained before 2021 has never seen prices like those of 2022; a model trained on 2022 overestimates today.

$$\bar{P}_d = \frac{1}{24} \sum_{h=0}^{23} P_{d,h}$$



Source: energy-charts.info (Fraunhofer ISE), from exchange and ENTSO-E data. Data: 1 Jan 2015 to 31 Oct 2026, daily means of hourly prices.

- Mean 2019: 50.41 EUR/MWh
  - Mean 2022: 265.31 EUR/MWh
  - Highest day 738.28 on 30 Aug 2022
  - Correlation of daily means RO–DE-LU: 0.94.

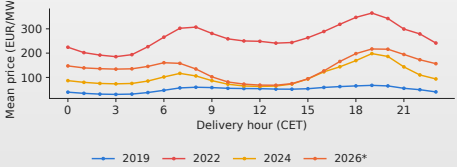
- **In practice:** choose the training window deliberately: three years, as in Lago et al. (2021), here.

NN_L13_price_history

Solar power has turned the midday peak into a trough

- **In one sentence:** the hourly shape is the strongest regularity a model can learn, and it has changed within six years.

$$\bar{P}^{(y)}_{\cdot,h} = \frac{1}{|D_y|} \sum_{d \in D_y} P_{d,h}$$



Source: energy-charts.info (Fraunhofer ISE), from exchange and ENTSO-E data. Data: 2019, 2022, 2024 and 2026 (*January-October), Romania.

- 2026 (January–October): lowest hour 13 at 67.90 EUR/MWh, highest hour 19 at 217.05 EUR/MWh.

- **In practice:** a forecaster must predict 24 different hours, not one daily number: hence 24 outputs.

NN_L13_hourly_profile

Information set at gate closure: what the model may use

- **In one sentence:** a forecast for day d may use only what is known before 12:00 CET on day $d-1$.

$$\hat{\mathbf{y}}_d = f_{\boldsymbol{\theta}}(\mathbf{x}_d), \quad \mathbf{x}_d \in \mathcal{I}_{d-1}$$

$$\mathbf{x}_d = (\mathbf{y}_{d-1}, \mathbf{y}_{d-2}, \mathbf{y}_{d-3}, \mathbf{y}_{d-7}, \mathbf{y}^{\mathrm{DE}}_{d-1}, \boldsymbol{\ell}_{d-2}, \mathbf{c}_d)$$

- **where**
  - $\mathbf{y}_d \in \mathbb{R}^{24}$ = Romanian prices of day d
  - $\mathbf{y}^{\mathrm{DE}}$ = DE-LU prices
  - $\boldsymbol{\ell}$ = Romanian load (complete two days back)
  - $\mathbf{c}_d$ = day-of-week dummies
  - $\mathcal{I}_{d-1}$ = information at gate closure.

- **151 inputs**, 3 819 daily samples since 2016.
- **Rolling origin**: re-estimate on the last 1 092 days, forecast the next day, move one day.
- **Test once**: 364 days never used for fitting or tuning.

- **In practice:** load of day $d-1$ is not complete at noon
  - Using it would leak the future into the forecast.

Experiment yourself: notebook §1

Error measures and the benchmark model

- **In one sentence:** report errors in EUR/MWh and relative to the naive forecast, never as a percentage of a price that can be zero.

$$\mathrm{MAE} = \frac{1}{24|\mathcal{T}|} \sum_{d \in \mathcal{T}} \sum_{h} |y_{d,h} - \hat{y}_{d,h}|, \quad \mathrm{rMAE} = \frac{\mathrm{MAE}}{\mathrm{MAE}^{\mathrm{naive}}}$$

$$\hat{y}^{\mathrm{naive}}_{d,h} = \begin{cases} y_{d-7,h}, & \text{Mon, Sat, Sun} \\ y_{d-1,h}, & \text{Tue–Fri} \end{cases}$$

- **where**
  - $\mathcal{T}$ = test days
  - $\mathrm{rMAE} < 1$ = better than naive
  - The root mean squared error (RMSE) weights spikes more.

- **Test year**: mean 134.59 EUR/MWh, standard deviation 77.19, range −97.55 to 954.59
  - 195 negative hours.
- **Naive**: mean absolute error (MAE) 27.74 EUR/MWh.

- **In practice:** a model with rMAE above 1 should not be deployed, however deep.

Experiment yourself: notebook §1

Outline

- The forecasting problem and the data
- A strong linear benchmark and a deep network
- Deep forecasting architectures
- Time-series foundation models
- Probabilistic forecasts and conformal calibration
- Foundation models for financial risk

Variance-stabilising transformation: price spikes and negative prices

- **In one sentence:** the area hyperbolic sine behaves like a log for large prices and like the identity near zero, so it accepts negative prices.

$$z = \operatorname{asinh}\left(\frac{y - \tilde{m}}{\hat{\sigma}}\right), \quad \hat{\sigma} = \frac{\mathrm{MAD}}{0.6745}, \quad y = \tilde{m} + \hat{\sigma}\sinh(\hat{z})$$

- **where**
  - $\tilde{m}$ = median of the training window
  - Median absolute deviation (MAD) scaled to a standard deviation
  - $\hat{z}$ = forecast on the transformed scale.

- **Fitted on the window only**: no test information enters $\tilde{m}$, $\hat{\sigma}$.
- **Same transform for LEAR and the DNN**, so differences come from the model, not from the scale.

- **In practice:** transform before fitting, invert after forecasting, evaluate in EUR/MWh.

