

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Spatial-temporal modeling of the electrical load in Portugal

Pedro Rodrigo Soares da Silva



Mestrado em Engenharia Eletrotécnica e de Computadores

Supervisor: Ricardo Bessa

Co-Supervisor: Carla Gonçalves

October 1, 2026

Resumo

A previsão da carga elétrica é uma ação central na operação e no planeamento dos sistemas de energia. Com a crescente disponibilidade de dados de consumo geograficamente detalhados, a previsão deixou de estar limitada à procura agregada ao nível do sistema, tornando-se possível estudar como a carga elétrica evolui no tempo e no espaço, em particular ao nível da distribuição, onde a procura é mais volátil, heterogénea e dependente do local.

Esta dissertação investiga a modelação espaço-temporal e a previsão de curto prazo da carga elétrica em Portugal continental, a partir de dados reais de consumo horário do portal de dados abertos da E-REDES. Muitas abordagens de previsão assentam sobretudo em informação temporal ou na procura agregada, deixando pouco explorado o contributo das relações espaciais e hierárquicas entre áreas ao nível da distribuição, pelo que o objetivo principal é avaliar se a informação entre localizações e entre níveis fornece valor preditivo útil para além de referências puramente temporais.

Para tal, é construído um conjunto de dados hierárquico a partir das séries de carga por código postal de 4 dígitos, agregadas em regiões de 2 e 1 dígitos e numa série nacional. Os dados são pré-processados para corrigir inconsistências temporais e analisados por decomposição multi-sazonal, separando tendência, sazonalidade diária, sazonalidade semanal e resíduo, sobre o qual é conduzida uma análise exploratória das dependências espaciais entre áreas.

Os resultados exploratórios mostram forte sazonalidade diária e semanal, heterogeneidade substancial entre áreas de código postal e uma estrutura espacial residual relevante após remover as componentes temporais dominantes. As dependências de curto prazo entre localizações concentram-se nos desfasamentos de uma e duas horas e enfraquecem consideravelmente em desfasamentos mais longos, revelando ainda o agrupamento que um grande subconjunto de áreas partilha dinâmicas de carga residual semelhantes.

Com base nestas conclusões, é formulada uma estrutura de previsão espaço-temporal que combina referências temporais com modelos LASSO-VAR-X, cuja regularização seleciona as relações espaciais mais informativas mantendo a abordagem tratável à escala da rede de códigos postais. As previsões são depois integradas por reconciliação hierárquica e a comparação é estendida à previsão probabilística através de regressão quantílica com modelos de *gradient boosting*.

Os resultados de previsão mostram que o valor desta informação é condicional e não uniforme. A informação entre localizações melhora a previsão ao nível de agregação intermédio, mas degrada-a ao nível mais desagregado e entre os poucos agregados regionais de maior dimensão, enquanto a informação entre níveis se comporta de forma inversa e se torna valiosa no topo da hierarquia. A reconciliação melhora a exatidão em todos os níveis para além de restaurar a coerência, dependendo o esquema de ponderação preferido das previsões de base e não da hierarquia, ao passo que na previsão probabilística a vantagem da informação entre localizações desaparece e o fator dominante passa a ser o regime de treino. Demonstra-se assim que os dados ao nível da distribuição contêm estrutura espacial explorável, cujo aproveitamento melhora a exatidão de curto prazo na maioria dos níveis da hierarquia.

Abstract

Electrical load forecasting is a central task in power system operation and planning. With the increasing availability of geographically detailed consumption data, forecasting is no longer limited to highly aggregated system-level demand. Instead, it becomes possible to study how electrical load evolves across both time and space, particularly at the distribution level, where demand is more volatile, heterogeneous, and locally dependent.

This dissertation investigates the spatio-temporal modelling and short-term forecasting of electrical load in mainland Portugal using real hourly consumption data made available through the E-REDES open data portal. Many load forecasting approaches rely primarily on temporal information or on aggregated demand, leaving the potential contribution of spatial and hierarchical relationships among distribution-level areas underexplored. The main objective is therefore to assess whether cross-location and cross-level information can provide useful predictive value beyond purely temporal baselines.

To support this analysis, a hierarchical dataset is constructed from 4-digit postal-code electrical load series and aggregated into 2-digit regions, 1-digit regions, and a national-level series. The data are preprocessed to correct temporal inconsistencies and analysed through multi-seasonal decomposition, separating trend, daily seasonality, weekly seasonality, and residual components, on which an exploratory analysis of the spatial dependencies between areas is then conducted.

The exploratory results show strong daily and weekly seasonal behaviour, substantial heterogeneity across postal-code areas, and relevant residual spatial structure after removing dominant temporal components. Short-term cross-location dependencies are mainly concentrated at one- and two-hour lags, while longer lagged relationships become considerably weaker, and a large subset of postal-code areas shares similar residual load dynamics.

Based on these findings, a spatio-temporal forecasting framework is formulated combining temporal baselines with LASSO-VAR-X models, whose regularisation selects the most informative spatial relationships while keeping the approach tractable at the scale of the postal-code network. The resulting forecasts are integrated through a hierarchical reconciliation step, and the comparison is extended to probabilistic forecasting through quantile regression with gradient boosting models.

The forecasting results show that the value of this information is conditional rather than uniform. Cross-location information improves the forecast at the intermediate aggregation level, while degrading it at the most disaggregated level and among the few large regional aggregates, whereas cross-level information behaves in the opposite manner and becomes valuable at the upper levels of the hierarchy. Reconciliation improves accuracy at every level beyond restoring coherence, with the preferred weighting scheme depending on the base forecasts rather than on the hierarchy, while in the probabilistic setting the advantage of cross-location information disappears and the dominant factor becomes the training regime. Distribution-level data is thereby shown to contain exploitable spatial structure whose use improves short-term accuracy at most levels of the hierarchy.

UN Sustainable Development Goals

The United Nations Sustainable Development Goals (SDGs) [30] provide a global framework to achieve a better and more sustainable future for all, comprising 17 goals that address challenges such as poverty, inequality, climate change and environmental degradation.

This dissertation contributes to a more sustainable and efficient electricity system by improving the short-term forecasting of electrical load at the distribution level. By exploiting the spatial and hierarchical structure of geographically detailed consumption data, it produces more accurate and coherent forecasts across aggregation levels, supporting the reliable and efficient operation of distribution networks, the integration of variable renewable generation, and a lower reliance on carbon-intensive reserves.

The specific Sustainable Development Goals addressed in this work are:

SDG 7 Ensure access to affordable, reliable, sustainable and modern energy for all.

SDG 9 Build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation.

SDG 13 Take urgent action to combat climate change and its impacts.

SDG	Target	Contribution	Performance Indicators and Metrics
7	7.2	More accurate demand forecasts facilitate the integration of variable renewable generation.	Relative error (RMSE%) reduced over the univariate baselines.
	7.3	Coherent hierarchical forecasts cut the reserves and losses needed to secure the system.	Consistent MAE, RMSE and MASE gains across levels.
9	9.1	Spatially informed forecasting improves awareness of the regional distribution grid, supporting resilient operation.	Calibrated uncertainty estimates (pinball loss, CRPS, coverage).
	9.5	Advances forecasting methods exploiting cross-location and cross-level dependencies via LASSO-VAR-X and reconciliation.	Interpretable spatial dependence recovered by the coefficients.
13	13.2	Improved accuracy supports network planning oriented to renewable integration and decarbonisation.	Short-term accuracy gains from cross-level information.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Ricardo Bessa, for the guidance, the availability and the trust placed in this work from the outset. His ability to identify what mattered at each important stage, and to say so plainly, kept the dissertation focused at moments when the number of possible directions could easily have made it otherwise.

I am deeply grateful to my co-supervisor, Carla Gonçalves, whose technical guidance and attention to detail were decisive throughout. Just as important, and in a different way, was the friendly and encouraging presence she brought to every weekly meeting, which had a real effect on the work and on the way I carried it out.

I also thank INESC TEC for hosting this dissertation and for providing the conditions under which it was carried out, from the workspace to the data and the computational resources without which the experiments reported here would not have been possible.

To my friends, thank you for the patience during the months in which I was less present, and for the reminders that there was a world outside this document.

Finally, to my family, thank you for the unconditional support and for a confidence in me that never wavered, even at the moments when my own did. This work owes more to you than these lines can convey.

Pedro Silva

Contents

Acknowledgements	iv
1 Introduction	1
1.1 Context and Motivation	1
1.2 Research Questions and Contributions	2
1.3 Document Structure	3
2 Background and State-of-the-Art	5
2.1 Characteristics of Electrical Load	5
2.1.1 Temporal Dependence and Seasonality	6
2.1.2 Signal Decomposition and Residual Dynamics	8
2.1.3 Spatio-temporal Dynamics	10
2.1.4 Aggregation, Hierarchy and Coherence	12
2.2 Electrical Load Forecasting	14
2.2.1 Forecasting Horizons and Operational Context	15
2.2.2 Statistical Methods	17
2.2.3 Machine Learning and Deep Learning	19
2.2.4 Probabilistic Forecasting	20
2.2.5 Hierarchical and Geographically Aggregated Forecasting	21
2.3 Evaluation Metrics	23
2.3.1 Point Forecasting Metrics	23
2.3.2 Probabilistic Forecasting Metrics	24
2.4 Research Gaps and Critical Discussion	25
3 Data Description and Exploratory Analysis	27
3.1 Data Source and Scope	28
3.2 Data Preprocessing	29
3.2.1 Data Cleaning and Correction	29
3.2.2 Time Series Decomposition	31
3.3 Spatio-temporal Analysis	34
3.3.1 Correlation Analysis	35
3.3.2 Partial Cross-Correlation Analysis	37
3.3.3 Clustering Analysis	41
3.4 Construction of the Hierarchical Dataset	46
3.5 Main Data Challenges	50
3.6 Chapter Summary	52

4	Problem Formulation and Methodology	53
4.1	Problem Definition	53
4.2	Forecasting Scenarios	54
4.2.1	Predictor Features	55
4.2.2	Scenario Design	56
4.3	Forecasting Approaches	58
4.3.1	Point Forecasting	58
4.3.2	Probabilistic Forecasting	66
4.4	Evaluation Metrics and Comparison Criteria	72
4.5	Chapter Summary	72
5	Results and Discussion	74
5.1	Base Point Forecast Results	74
5.1.1	Overview	75
5.1.2	Impact of Intra- and Cross-Level Dynamics	79
5.1.3	Precision and Dispersion	82
5.2	Reconciliation Results	87
5.3	Probabilistic Forecast Results	92
5.3.1	Effect of the Training Window	92
5.3.2	Univariate and Multivariate Forecasts	94
5.3.3	Calibration Across the Hierarchy	96
5.3.4	Temporal Behaviour of the Predictive Intervals	98
5.4	Chapter Summary	100
6	Conclusions	101
6.1	Main Findings	101
6.2	Limitations	103
6.3	Future Work	104
	References	105
A	Detailed Point Forecast Results	108
A.1	Validation Errors	108
A.2	Coefficient Inventories	109
B	Complete Probabilistic Forecast Metrics	110

List of Figures

2.1	Effect of aggregation level on residential load profiles. Source: [12].	7
2.2	Illustration of multiple seasonal-trend decomposition using LOESS (MSTL). Source: [22].	9
2.3	Regional-level spatio-temporal load patterns and graph abstraction. Source: [14]	11
2.4	Example of a Hierarchical Time Series (HTS) structure with unique nesting levels. Source: [16].	13
2.5	Example of a Grouped Time Series (GTS) structure with overlapping aggregation paths. Source: [16].	13
2.6	Classification of electrical load forecasting horizons. Source: [32]	16
2.7	Applications in short-term load forecasting. Source: [6]	17
3.1	Original hourly electrical load series for the four-digit postal code 4450.	30
3.2	Original 4450 postal-code electrical load profile around the daylight saving time transition.	30
3.3	Original and corrected 4450 postal-code electrical load profiles around the daylight saving time transition.	31
3.4	MSTL decomposition for the electrical load series of postal code 4450.	32
3.5	Average daily seasonal profile for the electrical load series of postal code 4450.	33
3.6	Average weekly seasonal profile for the electrical load series of postal code 4450.	33
3.7	Autocorrelation (left) and partial autocorrelation (right) functions for lags up to 168 hours.	34
3.8	Distribution of Pearson correlations between the trend component of the four-digit postal-code series and temperature.	35
3.9	Pearson correlation heatmap between residual electrical load series at the four-digit postal-code level.	36
3.10	Partial cross-correlation function between two selected four-digit postal-code residual series.	37
3.11	Partial cross-correlation heatmap between four-digit postal-code residual series for a one-hour lag.	39
3.12	Partial cross-correlation heatmap between four-digit postal-code residual series for a two-hour lag.	40
3.13	Hierarchical clustering dendrogram of four-digit postal codes (Pearson threshold = 0.6): (a) left half, (b) right half.	43
3.14	Representative electrical load series from a large high-correlation cluster and three examples of low-correlation postal-code series.	45
3.15	Spatial hierarchy used to aggregate the electrical load data: (a) four-digit postal-code level, (b) two-digit postal-code level, (c) one-digit postal-code level, and (d) national aggregate.	47

3.16	Example of an aggregation path across the spatial hierarchy: postal code 4450, two-digit region 44, one-digit region 4, and national aggregate.	49
4.1	Temporal train/test split of the eleven-month dataset.	59
4.2	LOMO cross-validation scheme for generating OOF predictions over the training period.	60
4.3	Procedure per LOMO fold: inner HP search, model fitting, and OOF prediction. .	60
4.4	Generalised LASSO-VAR-X pipeline for the base forecasts at a given hierarchical level.	62
4.5	Sliding training window at three instants of the test period, for $W = 210$ days. . .	66
4.6	Quantile levels of the predictive distribution and inheritance of hyperparameters from the pivot quantiles.	68
4.7	Validation block used for predictor selection within the training period.	68
4.8	Leave-one-month-out scheme used for the hyperparameter search.	69
4.9	Two-stage pipeline of the probabilistic forecasts at a given hierarchical level. . . .	70
5.1	Distribution of the per-series relative MAE by scenario, at the three levels with more than one series.	78
5.2	Effective contribution of the load-based information sources to the forecast, for the best spatio-temporal and the best hierarchical scenario of each level.	79
5.3	Retained cross-series coefficients of the four-digit scenario S3 model against the partial cross-correlation of the same pair of series at the same lag, at the one and two hour lags. Colour encodes the number of pairs per cell.	81
5.4	One representative series per level over the test period, in daily means, with the observed load, the best model of each level, and the univariate baseline B2	82
5.5	Per-series relative error (RMSE%) of the best model of each level against the univariate baseline (B2). The triangle marks one four-digit series beyond the axis range.	83
5.6	Difference in mean relative error between the best multivariate scenario of each level and the baseline B2 , as the worst-predicted multivariate series are removed, worst first.	84
5.7	Mean relative error of the four-digit multivariate (S3) and univariate (B2) forecasts by month of the test period.	85
5.8	The four-digit series 4835 and its forecasts (top), with two of its predictor series below. August is shaded.	86
5.9	The four-digit series 5225 and its forecasts (top), with two of its predictor series below. August is shaded.	87
5.10	Monthly CRPS of the national forecast for the three window lengths.	93
5.11	CRPS gap of scenario S2 to scenario B2 as the worst multivariate series are removed. .	96
5.12	Empirical against nominal coverage of the scenario S2 forecasts.	97
5.13	Calibration deviation per quantile level and relative CRPS of the two-digit series, for scenario S2	98
5.14	Predictive intervals of scenario S2 for postal code 47 over a June and an August week.	99

List of Tables

3.1	Summary of the E-REDES electrical load dataset characteristics.	29
3.2	Hierarchical clustering results for different Pearson correlation thresholds.	44
3.3	Structure of the hierarchical electrical load dataset.	48
4.1	Forecasting scenarios evaluated in this work. The scenarios progressively incorporate temporal, spatial and hierarchical information to assess the contribution of each source of predictive information.	57
4.2	Search space of the LASSO-VAR-X hyperparameters optimised by Bayesian optimisation.	64
4.3	Search space of the LightGBM hyperparameters tuned by the tree-structured Parzen estimator.	71
5.1	Point forecast accuracy across aggregation levels and forecasting scenarios. The model type that attains the best results at each level is shaded. The hierarchical scenarios (H1 to H3) do not apply at the four-digit level.	76
5.2	Mean absolute aggregation gap of the base forecasts relative to aggregate load.	88
5.3	Accuracy of the reconciled forecasts at every hierarchical level, set against the most accurate scenario obtained at that level without reconciliation. Within each base, the better weighting scheme is shaded.	89
5.4	Mean MASE over the four hierarchical levels, each level weighted equally, for every base-forecast set and reconciliation method.	91
5.5	Improvement of scenario B2 over the 90-day window, per hierarchical level.	93
5.6	Improvement of the sliding regime over a single static fit, for scenario B2 at the longest window.	94
5.7	Univariate and multivariate forecasts at the longest window, with the best value of each metric in bold.	95
A.1	Leave-one-month-out validation error of the spatio-temporal and hierarchical scenarios, by training month and hierarchical level. The lowest mean of each level is in bold.	108
A.2	Ten largest coefficients of three four-digit series in different regions, with the partial cross-correlation of each pair at the matching lag.	109
B.1	Probabilistic forecast metrics for every level, scenario and training window length W , averaged over the series of each level. The pinball loss, the CRPS and the interval width are expressed in kWh. The national level admits no multivariate scenario.	110

List of Acronyms

ACF	Autocorrelation Function
ADMM	Alternating Direction Method of Multipliers
ANN	Artificial Neural Network
API	Application Programming Interface
ARIMA	Autoregressive Integrated Moving Average
BU	Bottom-Up
CCF	Cross-Correlation Function
CNN	Convolutional Neural Network
CRPS	Continuous Ranked Probability Score
DL	Deep Learning
DSO	Distribution System Operator
E-REDES	Portuguese electricity distribution system operator
EMD	Empirical Mode Decomposition
GBDT	Gradient Boosting Decision Trees
GECAD	Research Group on Intelligent Engineering and Computing for Advanced Innovation and Development
GEFCom	Global Energy Forecasting Competition
GRU	Gated Recurrent Unit
GTS	Grouped Time Series
HP	Hyperparameters
HTS	Hierarchical Time Series
INESC TEC	Institute for Systems and Computer Engineering, Technology and Science
LASSO	Least Absolute Shrinkage and Selection Operator
LASSO-VAR	LASSO-regularised Vector Autoregression
LASSO-VAR-X	LASSO-regularised Vector Autoregression with exogenous inputs
LOESS	Locally Estimated Scatterplot Smoothing
LOMO	Leave-One-Month-Out
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MASE	Mean Absolute Scaled Error
MinT	Minimum Trace
ML	Machine Learning
MO	Middle-Out
MSTL	Multiple Seasonal-Trend decomposition using LOESS
NWP	Numerical Weather Prediction
OLS	Ordinary Least Squares

OOF	Out-Of-Fold
PACF	Partial Autocorrelation Function
PCCF	Partial Cross-Correlation Function
PI	Prediction Interval
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
SDG	Sustainable Development Goal
STL	Seasonal-Trend decomposition using LOESS
STLF	Short-Term Load Forecasting
SVR	Support Vector Regression
TCN	Temporal Convolutional Network
TD	Top-Down
TPE	Tree-structured Parzen Estimator
UN	United Nations
VAR	Vector Autoregression
WLS	Weighted Least Squares

Chapter 1

Introduction

This chapter introduces the relevance of load forecasting in modern power systems and motivates the need to analyse electrical demand jointly across time and space. It then presents the research questions and contributions of the dissertation, followed by an overview of the remaining chapters.

1.1 Context and Motivation

Electrical load forecasting is a key component in the planning and operation of power systems, supporting the continuous balance between energy supply and demand. At the distribution level, this task has become increasingly challenging due to evolving consumption patterns, the integration of renewable energy sources, and the need for more efficient and adaptive grid management. As a result, improving forecasting methodologies remains an active and relevant research topic.

Traditional forecasting approaches mainly focus on the temporal behaviour of electrical consumption, exploiting historical demand patterns, calendar effects, and exogenous variables such as weather conditions. Load time series are characterised by strong daily, weekly, and seasonal patterns, which are commonly incorporated into predictive models. However, electrical load is not governed solely by temporal factors. Geographically close or socio-economically similar regions often exhibit correlated demand behaviour, indicating the presence of spatial interactions that evolve over time.

The spatial dimension of electrical consumption is, however, frequently overlooked in the forecasting literature. Demand at geographically close or socio-economically similar regions tends to co-evolve in ways that purely temporal models cannot fully account for. Spatio-temporal dependencies, arising from shared climatic exposures, synchronised behavioural routines, or common network infrastructure, introduce structured patterns across regions that, if left unmodelled, represent a source of systematic information loss.

Distribution-level data are, moreover, naturally organised into a spatial hierarchy, where fine-grained local series can be aggregated into progressively coarser regional levels up to a system-wide total. This hierarchical structure is usually treated only as a coherence constraint, ensuring that forecasts add up consistently across levels. Yet it is also a potential source of predictive

information: lower levels carry local detail that may help forecast aggregate demand, while higher levels provide more stable contextual signals for disaggregated areas. Exploiting these cross-location and cross-level relationships is therefore not merely a methodological refinement, but a step towards forecasting models that are genuinely grounded in how demand propagates across a distribution network.

Motivated by these observations, this dissertation investigates the short-term forecasting of distribution-level electrical load in mainland Portugal, using real hourly consumption data made openly available through the E-REDES open data portal and organised as a postal-code hierarchy. The central question is whether spatial and hierarchical information can provide predictive value beyond purely temporal baselines. To this end, the work first characterises the temporal, spatial, and hierarchical structure of the data. It then evaluates forecasting approaches that incorporate cross-location and cross-level information within a regularised, interpretable framework. Coherence across aggregation levels and the quantification of forecast uncertainty are also addressed.

1.2 Research Questions and Contributions

The literature reviewed in the following chapter establishes that spatial dependence between load series is measurable and that hierarchical structures are routinely available at distribution level, while leaving largely open how much either of them is worth once a competent temporal baseline is already in place. Taking that gap as its starting point, this dissertation is organised around four research questions.

1. Does information drawn from other locations at the same aggregation level improve short-term load forecasting beyond purely temporal baselines, and does the answer depend on the level at which the forecast is produced?
2. Does information propagated from an adjacent level of the hierarchy carry predictive value, and does the form in which it is supplied, whether as individual forecasts, as an aggregate forecast or as observed lagged loads, alter that value?
3. Once base forecasts have been produced independently at each level, does hierarchical reconciliation improve accuracy beyond restoring coherence, and is the choice of weighting scheme governed by the hierarchy or by the base forecasts it is applied to?
4. Do the conclusions established for point forecasts carry over to probabilistic forecasts, where the quality of the entire predictive distribution is at stake rather than the accuracy of its centre?

In answering these questions, this dissertation makes the following contributions to the field of spatio-temporal electrical load forecasting:

- **Empirical characterisation of the residual spatial structure of Portuguese distribution-level load:** A dependence analysis conducted on the residuals of a multi-seasonal decomposition rather than on the raw series, so that the relationships reported between postal-code

areas are not artefacts of the calendar behaviour they share, establishing that the dependence is broad in space but confined to delays of one and two hours and that a substantial part of the network shares common residual dynamics while many areas remain isolated.

- **Unified and leakage-free evaluation framework across a four-level hierarchy:** A comparison of eight forecasting scenarios under a single estimator, at every aggregation level and with one validation protocol, in which every forecast used as an input feature at one level is generated out of fold at the level below, so that any observed difference is attributable to the information made available to the model rather than to the model itself.
- **Evidence that the value of cross-location information is not monotonic in aggregation:** A demonstration that this information yields a moderate loss of accuracy at the most disaggregated level, a decisive gain at the intermediate level and a pronounced loss among the few large regional aggregates, which identifies the number of series sharing a level and their mutual homogeneity, rather than the degree of aggregation itself, as the factors that govern its usefulness.
- **Characterisation of cross-level information as a function of its form and of its source:** A demonstration that information propagated from an adjacent level becomes valuable precisely where cross-location information fails, provided that the supplying level is regular enough to carry signal rather than error, and that the form in which it is supplied is as decisive as its availability, with the preferred form itself varying from one level to the next.
- **Reconciliation study conditioned on the base forecasts:** A comparison of the weighted least squares and minimum trace estimators over several sets of base forecasts, establishing that the preferred weighting scheme is a property of the base rather than of the hierarchy, that modelling the error correlations repays their estimation only where the base deliberately ignores the spatial structure of the load, and that the projection reverses which base is worth starting from at most levels.
- **Extension of the comparison to probabilistic forecasting:** An evaluation of the same hierarchy under quantile regression, establishing that the ranking obtained for point forecasts does not survive there and that the training regime, namely how much history each fit sees and how often it is re-estimated, outweighs every variation of the feature configuration, together with a calibration analysis that traces the residual miscalibration back to the seasonal coverage of the training window.

