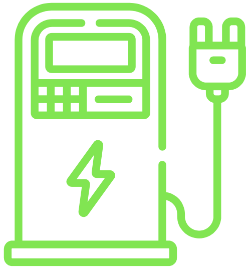
JDS 2025

Deep Learning applied to hierarchical forecasting

Yvenn Amara-Ouali

4th June 2025

OUTLINE



01 Forecasting Hierarchical Data

02 Problem Formulation

03 Data and Experimental Setting

04 Proposed Models and Results

05 Conclusion and Perspectives

OUTLINE

01 Forecasting Hierarchical Data

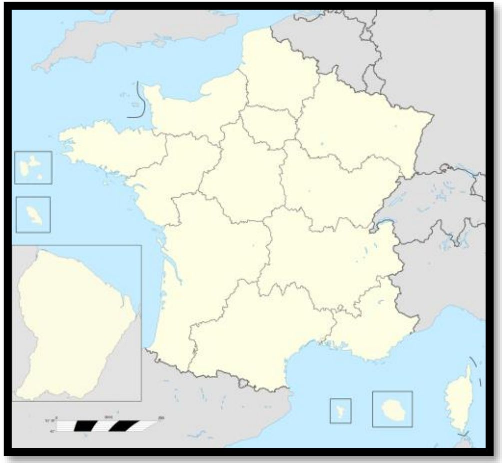
02 Problem Formulation

03 Data and Experimental Setting

04 Proposed Models and Results

05 Conclusion and Perspectives

What do we mean by « Hierarchichal Data »?



Spatial
Cities → Regions → Nation

Temporal
Days → Months → Year(s)

Functional
Total Load →
Industrial vs Residential

Different Basic Approaches

Direct Forecast Indirect Forecast

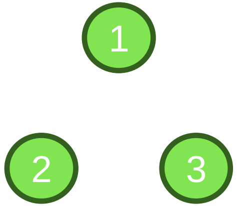
D: dataset
(training or validation)

Agnostic

Ignore the hierarchy and hope for the best

Example

- *Independent forecasts for each node* - $\{\hat{y}_k^i\}_{i \in D, k \in \{1,2,3\}}$

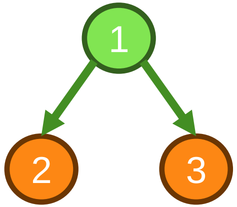


Top-Down Approaches

Forecast at the top level(s) and disaggregate

Example

- *Average Historical Proportions* - $\hat{y}_2^i = p\hat{y}_1^i$, $\hat{y}_3^i = (1-p)\hat{y}_1^i$

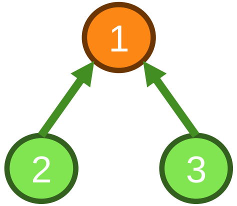


Bottom-up

Forecast at the bottom level(s) and aggregate

Classical methods

- *Summing all bottom nodes* - $\hat{y}_1^i = \hat{y}_2^i + \hat{y}_3^i$



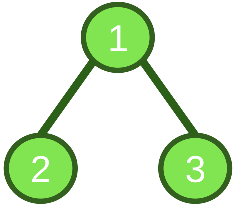
Coherence

Bottom-Up & **Top-Down** approaches ensure **Forecast Coherence**

Definition - Forecast Coherence

The forecasted values respect the natural hierarchy of the data. For instance, the sum of the forecasts for each French Regions are indeed equal to the French level forecast).

More generally, it ensures that forecasts at lower levels aggregate up to higher levels.



A little note on forecast reconciliation (not the core of this talk)

Is there a way to get **coherent forecasts** from **agnostic forecasts** (while **improving forecasting accuracy)?**

Theorem – MinT (Minimum Trace) Reconciliation

1. For **agnostic** or **base forecast** $\hat{y}^i = \{\hat{y}_k^i\}_{k \in \{1 \dots K\}}$ we define the **reconciled forecast** as follows :

$$\tilde{\boldsymbol{y}}^i = \boldsymbol{S}\boldsymbol{P}\hat{\boldsymbol{y}}^i$$