(Uniejewski, Weron & Ziel, 2018)

Experiment yourself: notebook §2

LEAR: 24 Lasso regressions, one per hour

- **In one sentence:** a linear model with many lags, where the Lasso keeps only the lags that matter for each hour.

$$\hat{\beta}_h = \arg\min_{\beta} \sum_{d \in W} \left(z_{d,h} - \mathbf{x}_d^\top \beta \right)^2 + \lambda_h \|\beta\|_1$$

- **where**
  - W = last 1 092 days
  - λ_h chosen by the Akaike information criterion (AIC) on the Lasso path
  - Re-estimated every day: 364 × 24 fits.

- **Lasso-estimated autoregressive model (LEAR)**: MAE 22.79 EUR/MWh, rMAE 0.822.
- **Why it is hard to beat**: estimated in seconds, interpretable, adapts every day.

- **In practice:** always report LEAR next to any network: it is the open benchmark of the field.

(Lago et al., 2021)

Experiment yourself: notebook §2

The DNN: one network, 24 outputs, trained on the same inputs

- **In one sentence:** a multilayer perceptron (MLP) forecasts the whole day at once, so hours share what they learn.

$$\hat{\mathbf{z}}_d = \mathbf{W}_3\ \mathrm{ReLU}\left(\mathbf{W}_2\ \mathrm{ReLU}(\mathbf{W}_1\mathbf{x}_d + \mathbf{b}_1) + \mathbf{b}_2\right) + \mathbf{b}_3$$

$$\mathcal{L} = \frac{1}{24|W'|}\sum_{d\in W'} \|\mathbf{z}_d - \hat{\mathbf{z}}_d\|_1, \quad z_{d,h} = \operatorname{asinh}\frac{y_{d,h} - m_d}{s_d}$$

- **where**
  - m_d = median price of day $d-1$
  - s_d = MAD of the lagged prices $+ 1$
  - Rectified linear unit (ReLU)
  - W' = first 80 % of the window.

- **Deep neural network (DNN)**: MAE 23.06 EUR/MWh, mean of 4 seeds (single seeds 23.57–24.03).
- **Size**: 256 and 128 units, 74 904 parameters, dropout 0.1, Adam, early stopping on the last 20 %.
- **Per-day scaling** keeps the network usable when the price level jumps.

- **In practice:** average several seeds
  - One seed is one draw of a random estimator.

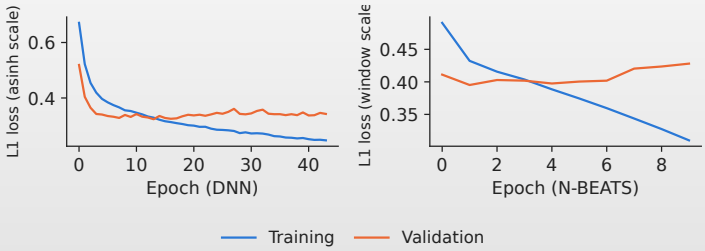
(Lago et al., 2021)

Experiment yourself: notebook §3

Early stopping picks the epoch on data the network did not train on

- **In one sentence:** training loss keeps falling; the validation loss tells when to stop.

$$e^* = \arg\min_e \mathcal{L}_{\text{val}}(\theta_e)$$



Source: energy-charts.info (Fraunhofer ISE), from exchange and ENTSO-E data. Data: training windows before 1 Nov 2025, Romania.

- DNN: best epoch 14 of 44
  - N-BEATS: best epoch 2 of 10 (patience rule).

NN_L13_training_curves

Outline

- The forecasting problem and the data
- A strong linear benchmark and a deep network
- Deep forecasting architectures
- Time-series foundation models
- Probabilistic forecasts and conformal calibration
- Foundation models for financial risk

A probabilistic recurrent network for forecasting

- **In one sentence:** the recurrent state summarises the past; the network outputs the parameters of tomorrow's distribution.

$$\mathbf{h}_t = \mathrm{LSTM}(\mathbf{h}_{t-1}, y_{t-1}, \mathbf{c}_t), \quad y_t \mid \mathbf{h}_t \sim \mathcal{N}\big(\mu(\mathbf{h}_t), \sigma^2(\mathbf{h}_t)\big)$$

$$\mathcal{L} = -\sum_t \log p\big(y_t \mid \mu(\mathbf{h}_t), \sigma(\mathbf{h}_t)\big)$$

- **Deep autoregressive recurrent network (DeepAR)**: one global model for many series.
- **Its idea lives on** in the foundation models below: global training, probabilistic output.

- **where**
  - Long short-term memory network (LSTM, Lecture 5)
  - $\mathbf{c}_t$ = covariates
  - Trained on many related series at once
  - Multi-step forecasts by sampling y_t and feeding it back.

- **In practice:** useful when you have thousands of related series (stores, meters, assets), not one.

(Salinas et al., 2020)

N-BEATS: stacked blocks that explain the input and forecast the rest

- **In one sentence:** each block removes what it can explain from the input (backcast) and adds its part to the forecast.

$$\mathbf{h}_\ell = \mathrm{FC}_\ell(\mathbf{r}_{\ell-1}), \quad \hat{\mathbf{r}}_\ell = \mathbf{V}^b_\ell \mathbf{h}_\ell, \quad \hat{\mathbf{y}}_\ell = \mathbf{V}^f_\ell \mathbf{h}_\ell$$

$$\mathbf{r}_\ell = \mathbf{r}_{\ell-1} - \hat{\mathbf{r}}_\ell, \qquad \hat{\mathbf{y}} = \sum_{\ell=1}^{L} \hat{\mathbf{y}}_\ell$$

- **where**
  - $\mathbf{r}_0$ = last 168 hours, standardised per window
  - FC_ℓ = 4 layers of width 256
  - $L = 6$ blocks
  - 1 739 904 parameters
  - Output 24 hours.