1.3 Document Structure

The remainder of this dissertation is organised into five chapters.

Chapter 2 reviews the literature required to frame the problem, covering the statistical characteristics of electrical load, its decomposition into trend and seasonal components, the nature of

spatial and spatio-temporal dependence between series, and the aggregation constraints that govern hierarchical structures. It then surveys the main families of forecasting methods, from statistical models to machine-learning and probabilistic approaches, and the criteria used to evaluate them. Gaps in the literature are identified to motivate the methodology introduced in this dissertation.

Chapter 3 describes the dataset used throughout the work and analyses its temporal, spatial and hierarchical properties. The chapter details the data source and its scope, the preprocessing applied to correct temporal inconsistencies, the exploratory evidence of dependence between postal-code areas obtained through residual correlation, partial cross-correlation and hierarchical clustering, and the construction of the four-level hierarchy on which every subsequent experiment is performed. It concludes with the data limitations that constrain the experimental design and with the mitigation adopted for them.

Chapter 4 formalises the forecasting problem and presents the methodology. The chapter defines the eight forecasting scenarios under comparison, from univariate baselines to spatio-temporal and cross-level configurations, and details the two forecasting approaches end to end, namely the LASSO-VAR-X pipeline that produces the point forecasts together with the reconciliation step applied to them, and the quantile regression pipeline that produces the probabilistic forecasts. It closes with the metrics and comparison criteria adopted throughout.

Chapter 5 reports and discusses the experimental results over the held-out test period. The eight scenarios are first compared at every level and the outcome is traced back to the estimated coefficients and to the behaviour of the individual series. The reconciliation of those forecasts is then evaluated over several sets of base forecasts, and the probabilistic forecasts are analysed in terms of the training regime, the comparison between univariate and multivariate scenarios, and the calibration of the predictive distributions across the hierarchy. A critical assessment of what each source of information is and is not worth is provided throughout.

Chapter 6 concludes the dissertation with a synthesis of the main findings, relating each of them back to the research questions stated above. It states the limitations that bound the scope of the results, notably the temporal extent of the dataset and the one-step-ahead horizon, and outlines the directions in which the work points next.

Supporting material that documents the results without altering their reading is gathered in the appendices, with Appendix A reporting the complete validation errors used for scenario selection together with the coefficient inventories of the case-study series, and Appendix B reporting the full set of probabilistic metrics for every level, scenario and training window length.

Chapter 2

Background and State-of-the-Art

The forecasting of electrical load has evolved substantially with the increasing availability of geographically detailed consumption data. Traditional forecasting studies often treat electrical demand as a single time series observed at a high aggregation level, such as the national system or a transmission area. In that setting, the main modelling challenge is to represent temporal regularities, including short-term persistence, daily and weekly seasonality, weather sensitivity, and calendar effects. However, distribution-level data introduce an additional dimension: electrical load is not only temporally dependent, but also spatially structured.

At finer spatial resolutions, such as substations, feeders, or postal-code regions, demand becomes more volatile and heterogeneous. Local consumption patterns may differ substantially depending on residential density, commercial activity, industrial presence, climatic exposure, and behavioural routines. At the same time, nearby or functionally similar areas may exhibit related load dynamics. This means that the forecasting problem can no longer be fully described by the past values of a single series. Instead, it becomes a spatio-temporal problem in which information from other locations and from different aggregation levels may contain predictive value.

This chapter reviews the literature required to frame this problem. Section 2.1 discusses the main statistical characteristics of electrical load, focusing on temporal dependence, seasonality, decomposition, spatial dependence, and hierarchical aggregation. Section 2.2 reviews the main families of load forecasting methods, including statistical models, machine-learning and deep-learning approaches, and probabilistic forecasting. Section 2.3 presents the evaluation criteria commonly used to compare forecasting models. Finally, Section 2.4 identifies the research gaps addressed in this dissertation and positions the proposed work within the broader literature on spatio-temporal and hierarchical load forecasting.

2.1 Characteristics of Electrical Load

Load demand is a signal whose behaviour depends on both time and space. From a temporal perspective, load observations are not independent, but exhibit persistence, recurrent seasonal patterns, and sensitivity to external drivers such as weather conditions and calendar effects. From a

spatial perspective, electrical consumption is observed across multiple geographical or network-related units, whose trajectories may evolve in a coordinated way due to shared climatic exposure, similar socio-economic characteristics, common behavioural routines, or common network structures. These properties make electrical load forecasting a spatio-temporal problem, particularly at the distribution level, where demand is more volatile, heterogeneous, and locally dependent.

This section reviews the main characteristics of electrical load that are relevant to the present dissertation. Section 2.1.1 discusses temporal dependence and multi-seasonality. Section 2.1.2 introduces signal decomposition and explains why residual dynamics are useful for dependency analysis. Section 2.1.3 discusses spatial and spatio-temporal dependence between load series. Finally, Section 2.1.4 reviews aggregation, hierarchy, and coherence across spatial levels.

2.1.1 Temporal Dependence and Seasonality

From a temporal perspective, load observations are inherently dependent. Consecutive observations are usually strongly related, since consumption patterns reflect persistent socio-technical drivers such as human routines, commercial and industrial schedules, and the gradual evolution of external conditions. This temporal memory is visible through strong autocorrelation at short lags and recurrent behaviour at longer lags. As a result, historical load values are among the most informative predictors in short-term operational forecasting, either through explicit autoregressive terms or through lag-based features engineered from past observations [5].

A defining aspect of electrical load time series is the coexistence of multiple seasonal patterns. At both distribution and system levels, daily and weekly cycles are typically prominent, with systematic differences between working days and weekends. At longer horizons, annual modulation also affects the level and shape of load profiles, particularly through weather conditions and seasonal changes in social and economic activity. Portuguese operational data provide evidence of these regularities, showing annual, weekly, and daily seasonality, as well as marked deviations during holiday periods [21].

This multi-seasonal structure is not only descriptive, but also methodological. Forecasting models must be able to represent multiple seasonal cycles simultaneously. For example, triple-seasonal formulations have been proposed to capture intraday, intraweek, and intrayear periodicities within a unified stochastic framework, rather than treating these scales in isolation [28]. This is particularly relevant for hourly electrical load, where daily and weekly cycles overlap and jointly determine the regular structure of the signal.

The temporal dependence of a load series is commonly analysed using the autocorrelation function (ACF) and the partial autocorrelation function (PACF). The ACF measures the total association between the current value and past values at different lags, and is useful for identifying persistence at short lags and recurrent patterns at seasonal lags, such as 24 hours for daily seasonality or 168 hours for weekly seasonality in hourly data. However, the dependence measured by the ACF may include both direct effects and indirect effects transmitted through intermediate lags. The PACF provides a complementary view by isolating the direct contribution of each lag after accounting for the effect of the intervening ones. This distinction is important because a high

autocorrelation at a given lag does not necessarily imply an independent predictive contribution from that lag. In load forecasting, ACF and PACF analyses can therefore support lag selection and feature design, helping to identify which past observations should be included as model inputs while reducing the risk of redundant temporal features [15, 19, 33].

The observed behaviour of electrical load also depends on the temporal resolution at which it is measured. Temporal aggregation, such as moving from minute-level to hourly, daily, or weekly observations, acts as a filter that attenuates high-frequency fluctuations and emphasises low-frequency components such as trends and long-term cycles. This scale dependence is particularly useful in practice because operational datasets often exist at multiple resolutions, each revealing different and complementary aspects of the same underlying process [19].

In addition to temporal aggregation, aggregation across consumers also has a pronounced effect on the observed behaviour of electrical load. As the number of aggregated consumers decreases, the smoothing effect induced by aggregation diminishes, and short-term variability becomes increasingly dominant. This phenomenon is illustrated in Figure 2.1, which shows residential load profiles aggregated over different numbers of households. While high aggregation levels produce smooth and regular daily patterns, lower aggregation reveals irregular, spiky behaviour driven by individual consumption events [12].

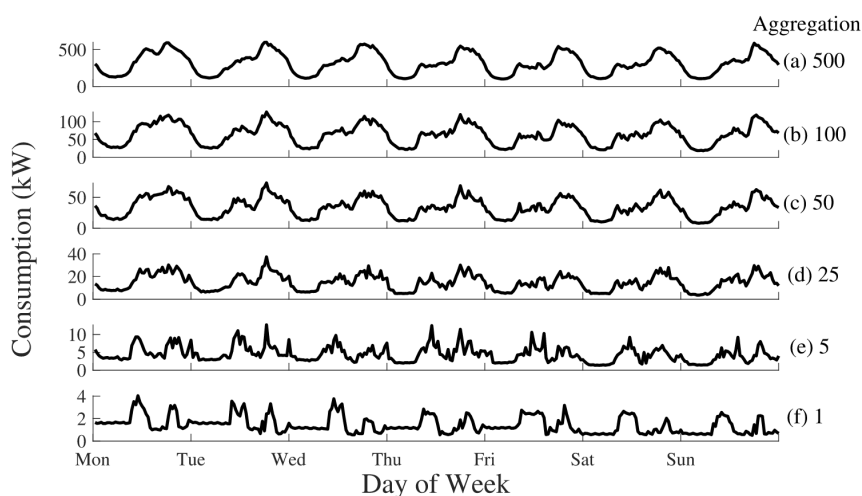


Figure 2.1: Effect of aggregation level on residential load profiles. Source: [12].

Overall, electrical load exhibits short-term persistence, multiple interacting seasonalities, and scale-dependent behaviour. These characteristics are essential for forecasting, but they also create a challenge for spatio-temporal analysis: if common trend and seasonal patterns dominate the signal, different locations may appear strongly related even when the relationship is mainly driven by shared temporal structure. This motivates the use of signal decomposition and residual analysis, discussed next.

2.1.2 Signal Decomposition and Residual Dynamics

The temporal structure of electrical load time series can be understood as the combination of long-term evolution, recurrent seasonal patterns, and short-term irregular fluctuations. Although these components jointly describe the observed signal, their coexistence often increases non-stationarity and obscure the dynamics that are most relevant for short-term forecasting and dependency analysis. Signal decomposition addresses this issue by separating the original series into interpretable components associated with different temporal scales, thereby enabling a clearer representation of underlying behaviour and facilitating subsequent analysis [7, 17].

A widely adopted decomposition framework is Seasonal-Trend decomposition using LOESS (STL) [18]. In its classical additive form, STL represents the observed series as an additive combination of three components,

$$y_t = T_t + S_t + R_t, \quad (2.1)$$

where T_t denotes the long-term trend component, S_t the seasonal component associated with a specified period, and R_t denotes the residuals or remainder component. STL estimates these components using locally weighted regression (LOESS) providing a flexible and non-parametric way to capture smooth trend and seasonal variation. Its robustness mechanism also reduces the influence of outliers, which is useful in real-world load data where irregular events and measurement anomalies may occur [17].

Despite these advantages, classical STL is limited by the assumption of a single dominant seasonal cycle. This is particularly restrictive for hourly electrical load, where daily and weekly patterns commonly coexist. To address this issue, extensions such as Multiple Seasonal-Trend decomposition using LOESS (MSTL) generalise the STL framework by explicitly extracting several seasonal components associated with different periods [18]. In this case, the series is represented as

$$y_t = T_t + \sum_{p=1}^m S_t^{(p)} + R_t, \quad (2.2)$$

where each $S_t^{(p)}$ corresponds to a distinct seasonal cycle, and m denotes the number of seasonal components considered. The seasonal components are estimated sequentially, typically from shorter to longer periods, while preserving the non-parametric and robust characteristics of the original STL approach [17].

Figure 2.2 illustrates the principle of MSTL decomposition, where the observed load is separated into a long-term trend, multiple seasonal components, and a residual term. By isolating periodic structures associated with different temporal scales, MSTL avoids conflating distinct seasonal effects into a single component and produces a residual series that more accurately reflects short-term variability. This property is particularly relevant for applications that aim to analyse dependencies or dynamics beyond dominant periodic behaviour [17].

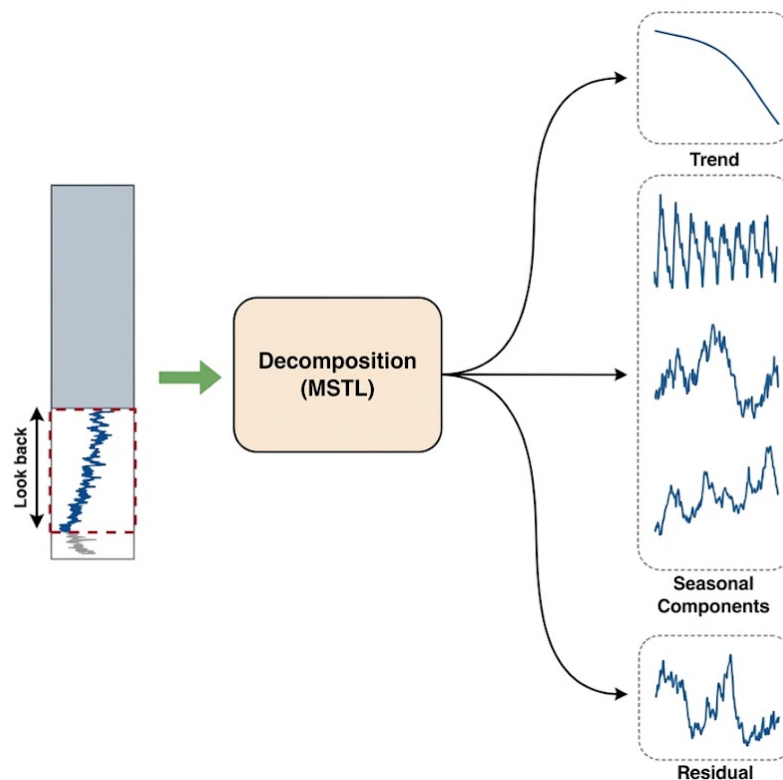


Figure 2.2: Illustration of multiple seasonal-trend decomposition using LOESS (MSTL). Source: [22].

Beyond STL-based methods, a range of alternative decomposition techniques has been proposed to handle the non-stationary and multi-scale nature of electrical load. Empirical Mode Decomposition (EMD) and its extensions decompose a signal into intrinsic mode functions that capture oscillatory behaviour at different characteristic frequencies without requiring predefined seasonal periods. Improved variants, such as ensemble-based formulations, aim to reduce mode mixing and enhance stability in the presence of noise [23]. These approaches are particularly suited to highly irregular signals but often sacrifice interpretability compared to seasonal-trend frameworks.

Complementary to data-adaptive decompositions, frequency-domain approaches explicitly model the spectral structure of electrical load signals. Wavelet-based methods perform a localised time-frequency analysis by decomposing the original series into components associated with different frequency bands, enabling the separation of smooth low-frequency behaviour from high-frequency fluctuations and transient events. Recent architectures combine wavelet transforms with deep learning models, allowing frequency-specific dynamics to be processed explicitly within forecasting frameworks. By isolating volatility patterns and abrupt changes that are difficult to capture in the time domain alone, wavelet-based representations can enhance modelling flexibility under complex temporal variability [29]. However, as with other frequency-driven methods, this gain in representational power typically comes at the cost of reduced interpretability when compared to

seasonal–trend decomposition approaches.

For the purposes of this dissertation, decomposition is especially relevant because the objective is not only to forecast load, but also to investigate spatial and spatio-temporal dependencies between locations. If dominant trend and seasonal components are left in the data, correlations between locations may be inflated by common temporal patterns. By contrast, analysing residual components after decomposition allows cross-location relationships to be studied on a signal where the main intra-series temporal structure has already been reduced. This motivates the use of multi-seasonal decomposition before the exploratory analysis of spatial dependencies.

2.1.3 Spatio-temporal Dynamics

In distribution-level settings, electrical load is typically observed as a collection of time series indexed by location, such as substations, transformers, feeders, or geographical zones. Although each individual series has its own temporal structure, operational datasets are inherently multi-variate: load trajectories across locations often exhibit coordinated behaviour rather than evolving independently. From a spatio-temporal perspective, electrical load is therefore treated as a coupled process in which temporal dependence coexists with spatial dependence, meaning that observations at one location may contain information about the behaviour of other locations at the same time or at nearby time steps [12, 21].

Spatial dependence is frequently linked to geographic proximity. Nearby areas tend to share exposure to weather, social routines, commuting patterns, and social and economic activity. In practice, this may translate into correlated demand patterns across nearby network nodes, particularly at regional and transformer aggregation levels. However, spatial dependence should not be interpreted only in terms of Euclidean distance. Similar load behaviour may also occur between geographically distant locations if they share similar consumption composition, land-use characteristics, or socio-economic routines. Conversely, adjacent areas can exhibit markedly different dynamics when their consumer mix is heterogeneous. These observations motivate modelling approaches in which spatial relationships are defined through an interaction topology that captures functional similarity and behavioural coupling, rather than physical distance alone [12].

A convenient abstraction is therefore to represent the spatial domain as a weighted graph. In this representation, nodes correspond to locations and edges encode dependence strength. The corresponding adjacency or weight matrix determines how information should be shared across nodes. This graph-based view is particularly suitable for distribution networks, where the relevant notion of neighbourhood can reflect physical proximity, network structure, or statistical similarity in observed load behaviour [14].

Figure 2.3 illustrates this spatio-temporal perspective at the regional or transformer level. The schematic layout in Figure 2.3(a) conceptualises a set of neighbouring regions, while Figure 2.3(b) shows that adjacent nodes may exhibit comparable load patterns, reflecting local coupling and shared driving conditions. Finally, Figure 2.3(c) connects these observations to a graph-based abstraction, in which the spatial interaction structure can be represented and allowed to evolve over

time. This type of representation supports the notion that spatial dependence is not merely descriptive, but a structural property that can be explicitly encoded before being exploited in forecasting formulations [14].

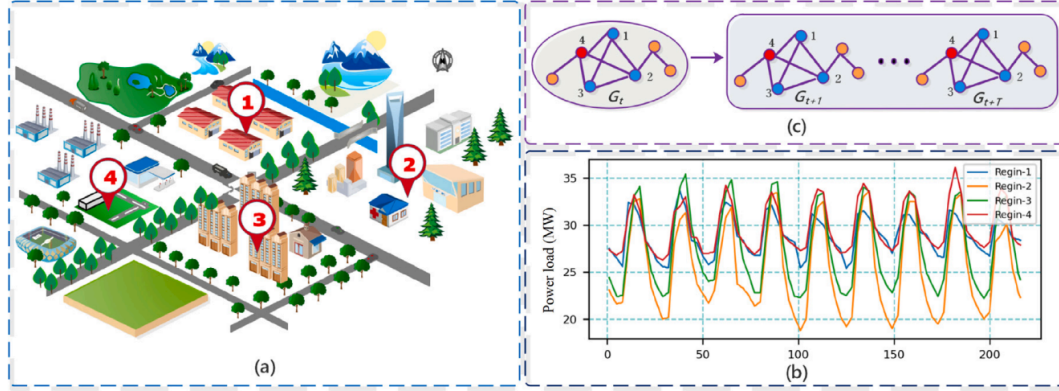


Figure 2.3: Regional-level spatio-temporal load patterns and graph abstraction. Source: [14]

Recent forecasting approaches increasingly combine spatial and temporal modelling. Graph-based neural networks encode the spatial domain through an adjacency structure, while recurrent, convolutional, or attention-based mechanisms model temporal evolution. In these approaches, the definition of the adjacency matrix is a central modelling choice: it may be fixed using distance-based criteria, enriched with data-driven similarity measures, or learned adaptively to capture changing interaction patterns over time [7, 14].

Before cross-location relationships can be interpreted, it is necessary to account for the dominant temporal structure within each individual series. Load data typically contain strong seasonal patterns, slow-varying components, and short-lag persistence. If these effects are left untreated, the similarity observed between different locations may be driven mainly by common temporal behaviour rather than by meaningful spatial or spatio-temporal relationships. For this reason, cross-series dependence is more informative when analysed after the main intra-series dynamics, such as trend and seasonality, have been removed or explicitly controlled. This motivates the use of decomposition-based residual analysis before studying cross-location relationships [3, 8, 19].

Autocorrelation analysis, using the ACF and PACF described in Section 2.1.1, provides the ability to measure how strongly current values relate to past values across different lags, and can further support the diagnosis of within-series temporal persistence before cross-location dependence is examined.

Once the dominant within-series structure has been controlled and the remaining temporal dependence has been diagnosed, cross-correlation analysis can be used to investigate dependencies between load series observed at different locations. The partial cross-correlation function (PCCF) indicates whether two series tend to co-move contemporaneously or exhibit lead-lag patterns, which is particularly relevant when changes in one area may precede similar changes elsewhere. In distribution-level studies, correlation measures, often based on Pearson-type coefficients, provide empirical evidence that spatial dependence is present and operationally meaningful [34].

Nevertheless, correlation-based results must be interpreted with caution. High correlation values may arise from shared external drivers, such as weather or calendar effects, rather than from direct interaction between locations. Consequently, cross-correlation is best understood as an exploratory tool for identifying candidate inter-location relationships, which can then be examined within a broader modelling framework that accounts for common influences, network structure, and the chosen spatial interaction topology [19].

Overall, the spatio-temporal dimension reframes electrical load analysis from a single-series problem into a coupled prediction setting, where meaningful structure may exist both across locations and across time. This motivates establishing empirical evidence of cross-location dependence and understanding its temporal behaviour before selecting predictive models or specifying interaction structures.

2.1.4 Aggregation, Hierarchy and Coherence

The spatial organisation of electrical load data within modern power systems naturally creates multi-level structures. These structures reflect both the physical topology of the distribution network and administrative geographical boundaries [21]. Understanding how load behaves across aggregation levels is important because the statistical properties of the signal change substantially with spatial scale [21].

A primary characteristic of multi-level electrical load data is the spatial smoothing effect, which is closely linked to the law of large numbers. At highly granular levels, such as individual secondary substations or local distribution zones, the electrical load is highly volatile and characterised by significant stochastic noise. This high variance reflects the uncoordinated and discrete behaviours of individual consumers, localized residential habits, or specific industrial shifts. As these individual profiles are aggregated into higher levels of the network, such as municipalities, districts and ultimately the national total, these localised stochastic fluctuations tend to cancel out. Consequently, higher aggregation levels are typically easier to forecast in relative terms, while lower aggregation levels preserve more local detail but are more difficult to model.

To formalise the relationship between series observed at different aggregation levels, the literature distinguishes between hierarchical time series (HTS) and grouped time series (GTS) [16]. A hierarchical time series has a strictly nested structure, where each lower-level unit belongs to exactly one parent at the level above. This produces a tree-like organisation. In an energy context, this may correspond to local consumption units aggregated into municipalities, municipalities into regions, and regions into a national total. Figure 2.4 illustrates this type of strictly nested hierarchy [16].

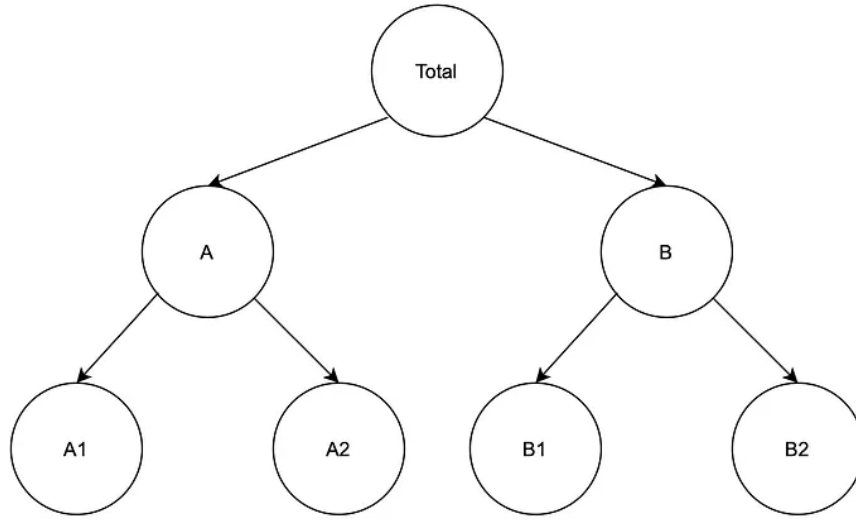


Figure 2.4: Example of a Hierarchical Time Series (HTS) structure with unique nesting levels. Source: [16].