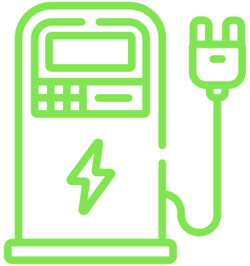
With $\boldsymbol{S}$ the hierarchy structure matrix and $\boldsymbol{P}$ projection (or reconciliation) matrix

2. Now let $\mathbf{W_h}$ be the positive definite covariance matrix of the h-step-ahead base forecast errors. Then the optimal reconciliation matrix, which minimizes $\boldsymbol{tr}[\boldsymbol{S}\boldsymbol{P}\boldsymbol{W}_{\boldsymbol{h}}\boldsymbol{P}^{\boldsymbol{T}}\boldsymbol{S}^{\boldsymbol{T}}]$ such that $\boldsymbol{SPS} = \boldsymbol{S}$, is given by

$$\boldsymbol{P} = \left(\boldsymbol{S}^{\boldsymbol{T}}\boldsymbol{W}_{\boldsymbol{h}}^{-1}\boldsymbol{S}\right)^{-1}\boldsymbol{S}^{\boldsymbol{T}}\boldsymbol{W}_{\boldsymbol{h}}^{-1}$$

Wickramasuriya, Shanika L., George Athanasopoulos, and Rob J. Hyndman. "Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization." *Journal of the American Statistical Association* 114.526 (2019): 804-819.

OUTLINE



01 Forecasting Hierarchical Data

02 **Problem Formulation**

03 Data and Experimental Setting

04 Proposed Models and Results

05 Conclusion and Perspectives

Motivations

- **Electricity demand must be forecasted accurately** to maintain balance between supply and demand, as electricity cannot be stored efficiently and must match real-time consumption.

- **New challenges arise** from decentralized networks, economic shifts like Covid, and non-stationary consumption patterns, increasing forecast complexity.

- **The challenge focuses on multi-scale forecasting**, leveraging granular consumption and geolocalized data to reduce uncertainty across different levels.

The aim of this Deep Learning Challenge is therefore to focus on methods that can propose a forecast at **different levels of granularity**

1. **Cities**
2. **Regions**
3. **France**

Formal metric

Cost function to optimize

$$C(\hat{y}) = \sum_{k=1}^{K} \sqrt{\frac{1}{T} \sum_{i=1}^{T} \left(y_k^i - \hat{y}_k^i\right)^2}$$

Essentially, it's the sum of the **RMSE** over each node of the hierarchy

Two main modelling approaches with deep learning

1. One Model per Node

$$f_k(x_k) = \widehat{y_k} \quad \forall\, k \in \{1 \dots K\}$$

Advantages

- **Specialization** - Each model can capture unique dynamics (seasonality, demand shifts, anomalies)
- **Robust to heterogeneity** - Works well when series differ significantly (e.g., regions with various climates)

Limitations

- **Scalability -** Training and maintaining many models is computationally expensive
- **Data sparsity -** Many bottom-level nodes may have little or noisy data → overfitting or poor generalization

2. One Global Model

$$f(x) = \hat{y} = \{\hat{y}_k\}_{k \in \{1 \dots K\}}$$

Advantages

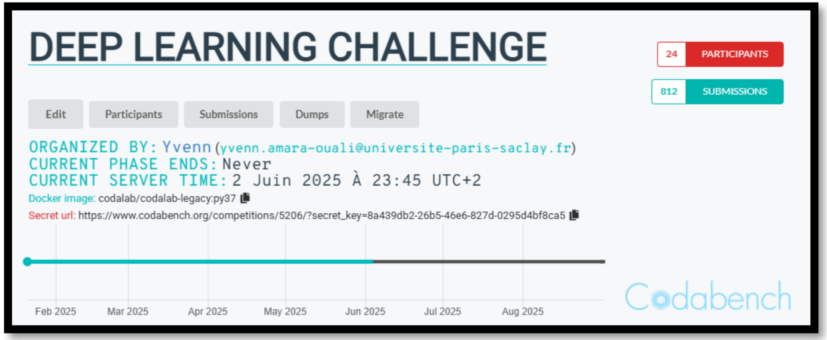
- **Scalability** - One model handles thousands of series efficiently, especially with GPUs or TPUs
- **Cross-series learning -** Learns shared patterns across time series

Limitations

- **Loss of specialization -** May struggle to model highly heterogeneous behavior in a specific series
- **Risk of underfitting large series -** Large (important) nodes may be dominated by the loss from small nodes

These represent the two main extremes, though many intermediate approaches are possible.