- **Neural basis expansion analysis (N-BEATS)**: prices only, trained once before the test year on every hourly origin.
- **Test MAE** 23.23 EUR/MWh (3 seeds: 23.69 to 24.12).

- **In practice:** a univariate network does not see German prices or load: compare it with LEAR, which does.

(Oreshkin et al., 2020)

Experiment yourself: notebook §3

Transformers for time series: patches as tokens

- **In one sentence:** cut the series into short patches, embed each patch like a word and let attention compare them.

$$\mathbf{p}_j = (y_{(j-1)P+1}, \ldots, y_{jP}), \quad \mathbf{e}_j = \mathbf{W}_e \mathbf{p}_j + \mathbf{q}_j$$

$$\mathrm{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}$$

- **where**
  - P = patch length (e.g. 16 or 32 hours)
  - $\mathbf{q}_j$ = position encoding
  - $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ = projections of $\mathbf{e}_j$ (Lecture 7).

- **Patch time-series transformer (PatchTST)**: 2,048 hours become 128 tokens of 16: attention cost falls 256 times.
- **Foundation models** below use exactly this design, pre-trained on huge collections of series.

- **In practice:** patching is what makes long contexts (weeks of hourly data) affordable.

(Nie et al., 2023)

Outline

- The forecasting problem and the data
- A strong linear benchmark and a deep network
- Deep forecasting architectures
- Time-series foundation models
- Probabilistic forecasts and conformal calibration
- Foundation models for financial risk

A foundation model forecasts a series it has never seen

- **In one sentence:** one network is pre-trained on many series; at use time it only reads our context, with no gradient step.

$$\hat{q}^{\tau}_{T+k} = g_{\theta^*}\left(y_{T-C+1:T}\right)_{k,\tau}$$

$$\theta^* = \arg\min_{\theta} \sum_{s\in\mathcal{S}} \sum_{t,\tau} \rho_\tau\left(y^{(s)}_t - \hat{q}^{\tau,(s)}_t\right)$$

- **where**
  - $\mathcal{S}$ = pre-training corpus (public and synthetic series)
  - C = context length (2 048 hours here)
  - ρ_τ = quantile (pinball) loss
  - $k = 1, \ldots, 24$ = horizon
  - θ^* is frozen.

- **Three models tested, zero-shot**
  - Chronos-2 (Amazon, 2025): covariates and many series together
  - Chronos-Bolt base (Amazon): direct quantile forecasts
  - Time-series Foundation Model (TimesFM) 2.5 (Google): decoder-only

- **In practice:** zero-shot means no training cost and no data pipeline, but also no adaptation to our market.

(Ansari et al., 2024, 2025; Das et al., 2024)

Experiment yourself: notebook §4

Chronos: tokenisation of time series

- **In one sentence:** scale the series, cut its values into bins, and train a language model to predict the next bin.

$$\tilde{y}_t = \frac{y_t}{s}, \quad s = \frac{1}{C}\sum_t |y_t|, \qquad v_t = \mathrm{bin}(\tilde{y}_t) \in \{1, \dots, B\}$$

$$\mathcal{L} = -\sum_t \log p_\theta(v_{t+1} \mid v_{1:t})$$

- **Scaling makes series comparable**: a price in EUR/MWh and a sales count become the same kind of input.
- **Quantile loss** gives probabilistic forecasts directly (next section).

- **where**
  - s = mean absolute scale of the context
  - B = 4,096 bins in the original Chronos
  - The model is a T5 language model
  - Chronos-Bolt and Chronos-2 replace the bins by patches and quantile heads.

- **In practice:** the model knows shapes (seasonality, mean reversion), not the Romanian market.

(Ansari et al., 2024)

Test-year accuracy: all nine forecasts on the same days

Model	MAE	RMSE	rMAE
Chronos-2 with covariates	20.41	33.88	0.736
Chronos-2	20.72	34.23	0.747
the ensemble	20.77	34.27	0.749
Chronos-Bolt	21.67	34.51	0.781
TimesFM	22.37	36.08	0.806
LEAR	22.79	36.27	0.822
DNN	23.06	38.42	0.831
N-BEATS	23.23	36.64	0.837
the naive benchmark	27.74	44.57	1.000

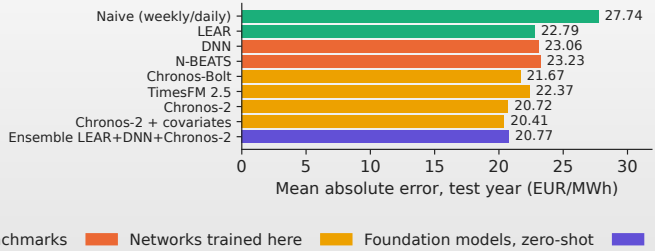
(own computation, EUR/MWh)

- **Lowest MAE**: Chronos-2 with covariates, 20.41 EUR/MWh.
- **Best foundation model**: Chronos-2 with covariates, 20.41, better than LEAR (22.79) without any training on our data.
- **All models** beat the naive forecast: rMAE below 1.
- Green = lower error
  - Significance on the next slides.

Experiment yourself: notebook §4

Trained networks versus zero-shot foundation models

- **In one sentence:** the same 8 736 hours, the same information set, nine forecasting strategies.



Source: energy-charts.info (Fraunhofer ISE), from exchange and ENTSO-E data. Data: 1 Nov 2025 to 31 Oct 2026, 360 days x 24 hours, Romania.

- Ensemble = simple mean of LEAR, DNN and Chronos-2: MAE 20.77 EUR/MWh.