In contrast, grouped time series (GTS) refer to structural arrangements where the aggregation paths are crossed rather than strictly nested, as shown in Figure 2.5 [16]. In a GTS framework, the electrical load data can be sliced and aggregated simultaneously across multiple overlapping attributes or features. For instance, the data can be grouped by physical geographic region and, concurrently, decomposed by consumer category (such as residential, commercial, or industrial) or tariff class [4, 20].

Regardless of whether the structure is strictly hierarchical or grouped, aggregation imposes linear constraints between the bottom-level series and the aggregate series. These constraints can be represented using a structural summing matrix $\mathbf{S}$. Let $\mathbf{b}_t$ denote the vector of bottom-level series at time t , and let $\mathbf{y}_t$ denote the vector containing all series across all aggregation levels. The

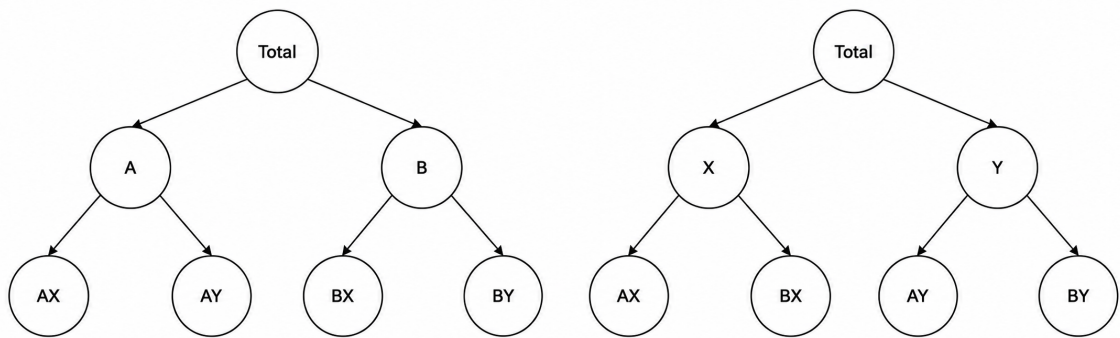


Figure 2.5: Example of a Grouped Time Series (GTS) structure with overlapping aggregation paths. Source: [16].

relationship between them can be written as

$$\mathbf{y}_t = \mathbf{S}\mathbf{b}_t. \quad (2.3)$$

This expression states that all aggregate series are obtained as linear combinations of the bottom-level series [19].

The existence of such constraints has important implications for forecasting. If models are trained independently at each level, their forecasts may become incoherent: the forecast for an aggregate series may not equal the sum of the forecasts of its components. Hierarchical forecasting methods address this issue through reconciliation, adjusting independently generated base forecasts so that they satisfy the aggregation constraints. However, hierarchy can also be used before reconciliation, by treating information from adjacent levels as potential explanatory input for forecasting models.

In this dissertation, both perspectives are explored. First, cross-level information is incorporated as explanatory input to higher-level models, using lower-level lagged observations and lower-level forecasts as features alongside temporal covariates, so that information flows upward from finer to coarser aggregation levels. Second, once independent base forecasts have been generated across all levels, a reconciliation step is applied in the opposite direction: higher-level forecasts are used to adjust the bottom-level series, exploiting the coherence constraints to correct lower-level predictions and propagate the improvement upward through the hierarchy.

2.2 Electrical Load Forecasting

Accurate electrical load forecasting is essential for both operational and planning decisions in power systems, as deviations between predicted and actual demand may lead to significant technical and economic consequences. Overestimating electrical load can result in inefficient resource utilisation and unnecessary operational costs, while underestimating demand may compromise system reliability and increase the risk of service disruptions. The literature consistently highlights the economic sensitivity of power system operation to forecasting accuracy. For instance, [2] reports that a 1% improvement in the mean absolute percentage error (MAPE) for a 10 000 MW plant could save up to \$1.6 million per year.

The forecasting problem is not uniform across power-system contexts. It depends strongly on the prediction horizon, the spatial aggregation level, and the operational decision being supported. At highly aggregated levels, such as national or transmission-level demand, individual consumption fluctuations tend to cancel out, producing smoother and more predictable profiles. At lower aggregation levels, particularly in distribution networks, demand becomes more volatile and heterogeneous, since local consumption behaviour, individual decisions, and area-specific conditions become more visible in the load signal. Therefore, forecasting methods that perform well at system level may not transfer directly to finer spatial resolutions.

In the context of this dissertation, the focus is on short-term forecasting of distribution-level electrical load. This setting is especially relevant because distribution-level forecasts support local operational decisions, such as network monitoring, flexibility management, congestion prevention, asset operation, and demand response. At the same time, the availability of geographically detailed consumption data creates the possibility of exploiting spatial and hierarchical relationships between load series. This section reviews the main forecasting concepts and methods relevant to this problem. Section 2.2.1 discusses forecasting horizons and the operational role of short-term load forecasting. Section 2.2.2 reviews statistical forecasting methods. Section 2.2.3 discusses machine-learning and deep-learning approaches. Section 2.2.4 introduces probabilistic forecasting. Finally, Section 2.2.5 reviews hierarchical and geographically aggregated forecasting.

2.2.1 Forecasting Horizons and Operational Context

Electrical load forecasting problems are commonly classified according to the forecasting horizon, which defines both the temporal scope of the prediction and the type of decisions it supports. Although definitions vary slightly across the literature, four main forecasting horizons are typically distinguished:

- **Very short-term load forecasting** spans from minutes to up to one hour ahead and is mainly associated with real-time system control and rapid operational adjustments.
- **Short-term load forecasting** generally covers horizons ranging from several hours to a few days ahead and plays a central role in day-to-day system operation, supporting decisions related to scheduling, dispatch, and operational security.
- **Medium-term load forecasting** extends from weeks to months and is primarily used for planning-oriented activities such as maintenance scheduling, resource allocation, and financial planning.
- **Long-term load forecasting** spans from several years to decades and supports strategic decisions related to capacity expansion, infrastructure development, and long-term system planning.

This horizon-based classification highlights the close relationship between the forecasting horizon and the operational or planning context in which forecasts are used. As illustrated in Figure 2.6, increasing horizons are associated with a gradual shift from real-time and operational decision-making towards medium- and long-term planning activities, whereas shorter horizons are more directly linked to operational control and short-term management tasks.

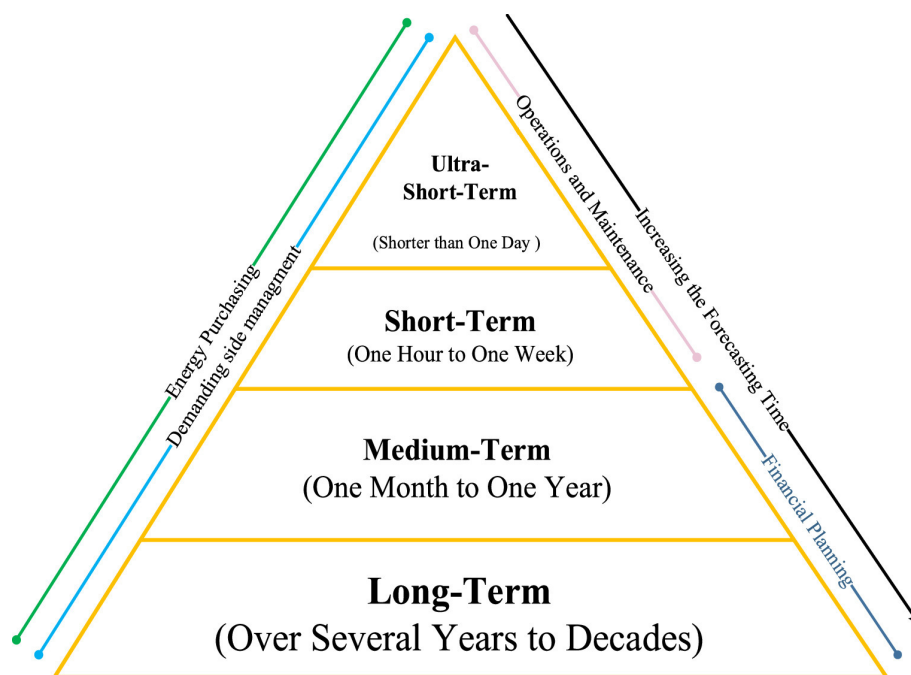


Figure 2.6: Classification of electrical load forecasting horizons. Source: [32]

Short-term load forecasting (STLF) is the most relevant horizon for the present dissertation. At this horizon, forecasts must capture short-lag persistence, daily and weekly recurrence, weather sensitivity, and calendar effects. These requirements become more demanding at distribution level, where reduced aggregation exposes local fluctuations that are often smoothed out in national or transmission-level demand. Consequently, STLF at finer spatial resolutions must balance two modelling objectives: exploiting stable temporal regularities while remaining flexible enough to represent local variability.

Beyond core operational scheduling, STLF also underpins a broad set of applications across different power system contexts. As illustrated in Figure 2.7, STLF connects system-level operational requirements with more specific forecasting and control tasks, including residential and building load forecasting, price forecasting, and demand response management.

The aggregation level also influences the relative difficulty of STLF. At highly aggregated levels, collective behaviour dominates and forecasting errors are generally lower in relative terms. At lower aggregation levels, local consumption patterns, behavioural heterogeneity, and sensitivity to local conditions increase the stochastic component of the signal. This motivates forecasting methodologies that can use temporal information effectively while also testing whether spatial or hierarchical information provides additional predictive value.

The previous discussion defines the operational context of the forecasting problem, but the choice of forecasting method depends on how the relevant information sources are represented. At short-term horizons, models usually rely on combinations of lagged load values, calendar variables, weather information, and, in multivariate settings, information from related series. Different forecasting families encode these inputs in different ways. Statistical models represent temporal

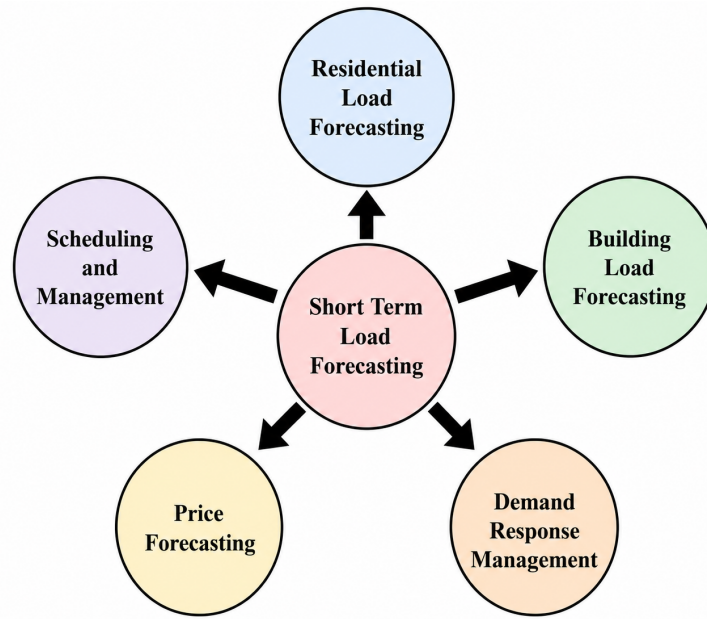


Figure 2.7: Applications in short-term load forecasting. Source: [6]

and cross-series dependence through explicit stochastic structures; machine-learning models use engineered features to capture non-linear relationships; deep-learning models learn temporal or spatio-temporal representations more directly from data; and probabilistic models extend the objective from predicting a single value to characterising predictive uncertainty.

The following subsections review these methodological families with emphasis on their relevance for distribution-level short-term load forecasting. The objective is not to provide an exhaustive catalogue of forecasting algorithms, but to identify the modelling principles that are most relevant to the comparison developed in this dissertation: purely temporal baselines, models enriched with exogenous variables, and approaches capable of exploiting spatial or hierarchical information.

2.2.2 Statistical Methods

Statistical approaches characterise load through an explicit stochastic structure, typically relying on a limited number of parameters estimated from historical data. Compared with machine-learning models, they impose stronger inductive assumptions, which improves transparency and stability when seasonality and calendar effects are dominant, often at lower computational cost. Their main limitation lies in reduced flexibility when load dynamics become highly non-linear or regime-dependent [5, 20].

The simplest forecasting approach is the persistence model, which assumes that future load equals the most recently observed value. Seasonal variants extend this idea by replacing the last observation with the value recorded at the same time in a previous cycle, such as the same hour of

the previous day or the same hour of the previous week. Despite their simplicity, persistence models serve as essential baselines: a forecasting method is only considered useful if it consistently outperforms them [20].

Classical univariate specifications, such as AutoRegressive Integrated Moving Average (ARIMA) models and exponential smoothing formulations, describe demand through autoregressive persistence and structured trend–seasonal components. Multi-seasonal extensions are particularly relevant for short-term horizons, where daily and weekly cycles coexist and must be represented jointly rather than sequentially. These models provide strong baselines and remain competitive when the temporal structure is stable and well defined [21, 28].

Regression-based statistical models extend this framework by incorporating exogenous variables, notably weather and calendar indicators. In contrast to purely data-driven ML models, they retain interpretability by separating explainable variation from serial dependence, which is advantageous for scenario analysis and operational planning, albeit at the cost of limited non-linear expressiveness [20].

In multivariate contexts, vector autoregressive models (VAR) allow load at each location to depend on its own past and on lagged observations from other locations, thereby embedding temporal and spatial dependence within a single statistical structure. However, the rapid growth in parameter dimensionality restricts their direct applicability at finer spatial resolutions [12].

This scalability issue is commonly addressed through regularisation. In LASSO-regularised VAR settings, coefficients are estimated by solving

$$\hat{\mathbf{B}} = \arg \min_{\mathbf{B}} \left(\frac{1}{2} \|\mathbf{Y} - \mathbf{ZB}\|_2^2 + \lambda \|\mathbf{B}\|_1 \right), \quad (2.4)$$

where $\mathbf{Y}$ denotes the matrix of observed series, $\mathbf{Z}$ contains lagged observations from all locations, $\mathbf{B}$ is the coefficient matrix, and $\lambda > 0$ is a scalar penalty parameter controlling sparsity. The first term measures the model fitting error, while the second introduces the LASSO penalty that shrinks small coefficients toward zero. Relative to both unregularised VARs and black-box ML models, this approach offers a compromise between interpretability and flexibility, retaining only the most relevant temporal and cross-location dependencies while remaining computationally tractable in high-dimensional settings [19, 35]. When the predictor set is extended with variables that the system does not itself model, such as calendar indicators, weather, or quantities originating from another aggregation level, the formulation becomes a vector autoregression with exogenous inputs, denoted VAR-X, and its regularised form LASSO-VAR-X.

These properties motivate the use of statistical models as one of the core forecasting frameworks in this dissertation. Due to their simple structure, linear statistical models tend to perform well at short forecasting horizons, and their computational efficiency and interpretability make them particularly well suited to a setting where understanding the spatio-temporal dynamics of electrical load is as important as predictive accuracy.

2.2.3 Machine Learning and Deep Learning

The landscape of electrical load forecasting has increasingly shifted towards data-driven paradigms, where machine learning (ML) and deep learning (DL) models are utilised to approximate complex, non-linear functions without the rigid structural assumptions of classical statistics [32]. These approaches are particularly effective in capturing high-dimensional feature interactions and accommodating the stochasticity inherent in distribution-level data [5, 6].

Among ML techniques, tree-based ensemble methods, specifically those based on Gradient Boosting Decision Trees (GBDT), have emerged as a dominant choice for short-term electrical consumption forecasting. Their popularity is largely attributed to their robustness against outliers, ability to handle missing values, and high predictive accuracy in competitive settings. In GBDT, the model is built as an additive expansion of weak learners, typically decision trees, where each subsequent tree is trained to minimise the residual loss of the ensemble [25]. Formally, the ensemble prediction for observation i at iteration t is expressed as:

$$\hat{y}_i^{(t)} = \sum_{k=1}^t f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i), \quad (2.5)$$

where $f_t(x_i)$ represents the new tree added at step t . To determine the optimal f_t , the model is trained by minimising an objective function that balances the loss over n observations with a regularisation term Ω to prevent overfitting:

$$\mathcal{L}^{(t)} = \sum_{i=1}^n \ell(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t), \quad (2.6)$$

where $\ell(\cdot)$ measures the prediction error.

Frameworks such as XGBoost, LightGBM, and CatBoost have become benchmarks in the field, consistently outperforming other algorithms in energy forecasting competitions such as the Global Energy Forecasting Competition (GEFCom) and more recent challenges like the GECAD 2025 Competition [27, 32].

Beyond ensemble trees, the literature explores a diverse range of ML and DL architectures. Shallow models, including Support Vector Regression (SVR) and Artificial Neural Networks (ANN), laid the foundation for more complex sequential learners [25, 33]. DL has further advanced the state-of-the-art through architectures capable of automated feature extraction from temporal sequences. Recurrent Neural Networks (RNN), particularly Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU), are widely employed to capture long-term temporal dependencies [6, 23]. Convolutional Neural Networks (CNN) and Temporal Convolutional Networks (TCN) offer a computationally efficient alternative, capturing local temporal patterns through hierarchical convolution operations [23]. More recently, attention-based models such as Transformers have been adapted for time series, with variants like the Informer and MFformer providing superior performance in multi-scale feature aggregation and global dependency modelling [29]. Graph-based operators are increasingly integrated into these architectures to explicitly

account for structured spatial dependencies in multivariate load datasets [14]. While DL models offer significant representational power, hybrid frameworks, combining signal decomposition with ML/DL predictors, often yield the most robust results in volatile distribution-level scenarios [17, 23].

Despite these advances, deep learning models present important limitations. Their capacity to model complex non-linearities comes at the cost of interpretability: the internal representations learned by LSTM, Transformer, or graph-based architectures are difficult to inspect and rarely provide actionable insight into the drivers of load behaviour. Training can also be unstable, requiring careful hyperparameter tuning, large amounts of data, and regularisation strategies to avoid overfitting, particularly at distribution-level granularities where data may be scarce or highly heterogeneous. Furthermore, the computational overhead of deep architectures limits their practicality in operational settings with strict latency constraints. These trade-offs motivate the continued relevance of statistical and simpler machine-learning models, especially at short forecasting horizons where the added complexity of deep learning does not consistently translate into accuracy gains [5, 20].

2.2.4 Probabilistic Forecasting

The methodological evolution presented throughout this chapter reflects a transition from rigid statistical frameworks towards increasingly flexible data-driven models. While some of the techniques previously discussed possess the architectural capacity to be adapted for uncertainty quantification, many implementations, and particularly classical structures such as Vector Autoregressive (VAR) models, remain fundamentally deterministic in nature, providing a single expected value rather than a full distribution [19, 25].

In the context of distribution-level electrical load, relying solely on these point estimates becomes increasingly problematic due to the high volatility and stochastic noise characteristic of lower aggregation levels [21, 31]. Probabilistic forecasting addresses this limitation by providing a complete description of potential future states, which is essential for risk management and operational decision-making in modern power systems [32]. This uncertainty is typically represented through prediction intervals (PIs), specific quantiles, or full probability density functions, allowing operators to move beyond a single “best guess” and account for the full range of possible outcomes [2, 13].

The development of probabilistic frameworks follows three main methodological trajectories [31]:

- **Scenario-generation methods** use multiple realisations of exogenous variables, such as weather forecasts, to generate an ensemble of electrical load trajectories through deterministic models [2, 13].
- **Natively probabilistic architectures** are designed to output distributional parameters or multiple quantiles directly, often by employing specific loss functions within the tree-based

ensembles or deep learning networks discussed in the previous section. Examples include quantile regression forests, models with Gaussian or Beta distributional outputs, and Bayesian neural networks [24, 32].

- **Residual-based and ensemble methods** derive uncertainty from the empirical distribution of a model's past errors or by aggregating forecasts from diverse heterogeneous models to capture a broader spectrum of uncertainty. Specific approaches include bootstrap resampling, conformal prediction, and copula-based methods, which model the joint dependence structure of forecast errors across locations or horizons [20, 31].

The evaluation of probabilistic forecasts requires metrics that assess both calibration and sharpness. A forecast distribution should be calibrated, meaning that observed values should fall within predicted intervals with the expected frequency, but it should also be sharp, meaning that uncertainty intervals should not be unnecessarily wide. This creates a more demanding evaluation problem than point forecasting, since the quality of the full predictive distribution must be assessed rather than only the central forecast [4].

Probabilistic forecasting is also challenging in hierarchical settings. When forecasts are produced across several spatial or temporal aggregation levels, uncertainty estimates should ideally remain coherent across the hierarchy. For example, predictive quantiles or scenarios at lower levels should be consistent with the corresponding aggregate-level uncertainty. Maintaining this probabilistic coherence is more complex than reconciling point forecasts and remains an active research topic [4, 19].

In this dissertation, probabilistic forecasting is addressed through quantile regression applied within a gradient boosting framework, extending the deterministic point forecasts produced by the LASSO-VAR-X models. This approach falls within the natively probabilistic category, generating prediction intervals directly without requiring distributional assumptions. The hierarchical structure of the data introduces an additional layer of complexity, as uncertainty estimates must remain meaningful across aggregation levels. This motivates the evaluation of probabilistic coherence alongside point forecast accuracy throughout the experimental analysis.

2.2.5 Hierarchical and Geographically Aggregated Forecasting

As discussed in Section 2.1.4, electrical load data often have a multi-level structure. Consumption may be observed or aggregated across different geographical, administrative, or network levels, from local units to regional and national totals. This structure creates both an opportunity and a challenge for forecasting. On one hand, lower levels preserve local information, while higher levels provide smoother and more stable signals. On the other hand, forecasts produced independently at different levels may be incoherent, meaning that aggregate forecasts may not equal the sum of their component forecasts.

Let $\hat{\mathbf{y}}_t$ denote a vector of independently generated base forecasts for all levels of a hierarchy, and let $\hat{\mathbf{b}}_t$ denote the corresponding bottom-level forecasts. If these forecasts are produced

independently, they will generally fail to satisfy the aggregation constraints defined by the summing matrix $\mathbf{S}$, so that $\hat{\mathbf{y}}_t \neq \mathbf{S}\hat{\mathbf{b}}_t$ [16, 19]. Hierarchical forecasting methods address this issue by producing or adjusting forecasts so that they remain coherent across aggregation levels.

Traditional approaches achieve coherence by restricting the forecasting task to a specific layer of the hierarchy. The Bottom-Up (BU) approach generates base forecasts exclusively for the most disaggregated series and then aggregates them upwards using the summing matrix, guaranteeing coherence [20]. While the BU method effectively captures localised dynamics, it is highly susceptible to the accumulation of stochastic noise and fails to exploit the clearer signal properties found at higher aggregation levels [21]. Conversely, Top-Down (TD) approaches model the total aggregate system directly and disaggregate the resulting forecast to lower levels using allocation proportions, such as historical relative contributions [16]. Although the TD method benefits from the structural stability of aggregated electrical load, it assumes homogeneous consumption trends across all branches, thereby misrepresenting localised fluctuations and sub-aggregate variations [20]. The Middle-Out (MO) approach acts as a hybrid alternative for strictly nested hierarchies, selecting an intermediate level to forecast directly, then applying bottom-up aggregation for higher levels and top-down disaggregation for lower ones [16].

To overcome the biases and information loss of these classical unilateral methods, the state of the art leverages the optimal forecast reconciliation framework. This approach does not alter the underlying forecasting algorithms but functions as a structural post-processing step that adjusts the independent base forecasts $\hat{\mathbf{y}}_t$ into a coherent set of reconciled forecasts $\tilde{\mathbf{y}}_t$ through a linear mapping matrix $\mathbf{P}$ [16, 19]:

$$\tilde{\mathbf{y}}_t = \mathbf{S}\mathbf{P}\hat{\mathbf{y}}_t. \quad (2.7)$$

The role of the mapping matrix $\mathbf{P}$ is to extract and recombine the information from all base forecasts to minimise the reconciliation error variance. Under the assumption of unbiased base forecasts, the optimal reconciliation matrix that minimises the trace of the reconciled error covariance matrix, known as the Minimum Trace (MinT) framework, is formulated by defining the weighting matrix based on the covariance structure of the base forecast errors $\mathbf{W}_t$ [4, 20]. Variants such as Ordinary Least Squares (OLS), Weighted Least Squares (WLS), and structural scaling provide different approximations of $\mathbf{W}_t$ to balance computational complexity and estimation accuracy in large-scale networks [20]. Recent advancements have further extended this linear framework to probabilistic settings, ensuring that entire predictive quantiles remain coherent, and to cross-temporal frameworks that simultaneously reconcile spatial boundaries with multiple temporal aggregation scales [19, 20].