Data Challenge



12 teams of students
Master 1 – Maths & AI

université PARIS-SACLAY

Evaluation Criterias

1. **Public & Private Scores**
2. **Presentation**
3. **Technical Report**
4. **Git Repository**

OUTLINE

Electricity Demand and Weather Data

Electricity Data – Source RTE

Perimeter
- France
- 12 administrative regions (mainland)
- 12 major cities

Step
- 30 minutes

Target variable
- Load [MW]

Weather Data – Source SYNOP (WMO)

Perimeter
- 69 weather stations across France

Step
- 3 hours

Features
- Temperature
- Humidity
- Wind (direction & speed)
- …

Experimental setting

Private/Public set
(chosen at 50/50 random)

Training Data

Validation Data

2017

2022

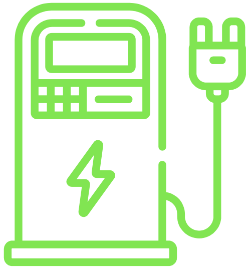
2023

Features – Calendar and Weather Data
Target – Load for 25 nodes

Mid-Term Forecasting Task
(using weather observed data)

```
date,pred_France,pred_Auvergne-Rhône-Alpes,[...],pred_Île-de-France,pred_Montpellier Méditerranée
Métropole,[...],pred_Toulouse Métropole
2022-01-01 00:00:00+01:00,69135.70,9321.05,1443.35,1078.94,1676.82
2022-01-01 00:30:00+01:00,67406.70,9114.47,2780.58,3579.17,2850.35
2022-01-01 01:00:00+01:00,64773.11,8833.41,2721.82,3328.47,2739.76
2022-01-01 01:30:00+01:00,64469.11,8828.70,2755.88,5793.47,6824.76
```

OUTLINE



01 Forecasting Hierarchical Data

02 Problem Formulation

03 Data and Experimental Setting

04 **Proposed Models and Results**

05 Conclusion and Perspectives

Base Model

Mean per month, day of week and timestep

```
date,pred_France,pred_Auvergne-Rhône-Alpes,[...],pred_Île-de-France,pred_Montpellier Méditerranée
Métropole,[...],pred_Toulouse Métropole
2022-01-01 00:00:00+01:00,69135.70,9321.05,1443.35,1078.94,1676.82
2022-01-01 00:30:00+01:00,67406.70,9114.47,2780.58,3579.17,2850.35
2022-01-01 01:00:00+01:00,64773.11,8833.41,2721.82,3328.47,2739.76
2022-01-01 01:30:00+01:00,64469.11,8828.70,2755.88,5793.47,6824.76
```

A little note on scaling the output

Computing the **MSE** on the **scaled outputs** gives equal weight to each output node in the loss.

However, this doesn't reflect the behavior of our target metric, where various nodes in the hierarchy can contribute differently to the total error (in MW).

Target Metric (RMSE)

$$\sum_{k=1}^{K} \sqrt{\frac{1}{T} \sum_{i=1}^{T} \left(y_k^i - \hat{y}_k^i\right)^2}$$

Scaling per node (e.g., min-max)

$$z_k^i = \frac{y_k^i - \min_{i \in D_{train}} y_k^i}{\max_{i \in D_{train}} y_k^i - \min_{i \in D_{train}} y_k^i}$$