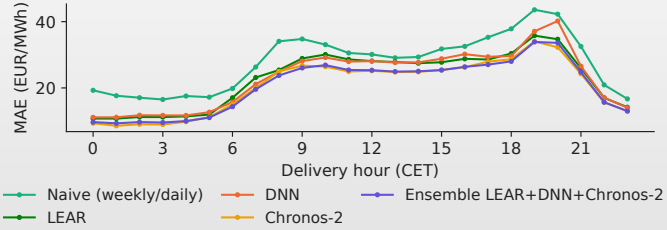
- **In practice:** start with the benchmark and a foundation model: both cost minutes, not weeks.

NN_L13_mae_models

Errors are largest in the evening and around solar hours

- **In one sentence:** the average hides hours where all models struggle together.

$$\mathrm{MAE}_h = \frac{1}{|\mathcal{T}|} \sum_{d \in \mathcal{T}} |y_{d,h} - \hat{y}_{d,h}|$$



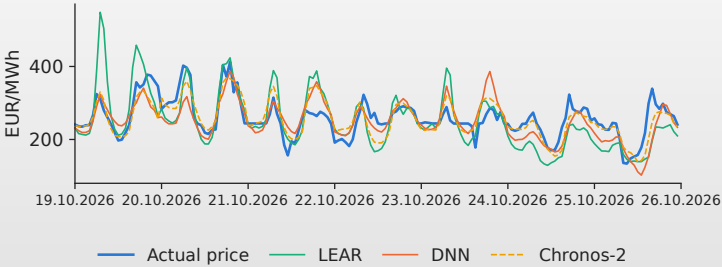
Source: energy-charts.info (Fraunhofer ISE), from exchange and ENTSO-E data. Data: 1 Nov 2025 to 31 Oct 2026, 360 days x 24 hours, Romania.

- LEAR: worst hour 19 (35.78), best hour 1 (10.75)
  - Chronos-2: worst hour 19 (34.05).

NN_L13_mae_by_hour

Forecasts for one week of the test year

- **In one sentence:** all three models follow the daily shape; none of them anticipates the largest evening spikes in full.



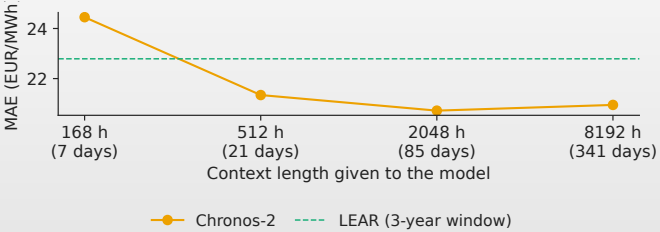
Source: energy-charts.info (Fraunhofer ISE), from exchange and ENTSO-E data. Data: 19 Oct 2026 to 25 Oct 2026, hourly, Romania.

- Week 19 Oct 2026 to 25 Oct 2026: MAE LEAR 43.34, DNN 32.33, Chronos-2 25.41 EUR/MWh.

NN_L13_forecast_week

Effect of context length on the foundation model

- **In one sentence:** the model sees only the context: one week shows the daily shape, three months also the level.



Source: energy-charts.info (Fraunhofer ISE), from exchange and ENTSO-E data. Data: 1 Nov 2025 to 31 Oct 2026, 360 days x 24 hours, Romania.

- Chronos-2 MAE: 168 h 24.45, 512 h 21.34, 2,048 h 20.72, 8,192 h 20.95 EUR/MWh.

- **In practice:** the context length is the main tuning knob of a zero-shot model: choose it on past data.

NN_L13_context_length

Covariates in a foundation model: German prices and Romanian load

- **In one sentence:** Chronos-2 reads several aligned series and lets attention share information across them.

$$\hat{\mathbf{q}}_{T+1:T+24} = g_{\theta^*}\left(y_{1:T},\ x^{\mathrm{DE}}_{1:T},\ x^{\mathrm{HU}}_{1:T},\ \ell_{1:T}\right)$$

- **where**
  - Past-only covariates: DE-LU and HU prices and Romanian load up to the forecast origin (their future values are not known at gate closure).

- **Without covariates**: MAE 20.72
  - **with covariates**: 20.41 EUR/MWh.
- **Covariates helped** here
  - Test, do not assume.

- **In practice:** a covariate known in advance (a weather or load forecast) is worth more than one known only in the past.

(Ansari et al., 2025)

Experiment yourself: notebook §5

The Diebold–Mariano test on daily losses

- **In one sentence:** a lower average error can be luck; test whether the daily loss differential has mean zero.

$$\Delta_d = \|\mathbf{e}_d^A\|_1 - \|\mathbf{e}_d^B\|_1, \qquad \mathrm{DM} = \frac{\bar{\Delta}}{\hat{\sigma}_\Delta / \sqrt{|\mathcal{T}|}} \overset{a}{\sim} \mathcal{N}(0,1)$$

- **where**
  - $\mathbf{e}_d^A \in \mathbb{R}^{24}$ = errors of model A on day d
  - One statistic per pair (multivariate version)
  - One-sided H_1: B is more accurate than A
  - $|\mathcal{T}| = 360$ days.

- **Diebold–Mariano (DM)**: LEAR vs DNN, $p = 0.761$
  - LEAR vs Chronos-2, $p = <0.001$.
- **Daily, not hourly**: hours of one day share the same information, so they are not independent.

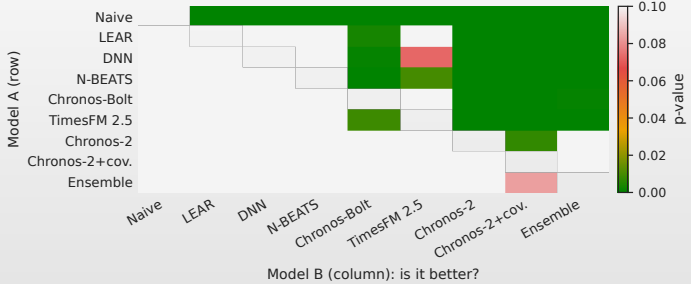
- **In practice:** claim that a model is better only with a test on the test period.