A concrete illustration of these challenges appears in the context of wind power forecasting. Sharma et al. [26] consider a hierarchy of geographically distributed wind farms aggregated into intermediate clusters and a system-level total, and show that independently generated base forecasts at each level are almost always incoherent. To address this, the authors apply MinT(Shrink), an estimator that incorporates spatial correlations among farms while shrinking the off-diagonal elements of the covariance matrix toward zero, guaranteeing invertibility regardless of network

size. Applied to 37 wind farms across multiple forecasting horizons, MinT(Shrink) consistently achieves lower forecast errors at all hierarchical levels compared to OLS and WLS.

In this dissertation, hierarchy is relevant not only as a post-processing reconciliation problem, but also as a source of predictive information. Forecasts at one level may benefit from explanatory variables derived from adjacent levels. Lower-level series can provide detailed local information for aggregate forecasting, while upper-level series can provide broader contextual signals for disaggregated forecasting. Therefore, hierarchical forecasting is connected to the central question of this work, whether cross-level information improves short-term load forecasting beyond purely temporal baselines.

2.3 Evaluation Metrics

The comprehensive assessment of electrical load forecasting models requires a multi-dimensional evaluation framework that accounts for deterministic accuracy, the reliability of uncertainty estimation, and the operational computational burden [20]. Depending on the specific objectives of the forecasting task, the literature defines several categories of metrics to benchmark model performance.

2.3.1 Point Forecasting Metrics

Deterministic or point forecasting metrics evaluate the error between a single expected value provided by the model and the actual observation. These classical metrics remain the standard for assessing the central tendency of the predictions [31].

Let y_t denote the observed electrical load at time t , $\hat{y}_t$ the corresponding point forecast, and n the number of forecasted observations. One of the most common metrics is the Root Mean Square Error (RMSE), defined as

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{t=1}^n (y_t - \hat{y}_t)^2}. \quad (2.8)$$

RMSE penalises large errors more strongly because the deviations are squared before averaging. This makes it useful when large forecasting errors are particularly undesirable, but also makes it sensitive to outliers.

The Mean Absolute Error (MAE) is defined as

$$\text{MAE} = \frac{1}{n} \sum_{t=1}^n |y_t - \hat{y}_t|. \quad (2.9)$$

Unlike RMSE, MAE treats errors linearly and is therefore less sensitive to extreme deviations. It is often easier to interpret because it is expressed in the same unit as the load variable, such as kWh or MWh.

Finally, the Mean Absolute Percentage Error (MAPE) is defined as

$$\text{MAPE} = \frac{1}{n} \sum_{t=1}^n \left| \frac{y_t - \hat{y}_t}{y_t} \right| \times 100. \quad (2.10)$$

MAPE expresses the error as a percentage of the observed value, which facilitates comparisons across series with different scales or magnitudes. However, it becomes numerically unstable when y_t is close to zero and tends to penalise over-forecasting more heavily than under-forecasting [20].

The Mean Absolute Scaled Error (MASE) expresses the forecasting error relative to a naïve reference forecast, making it scale-independent and directly comparable across series of very different magnitudes. It is defined as the MAE of the model divided by the MAE of a naïve benchmark computed on the training data:

$$\text{MASE} = \frac{\frac{1}{n} \sum_{t=1}^n |y_t - \hat{y}_t|}{\frac{1}{N-m} \sum_{t=m+1}^N |y_t - y_{t-m}|}, \quad (2.11)$$

where the numerator is the MAE of the model over the n forecasted observations, and the denominator is the MAE of a seasonal naïve forecast over the N training observations. This benchmark predicts each value using the observation from m steps earlier, where m is the seasonal period (for example, $m = 24$ for the daily cycle in hourly data, while $m = 1$ recovers the simple persistence forecast). A value below one indicates that the model outperforms the naïve benchmark, whereas a value above one indicates worse performance. This property makes MASE particularly informative in this work, where the central goal is to assess whether the proposed models improve upon simple temporal baselines [20].

In the context of distribution-level load forecasting, no single point metric fully characterises model performance. RMSE highlights large errors, MAE provides a robust average error magnitude, MAPE enables relative comparisons across scales, and MASE positions performance against a naïve benchmark. For this reason, forecasting studies often report multiple metrics to provide a more complete assessment of model behaviour.

2.3.2 Probabilistic Forecasting Metrics

As power systems increasingly require uncertainty quantification, proper scoring rules are necessary to evaluate the predictive densities and intervals, ensuring they are both well-calibrated and sharp [13].

For quantile forecasts, the standard evaluation criterion is the pinball loss, also known as the quantile score. For a quantile level $\tau \in (0, 1)$ and a predicted quantile $\hat{q}_t(\tau)$, the pinball loss is defined as

$$L^\tau(y_t, \hat{q}_t(\tau)) = \begin{cases} \tau(y_t - \hat{q}_t(\tau)), & \text{if } y_t \geq \hat{q}_t(\tau), \\ (1 - \tau)(\hat{q}_t(\tau) - y_t), & \text{if } y_t < \hat{q}_t(\tau). \end{cases} \quad (2.12)$$

This loss applies asymmetric penalties depending on whether the observation lies above or below the predicted quantile. It is therefore appropriate for evaluating whether a model captures specific percentiles of the load distribution and is widely used in probabilistic load forecasting and forecasting competitions [24].

When a full predictive distribution is available, the Continuous Ranked Probability Score (CRPS) is commonly used. CRPS compares the predicted cumulative distribution function with the observed outcome and can be interpreted as a probabilistic generalisation of the MAE. Lower CRPS values indicate better probabilistic forecasts, combining information about both calibration and sharpness [1].

2.4 Research Gaps and Critical Discussion

The literature reviewed in the preceding sections establishes two facts that this dissertation takes as its point of departure. Electrical load observed at fine spatial resolution is demonstrably dependent across locations, and distribution-level data is almost always organised into a hierarchy of aggregation levels. What the literature leaves considerably less settled is how much either property is worth once a competent temporal model is already in place, and it is in that gap that the present work is positioned.

A first limitation concerns how spatial dependence is established. Studies that document it overwhelmingly do so descriptively, through correlation matrices, graph adjacencies or spatial weight structures estimated on the observed series, and the dependence so measured is then assumed to be predictive. Two difficulties follow. Because load series across a country share the same calendar and broadly the same weather, correlations computed on raw observations are inflated by common temporal structure, so the dependence attributed to space may be little more than the daily and weekly cycles reappearing at every location [8, 19]. Furthermore, a dependence that is real is not necessarily useful, since a neighbouring series can be strongly correlated with the target and still carry no information that the target's own recent past does not already provide. Establishing the second claim requires the spatial model to be compared against a temporal baseline of comparable quality, which is done less consistently than the descriptive evidence would suggest [12, 21].

A second limitation concerns the resolution at which the question is answered. Comparisons between spatial and temporal models are typically reported at a single aggregation level, most often the one at which the data happen to be available, and the conclusion is then stated as a property of the method rather than of the level. Since aggregation changes the variance, the smoothness and the number of series available simultaneously, there is no reason to expect the balance between spatial and temporal information to be the same at a postal-code area and at a regional total. A comparison carried out at every level of the same hierarchy, with the model and the validation protocol held fixed, is therefore necessary before the usefulness of spatial information can be attributed to anything other than the particular scale examined.

A third limitation concerns the role assigned to the hierarchy itself. In the hierarchical forecasting literature the aggregation structure is treated almost exclusively as a coherence constraint, with the forecasting models remaining independent and the hierarchy entering only in a post-processing projection [16, 20]. The possibility that an adjacent level supplies explanatory information before any reconciliation takes place is comparatively unexplored in load forecasting, although it has been demonstrated in related domains, notably by Gilbert et al. [11], where turbine-level information improves wind farm forecasts. Whether the same mechanism operates over a postal-code hierarchy, and in which form the lower-level information should be supplied, remains open.

A fourth limitation concerns how reconciliation itself is evaluated. Comparative studies generally report a single preferred estimator, frequently the minimum trace with shrinkage, on the basis of one set of base forecasts [26]. Since the weighting matrix is estimated from the residuals of those base forecasts, the ranking between estimators is not independent of what produced them, and a base that already exploits the spatial structure of the load leaves a different residual covariance from one that ignores it. Reporting the comparison against a single base therefore risks presenting as a property of the method what may be a property of the input.

A fifth limitation concerns the extension of these questions to probabilistic forecasting. Coherent probabilistic forecasting across a hierarchy is recognised as difficult and remains an active topic [4, 19], but a more basic question is also unresolved, namely whether the cross-location information that helps a point forecast also improves the predictive distribution. Because the tails of that distribution are determined by infrequent observations, they are considerably more sensitive than a conditional mean to how much history each model sees and to how often it is re-estimated, and the training regime is seldom treated as a design variable on the same footing as the feature set.

Taken together, these limitations define the scope of this dissertation. The spatial dependence of the data is established on decomposed residuals rather than on raw series, so that the calendar structure cannot account for it. The comparison between temporal, spatial and cross-level information is then carried out at every level of one hierarchy, under a single estimator and a single leakage-free validation protocol, so that any difference is attributable to the information supplied rather than to the model or to the scale. The hierarchy is used in both of its capacities, as a source of explanatory features before forecasting and as a coherence constraint afterwards, and the reconciliation estimators are compared over several distinct sets of base forecasts rather than one. Finally, the same hierarchy is evaluated under probabilistic forecasting, where the training regime is varied alongside the feature configuration so that the two effects can be separated.

Chapter 3

Data Description and Exploratory Analysis

This chapter presents the dataset used in the dissertation and analyses its main temporal, spatial, and hierarchical properties. The objective is twofold. First, the chapter describes how the raw electrical load data are cleaned, temporally normalised, and organised into a coherent spatial hierarchy. Second, it provides exploratory evidence of the temporal and cross-location structures that motivate the forecasting methodology developed in Chapter 4.

The analysis focuses on hourly electrical load data from mainland Portugal, originally available at the four-digit postal-code level. Since the purpose of the dissertation is to assess whether spatial and hierarchical information can improve short-term load forecasting, the chapter places particular emphasis on three aspects of the data: the presence of strong temporal structure, the existence of residual dependencies between postal-code areas after removing dominant trend and seasonal components, and the construction of aggregation levels that allow information to be compared across spatial scales.

Therefore, this chapter is organised as follows. Section 3.1 describes the data source and the scope of the dataset, detailing the E-REDES open data portal and the geographical and temporal coverage of the available load measurements. Section 3.2 covers the preprocessing pipeline, including data cleaning, correction of temporal inconsistencies, and the decomposition of the load signal into trend, seasonal, and residual components. Section 3.3 presents the spatio-temporal analysis, examining the spatial correlation structure, the cross-location dependencies that remain after the dominant temporal components are removed, and the clustering of postal-code areas based on load behaviour. Section 3.4 describes the construction of the hierarchical dataset, from four-digit postal-code series up to the national aggregate. Section 3.5 discusses the main data challenges identified throughout this process and their implications for modelling. Finally, Section 3.6 summarises the chapter and relates the structure identified in the data to the forecasting methodology that follows.

3.1 Data Source and Scope

The empirical foundation of this dissertation is real electrical load data provided by E-REDES, the primary Distribution System Operator (DSO) in mainland Portugal. The data are obtained through the E-REDES Open Data Portal [9], specifically from the hourly consumption dataset aggregated by four-digit postal code [10]. The analysed period spans from November 2022 to September 2023, resulting in eleven months of hourly observations. This period captures relevant seasonal variations and intraday consumption dynamics across the network.

The dataset is particularly suitable for the objectives of this dissertation because it provides geographically disaggregated consumption information. Instead of considering only national or system-level demand, the data allow electrical load to be analysed across multiple local areas. This is essential for studying whether spatial and hierarchical relationships contain useful information for short-term forecasting, as discussed in Section 2.1.3. At the most disaggregated level available in the dataset, electrical load is reported for 465 distinct units at the four-digit postal-code level. Within this structure, 464 units correspond to precise geographical regions across mainland Portugal. The remaining unit is a distinct category designated as “OTHERS”, which aggregates electricity consumption from unidentified or non-georeferenced origins. Since this residual category does not correspond to any specific geographical region, it is excluded from the analysis throughout this dissertation.

Furthermore, the dataset exhibits profound heterogeneity in the magnitude of electrical load across the different spatial units. Average hourly consumption varies by multiple orders of magnitude, ranging from approximately 14 kWh (e.g., postal code 4099) to approximately 89 762 kWh in the most concentrated area, postal code 2840. This immense spatial variance is primarily driven by underlying demographic and socioeconomic factors, notably varying population densities and the predominant type of consumption, which can range from purely residential neighbourhoods to heavy industrial zones.

According to the methodological documentation provided by E-REDES, the dataset is primarily constructed using load diagrams collected through advanced metering infrastructure (smart meters). When smart meters data are incomplete or unavailable, E-REDES applies standardised consumption profiling methodologies to estimate the electrical load. While this procedure ensures the continuity of the time series across all spatial units, it introduces an inherent degree of estimation uncertainty that should be considered when interpreting distribution-level load patterns.

Table 3.1 summarises the core characteristics of the dataset used in this study. This high-resolution data forms the baseline for the hierarchical construction detailed in Section 3.4.

Table 3.1: Summary of the E-REDES electrical load dataset characteristics.

Feature	Description
Platform	E-REDES Open Data Portal
Specific Dataset	Hourly Consumption by Postal Code - 4 Digits
Spatial Coverage	Mainland Portugal
Spatial Resolution	465 four-digit postal-code units (464 postal-code regions and one “OTHERS” category)
Temporal Resolution	Hourly
Time Period	November 2022 to September 2023 (8 016 observations)

3.2 Data Preprocessing

Before constructing the spatial hierarchy and analysing cross-location relationships, the raw electrical load data were preprocessed to ensure temporal consistency and to separate the main temporal components of the signal. This step is essential because hourly load forecasting relies on a regular time index, and spatio-temporal analysis can be distorted if common seasonal or trend components dominate the original series.

The preprocessing procedure had two main objectives. First, the hourly time index was inspected and corrected, with particular attention to anomalies associated with the daylight-saving-time transition. Second, the cleaned series were decomposed into trend, seasonal, and residual components, allowing the subsequent exploratory analysis to focus on residual dynamics after the dominant temporal structure had been removed.

3.2.1 Data Cleaning and Correction

The electrical load data extracted from E-REDES were first analysed in their original hourly structure, indexed by timestamp and postal code. Before any additional transformations, it was necessary to verify whether the time series were complete, temporally consistent, and free from obvious anomalies caused by the timestamp conversion process.

This initial inspection was performed at the individual postal-code level. Figure 3.1 shows an example for postal code 4450, covering the period between November 2022 and September 2023. The series presents the expected hourly variability of electricity consumption, with recurrent intra-day fluctuations and visible changes in the overall consumption level throughout the year. In this example, load increases during the first months of the period, remains relatively high during winter, decreases after February and March, and then stabilises during the following months.

However, the same figure also reveals an abnormal isolated drop near the end of March. This drop is inconsistent with the surrounding consumption pattern and occurs around the daylight saving time transition, more specifically on 26 March 2023. This suggested that the anomaly

was not related to a real change in electricity consumption, but rather to the way the timestamp sequence was represented in the original data.

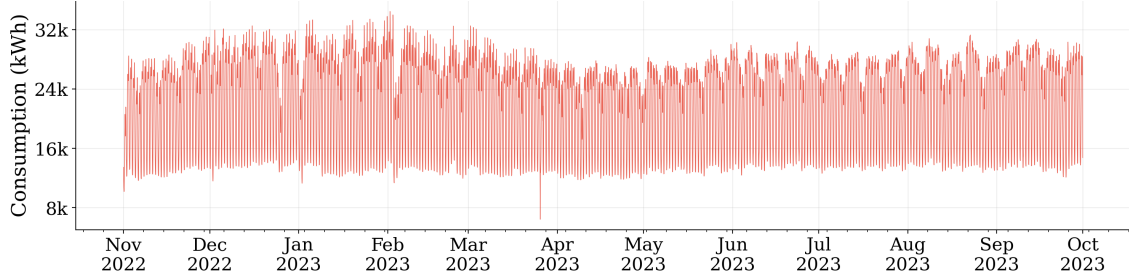


Figure 3.1: Original hourly electrical load series for the four-digit postal code 4450.

A closer view of this period is presented in Figure 3.2. In the original E-REDES data, the timestamp corresponding to 01:00 on 26 March 2023 is absent due to the daylight saving time change. In addition, the neighbouring observations around 00:00 and 02:00 are strongly distorted, producing an artificial drop in the consumption profile. These values do not follow the behaviour observed in the previous and following days and are therefore not representative of the actual load pattern.

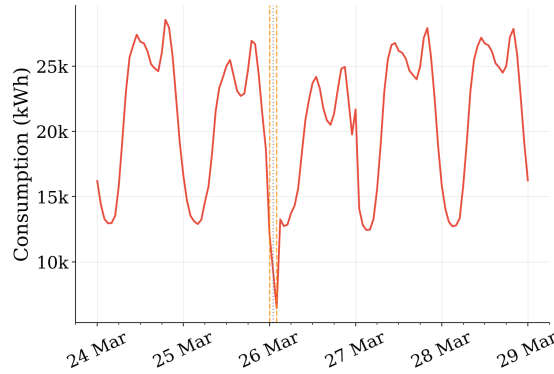


Figure 3.2: Original 4450 postal-code electrical load profile around the daylight saving time transition.

To correct this inconsistency, the missing timestamp corresponding to 01:00 on 26 March 2023 was explicitly inserted into the hourly index. Since no real observation was available for that timestamp, its consumption value was assigned as NaN. This preserves the regular hourly structure of the time index while explicitly representing the absence of an observed value.

The distorted values at 00:00 and 02:00 were then replaced by more realistic estimates based on the same hours in the neighbouring weeks. For each affected timestamp, the corrected value was computed as the average between the corresponding observation one week before and one week after:

$$\tilde{y}_t = \frac{y_{t-168} + y_{t+168}}{2}, \quad (3.1)$$

where y_t denotes the original consumption value at hour t , $\tilde{y}_t$ denotes the corrected value, and 168 corresponds to the number of hours in one week. This correction was chosen because electricity consumption typically presents strong weekly regularity, meaning that the same hour in the previous and following weeks provides a reasonable reference for the expected consumption level.

Figure 3.3 shows the effect of the correction. The unrealistic drop present in the original series is removed, while the corrected profile becomes more consistent with the surrounding daily load curves. The timestamp corresponding to 01:00 is now present in the time index with its consumption value set to NaN, whereas the affected values at 00:00 and 02:00 are replaced using the weekly average described above.

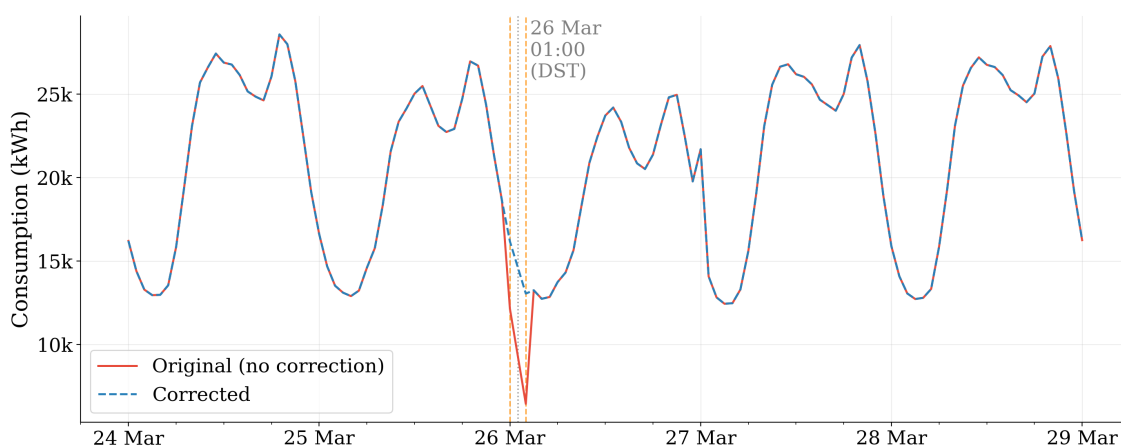


Figure 3.3: Original and corrected 4450 postal-code electrical load profiles around the daylight saving time transition.

3.2.2 Time Series Decomposition

After correcting the main temporal inconsistencies, the hourly electrical load series were analysed through time series decomposition. This step was used to separate the original signal into interpretable components and to better understand the temporal structure of the data.

The decomposition was performed using a multiple seasonal decomposition approach (see Section 2.1.2), separating the original series into trend, daily, and weekly seasonality, and a residual component. Two seasonal periods were considered: a daily period of 24 hours, represented by S_{24} , and a weekly period of 168 hours, represented by S_{168} . These periods correspond, respectively, to the number of hours in one day and in one week.

Figure 3.4 presents the decomposition obtained for postal code 4450 over the full analysed period. The original series shows recurrent oscillations throughout the year, reflecting the strong temporal structure of the electrical load profile. The trend component captures the slower evolution of the average load level: it increases during the first months of the period, decreases around the middle of the year, and rises again towards the final months. This behaviour indicates that, in addition to short-term fluctuations, the series also contains a relevant long-term component.

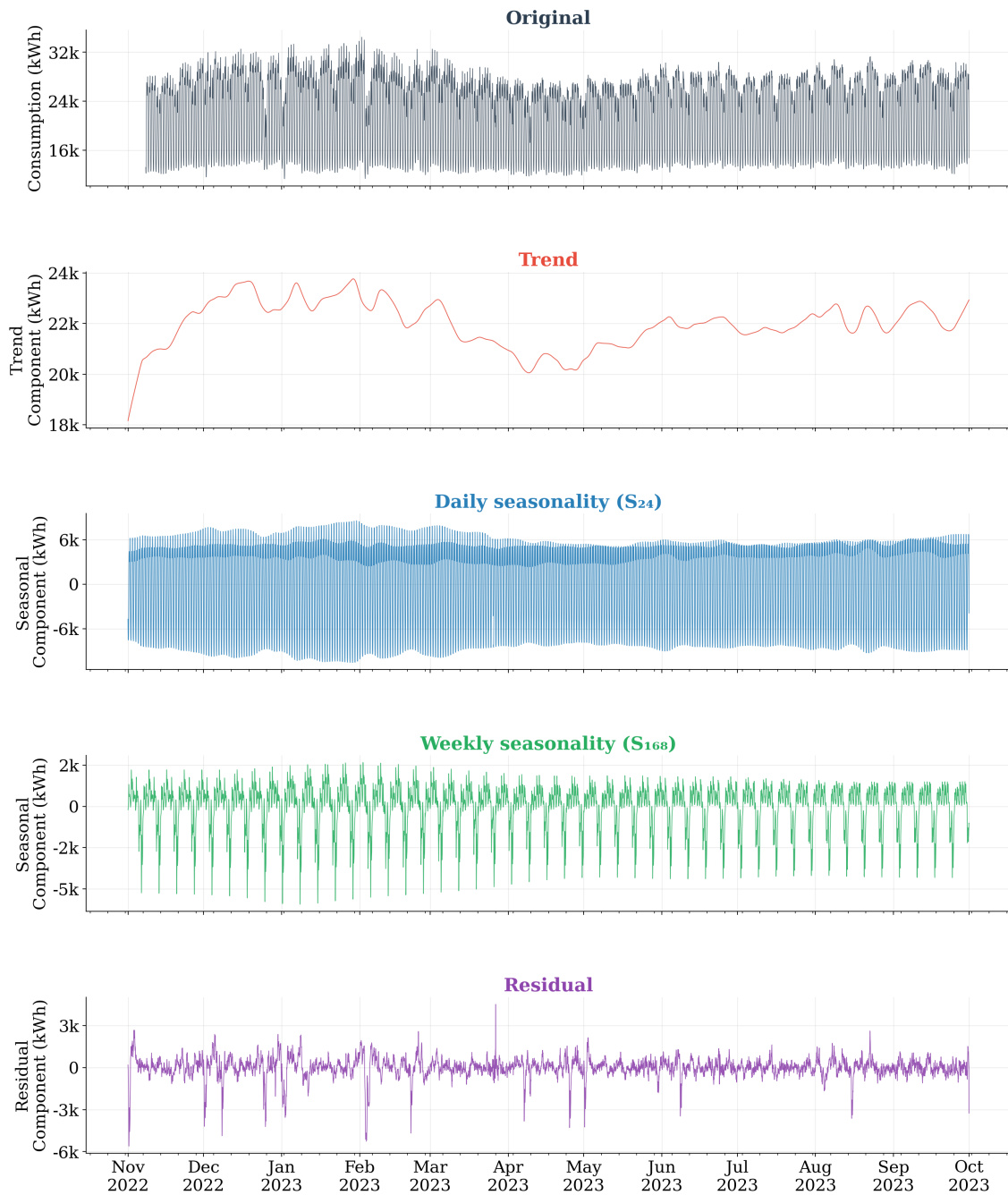


Figure 3.4: MSTL decomposition for the electrical load series of postal code 4450.