Actual Metric (MSE)

$$\sum_{k=1}^{K} \frac{1}{T} \sum_{i=1}^{T} (z_k^i - \hat{z}_k^i)^2$$

3rd – Jie & Youwei

Final Scores
Public : 7386 MW
Private : 7404 MW

Features

- StandardScaler
- Fourier basis
- Linear interpolation of weather data

Architecture

MLP with attention

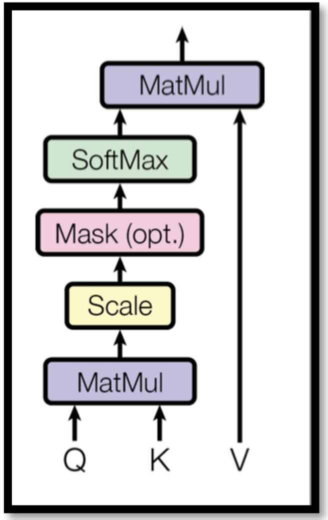
- 1 Attention Layer
- 3 MLP Layers
- BatchNorm
- Dropout - 0.2
- Hidden dimension - 512
- ReLU

Basic **Feature Attention** Mechanism

For inputs $x \in \mathbb{R}^d$:

1. Compute attention scores: $s = W_{att}x + b$
2. Convert to weights: $\alpha = softmax(s)$
3. Apply weights: $x_{weighted} = x \odot \alpha$
4. Feed into regression head: $\hat{y} = f(x_{weighted})$

x		s
Features	**Value**	**Attention Weights**
Temperature	25°	0.60
nebu	Cloud Cover	0.20
is_weekend	Monday	0.10
region_id	Bretagne	0.10