(Diebold & Mariano, 1995; Ziel & Weron, 2018)

Experiment yourself: notebook §6

Pairwise Diebold–Mariano tests for all models

- **In one sentence:** green cells: the column model is significantly more accurate than the row model.



Source: energy-charts.info (Fraunhofer ISE), from exchange and ENTSO-E data. Data: 1 Nov 2025 to 31 Oct 2026, 360 days x 24 hours, Romania.

- Ensemble vs LEAR: $p = <0.001$
  - Ensemble vs Chronos-2: $p = 0.565$.

NN_L13_dm_pvalues

Outline

- The forecasting problem and the data
- A strong linear benchmark and a deep network
- Deep forecasting architectures
- Time-series foundation models
- Probabilistic forecasts and conformal calibration
- Foundation models for financial risk

Quantile forecasts are scored with the pinball loss

- **In one sentence:** a quantile forecast is good when the right share of outcomes falls below it and the band is narrow.

$$\rho_\tau(u) = u\left(\tau - \mathbb{1}\{u < 0\}\right), \quad u = y - \hat{q}^\tau$$

$$\mathrm{PL} = \frac{1}{9} \sum_{\tau \in \{0.1, \ldots, 0.9\}} \frac{1}{24|\mathcal{T}|} \sum_{d,h} \rho_\tau\left(y_{d,h} - \hat{q}^\tau_{d,h}\right)$$

- **where**
  - τ = quantile level
  - The average over the nine deciles approximates the continuous ranked probability score.

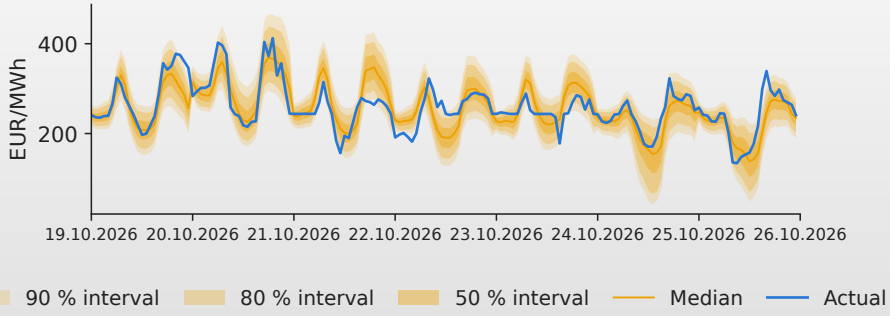
- **In practice:** a battery or a risk desk needs quantiles
  - Score them with pinball, not with MAE.

- **Average pinball loss, test year** (EUR/MWh)
  - Chronos-2 8.25; with covariates 8.11
  - Chronos-Bolt 8.62; TimesFM 8.97
  - LEAR + empirical errors of the last 60 days 9.12

Experiment yourself: notebook §7

A week of Chronos-2 quantiles: the band widens where prices jump

⊡ **In one sentence:** the model is less sure in the evening hours, exactly where the errors are largest.



Source: energy-charts.info (Fraunhofer ISE), from exchange and ENTSO-E data. Data: 19 Oct 2026 to 25 Oct 2026, hourly, Romania.

NN_L13_chronos_fan

Split-conformal correction of prediction intervals

- **In one sentence:** measure how far the outcomes fell outside the raw band in the recent past, and widen (or shrink) the band by that amount.

$$s_i = \max\left(\hat{q}_i^{\alpha/2} - y_i,\ y_i - \hat{q}_i^{1-\alpha/2}\right), \qquad \hat{Q} = s_{(\lceil (n+1)(1-\alpha) \rceil)}$$

$$\hat{C}_{d,h} = \left[\hat{q}_{d,h}^{\alpha/2} - \hat{Q},\ \hat{q}_{d,h}^{1-\alpha/2} + \hat{Q}\right], \qquad \Pr\left(y \in \hat{C}\right) \geq 1 - \alpha$$

- **where**
  - s_i = conformity scores on the last 60 days ($n = 24 \times$ days)
  - $s_{(k)}$ = k-th smallest
  - The guarantee holds under exchangeability, only approximately for time series.

- **Conformalised quantile regression (CQR)** applied to Chronos-2, 80 % band: raw coverage 76.5 %, corrected 79.3 %.
- **Width**: 60.03 → 63.89 EUR/MWh.

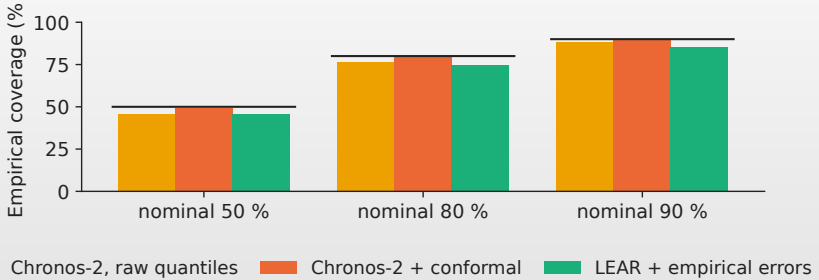
- **In practice:** a foundation model plus conformal calibration is a cheap, honest probabilistic forecaster.

(Romano, Patterson & Candès, 2019)

Experiment yourself: notebook §7

Conformal calibration moves coverage towards the nominal level

- **In one sentence:** black lines: the nominal level; bars: the share of test hours inside the band.



Source: energy-charts.info (Fraunhofer ISE), from exchange and ENTSO-E data. Data: 1 Nov 2025 to 31 Oct 2026, 360 days x 24 hours, Romania.