The daily and weekly seasonal components capture repeated patterns at different temporal scales. The daily component captures the intra-day variation associated with regular activity cycles, while the weekly seasonality captures systematic differences between weekdays and weekends. In contrast, the residual component is more irregular, with isolated peaks and short-term deviations that are not explained by the trend or seasonal components. This residual series is therefore used as a more appropriate signal for studying short-term irregular variation and cross-location relationships.

Although the seasonal components are visible in the decomposition, their interpretation becomes clearer when their average profiles are analysed separately. Figure 3.5 presents the average daily seasonal profile. The component shows a pronounced low-load period during the early morning, followed by a progressive increase throughout the day. Higher values are observed during the afternoon and evening, corresponding to periods of greater activity. The relatively narrow standard deviation band indicates that this daily profile is stable across the analysed period.

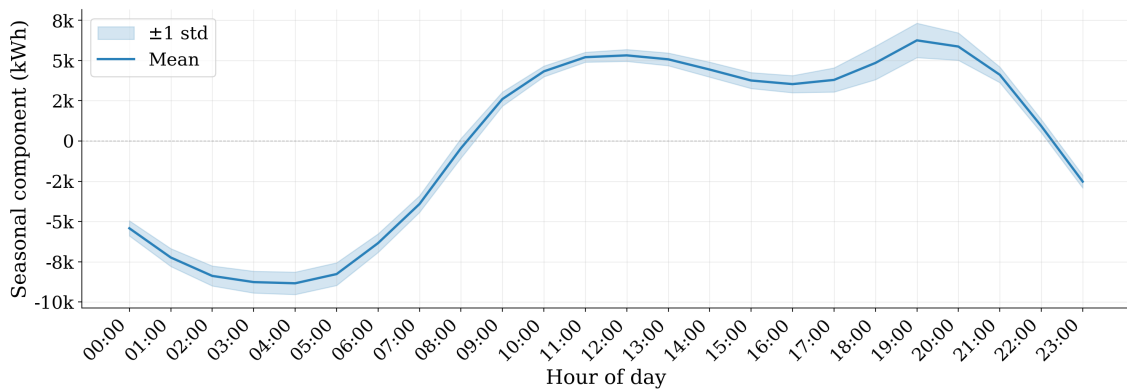


Figure 3.5: Average daily seasonal profile for the electrical load series of postal code 4450.

The average weekly seasonal profile is shown in Figure 3.6. From Monday to Friday, the weekly component is generally positive or close to zero, indicating load levels near or above the weekly average after accounting for the daily cycle. During Saturday and especially Sunday, the component becomes clearly negative, showing a systematic reduction in electrical load during weekends. This confirms that the weekly structure is an important component of the series and should be treated separately from the remaining residual behaviour.

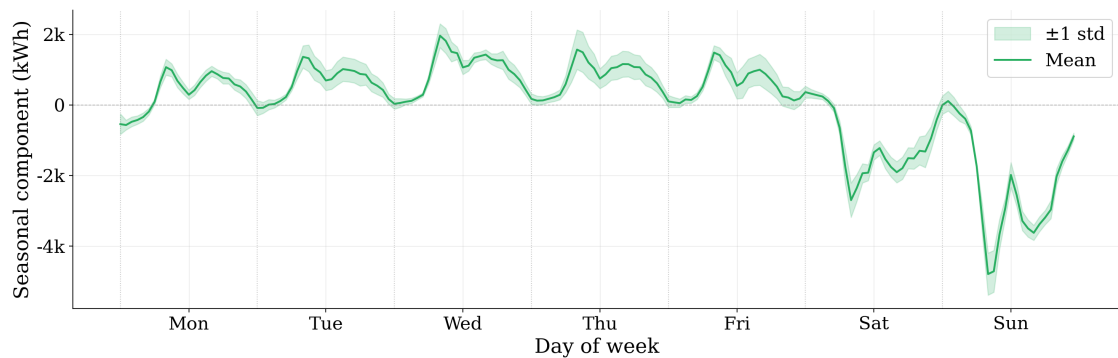


Figure 3.6: Average weekly seasonal profile for the electrical load series of postal code 4450.

To further characterise the temporal dependence of the series, the autocorrelation function (ACF) and partial autocorrelation function (PACF) were computed for lags up to 168 hours, corresponding to one full week, as shown in Figure 3.7. The ACF shows a strong positive correlation at lag 1, confirming high short-term persistence between consecutive hourly observations. The

correlation then oscillates with a clear periodic pattern, reaching pronounced positive peaks at every multiple of 24 hours, which reflects the daily seasonality of load demand. Secondary peaks are also visible at intermediate multiples of 12 hours, consistent with a half-day sub-cycle often present in consumption data. At lag 168, the ACF exhibits a particularly strong positive correlation, indicating that the load observed at a given hour is closely related to the load observed at the same hour exactly one week earlier. This confirms the presence of a weekly seasonal component alongside the daily one.

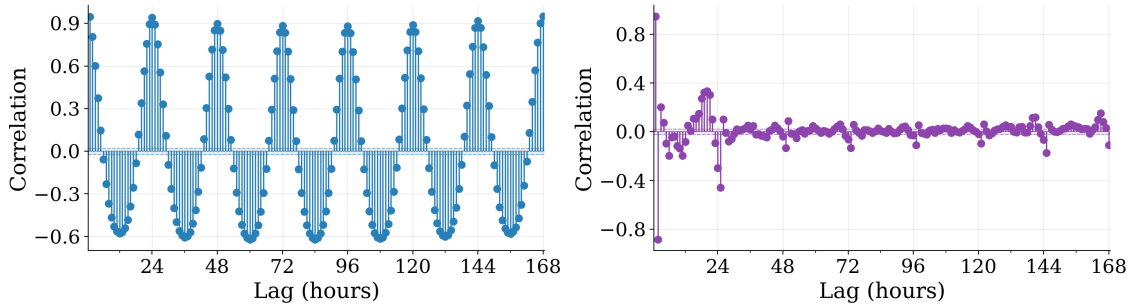


Figure 3.7: Autocorrelation (left) and partial autocorrelation (right) functions for lags up to 168 hours.

The partial autocorrelation function complements this analysis by measuring which lags retain a direct relationship with the current value after accounting for intermediate lags. The first lag presents a strong direct positive effect, confirming the importance of the immediately preceding hour. The second lag also shows a strong direct influence, although negative, suggesting a correction effect once the first lag is accounted for. Significant partial autocorrelations persist up to approximately lag 25, after which the values remain close to zero throughout the remaining horizon, indicating that the direct temporal dependencies are largely concentrated within the first day.

Overall, the decomposition and autocorrelation analyses show that the electrical load series contains strong daily and weekly components, a slowly varying trend, and a residual component with smaller and more irregular fluctuations. These results are important for the following exploratory analyses, since they show that temporal components such as trend and seasonality can dominate the original series and should be separated when the objective is to study residual relationships between postal-code areas.

3.3 Spatio-temporal Analysis

After preprocessing and decomposition, the next step is to analyse whether the electrical load series contain spatial and spatio-temporal structure that can be useful for forecasting. The objective of this section is not only to describe the dataset, but also to identify empirical patterns that motivate the forecasting scenarios developed later in this dissertation.

The analysis focuses on the residual components of the four-digit postal-code series. As discussed in Section 3.2.2, the original load series contain strong trend and seasonal components. If these components are kept in the data, correlations between postal-code areas may be dominated by common calendar-driven behaviour rather than by meaningful residual relationships. For this reason, the exploratory analysis is performed after removing the dominant trend and seasonal components.

Three complementary analyses are considered. First, residual correlations are used to assess whether postal-code areas exhibit similar behaviour at the same timestamp. Second, partial cross-correlation analysis is used to investigate whether residual dependencies persist across short temporal delays. Third, hierarchical clustering is used to identify groups of postal-code areas with similar residual load dynamics.

3.3.1 Correlation Analysis

The first exploratory analysis examines the correlation structure between the four-digit postal-code series. In this stage, only contemporaneous relationships were considered, meaning that all correlations were computed without temporal lags. The objective was to assess whether different postal-code areas present similar electrical load behaviour at the same timestamps.

As discussed in Section 3.2.2, the electrical load series contain strong daily and weekly seasonal components. If these components are kept in the data, the resulting correlations may be dominated by common calendar-driven patterns, rather than by spatial relationships between postal-code areas. For this reason, the seasonal components were removed before computing the correlation between series.

The trend component was also analysed before being included or excluded from the correlation study. As observed in the time series decomposition, the trend captures slow variations in the average load level over time. These variations may be associated with external factors affecting several regions simultaneously, such as temperature. To assess this relation, the Pearson correlation between the trend component of each four-digit postal-code series and the corresponding temperature series was computed. The resulting distribution is shown in Figure 3.8.

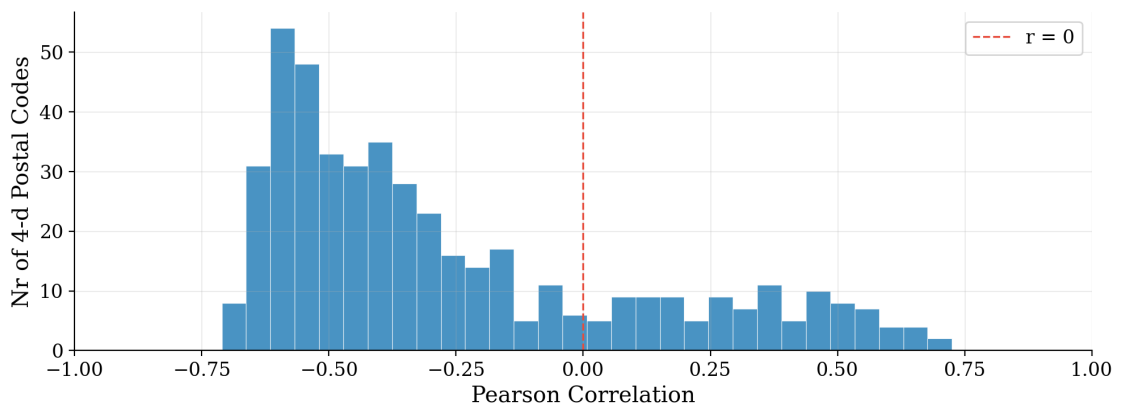


Figure 3.8: Distribution of Pearson correlations between the trend component of the four-digit postal-code series and temperature.

The histogram shows that a large number of four-digit postal-code series present a negative correlation between the trend component and temperature. This behaviour is consistent with the expected increase in electrical load during colder periods. Although positive correlations are also present, indicating heterogeneity across postal-code areas, the distribution suggests that the trend component is substantially related to temperature-driven effects. Therefore, the trend was excluded from the subsequent correlation analysis, in order to focus on relationships between postal-code series that are not primarily explained by broad seasonal or temperature-related variations.

After removing both the daily seasonal component, weekly seasonal component, and trend, the remaining residual component was used to compute the correlation matrix between all four-digit postal-code series. The resulting heatmap is presented in Figure 3.9. Since no lag is applied, each entry represents the Pearson correlation between two residual series at the same timestamp.

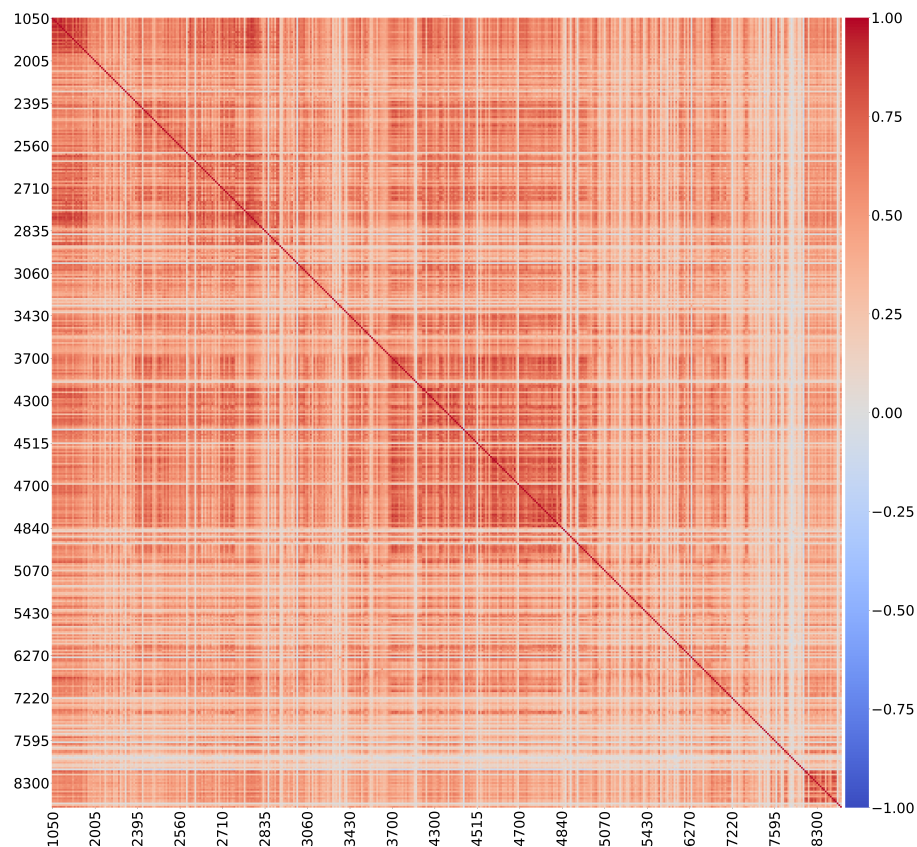


Figure 3.9: Pearson correlation heatmap between residual electrical load series at the four-digit postal-code level.

The heatmap shows predominantly positive correlations, with reddish tones covering most of the matrix. This indicates that, even after removing trend and seasonal components, many postal-code areas still exhibit similar residual electrical load behaviour. Stronger correlation blocks are especially visible around the diagonal, where neighbouring or structurally similar postal-code areas tend to be grouped. A particularly intense block can also be observed near the upper-left

region of the heatmap, suggesting the presence of a subset of postal-code areas with highly related residual dynamics.

These results indicate that the electrical load dataset contains relevant spatial structure. The existence of broad positive correlations and visible correlation blocks suggests that some postal-code areas share common dynamics that are not fully explained by seasonality or long-term trend effects. This provides an initial indication of spatial dependence between series and motivates the following analyses, where lagged cross-correlations and clustering structures are investigated in more detail.

3.3.2 Partial Cross-Correlation Analysis

The previous analysis considered only contemporaneous relationships. However, spatio-temporal dependence may also occur across short temporal delays, meaning that residual variations in one postal-code area may be associated with residual variations in another area one or more hours later. To investigate this possibility, partial cross-correlation analysis was applied to the residual four-digit postal-code series.

As in the distinction between the autocorrelation and partial autocorrelation functions, discussed in Section 3.2.2, the purpose is to isolate the direct contribution of each lag. In this case, the analysis is extended to pairs of postal-code series, allowing the relevance of each lagged interaction to be assessed more independently. While standard cross-correlation measures the overall association between two series at a given temporal displacement, partial cross-correlation reduces the influence of neighbouring lagged effects and provides a clearer indication of the lags with stronger direct relevance.

Figure 3.10 presents an example of the partial cross-correlation function between the residual series of postal codes 4450 and 4560. This example illustrates how lagged dependencies can be identified between two residual electrical load series.

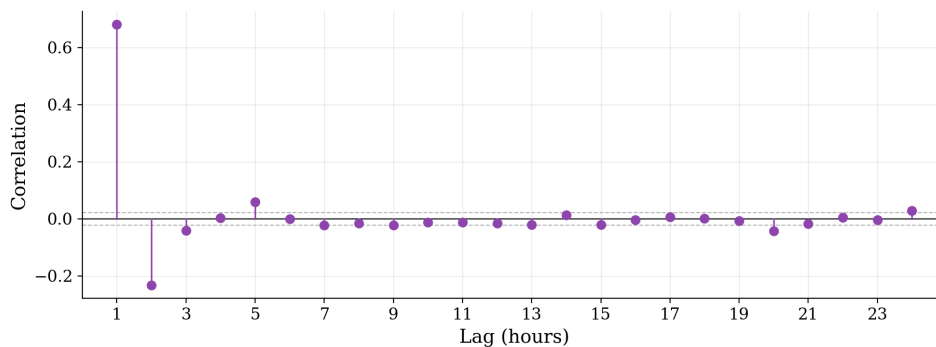


Figure 3.10: Partial cross-correlation function between two selected four-digit postal-code residual series.

The partial cross-correlation values are close to zero for most lags, indicating a weak direct relationship between the two series for most temporal displacements. The main exceptions occur at the first and second lags. At the first lag, the value is strongly positive, suggesting that short-term

variations in one postal-code area are directly associated with variations in the other area one hour later. This is consistent with the strong relationship already observed between these series without temporal displacement. At the second lag, however, the correlation becomes clearly negative, which may indicate a short-term compensating behaviour after approximately two hours, where an increase in one series is associated with a delayed decrease in the other, or vice versa. Since this interpretation cannot be fully established from a single pair of series, a broader view across all four-digit postal-code areas is required.

For this reason, partial cross-correlation heatmaps were computed for all pairs of four-digit postal-code residual series. In these heatmaps, the postal-code series represented on the horizontal axis are shifted in time relative to those on the vertical axis, meaning that each cell represents the direct lagged association between a reference series and another series delayed by the corresponding number of hours. Figure 3.11 presents the heatmap for a one-hour lag. Similarly to the correlation structure observed without temporal displacement, the matrix is dominated by reddish tones, indicating widespread positive relationships between postal-code areas. The strongest positive correlations are particularly visible inside the highlighted region, which includes postal codes approximately from the 1000 range to values close to the 5000 range.

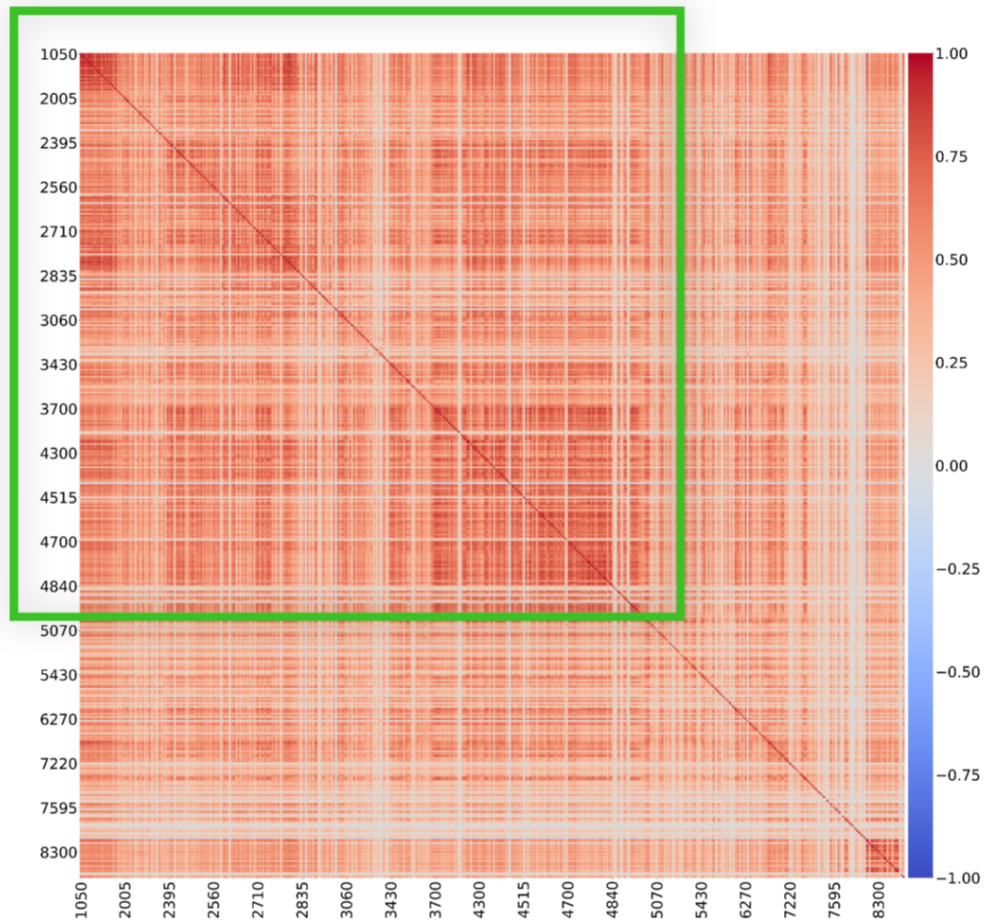


Figure 3.11: Partial cross-correlation heatmap between four-digit postal-code residual series for a one-hour lag.

Since the first digit of a postal code identifies a broad region of the country, as detailed later in Section 3.4, this highlighted area can be read directly from the code ranges, and it covers the regions associated with more urban and densely populated zones, namely the areas around Lisbon, Aveiro, Porto, and Braga. The stronger positive relationships observed in this block suggest that these postal-code areas share similar short-term residual electrical load dynamics. Outside this region, the correlations are generally weaker, which may indicate more specific local behaviour in areas with different activity patterns, such as the *Faro*, *Alentejo*, *Beira Interior*, or *Trás-os-Montes* regions.

The heatmap for a two-hour lag is shown in Figure 3.12. In this case, the most relevant structures are no longer predominantly positive. Instead, the highlighted region presents moderate negative partial cross-correlations, indicating that some relationships between postal-code areas become inverse when a two-hour delay is considered.

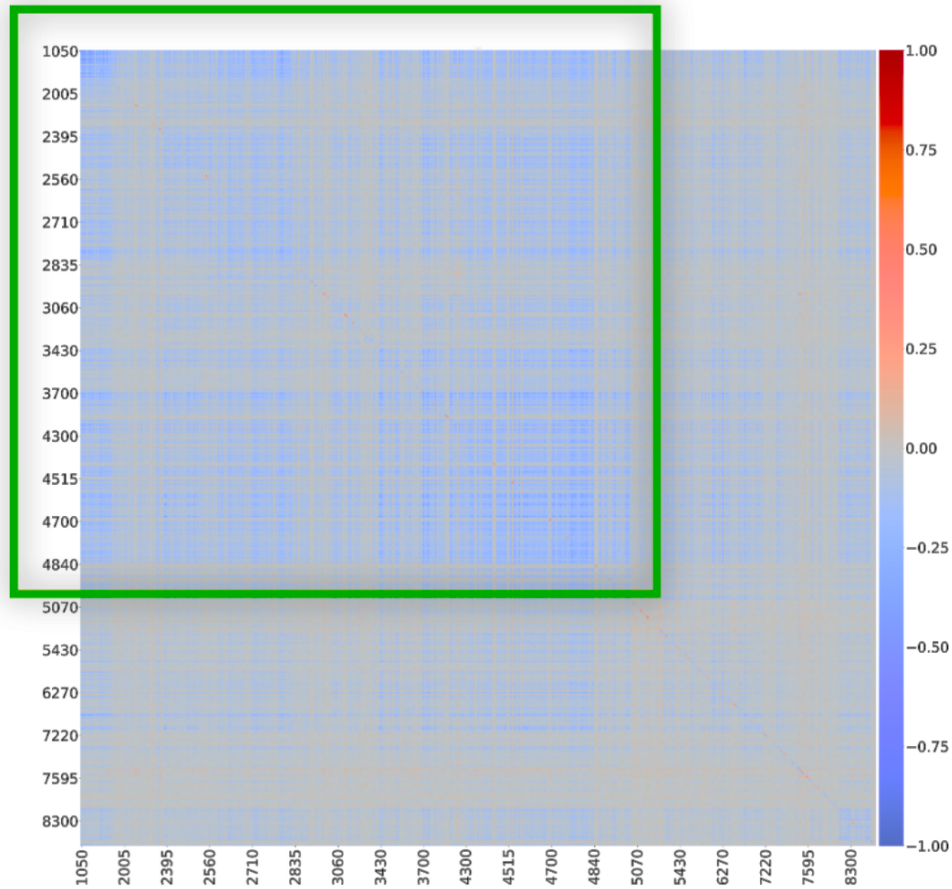


Figure 3.12: Partial cross-correlation heatmap between four-digit postal-code residual series for a two-hour lag.