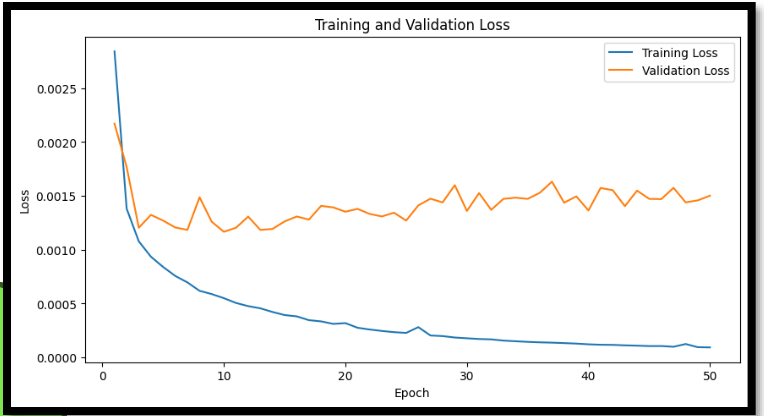


2nd – Jules & Mathias

Final Scores
Public : 7306 MW
Private : 7282 MW

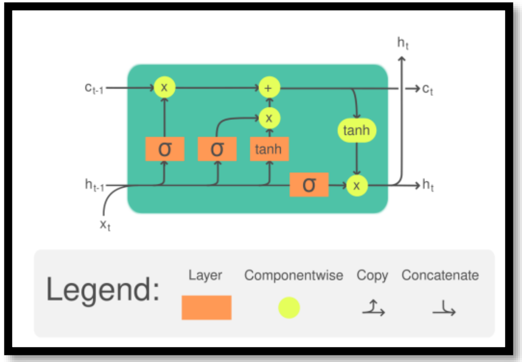
Features

- All available features
- Linear interpolation of weather data

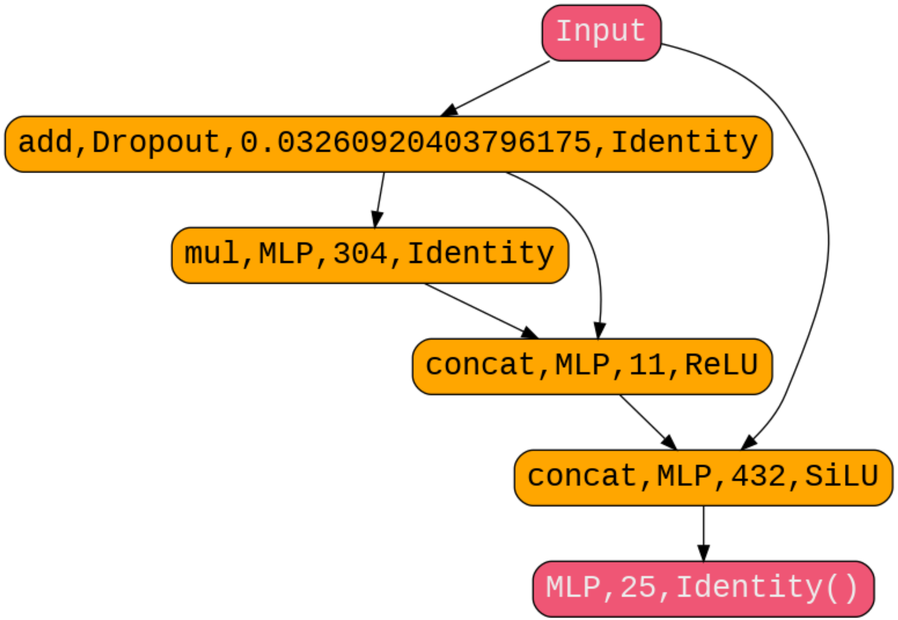


Architecture

- One model per Node
- 2 layer LSTM with 64 hidden neurons per layer
- Followed by a 32 neurons MLP
- Activation functions : ReLU
- Optimiseur: Adam
- Loss: MSE
- Learning rate: 0.01
- Batch size = 32
- Epochs : around 8 per zone



Honourable Mention – Felix et Lylian



Scores
Public : 8515 MW
Privé : 8566 MW

Architecture

- DRAGON (auto-DL) []
- algorithme de recherche : Mutant-UCB
- Limits: Overfitting

Keisler, Julie, et al. "Automated Deep Learning for load forecasting."
International Conference on Automated Machine Learning. PMLR, 2024.

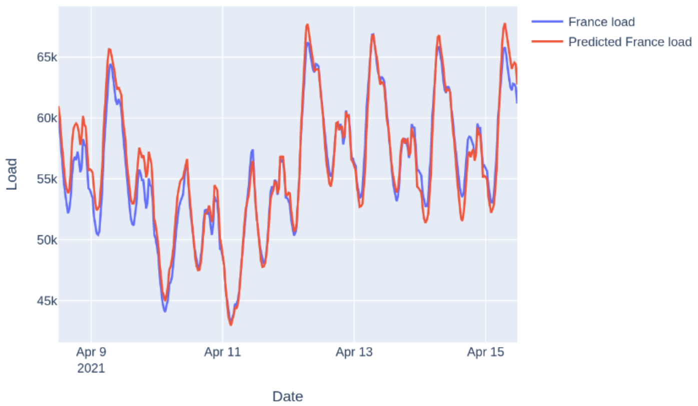
1st – Félix et Samuel

Final Scores
Public : 6068 MW
Private : 6063 MW

Features

- Sequence size: 48 (1 day)
- Keeping all weather stations
- CO_2 emissions (sobriety)

Predictions on the first week of the validation set



Architecture

- Single LSTM for France, Regions and Cities
- Learning rate: 0.0012
- Hidden layers: 256 neurons
- 10 epochs
- Batch size: 128

$$l(\hat{Y}, Y^{(B)}) = \sum_{j=1}^{m} \sqrt{\frac{1}{B_j} \sum_{i=1}^{B} 1_{Y_{ij} \neq -1} (\hat{Y}_{ij} - Y_{ij})^2}.$$

OUTLINE

Perspectives

Brégère, Margaux, and Malo Huard. "Online hierarchical forecasting for power consumption data." International Journal of Forecasting 38.1 (2022): 339-351.

Campagne, Eloi, et al. "Leveraging Graph Neural Networks to Forecast Electricity Consumption." European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases 2024.

Principato, Guillaume, et al. "Conformal Prediction for Hierarchical Data." arXiv e-prints (2024): arXiv-2411.

Appendix

MinT Assumptions

Assumption	If Violated	Solution
Unbiased base forecasts	Coherent forecasts will also be biased	Apply bias correction beforehand
Accurate estimation of **W**	Poor reconciliation or overfitting	Use shrinkage or WLS approximations
Fixed hierarchy **S**	Coherence becomes ill-defined	Rebuild or dynamically update the structure
Linear reconciliation	Can't model nonlinear dependencies	Use machine learning-based methods separately
Horizon-specific	Cross-horizon dependency ignored	Use multi-horizon models or scenario reconciliation