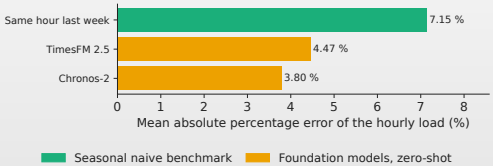
- 90 % band: raw 88.3 %, conformal 89.4 %, LEAR + empirical errors 85.0 %
  - TimesFM 80 %: raw 79.6 %, conformal 79.7 %.

NN_L13_coverage

Load forecasts of foundation models: errors of a few percent

- **In one sentence:** load is smooth and seasonal, the shapes a pre-trained model knows best.

$$\mathrm{MAPE} = \frac{100}{24|\mathcal{T}|} \sum_{d,h} \left| \frac{\ell_{d,h} - \hat{\ell}_{d,h}}{\ell_{d,h}} \right|$$



Source: energy-charts.info (Fraunhofer ISE), from exchange and ENTSO-E data. Data: 1 Nov 2025 to 31 Oct 2026, 360 days, Romanian load.

- Mean absolute percentage error (MAPE): last week 7.15 %, TimesFM 4.47 %, Chronos-2 3.80 % (MAE 214.4 MW, mean load 5 932 MW).

- **In practice:** percentage errors are fine for load, which is never close to zero.

NN_L13_load_mape

Outline

- The forecasting problem and the data
- A strong linear benchmark and a deep network
- Deep forecasting architectures
- Time-series foundation models
- Probabilistic forecasts and conformal calibration
- Foundation models for financial risk

Value-at-Risk is a conditional quantile, backtested by counting violations

⊡ **In one sentence:** a 1 % VaR should be breached on about 1 % of days, and the breaches should not cluster.

$$\Pr\left(r_{t+1} < \mathrm{VaR}^{\alpha}_{t+1} \mid \mathcal{F}_t\right) = \alpha, \qquad I_{t+1} = \mathbb{1}\{r_{t+1} < \mathrm{VaR}^{\alpha}_{t+1}\}$$

$$\mathrm{LR}_{uc} = -2 \ln \frac{(1-\alpha)^{n_0}\alpha^{n_1}}{(1-\hat{\pi})^{n_0}\hat{\pi}^{n_1}} \sim \chi^2_1$$

⊡ **where**

- r_t = daily log return (%)
- n_1 = violations, n_0 = other days, $\hat{\pi} = n_1/(n_0 + n_1)$
- Likelihood ratio (LR) test of Kupiec
- Christoffersen adds a test of independence of I_t.

⊡ **Three forecasters**

- Historical simulation (HS): empirical quantile of the last 250 days
- Generalised autoregressive conditional heteroskedasticity (GARCH): GARCH(1,1) with Student-t errors, re-estimated every 21 days
- Chronos-2 quantiles from the last 1,024 returns, zero-shot

⊡ **In practice:** the Basel traffic light turns on the number of 1 % exceptions in 250 days.

(Kupiec, 1995; Christoffersen, 1998)

Experiment yourself: notebook §9

A one-sided conformal shift turns a raw quantile into a calibrated VaR

- **In one sentence:** if the model's quantile was breached too often in the last year, lower it by exactly the amount that would have fixed it.

$$s_i = \hat{q}_i^{\alpha} - r_i, \qquad c_t = s_{(\lceil (n+1)(1-\alpha) \rceil)} \text{ over the last } n = 250$$

$$\widehat{\mathrm{VaR}}_{t+1}^{\alpha} = \hat{q}_{t+1}^{\alpha} - c_t$$

- **where**
  - $\hat{q}^{\alpha}$ = raw Chronos-2 quantile
  - $c_t > 0$ widens the VaR, $c_t < 0$ narrows it
  - The shift uses only past forecasts and returns.

- **S&P 500, 1 % level**: raw Chronos-2 22 violations, conformal 16, expected 17.2.
- **Same idea** as the energy bands, applied to one tail.

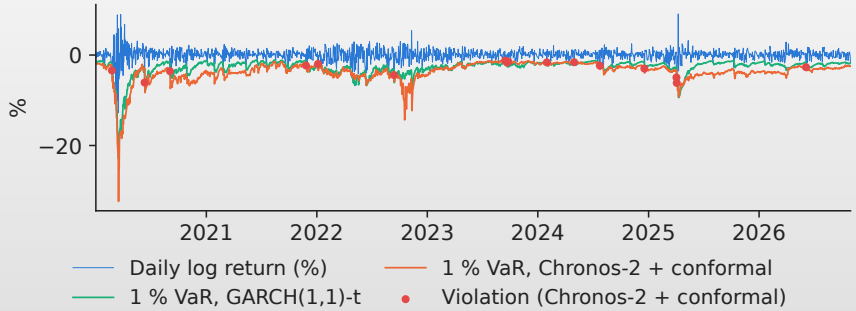
- **In practice:** recalibration is a model change: document it and backtest it out of sample.

(Romano et al., 2019)

Experiment yourself: notebook §9

S&P 500: the calibrated foundation-model VaR follows volatility regimes

- **In one sentence:** both lines widen in March 2020 and in April 2025; red dots are the days the conformal Chronos-2 VaR was breached.



Source: Yahoo Finance, daily quotes. Data: 2 Jan 2020 to 30 Oct 2026, trading days, S&P 500.