One possible explanation for these negative two-hour relationships is the temporal redistribution of activity between different types of areas. For instance, changes in electrical load may occur first in residential areas and later in urban, commercial, or industrial zones, or in the opposite direction depending on the time of day. This type of delayed response may be associated with daily mobility patterns, such as commuting between residential and work areas, where a decrease in load in one region may be followed by an increase in another region after a short delay. However, this interpretation should be treated as exploratory, since confirming it would require additional information on mobility, land use, and local activity patterns. Even so, the presence of structured negative correlations suggests that the two-hour lag captures relevant spatio-temporal behaviour rather than random variation.

To verify whether relevant direct relationships persist at longer temporal delays, the same analysis was repeated for delays of three and 24 hours. At three hours the structure observed for the first two lags largely dissipates, with the partial cross-correlations concentrated around zero, a mean of -0.013 across all pairs of distinct series and only 0.5% of the pairs exceeding

0.1 in absolute value, against 17.8% at the two-hour lag. Some residual traces of the negative relationships found at two hours nevertheless remain, located mostly within the same urban block identified at the shorter lags, among the postal codes starting with 1 and 4 that correspond to the Lisbon and Porto areas, where the weakest values reach about -0.20 . Given their small magnitude and the very small fraction of pairs involved, these remnants were not considered strong enough to justify including the three-hour lag among the cross-series predictors of the forecasting models.

At the 24-hour delay the matrix is closer still to uniform. Although the original electrical load series contain a strong daily cycle, this component was removed by the decomposition before the cross-correlation analysis, so the residual series retain no meaningful direct cross-dependence at the daily lag, with a mean partial cross-correlation of 0.009 and only four of the 215 760 pairs exceeding 0.1 in absolute value. Any role of the 24-hour lag in the forecasting models therefore reflects the daily seasonality shared by the series rather than a spatial link between areas.

Overall, the partial cross-correlation analysis shows that the main spatio-temporal dependencies between four-digit postal-code residual series are concentrated at short delays, particularly at one and two hours. The one-hour lag is mainly associated with positive relationships between areas, while the two-hour lag reveals structured negative associations in specific postal-code blocks. For longer lags, the direct cross-correlations become weak, suggesting that the most relevant spatio-temporal interactions in the residual electrical load series occur over short time scales.

3.3.3 Clustering Analysis

The previous correlation analyses indicate that several four-digit postal-code series share similar residual electrical load behaviour. This motivates a clustering analysis to identify groups of postal-code areas with comparable temporal dynamics. Since the dataset contains a large number of individual series, this type of analysis is also useful from a practical perspective: instead of analysing all four-digit postal codes independently in later stages, representative groups can be identified and used to reduce the complexity of the study.

Hierarchical clustering was applied to the residual electrical load series at the four-digit postal-code level. The similarity between each pair of series was measured using the Pearson correlation coefficient, and the corresponding dissimilarity was defined as

$$d_{ij} = 1 - r_{ij}, \quad (3.2)$$

where r_{ij} is the Pearson correlation between postal-code series i and j . With this definition, the distance ranges from 0 to 2, since the Pearson correlation coefficient ranges from -1 to 1 . A distance of 0 corresponds to perfectly correlated series, while a distance of 2 corresponds to perfectly anti-correlated series. Therefore, highly correlated series have small distances, whereas weakly correlated or negatively correlated series are placed further apart. The hierarchical clustering was performed using average linkage, meaning that the distance between two clusters is computed as the average distance between all pairs of elements belonging to the two groups.

A Pearson correlation threshold of 0.6 was selected as the main clustering level. Since the dendrogram distance is defined as $1 - r$, this corresponds to cutting the tree at a distance of 0.4. This threshold provides a compromise between cluster cohesion and the number of resulting groups: it avoids merging almost all series into a single dominant cluster, while also preventing excessive fragmentation into very small groups.

The dendrogram obtained for this threshold is presented in Figure 3.13. Due to the large number of four-digit postal-code series, the dendrogram was split into two panels for readability. Panel (a) shows the left half of the dendrogram, while panel (b) shows the right half. The branch colours encode the clustering structure at the selected cut: each non-blue colour identifies a distinct cluster, that is, a group of series that merge *below* the cut-off distance of 0.4 (correlation ≥ 0.6) and are therefore considered sufficiently similar to be grouped together. Branches drawn in blue correspond to merges that occur *above* the cut-off distance, joining series or groups whose residual correlation is below 0.6. Therefore, these are not grouped at this threshold, so the long blue links near the top of the tree represent the weakly correlated series that remain isolated as singletons.

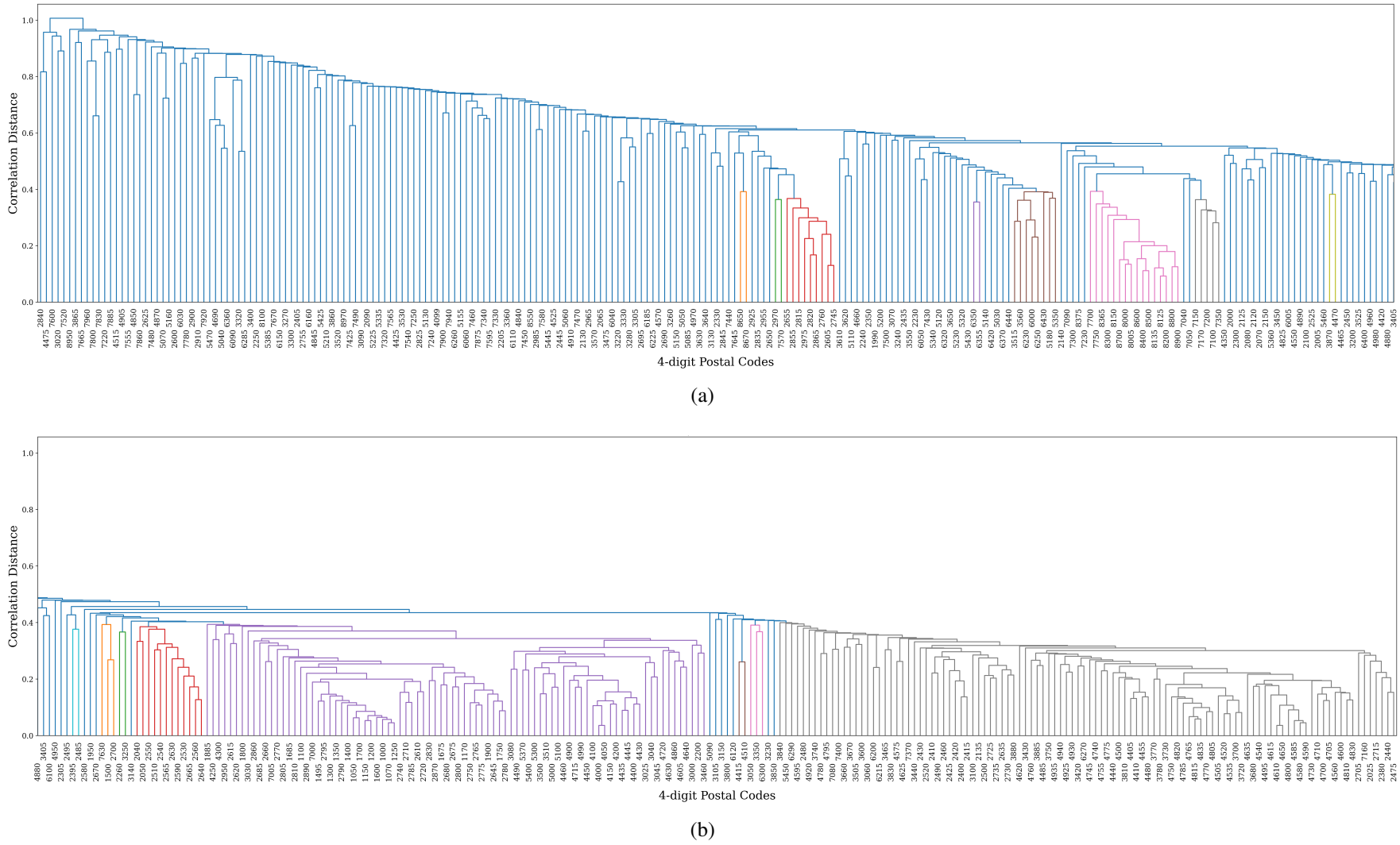


Figure 3.13: Hierarchical clustering dendrogram of four-digit postal codes (Pearson threshold = 0.6): (a) left half, (b) right half.

At this correlation level, the hierarchical clustering produces 16 clusters and 202 singletons. The largest cluster contains 106 postal-code series, indicating that a substantial part of the dataset still shares a common residual load structure even after the trend and the seasonal components have been removed. At the same time, the large number of singletons (202 out of 464 series) shows that many postal-code areas have residual dynamics that are not sufficiently similar to any other group at the selected threshold.

The dendrogram also shows that some postal-code series merge at very low distances, forming compact branches. These groups correspond to areas with high mutual correlation and therefore similar residual load behaviour. Other branches only merge at higher distances, indicating weaker similarity between the corresponding series. This structure confirms that the dataset is not homogeneous: instead, it contains groups of series with clearly different levels of similarity.

To assess the sensitivity of the clustering structure to the selected correlation threshold, several threshold values were tested. Table 3.2 summarises the number of clusters, singletons, and the size of the largest cluster obtained for each value.

Table 3.2: Hierarchical clustering results for different Pearson correlation thresholds.

Pearson threshold	Clusters	Singletons	Largest cluster
0.4	10	103	323
0.5	12	148	244
0.6	16	202	106
0.7	19	271	56
0.8	21	352	20
0.9	10	433	8

For lower thresholds, such as 0.4 or 0.5, the clustering structure is dominated by one very large group, containing between 244 and 323 postal-code series. This indicates that, even after removing the trend and seasonal components, most residual series still share at least a moderate level of correlation. However, these thresholds are too permissive to reveal more detailed differences between groups.

As the threshold increases, the clustering structure becomes progressively more fragmented. At a threshold of 0.8, the number of clusters reaches 21 and the number of singletons rises to 352, while the largest remaining cluster contains only 20 series. At 0.9, the requirement of near-perfect residual correlation leaves only 10 small clusters, with 433 of the 464 series isolated as singletons and the largest cluster containing just 8 series. This means that only a very limited number of postal-code series remain grouped when near-perfect correlation is required.

The selected threshold of 0.6 therefore provides a balanced representation of the dataset. It retains a meaningful number of clusters, while avoiding both excessive aggregation and excessive fragmentation. From an exploratory perspective, this result is particularly useful because it shows that several four-digit postal-code series can be grouped according to similar residual electrical

load behaviour. These groups can later support the selection of representative series or the design of experiments using a smaller but more informative subset of the full dataset.

To complement the clustering results, Figure 3.14 provides a visual comparison between one highly representative postal-code series and three low-correlation series.

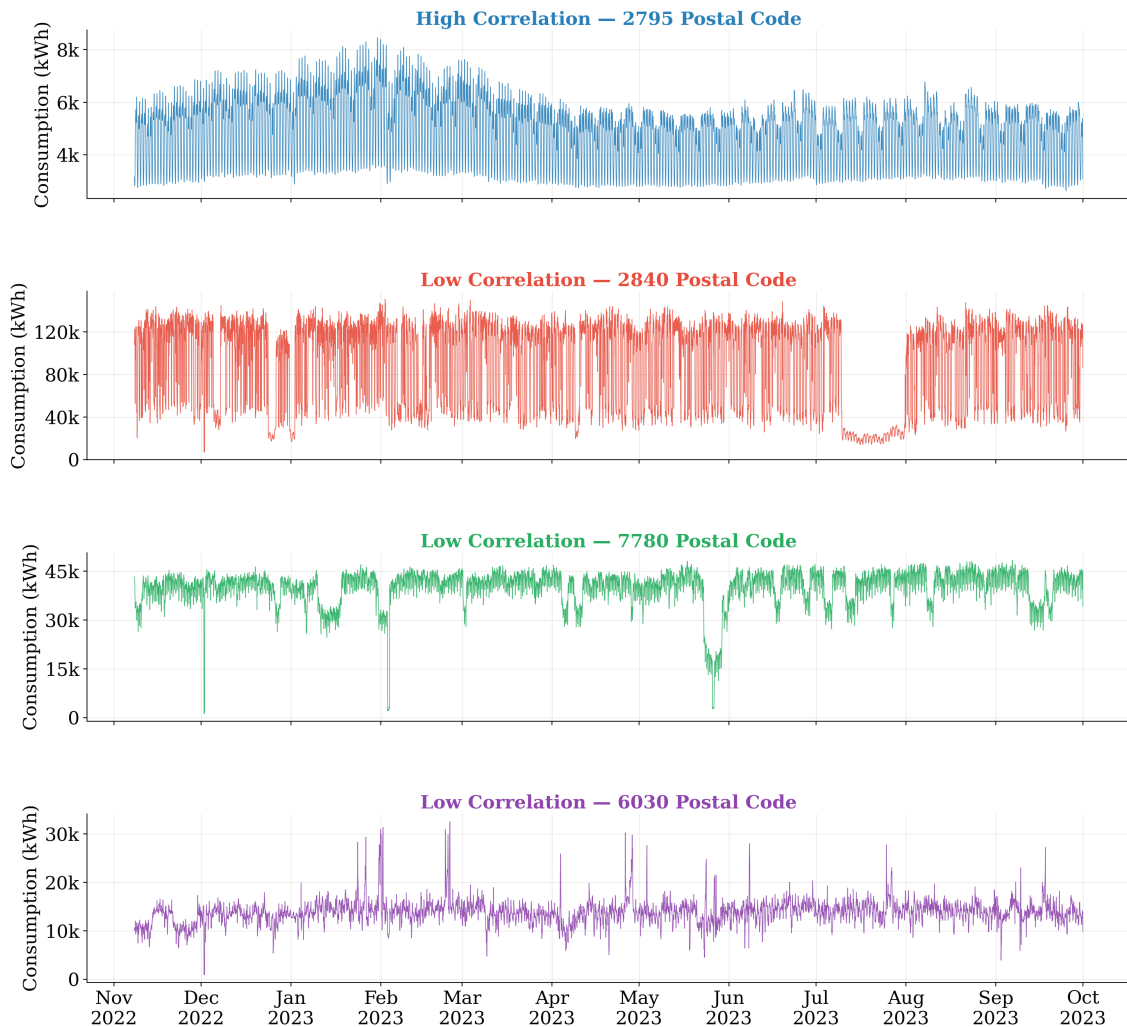


Figure 3.14: Representative electrical load series from a large high-correlation cluster and three examples of low-correlation postal-code series.

The representative series, postal code 2795, Carnaxide and Oeiras, is one of the most strongly correlated series within the metropolitan Lisbon cluster, one of the two largest groups obtained at the Pearson threshold of 0.6, which contains 86 postal-code series. On the residual series, its average correlation with the remaining members of this cluster is 0.74 (the 10th highest among the 86 members), and its correlation with the mean residual profile of the cluster reaches 0.90, indicating that it closely follows the typical residual behaviour of this group. Its average electrical load is approximately 4 845 kWh, with a standard deviation of 1 211 kWh, and its average residual correlation with all other postal-code series is 0.52, among the highest values in the dataset

(31st out of 464). Visually, the profile is regular and well structured, with clear daily and weekly patterns, a visible seasonal evolution, and no major irregular disruptions.

The remaining panels illustrate three different forms of low-correlation behaviour. All three correspond to series that remain isolated as singletons at the selected threshold and that rank among the least correlated postal codes in the entire dataset. The first low-correlation example, postal code 2840, Seixal, has an average electrical load of approximately 89 762 kWh and a standard deviation of 40 732 kWh, making it much larger and more variable than the representative series. Its average residual correlation with the rest of the dataset is essentially zero (-0.001), the 6th lowest value among all 464 series, indicating that its residual behaviour is not aligned with the dominant patterns. This is also visible in Figure 3.14, where the profile shows large amplitude variations, abrupt level changes, and a markedly irregular structure. Such behaviour is consistent with the presence of strong localised load drivers, such as industrial or highly intermittent demand.

A different type of atypical behaviour is observed in postal code 7780, Mértola. In this case, the average electrical load is approximately 39 993 kWh, while the standard deviation is only 5 601 kWh, indicating a comparatively stable profile despite the high load level. Even so, its average residual correlation with the rest of the dataset is only 0.08, the 21st lowest value among the 464 series, again showing weak alignment with the dominant pattern. The series remains relatively stable during most of the year, but contains several pronounced downward deviations. This suggests that low correlation does not necessarily result from high volatility: it may also arise from a stable but atypical temporal structure.

The third low-correlation example, associated with postal code 6030, Castelo Branco, presents an average electrical load of approximately 13 828 kWh and a standard deviation of 2 479 kWh. Its average residual correlation with all other series is only 0.12, the 32nd lowest value among the 464 series. Although this series is less extreme in magnitude than the previous two examples, its residual structure does not follow the dominant recurrent pattern observed in the representative cluster. The presence of isolated spikes and a less clearly defined seasonal pattern helps explain why it remains weakly connected to the rest of the dataset.

Overall, these examples show that low-correlation postal-code series are not all atypical in the same way. Some differ because of very large and irregular load variations, others because of unusually stable profiles, and others because their temporal structure does not align with the dominant daily and weekly patterns.

3.4 Construction of the Hierarchical Dataset

After the preprocessing stage, the electrical load data were organised into a spatial hierarchy based on the Portuguese postal-code structure. The original numerical series correspond to four-digit postal codes, which represent the most disaggregated spatial level available in the dataset. From this base level, higher aggregation levels were constructed by progressively grouping postal codes according to their prefix.

The resulting hierarchy contains four spatial aggregation levels: four-digit postal codes, two-digit postal-code regions, one-digit postal-code regions, and the national aggregate. The intermediate three-digit postal-code level was not included, since in the available data it mostly grouped only two or three four-digit postal-code series, providing little additional aggregation between the four-digit and two-digit levels. The adopted hierarchy is illustrated in Figure 3.15: panel (a) shows an approximation of the most disaggregated four-digit postal-code level, panel (b) shows the two-digit aggregation, panel (c) shows the one-digit aggregation, and panel (d) represents the national aggregate.

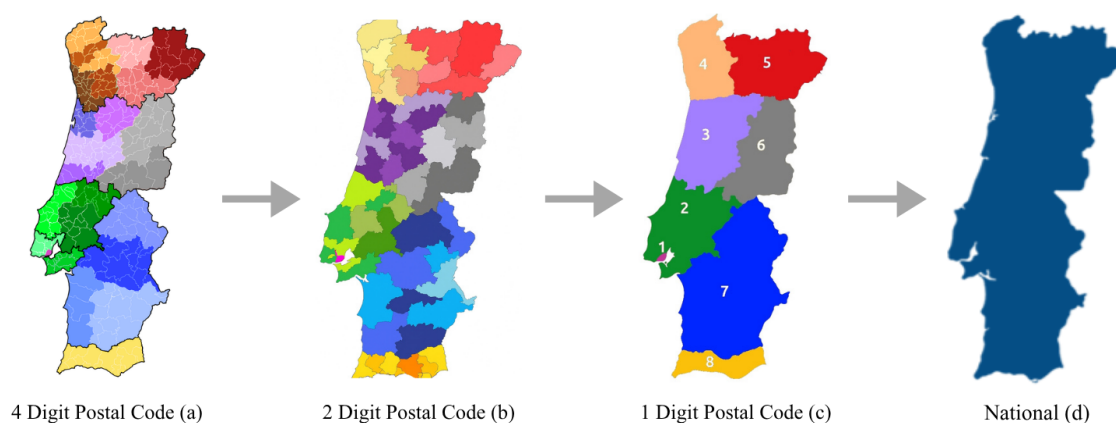


Figure 3.15: Spatial hierarchy used to aggregate the electrical load data: (a) four-digit postal-code level, (b) two-digit postal-code level, (c) one-digit postal-code level, and (d) national aggregate.

Although the original data extraction produced 465 parquet files, one of them corresponded to the category “OTHERS”. This category was excluded from the hierarchical construction because it does not correspond to a numerical postal-code region and cannot be assigned consistently to any spatial aggregation level. Therefore, the base level of the hierarchy contains 464 numerical four-digit postal-code series.

The aggregation was performed by summing, at each timestamp, the hourly electrical load values of all lower-level series belonging to the same postal-code prefix. Thus, each two-digit postal-code region was obtained from the set of four-digit postal codes sharing the same first two digits, each one-digit postal-code region was obtained from the corresponding two-digit regions, and the national aggregate was obtained from all numerical four-digit postal-code series included in the dataset.

This procedure ensures that the hierarchy is internally coherent: at any timestamp, each upper-level series corresponds to the sum of the lower-level series assigned to it. Since the “OTHERS” category was excluded, the national aggregate should be interpreted as the aggregate of all numerical postal-code series used in this work, rather than as an official total load series for mainland Portugal.

Table 3.3 summarises the number of time series available at each level of the hierarchy. The final hierarchical dataset contains 542 hourly electrical load series.

Table 3.3: Structure of the hierarchical electrical load dataset.

Level	Granularity	Number of series
1	four-digit postal codes	464
2	two-digit postal-code regions	69
3	one-digit postal-code regions	8
4	National aggregate	1
Total		542

The number of lower-level series included in each aggregate is not uniform across the country. At the two-digit postal-code level, the aggregates are formed from groups of four-digit-level series. These groups range from a single series to 20 series, with an average size of approximately 6.7.

At the one-digit postal-code level, each aggregate contains between 5 and 10 regions from the two-digit level. In terms of the original four-digit-level series, the size of these aggregates ranges from 21 to 120 series. The eight regions at this level, identified by the first postal-code digits from 1 to 8, cover mainland Portugal.

The four-digit level presents substantial heterogeneity in electrical load magnitude. The average hourly electrical load ranges from approximately 14 kWh in almost residual series to approximately 89 762 kWh in the largest postal-code area. The median average hourly electrical load is approximately 7 023 kWh, while the mean is approximately 10 089 kWh. This difference between the mean and the median indicates a strongly right-skewed distribution, where a limited number of high-load postal-code areas account for a disproportionate share of the total electrical load. These areas are likely associated with dense urban, commercial, or industrial zones.

At the national level, the average hourly electrical load is approximately 4 681 MWh. The two largest one-digit postal-code regions are those identified by the first digits 2 and 4, with average hourly electrical loads of approximately 1 409 MWh and 1 361 MWh, respectively. Together, these two regions represent 59% of the national electrical load, highlighting the spatial concentration of demand in the Lisbon/Setúbal and Northern/Greater Porto regions.

To illustrate how electrical load series change across the hierarchy, Figure 3.16 presents one representative aggregation path: postal code 4450, two-digit postal-code region 44, one-digit postal-code region 4, and the national aggregate.

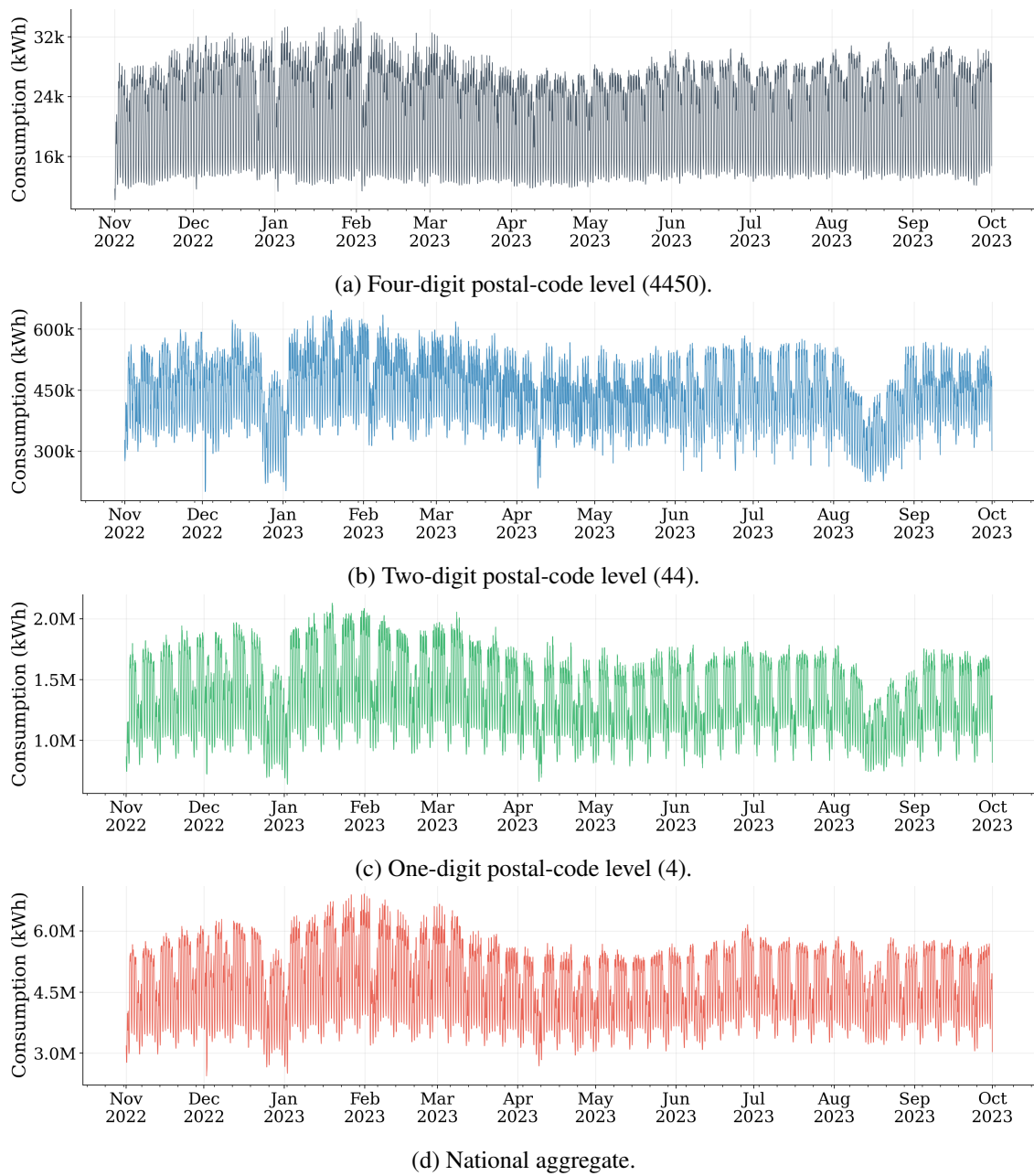


Figure 3.16: Example of an aggregation path across the spatial hierarchy: postal code 4450, two-digit region 44, one-digit region 4, and national aggregate.

At the four-digit postal-code level, the series presents clear daily variability, with recurrent peaks and valleys over the full period. The profile is relatively irregular, reflecting the fact that this level captures a local electrical load pattern with no spatial aggregation. As a result, local fluctuations and isolated changes are directly visible in the series.

At the two-digit postal-code level, the main daily and weekly patterns remain visible. However, in this particular aggregation path, the profile does not become smoother when compared with the four-digit level. Instead, the series appears more irregular, which may result from the aggregation of postal-code areas with different electrical load behaviours, such as residential, commercial,

and industrial profiles. A marked low-load period is also visible near the end of December at this level, and this feature remains clearly identifiable at higher aggregation levels. Therefore, although spatial aggregation can often reduce the impact of local fluctuations, intermediate aggregation levels may also combine heterogeneous load patterns and reveal broader regional effects that are not evident at the most disaggregated level.

At the one-digit postal-code level, the electrical load profile corresponds to a broader regional aggregate. The daily cycle remains well defined, but individual local effects are further diluted. Compared with the previous levels, the series exhibits broader seasonal movements, including higher load levels during part of the winter period and a reduction during the summer period. The low-load period observed near the end of December also remains visible, indicating that this feature is not restricted to the two-digit aggregation level.

Finally, for the national level, the profile combines all numerical four-digit postal-code series included in the dataset. Local effects become almost imperceptible, while macro-level electrical load patterns dominate the series. The national aggregate shows a clear daily rhythm, recurrent weekly behaviour, higher load levels during part of the winter period, and a visible reduction around August. The low-load period near the end of December is also present at this level, suggesting that it reflects a broader system-level pattern rather than an isolated local fluctuation.

Overall, the constructed hierarchy provides a coherent representation of electrical load at different spatial scales. The four-digit level preserves local detail and spatial heterogeneity, while the national level captures the aggregate behaviour of the numerical postal-code series considered in this work. The intermediate two-digit and one-digit levels provide a structured transition between these two extremes, allowing the analysis of how electrical load patterns change with spatial aggregation.

3.5 Main Data Challenges

The dataset used in this dissertation presents several challenges that must be considered before the modelling stage. These challenges are mainly related to the limited temporal coverage of the data and to the particular behaviour observed during the summer months. Both aspects affect the interpretation of the exploratory analysis and have direct implications for the design of the validation strategy.

The first limitation is the reduced amount of historical data available. The dataset covers the period from November 2022 to September 2023 and therefore does not include a complete annual cycle. This is particularly relevant in electrical load analysis, since load patterns are affected by seasonal effects, calendar structure, temperature variation, and changes in social and economic activity throughout the year. With less than one full year of observations, the models cannot be trained with examples from all months of the year.

This limitation becomes especially important in out-of-sample forecasting. When a given month is used as a test period, there may be no prior occurrences of that month in the training data. As a consequence, the model may need to forecast periods with specific seasonal characteristics

that were not directly observed during training. This can be problematic because each month may contain particular load behaviours associated with weather conditions, holidays, working routines, or school calendars. Therefore, the limited temporal coverage makes it harder to evaluate whether a model is learning robust electrical load patterns or simply adapting to the months available during the training period.

A second challenge follows from this limitation and concerns the behaviour of the summer months, particularly July, August, and September. These months tend to differ from the rest of the year due to higher temperatures, school breaks, holiday periods, and changes in residential, commercial, and industrial activity. These factors can introduce additional variability and make electrical load profiles less regular than during other periods. In the previous exploratory figures, this effect is already visible in the aggregated series. For example, Figures 3.16b and 3.16c show a marked decrease in electrical load around the end of August, which is likely associated with the summer holiday period and the corresponding reduction in economic and social activity.

These characteristics create a difficult evaluation context. Summer months may present electrical load patterns that are not fully represented in the rest of the dataset, while the short historical window limits the possibility of learning recurring annual behaviour. As a result, forecasting errors during these months should be interpreted with caution, since they may reflect not only model limitations but also the intrinsic difficulty of predicting periods with atypical or weakly represented patterns.

To mitigate these issues, the validation procedure must be designed to make efficient use of the limited historical period while avoiding leakage from the observations being evaluated. For this purpose, out-of-fold (OOF) predictions are used. These correspond to predictions generated for observations that were not seen by the model during the corresponding training step, allowing model behaviour to be assessed on data that were held out from fitting. Therefore, in this work, OOF predictions are obtained through a Leave-One-Month-Out (LOMO) cross-validation strategy. In each fold, one month is held out and predicted using a model trained on the remaining available months.

In addition, the results are analysed with explicit awareness of the summer-month effect. Rather than reporting a separate aggregate metric for the least affected weeks, which the short test window would render unreliable, the analysis of Chapter 5 examines how the error of each approach evolves within the test period, contrasting the ordinary weeks of June with the holiday period of August both for individual series and for the monthly aggregates. This allows the evaluation to distinguish between the general performance of a model and its behaviour during the months in which the load departs most from the training period. Overall, these data limitations do not prevent the development of forecasting models, but they require a careful validation strategy and a cautious interpretation of the results.

3.6 Chapter Summary

This chapter has presented the dataset on which the dissertation rests and has characterised its temporal, spatial and hierarchical properties. The eleven months of hourly consumption obtained from the E-REDES open data portal were corrected for the inconsistencies introduced by the daylight saving time transition and then decomposed into trend, daily seasonality, weekly seasonality and residual, so that the dependence between locations could be examined on a signal from which the dominant temporal structure had already been removed.

The exploratory analysis conducted on those residuals establishes the basis for the scenarios that follow. Since the trend proved to be substantially related to temperature, it was excluded from the dependence analysis, leaving a correlation matrix that remains predominantly positive and shows visible blocks around the denser urban regions of the country. That dependence is sharply confined in time, being strongly positive at one hour, structurally negative at two hours in specific groups of areas, and indistinguishable from zero at three and at twenty-four hours.

The hierarchical dataset built from these series contains 542 hourly load series, obtained by aggregating the four-digit postal codes into two-digit regions, one-digit regions and a national total. Two limitations of the data, discussed in Section 3.5, shape every experimental decision taken afterwards, since the eleven-month window contains no complete annual cycle and the months available for testing are the least regular of the year. Rather than being treated as an obstacle, both are addressed through a leakage-free validation strategy.

With the data characterised and the hierarchy established, Chapter 4 turns this evidence into a forecasting problem. The lags at which the dependence was found to concentrate become the predictor structure of the models, the levels constructed here become the scales at which every scenario is evaluated, and the clustering obtained from the residuals is reused to assign hyperparameter representatives. The question addressed from this point onwards is whether the structure documented in this chapter translates into predictive value beyond what a purely temporal model already extracts.

Chapter 4

Problem Formulation and Methodology

Throughout this chapter, electrical load is indexed by spatial aggregation level, spatial unit, and time. Let $\ell \in \{4, 2, 1, 0\}$ denote the aggregation level, corresponding respectively to 4-digit postal codes, 2-digit postal-code regions, 1-digit postal-code regions, and the national aggregate. The observed electrical load at time t for spatial unit i at level ℓ is denoted by $y_{i,t}^{(\ell)}$.

At each forecast origin t , forecasts are produced for a future target time $t + h$, where h denotes the forecast horizon. The corresponding point forecast is written as $\hat{y}_{i,t+h|t}^{(\ell)}$, indicating that the forecast for time $t + h$ is produced using only the information available up to time t . Probabilistic forecasts are represented through quantiles, $\hat{q}_{i,t+h|t}^{(\ell)}(\tau)$, where $\tau \in \{0.05, \dots, 0.95\}$ is the quantile level. This notation allows forecasts to be defined consistently across all spatial aggregation levels, while preserving the distinction between forecast origin, forecast horizon, and target time.

This chapter is organised as follows. Section 4.1 formulates the forecasting problem, defining the target variable, the one-step-ahead task and the hierarchical structure over which the forecasts are produced. Section 4.2 defines the features and the eight forecasting scenarios evaluated in this work, ranging from univariate baselines to spatio-temporal and cross-level configurations. Section 4.3 presents the two forecasting approaches, point and probabilistic, each of them beginning with the setup that governs the temporal split, the validation scheme and the generation of leakage-free training-period predictions, and continuing with the implemented pipelines, including the LASSO-VAR-X base forecasts and the subsequent hierarchical reconciliation. Section 4.4 presents the evaluation metrics and the criteria used to compare the scenarios and the approaches. Finally, Section 4.5 summarises the chapter and states the two properties of the experimental design on which the interpretation of the results depends.

4.1 Problem Definition

The forecasting problem addressed in this work follows from the spatial and spatio-temporal dynamics identified between postal-code areas in Section 3.3. After observing that residual load

dynamics are not independent across postal-code areas, the central question is whether these relationships can be exploited to improve electrical load forecasting. Therefore, the forecasting problem is not limited to using only the past values of the target series. Instead, the aim is to evaluate whether information from other postal-code series can provide additional predictive value.

For each spatial unit i at aggregation level ℓ , the forecasting task consists of estimating the future electrical load $\hat{y}_{i,t+h|t}^{(\ell)}$, using information available up to the forecast origin t . More generally, the forecasting model can be expressed as

$$\hat{y}_{i,t+h|t}^{(\ell)} = f_{\theta}(\mathbf{x}_{i,t}^{(\ell)}), \quad (4.1)$$

where $f_{\theta}(\cdot)$ denotes the forecasting model with parameters θ and $\mathbf{x}_{i,t}^{(\ell)}$ represents the predictor vector available at time t . The composition of this predictor vector depends on the forecasting scenario under consideration and may include temporal, spatial, and hierarchical information.

The simplest configuration uses only historical observations of the target series, corresponding to a conventional univariate forecasting problem. More advanced configurations progressively expand the predictor set by incorporating lagged observations from neighbouring spatial units at the same aggregation level and features derived from adjacent levels of the spatial hierarchy. The objective is to determine whether these additional sources of information lead to consistent improvements in forecasting accuracy.

The forecasting task is carried out across the four aggregation levels constructed in Section 3.4, allowing the effect of spatial information to be analysed at different scales.

Two modelling hypotheses are investigated throughout this dissertation:

1. Spatial dependencies between neighbouring or statistically related locations contain predictive information beyond the historical observations of the target series.
2. Cross-level information exchanged between adjacent levels of the spatial hierarchy provides additional predictive value for electrical load forecasting.

The second hypothesis is motivated by the hierarchical nature of distribution-level electrical load. Lower-level series can provide detailed information for forecasting an aggregate series, since upper-level load profiles are built from their lower-level components. This idea is related to Gilbert et al. [11], where turbine-level information is introduced as an explanatory input to improve wind farm-level probabilistic forecasts. Conversely, higher-level series may provide broader contextual signals for forecasting more disaggregated units. The problem is therefore both spatial and hierarchical: the relevant information for forecasting a given series may come from the target series itself, from other units at the same level, or from adjacent aggregation levels.

4.2 Forecasting Scenarios

This section defines the set of forecasting scenarios evaluated in this study. A scenario is characterised by the collection of features made available to the predictive model, ranging from purely

autoregressive configurations to richer settings that exploit the spatio-temporal dependencies and the hierarchical structure of the data. Specifically, these scenarios are constructed to address the following questions:

1. How well can electrical load be predicted using only the historical values of the target series?
2. Does incorporating spatial information from neighbouring series improve forecasting performance?
3. Does exploiting information from adjacent levels of the spatial hierarchy provide additional predictive value?

To answer these questions, eight forecasting scenarios are defined. Before describing them, the features considered throughout the work are introduced and grouped into three categories, namely temporal features, lagged features, and an exogenous feature.

4.2.1 Predictor Features

Three groups of predictors are considered throughout the experimental study: calendar features, lagged load observations, and exogenous weather features.

Calendar features. Two calendar features are considered, the hour of the day and the day of the week. Features spanning longer time scales, such as the month or the year, were not adopted. Because the forecasting task addressed in this work is one-step-ahead, predicting consumption in the single hour that immediately follows the last observation, such broad temporal indicators carry no useful information at a temporal resolution this fine. Furthermore, given that one of the predictive models employed is a multivariate statistical model (Section 4.3), a raw integer encoding of the retained features would introduce artificial discontinuities, namely the abrupt transition from hour 23 to hour 0, or from day 7 to day 1 of the week. To preserve the inherent periodicity of these features, both were additionally subjected to a cyclic transformation that maps each value onto the unit circle,

$$x_{\sin} = \sin\left(\frac{2\pi x}{P}\right), \quad x_{\cos} = \cos\left(\frac{2\pi x}{P}\right), \quad (4.2)$$

where x is the raw value and P is its period ($P = 24$ for the hour and $P = 7$ for the day of the week). This encoding guarantees that temporally adjacent values remain close in the feature space, regardless of the wrap-around boundary. Applied to the hour and to the day of the week, this yields the four features `hour_sin`, `hour_cos`, `dow_sin` and `dow_cos`.

Lagged features. The lagged features combine two complementary sources. The first are the short lags that exhibited the strongest spatio-temporal dependence in the partial cross-correlation analysis of Section 3.3.2, namely *lag 1* and *lag 2*, the values one and two hours before the instant being predicted. The second are the lags of 24 and 168 hours, one day and one week, which were identified as dominant in the temporal domain through the time series decomposition and

the autocorrelation analysis of Section 3.2.2. The predictor set therefore relies on lags 1, 2, 24, and 168. Depending on the scenario, these lags are drawn either exclusively from the series being predicted, in the univariate baseline scenarios, or from every series belonging to the same hierarchical level, in the multivariate spatio-temporal scenarios.

Exogenous feature. The single exogenous feature considered is temperature, a major driver of consumption behaviour, as established in Section 3.3.1. The temperature data are in fact temperature forecasts produced by the MeteoGalicía Numerical Weather Prediction (NWP) model, accessed through the Sentinel API provided by INESC TEC. For each four-digit postal code, the forecast was extracted at the corresponding geographical centroid, with a spatial grid resolution of 4 km for northern and central Portugal and 12 km for southern Portugal. From the hourly forecast series, two derived features were computed. The first, `temp_2d`, is a 48-hour trailing moving average, and the second, `temp_3d`, is the equivalent over 72 hours. These features capture thermal inertia at two distinct time scales and are supplied as input to the forecasting models.

4.2.2 Scenario Design

The forecasting scenarios are organised into two groups that differ fundamentally in whether they exploit the spatio-temporal dynamics of the dataset. The *baseline scenarios* deliberately ignore these dynamics. Each series is predicted solely from its own past, in a univariate fashion, so that they may serve as a reference against which the benefit of modelling cross-series interactions can be quantified. The *spatio-temporal scenarios*, in contrast, predict each series using the lagged values of all series within the same hierarchical level, in a multivariate fashion, thereby leveraging the interdependencies revealed in the exploratory analysis. Two baseline scenarios and six spatio-temporal scenarios are defined, for a total of eight, as listed in Table 4.1.

The two baseline scenarios are the following. The first uses only the autoregressive lags of each series to predict that same series, providing the most parsimonious reference possible. The second augments this configuration with the cyclic temporal features, hour and day of the week, thereby isolating the contribution of calendar information in the absence of any cross-series dependence.

The six multivariate scenarios all share the cross-series lag structure and differ in the additional features they incorporate. They fall into two groups. The three spatio-temporal scenarios progressively enrich the predictor set with the available temporal and exogenous features, the first adding the cyclic hour, the second adding the cyclic hour and the cyclic day of the week, and the third further adding the temperature features (`temp_2d` and `temp_3d`).

The three hierarchical scenarios build upon the best-performing configuration among the spatio-temporal ones (**S1** to **S3**) and extend it with information propagated from the immediately lower level of the hierarchy. The forecasts supplied by that level are themselves taken from the scenario with the lowest validation error there, considered among both the spatio-temporal and the hierarchical scenarios defined at that level, so that what is propagated upward is the most accurate forecast the level below can provide. In both cases the selection is performed strictly on validation data. As detailed in Section 4.3.1.1, each scenario generates out-of-fold (OOF) predictions over

Table 4.1: Forecasting scenarios evaluated in this work. The scenarios progressively incorporate temporal, spatial and hierarchical information to assess the contribution of each source of predictive information.

Scenario	Group	Own lags	Spatial lags	Calendar	T (°C)	Hierarchical
B1	Baseline	✓	✗	✗	✗	✗
B2	Baseline	✓	✗	HoD + DoW	✗	✗
S1	Spatio-temporal	✓	✓	HoD	✗	✗
S2	Spatio-temporal	✓	✓	HoD + DoW	✗	✗
S3	Spatio-temporal	✓	✓	HoD + DoW	✓	✗
H1	Hierarchical	✓	✓	as in S*	as in S*	LL forecasts
H2	Hierarchical	✓	✓	as in S*	as in S*	LL aggregate forecast
H3	Hierarchical	✓	✓	as in S*	as in S*	LL lags

✓=Yes ✗=No HoD=hour of the day sin and cos encoders DoW=day of the week sin and cos encoders
 LL=Lower-level, supplied by the scenario of that level with the lowest validation error, considered among
S1 to **S3** and **H1** to **H3** S*=the configuration of **S1** to **S3** selected on validation data

the seven training months through the leave-one-month-out procedure, and the scenario whose OOF predictions minimise the validation error metric (Section 4.4) is then retained as the best. Because this comparison relies exclusively on the training period, the test period remains entirely held out for the final evaluation.

Starting from this selected configuration, the three remaining scenarios each add a different form of cross-level information drawn from the immediately lower level of the hierarchy, and are presented below in order. Scenario **H1** incorporates the forecasts produced for the individual series of the lower level directly as features.

Scenario **H2** relies on a lower-level *aggregate forecast* which, rather than a naive bottom-up sum of the lower-level forecasts, is itself the output of a dedicated aggregation model. For a given level ℓ , this aggregation model is a separate LASSO-regularised regression, estimated with the same solver, whose predictor matrix is composed of the forecasts of all series at the level immediately below, $\ell - 1$, and whose target is the actual consumption of each series at level ℓ . The model therefore learns a data-driven mapping from the lower-level forecasts to the current-level consumption, assigning each lower-level forecast its own weight through the sparsity induced by the LASSO penalty, instead of combining them by a fixed summation. The forecast produced by this aggregation model for each current-level series is then added as one extra predictor to the direct model at level ℓ , the model that already relies on the cross-series lags of level ℓ together with the calendar and exogenous features of the selected **S1** to **S3** configuration. The direct model is thus given, alongside its autoregressive and calendar inputs, an independent estimate of the very quantity it is trying to predict, one derived exclusively from the forecasts of the level below. This allows the model to lean on the bottom-up signal whenever the information carried by the lags

alone is insufficient, effectively blending the level's own autoregressive dynamics with evidence propagated upward from the finer spatial resolution. This aggregation mechanism is applied consistently at every level of the hierarchy where a level below exists, the four-digit to two-digit case being one instance.

Finally, Scenario **H3** instead adds the lagged values of the lower-level series as features. Across these scenarios, to prevent information leakage, any forecast-derived feature used over the training period is computed out-of-fold, consistent with the procedure of Section 4.3.1.1. This applies to the individual lower-level forecasts of Scenario **H1**, as well as to the lower-level forecasts feeding the aggregation model and the predictions the aggregation model itself produces in Scenario **H2**.

4.3 Forecasting Approaches

Having defined the forecasting problem in Section 4.1 and the feature configurations that distinguish the eight scenarios in Section 4.2, this section presents the two forecasting approaches and details their implementation. The two approaches mirror the distinction drawn in the review of the state of the art (Section 2.2), with point forecasting producing a single estimate of consumption for each series and time step, and probabilistic forecasting instead characterising the predictive uncertainty by estimating a set of conditional quantiles.

Each approach is presented in the same order, beginning with the setup that establishes how the data are partitioned and how predictions are generated within the training period, and continuing with the generation of the forecasts themselves. For each of them the data flow is described end to end, from the construction of the predictor and target matrices, through normalisation and hyperparameter optimisation, to the production of the forecasts at every level of the hierarchy. The point-forecasting pipeline is described first, followed by the reconciliation step that enforces coherence across the hierarchical structure, and the probabilistic pipeline afterwards.

4.3.1 Point Forecasting

4.3.1.1 Forecasting Setup

This subsection describes the framework within which the point forecasting models are trained, validated and evaluated, covering the temporal structure of the study, the roles assigned to each data partition, and the procedure used to generate predictions within the training period without incurring information leakage. Although the temporal split established here is common to both forecasting approaches and is reused unchanged by the probabilistic one in Section 4.3.2.1, the roles assigned to the two partitions differ between them, which is why each approach opens with a setup of its own.

Temporal train/test split. The dataset spans eleven months, from November 2022 to September 2023. Given this limited temporal coverage, the choice of a train/test split must balance two

competing objectives: allocating sufficient data to the training period for robust model fitting and hyperparameter tuning, while reserving a test period representative enough for meaningful out-of-sample evaluation. A chronological split was adopted, consistent with the temporal ordering of time series data, where training must always precede the final test set. The training period covers the first seven months, from November 2022 to May 2023, and the test period corresponds to the remaining four months, from June to September 2023, as illustrated in Figure 4.1.

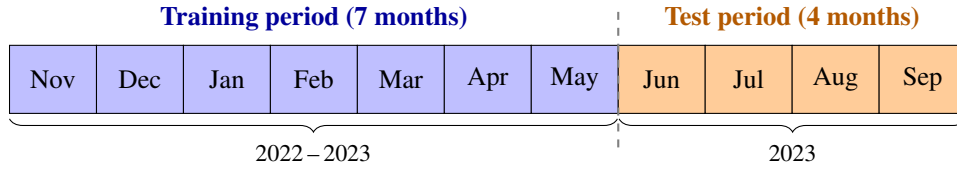


Figure 4.1: Temporal train/test split of the eleven-month dataset.

The primary motivation for this 7:4 partition is the short duration of the dataset itself. With only eleven months of observations, a conventional split would leave barely two months for evaluation. Moreover, the test period deliberately encompasses all four summer months — characterised by higher temperatures, school breaks, and vacation periods, as discussed in Section 3.5 — making it the most demanding segment for model generalisation.

The test period is treated as a strictly held-out partition: no hyperparameter tuning or modelling decision is informed by it prior to final evaluation. For the final evaluation, hyperparameters are selected through an inner leave-one-month-out search conducted exclusively on the seven training months, after which the model is fitted on the entire training set using the optimal configuration, and finally applied to the four test months.