NN_L13_var_sp500

Backtests, January 2020 to October 2026: S&P 500 and Bitcoin

Asset	Model	Viol. 1 %	p_{uc}	p_{ind}	Viol. 5 %	p_{uc}	Tick loss 5 %
S&P 500	HS (250 days)	31	0.003	0.002	89	0.729	0.1582
	GARCH(1,1)-t	32	0.001	0.607	132	0.000	0.1347
	Chronos-2, raw	22	0.262	0.285	122	0.000	0.1368
	Chronos-2 + conformal	16	0.774	0.138	93	0.434	0.1383
Bitcoin	HS (250 days)	33	0.123	0.458	126	0.912	0.3633
	GARCH(1,1)-t	22	0.543	0.189	128	0.770	0.3587
	Chronos-2, raw	31	0.242	0.402	134	0.404	0.3591
	Chronos-2 + conformal	21	0.413	0.170	120	0.657	0.3614

- Expected violations: S&P 500 17.2 (1 %) and 85.9 (5 %) over 1 717 days; Bitcoin 25.0 and 124.8 over 2 496 days (traded every day).
- Green p-value: coverage not rejected at 10 %; red: rejected. Tick loss = pinball loss at $\tau = 0.05$ (lower is better).

(own computation; Kupiec, 1995; Christoffersen, 1998)

Experiment yourself: notebook §9

Quantile forecasts from language models and time-series models

- **Large language model (LLM) route** (Lecture 12)
  - The returns are written as text in a prompt; a general model answers with a number
  - LLM-VaR and, for the expected shortfall (ES), LLM-ES, of our group: zero-shot, no fine-tuning (Pele et al., 2026)
- **Time-series foundation model route** (this lecture)
  - Numbers enter as scaled patches; the output is a set of quantiles
  - Cheaper per forecast, can be run locally, calibrated here with a conformal shift
- **Common rules**
  - Backtest out of sample, count violations, test independence, compare with GARCH and HS

Quiz 1: leakage

- A colleague adds the actual Romanian load of day d to the inputs of the price forecast for day d and the MAE falls
  - Should you deploy the model?

Answer

- No: the actual load of day d is unknown at 12:00 on $d-1$
  - Use the operator's load forecast or lags available at gate closure.

Quiz 2: conformal width

- The raw 80 % Chronos-2 band covered 76.5 % of test hours
  - What does the conformal step do to its width, and why?

Answer

- It changes it by $\hat{Q}$, the $(1-\alpha)$ quantile of recent out-of-band distances: wider if the band was too narrow, narrower if too wide (here 60.03 $\rightarrow$ 63.89 EUR/MWh).

Quiz 3: Kupiec

- Over 1 717 days a 1 % VaR had 3 times the expected number of violations
  - Is the Kupiec p-value large or small?

Answer

- Small: $\hat{\pi}$ is far from α, so LR_{uc} is large and coverage is rejected
  - The model underestimates risk.

Wrap-up

- Benchmarks first: naive 27.74, LEAR 22.79 EUR/MWh
  - Lowest MAE: Chronos-2 with covariates (20.41).
- Foundation models forecast prices and load zero-shot
  - The context length and covariates are their tuning knobs.
- Significance by Diebold–Mariano on daily losses
  - Several seeds for every network.
- Conformal calibration turns raw quantiles into honest bands and VaR
  - Backtest with Kupiec and Christoffersen.
- Next: multimodal models and the synthesis of the course (Lecture 14).

Research from our group

- **Four generations of energy price forecasting**: econometric models, machine learning, LSTM and transformers compared on European energy markets, with pre-training and transfer learning (Andrei, Velev, Toma, Pele & Lessmann, 2024, arXiv 2411.03372).
- **Early warning of price spikes**: explainable gradient boosting on Romanian hourly data
  - The residual load forecast is the main trigger of spike warnings (Andrei, Pele, Cramer, Amza & Diaconu, 2026, PICBE).
- **Multi-horizon explainability**: which drivers matter from 0–6 hours to two weeks ahead and in which price regime, Romania 2018–2024 (Andrei, Pele & Cramer, 2026, PICBE).
- **LLM-VaR and LLM-ES**: general-purpose language models produce Value-at-Risk and expected shortfall (ES) forecasts zero-shot from a prompt with past returns (Pele et al., 2026, Expert Systems with Applications).
- **Early warning for crypto 'zombies'**: machine learning predicts which crypto-assets will become inactive (Będowska-Sójka, Wójcik & Pele, 2026, North American Journal of Economics and Finance).
- Findings as stated in the papers' abstracts; numbers on the other slides come from our own data.

(Bucharest University of Economic Studies, ASE; Romanian Academy, Institute for Economic Forecasting, IPE)

Further reading and resources

- **Chronos** code and models: github.com/amazon-science/chronos-forecasting.
- **TimesFM**: github.com/google-research/timesfm.
- **epftoolbox**, the open benchmark of Lago et al. (2021): github.com/jeslago/epftoolbox.
- **Conformal prediction**, a gentle introduction: arXiv:2107.07511.
- **Forecasting: Principles and Practice**, 3rd ed.: otexts.com/fpp3.
- **This lecture**: notebook, seminar and Quantlets — github.com/danpele/NEURAL_BIZ.

References (1/2)

- Andrei, A.-V., Pele, D.-T., Cramer, A.-A., Amza, A. & Diaconu, D. (2026). Explainable AI-based early warning system for electricity price spikes: Evidence from the Romanian market. *Proceedings of the International Conference on Business Excellence*, 20(1), 4820–4836. doi:10.2478/picbe-2026-0352
- Andrei, A.-V., Pele, D.-T. & Cramer, A.-A. (2026). Multi-horizon explainability in energy price forecasting: Disentangling market dynamics and renewable intermittency in Romania's liberalized electricity market. *Proceedings of the International Conference on Business Excellence*, 20(1), 104–117. doi:10.2478/picbe-2026-0014
- Andrei, A.-V., Velev, G., Toma, F.-M., Pele, D. T. & Lessmann, S. (2024). Energy price modelling: A comparative evaluation of four generations of forecasting methods. arXiv:2411.03372
- Ansari, A. F. et al. (2024). Chronos: Learning the language of time series. *Transactions on Machine Learning Research*. arXiv:2403.07815
- Ansari, A. F. et al. (2025). Chronos-2: From univariate to universal forecasting. arXiv:2510.15821
- Będowska-Sójka, B., Wójcik, P. & Pele, D. T. (2026). Early warning systems for cryptocurrency markets: Predicting 'zombie' assets using machine learning. *The North American Journal of Economics and Finance*, 81, 102543. doi:10.1016/j.najef.2025.102543
- Christoffersen, P. F. (1998). Evaluating interval forecasts. *International Economic Review*, 39(4), 841–862. doi:10.2307/2527341
- Das, A., Kong, W., Sen, R. & Zhou, Y. (2024). A decoder-only foundation model for time-series forecasting. *Proceedings of the 41st International Conference on Machine Learning*. arXiv:2310.10688
- Diebold, F. X. & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3), 253–263. doi:10.1080/07350015.1995.10524599
- Kupiec, P. H. (1995). Techniques for verifying the accuracy of risk measurement models. *The Journal of Derivatives*, 3(2), 73–84. doi:10.3905/jod.1995.407942

References (2/2)

- Lago, J., Marcjasz, G., De Schutter, B. & Weron, R. (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
- Nie, Y., Nguyen, N. H., Sinthong, P. & Kalagnanam, J. (2023). A time series is worth 64 words: Long-term forecasting with transformers. *International Conference on Learning Representations*. arXiv:2211.14730
- Oreshkin, B. N., Carpov, D., Chapados, N. & Bengio, Y. (2020). N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. *International Conference on Learning Representations*. arXiv:1905.10437
- Pele, D. T., Bolovăneanu, V., Lin, M.-B., Ren, R., Ginavar, A. T., Spilak, B., Andrei, A.-V., Toma, F.-M., Lessmann, S. & Härdle, W. K. (2026). In the beginning was the Word: LLM-VaR and LLM-ES. *Expert Systems with Applications*, 295, 128676. doi:10.1016/j.eswa.2025.128676
- Romano, Y., Patterson, E. & Candès, E. (2019). Conformalized quantile regression. *Advances in Neural Information Processing Systems*, 32. arXiv:1905.03222
- Salinas, D., Flunkert, V., Gasthaus, J. & Januschowski, T. (2020). DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3), 1181–1191. doi:10.1016/j.ijforecast.2019.07.001
- Uniejewski, B., Weron, R. & Ziel, F. (2018). Variance stabilizing transformations for electricity spot price forecasting. *IEEE Transactions on Power Systems*, 33(2), 2219–2229. doi:10.1109/TPWRS.2017.2734563
- Weron, R. (2014). Electricity price forecasting: A review of the state-of-the-art with a look into the future. *International Journal of Forecasting*, 30(4), 1030–1081. doi:10.1016/j.ijforecast.2014.08.008
- Ziel, F. & Weron, R. (2018). Day-ahead electricity price forecasting with high-dimensional structures: Univariate vs. multivariate modeling frameworks. *Energy Economics*, 70, 396–420. doi:10.1016/j.eneco.2017.12.016

Appendix A: the coverage guarantee of the conformal band

- Let $s_1, \ldots, s_n, s_{n+1}$ be exchangeable conformity scores
  - The rank of s_{n+1} among them is uniform on $\{1, \ldots, n+1\}$.
- Hence $\Pr(s_{n+1} \leq s_{(k)}) \geq k/(n+1)$
  - Choosing $k = \lceil (n+1)(1-\alpha) \rceil$ gives $\Pr(s_{n+1} \leq \hat{Q}) \geq 1-\alpha$.
- $s_{n+1} \leq \hat{Q}$ is equivalent to $y_{n+1} \in [\hat{q}^{\alpha/2} - \hat{Q}, \hat{q}^{1-\alpha/2} + \hat{Q}]$, which is the band.
- Hours of a price series are dependent and drift
  - The rolling window of 60 days trades exactness for adaptation.

(Romano, Patterson & Candès, 2019)

Appendix B: Christoffersen's test of independence

- Count transitions of the hit sequence: n_{ij} = number of days with $I_{t-1} = i$ and $I_t = j$
  - $\hat{\pi}_{01} = n_{01}/(n_{00} + n_{01})$, $\hat{\pi}_{11} = n_{11}/(n_{10} + n_{11})$.
- $\mathrm{LR}_{ind} = -2 \ln \frac{(1-\hat{\pi})^{n_{00}+n_{10}} \hat{\pi}^{n_{01}+n_{11}}}{(1-\hat{\pi}_{01})^{n_{00}} \hat{\pi}_{01}^{n_{01}} (1-\hat{\pi}_{11})^{n_{10}} \hat{\pi}_{11}^{n_{11}}} \sim \chi^2_1$.
- A small p-value means violations cluster: the model reacts too slowly to a volatility shock.

(Christoffersen, 1998)

Appendix C: implementation details

- LEAR: scikit-learn LassoLarsIC (criterion AIC), standardised inputs, daily re-estimation on 1 092 days.
- DNN and N-BEATS: PyTorch, L1 loss, Adam, early stopping
  - Seeds 20261108 and the next ones.
- Foundation models: amazon/chronos-2, amazon/chronos-bolt-base, google/timesfm-2.5-200m-pytorch from Hugging Face, inference only
  - The full run took 2.7 minutes on a laptop.
- Hours on the CET calendar
  - The doubled autumn hour is averaged
  - Days with a missing hour are excluded from training targets and from the test.