Out-of-fold prediction generation. In addition to model fitting and hyperparameter selection, the training period is used to generate out-of-fold (OOF) predictions. These predictions are essential because the proposed methodology uses forecasts from level ℓ as input features for the model at level $\ell + 1$. Therefore, if the cross-level features were produced by a model trained on the same observations being predicted, the result would be in-sample features — values unrealistically accurate compared to what the model encounters in deployment, causing leakage and inflated apparent performance. To avoid this, each training-period prediction used as a feature must come from a model that never observed the corresponding month, which is precisely what the LOMO strategy achieves, as illustrated in Figure 4.2.

In each fold k , month m_k is designated as the held-out period and a model is fitted on the remaining $K - 1$ months, which then generates predictions for m_k . Repeating this for every month yields a complete set of OOF predictions, each produced by a model that never observed the corresponding month during training. With seven training months, this results in seven folds and seven distinct model fits per hierarchical level. Hyperparameter optimisation is also performed within each fold, as detailed in Figure 4.3. Let M_k denote the set of $K - 1$ months available in fold k . For each candidate configuration HP, an inner leave-one-month-out loop is run over M_k : for

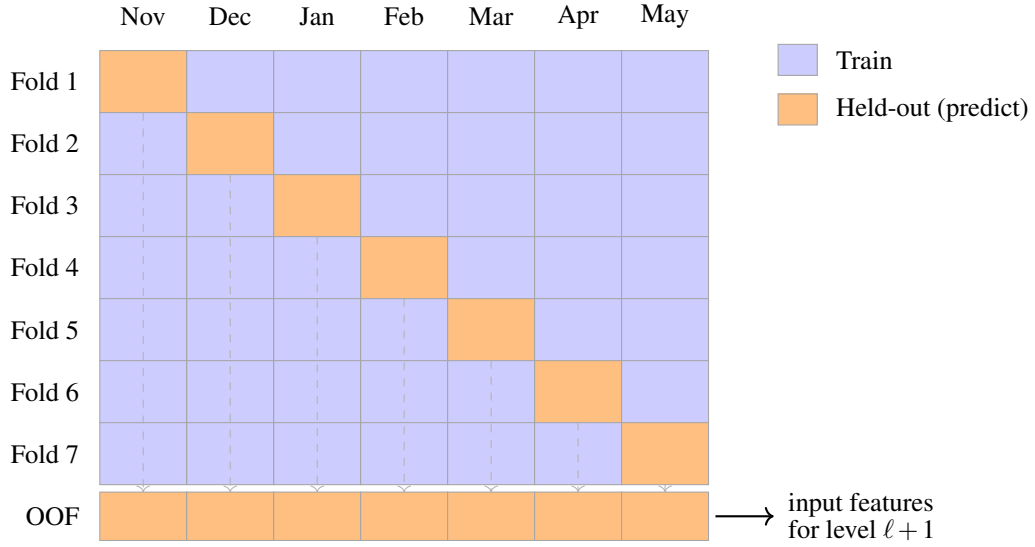


Figure 4.2: LOMO cross-validation scheme for generating OOF predictions over the training period.

each month $m \in M_k$, a model is fitted on $M_k \setminus \{m\}$ and evaluated on m , yielding a validation metric $L_m(\text{HP})$. The optimal configuration is then

$$\text{HP}^* = \arg \min_{\text{HP}} \sum_{m \in M_k} L_m(\text{HP}), \quad (4.3)$$

where $L_m(\text{HP})$ is a validation metric. The final model for fold k is then fitted on all months in M_k using HP^* before predicting the held-out month m_k .

Inside each LOMO fold (example: Fold 1, held-out: Nov)

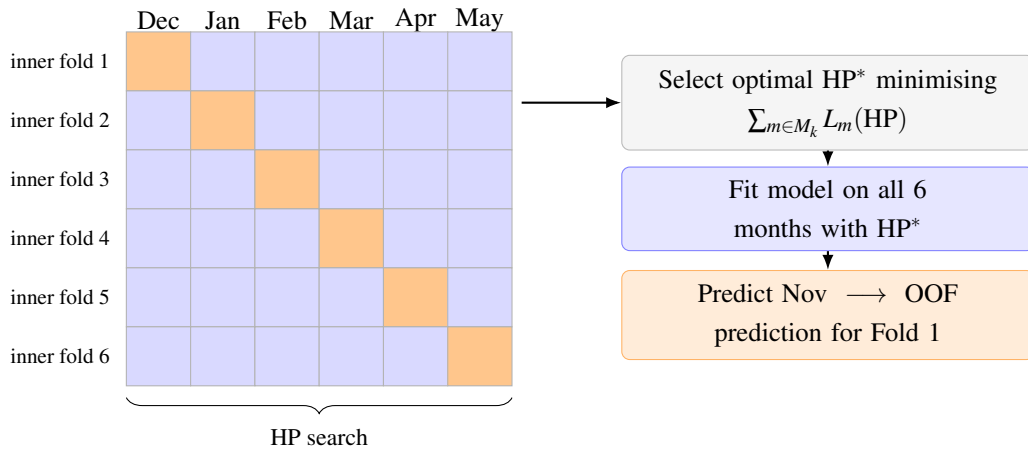


Figure 4.3: Procedure per LOMO fold: inner HP search, model fitting, and OOF prediction.

This procedure is applied consistently across every point-forecasting model and every scenario considered in this work. The LOMO strategy makes efficient use of a limited training period,

since every month contributes to the fitting stage of six folds and is the target of exactly one OOF prediction, and it accounts for the possibility that optimal model settings differ across months, a relevant consideration given the heterogeneous load dynamics described in Section 3.5.

4.3.1.2 Base Forecast Generation

The point forecasting methodology is based on the LASSO-VAR-X model introduced in Section 2.2.2. The formulation extends the classical vector autoregressive model in two directions, incorporating an ℓ_1 regularisation term that allows simultaneous multivariate forecasting and automatic feature selection, and admitting exogenous regressors alongside the lagged loads, namely the calendar features, the temperature and, in the hierarchical scenarios, the information propagated from the level below. Among the eight scenarios only the purely autoregressive baseline **B1** dispenses with those regressors entirely, so the exogenous block is what distinguishes one scenario from another. Both properties are advantageous in the present study, where the number of candidate predictors grows rapidly with the number of spatial series and lagged observations.

For each hierarchical level, the model predicts all consumption series simultaneously. The predictor matrix $\mathbf{Z}$ contains the lagged observations of every series at the same aggregation level together with the calendar and, where applicable, temperature features defined in Section 4.2. The target matrix $\mathbf{Y}$ contains the corresponding consumption values at the forecasting instant. The model is estimated by solving

$$\hat{\mathbf{B}} = \arg \min_{\mathbf{B}} \left(\frac{1}{2} \|\mathbf{Y} - \mathbf{ZB}\|_2^2 + \lambda \|\mathbf{B}\|_1 \right), \quad (4.4)$$

where the first term minimises the prediction error and the second promotes sparse solutions by shrinking the least informative coefficients towards zero. The optimisation problem is solved using the Alternating Direction Method of Multipliers (ADMM), whose behaviour is governed by the regularisation parameter λ and the augmented Lagrangian parameter ρ .

Figure 4.4 summarises, in a generalised form, the sequence of steps shared by all the scripts that produce the base forecasts, regardless of the hierarchical level or the scenario. The blocks of the diagram are detailed below, taking as an example the forecast at the four-digit level under scenario **S2**, which forecasts consumption from the cross-series lags and the calendar features.

Construction of $\mathbf{Z}$ and $\mathbf{Y}$. Let $y_i(t)$ denote the consumption of series i at hour t , with $i = 1, \dots, N$ and $N = 464$ series at the four-digit level. The target matrix $\mathbf{Y} \in \mathbb{R}^{T \times N}$ contains, in each row, the vector of consumption of all series at that instant, and is the same for every scenario. What distinguishes one scenario from another is the composition of the predictor matrix $\mathbf{Z}$, which is illustrated below for the three cases that arise in this work. The rows with missing values resulting from the construction of the lags are discarded in all of them.

In scenario **S2** each row of $\mathbf{Z}$ is formed by four lag blocks of all series, at $t - 1$, $t - 2$, $t - 24$ and $t - 168$, which were identified in Sections 3.3.2 and 3.2.2, followed by the cyclic components

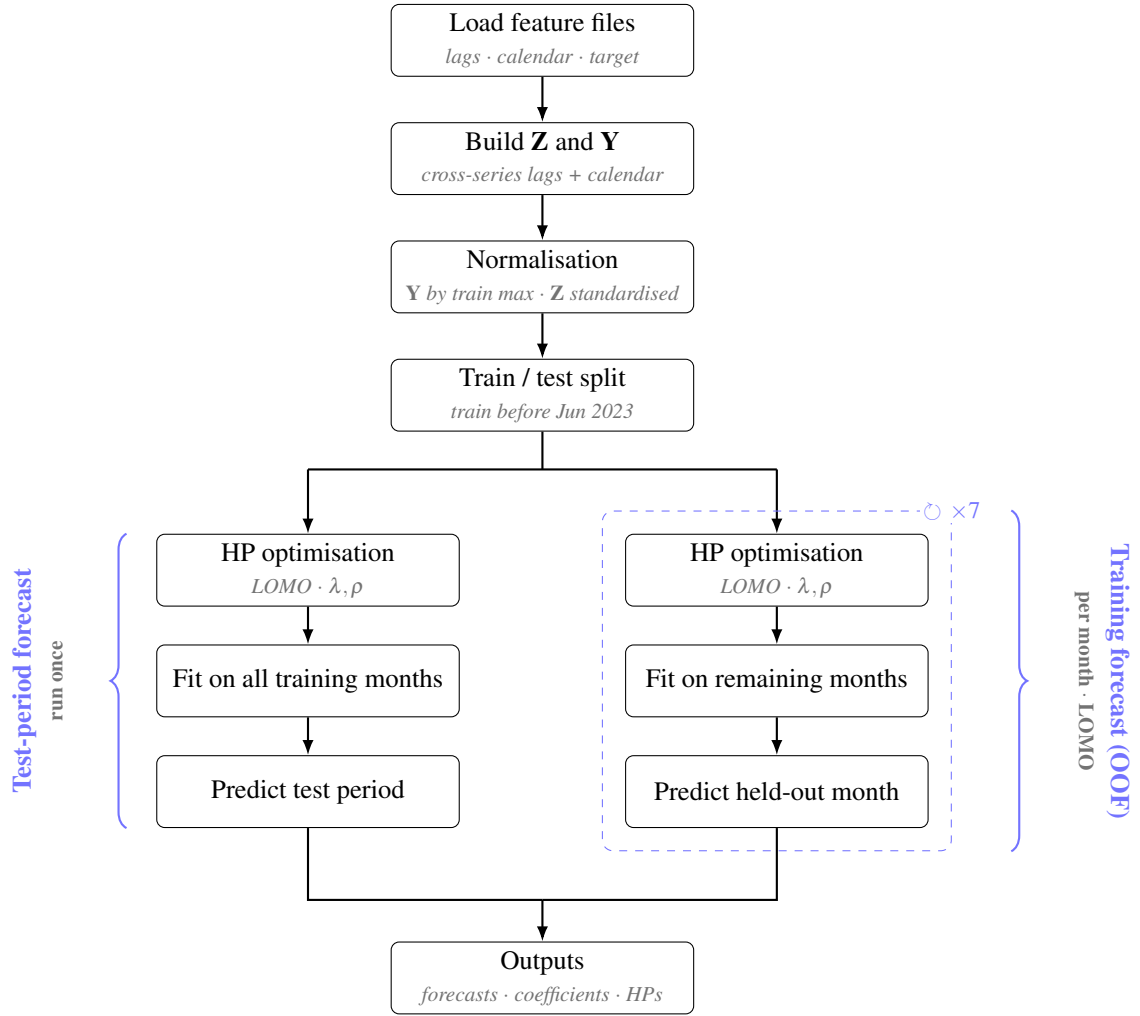


Figure 4.4: Generalised LASSO-VAR-X pipeline for the base forecasts at a given hierarchical level.

of the hour and of the day of the week defined in Equation 4.2. The predictor row is therefore

$$\mathbf{Z}_t = \left[\underbrace{y_1(t-1) \cdots y_N(t-1)}_{\text{lag 1}} \mid \underbrace{y_1(t-2) \cdots y_N(t-2)}_{\text{lag 2}} \mid \cdots \mid \underbrace{y_1(t-168) \cdots y_N(t-168)}_{\text{lag 168}} \mid \underbrace{s_t^h \ c_t^h \ s_t^w \ c_t^w}_{\text{calendar}} \right], \quad (4.5)$$

where s_t^h, c_t^h and s_t^w, c_t^w are the sine and cosine of the hour and of the day of the week. The dimension of $\mathbf{Z}$ is therefore $T \times (4N + 4)$, and the model is written compactly as $\mathbf{Y} = \mathbf{Z}\mathbf{B} + \mathbf{E}$, with $\mathbf{B} \in \mathbb{R}^{(4N+4) \times N}$.

The hierarchical scenarios keep this structure and append information propagated from the level immediately below. In scenario **H1** the forecasts produced for the individual series of that level are added as further columns, so the predictor row becomes

$$\mathbf{Z}_t = \left[\underbrace{\text{lag blocks of level } \ell}_{\text{as in scenario S2}} \mid \underbrace{s_t^h \ c_t^h \ s_t^w \ c_t^w}_{\text{calendar}} \mid \underbrace{\hat{y}_1^{(\ell-1)}(t) \cdots \hat{y}_M^{(\ell-1)}(t)}_{\text{level } \ell-1 \text{ forecasts}} \right], \quad (4.6)$$

with $\mathbf{Y}$ remaining the actual consumption of level ℓ . For the training period the forecasts $\hat{y}^{(\ell-1)}$ used are the out-of-fold ones, consistent with the procedure of Section 4.3.1.1, which is what prevents information leakage.

Scenario **H2** requires an intermediate model, and is the only case in which $\mathbf{Y}$ also changes. An aggregation model is first fitted as a LASSO-regularised regression estimated with the same solver, taking as predictors the forecasts of *all* series of the level below and as target the actual consumption of level ℓ ,

$$\mathbf{Z} = [\hat{y}_1^{(\ell-1)} \dots \hat{y}_M^{(\ell-1)}], \quad \mathbf{Y} = \mathbf{y}^{(\ell)}, \quad (4.7)$$

so that it learns a direct mapping from the lower-level forecasts to the upper-level consumption, weighting each one through the LASSO sparsity instead of combining them by a fixed sum. Its output is then appended as a single additional column to the matrix of Equation 4.5, giving the direct model of level ℓ one independent estimate of the quantity it is predicting.

Normalisation. Before fitting, both matrices are scaled. The targets are normalised by dividing each series by its maximum over the training period, which places series of widely different magnitudes, from rural to urban postal codes, on a comparable scale and prevents the higher-consumption series from dominating the fitting error. The predictors are standardised with a `StandardScaler` fitted on the training data only, so that the LASSO penalty in (4.4) acts evenly across all coefficients, regardless of the original units of each feature. In both cases, the normalisation statistics are estimated exclusively on the training data and then applied to the test period, avoiding any information leakage.

Test- and training-period forecasts. As illustrated in Figure 4.4, the forecasts are produced through two branches. In the test branch, the hyperparameters (λ, ρ) are optimised over the whole training period, the model is refitted with the optimal configuration on all training months, and finally applied to the test period. In the training branch, out-of-fold (OOF) predictions are generated through the leave-one-month-out (LOMO) scheme described in Figure 4.2. Each of the seven training months is left out in turn, the hyperparameters are re-optimised on the remaining six months, and the model predicts the left-out month. This inner procedure is therefore repeated seven times, once per training month, as indicated by the $\odot \times 7$ loop in Figure 4.4, and yields OOF predictions that span the entire training period. These predictions are indispensable to the later stages, namely the reconciliation and the scenarios that reuse forecasts from other levels, as they constitute training-period forecasts free from overfitting contamination.

Hyperparameter optimisation. The optimisation of (λ, ρ) , common to both branches, is carried out by Bayesian optimisation (Gaussian process, `gp_minimize`), with 30 random initial points followed by guided evaluations up to a total of 50. Each candidate (λ, ρ) is evaluated by the mean absolute error (MAE) of an inner leave-one-month-out validation over the training months available in the fold, as detailed in Figure 4.3. The search space is summarised in Table 4.2.

Table 4.2: Search space of the LASSO-VAR-X hyperparameters optimised by Bayesian optimisation.

Hyperparameter	Range	Scale
λ (LASSO penalty)	$[10^{-3}, 50]$	log-uniform
ρ (ADMM penalty)	$[10^{-3}, 50]$	log-uniform

Parallelisation. The procedure exploits parallelism at two points. In the test forecast, the evaluations of the Bayesian optimisation are distributed across all available cores. In the training forecast, the LOMO folds, one per month, are processed in parallel, each with its own sequential inner optimisation.

Outputs. On completion, each script writes a fixed set of files, illustrated here for the two-digit level.

- The test-period and OOF training-period forecasts, stored alongside their corresponding ground-truth values, with one column per series, in de-normalised (real) units.
- The estimated coefficient matrix $\hat{\mathbf{B}}$ and the auxiliary sparse matrix produced by the ADMM solver, saved with one row per predictor, the intercept followed by each lagged and calendar feature, and one column per target series. Their entries refer to the normalised and standardised quantities on which the model was fitted.
- A summary file recording the selected hyperparameters (λ, ρ) , the test and training error metrics (MAE, RMSE and MAPE) on the real scale, the execution times, and the mapping between each aggregate code and its constituent series.
- A per-fold validation record reporting, for each of the seven training months, the λ and ρ chosen in that fold together with the corresponding validation MAE. This validation MAE is expressed in normalised units, since it is computed on the normalised targets during the inner search, and is therefore not directly comparable in magnitude to the real-scale error metrics reported in the summary file.

The pipeline described so far, and summarised in Figure 4.4, applies unchanged to every hierarchical level and to every scenario. As shown above, the scenarios differ only in the composition of $\mathbf{Z}$, while the model, the normalisation, the hyperparameter optimisation and the train/test procedure remain identical throughout.

4.3.1.3 Reconciliation

Once the base forecasts have been produced, the need arises to make them coherent across the hierarchy, as discussed in Section 2.2.5. Since each level is forecasted independently, the resulting predictions are not aggregation-consistent: the forecast of a two-digit code does not, in

general, equal the sum of the forecasts of its constituent four-digit series. Reconciliation is the post-processing step that restores this consistency while, at the same time, exploiting the information shared across levels to improve accuracy.

The structure of the hierarchy, which runs from four-digit to two-digit to one-digit and finally to the national level, was described in Section 3.4. It is encoded in a summing matrix $\mathbf{S} \in \{0, 1\}^{m \times n}$, built deterministically from the postal-code prefixes, where $n = 464$ is the number of bottom-level four-digit series. The total number of nodes m is obtained by adding the series of all four levels, namely the 464 four-digit codes, the 69 two-digit codes, the 8 one-digit codes and the single national series, which gives $m = 464 + 69 + 8 + 1 = 542$. A coherent forecast vector is one that can be written as $\tilde{\mathbf{y}} = \mathbf{S}\tilde{\mathbf{b}}$ for some bottom-level vector $\tilde{\mathbf{b}}$, that is, one in which every aggregate equals the sum of its components.

Let $\hat{\mathbf{y}} \in \mathbb{R}^m$ denote the stacked base forecasts of all nodes, which are in general incoherent. Reconciliation projects $\hat{\mathbf{y}}$ onto the coherent subspace through

$$\tilde{\mathbf{y}} = \mathbf{S} \underbrace{(\mathbf{S}^\top \mathbf{W}^{-1} \mathbf{S})^{-1} \mathbf{S}^\top \mathbf{W}^{-1}}_{\mathbf{G}} \hat{\mathbf{y}}, \quad (4.8)$$

where $\mathbf{W} \in \mathbb{R}^{m \times m}$ is the covariance matrix of the base-forecast errors. The product $\tilde{\mathbf{b}} = \mathbf{G}\hat{\mathbf{y}}$ is a single, *corrected* set of bottom-level forecasts, and the remaining levels are recovered by summing this corrected base back up the hierarchy, $\tilde{\mathbf{y}} = \mathbf{S}\tilde{\mathbf{b}}$. The key point is that $\mathbf{G}$ builds the corrected base by drawing on the base forecasts of *every* level simultaneously, weighted by $\mathbf{W}^{-1}$: information from the upper levels flows downward to refine the bottom level, and the refined base then propagates upward, improving both the coherence and, in general, the accuracy of the whole hierarchy. In implementation, Equation (4.8) is solved numerically without ever forming the matrix inverse explicitly.

As presented in Section 2.2.5, the reconciliation method is determined by how the weighting matrix $\mathbf{W}$ is built, and three variants are commonly used, namely OLS, WLS and MinT(Shrink). Of these, only WLS and MinT(Shrink) are adopted in this work. OLS sets $\mathbf{W} = \mathbf{I}$ and therefore assumes the base-forecast errors to be homoscedastic, which is untenable in this setting, since the node magnitudes span several orders of magnitude and a forecast error at the national level cannot be weighted as heavily as one at a single four-digit code. WLS and MinT(Shrink) both account for this heteroscedasticity and differ only in whether the error correlations between nodes are modelled, which is precisely the aspect this work sets out to test. WLS ignores those correlations and relies solely on the error variance of each node, whereas MinT(Shrink) uses the full error covariance, regularised with a Schäfer–Strimmer shrinkage estimator so that it remains well-conditioned and invertible for the 542 nodes.

Finally, the matrix $\mathbf{W}$ is estimated from the out-of-fold residuals of the training period, $\mathbf{r} = \mathbf{y}_{\text{train}} - \hat{\mathbf{y}}_{\text{train}}$, computed from the OOF forecasts produced in the previous stage. This is precisely why those forecasts were required: an in-sample residual would understate the true forecast error and bias the weighting, whereas the OOF residuals provide an honest estimate of the error covariance. The covariance is estimated on the training period and then applied, unchanged, to reconcile

the base forecasts of the test period.

4.3.2 Probabilistic Forecasting

4.3.2.1 Forecasting Setup

The probabilistic approach adopts the same temporal division of the dataset established for the point forecasts in Section 4.3.1.1 and illustrated in Figure 4.1, but it assigns a different role to each partition. In point forecasting the models are estimated once over the entire training period and then applied unchanged to the four test months. In probabilistic forecasting the training period is used only to select the predictors and to tune the hyperparameters, and the forecasts themselves are produced by models that are re-estimated on a window which advances through the test period, one day at a time.

For every day of the test period a separate model is estimated on the W days that immediately precede it, and that model forecasts the twenty-four hours of that day. The window then advances by one day and the procedure is repeated, so that over the four test months 122 models are estimated for each series and each quantile. The amount of history used for training is known to affect the quality of probabilistic renewable energy forecasts, and has been reported to degrade them as the training period is shortened [1], so the window length W is treated here as a parameter of the method rather than fixed beforehand, and values from 90 to 210 days were evaluated. The upper limit of that interval is imposed by the extent of the dataset itself. The mechanism is illustrated in Figure 4.5, which shows three of the 122 forecast days, with the window progressively leaving the training months behind and coming to be dominated by observations from the test period.

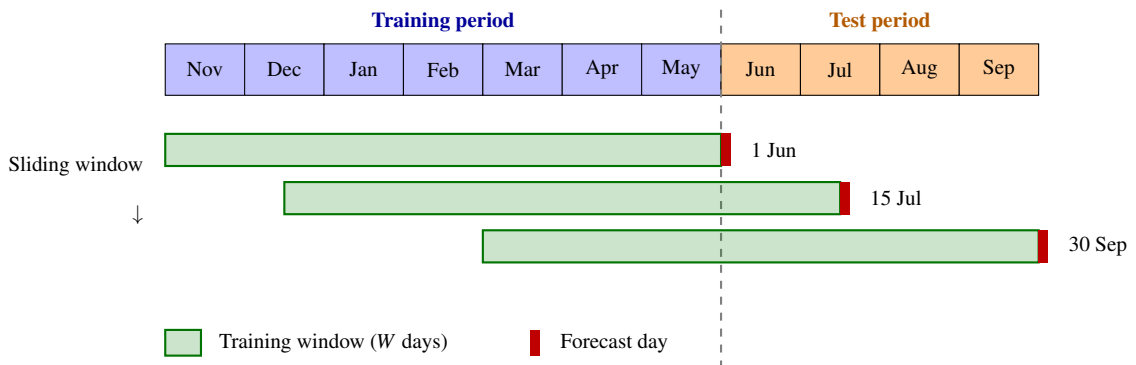


Figure 4.5: Sliding training window at three instants of the test period, for $W = 210$ days.

Because the window advances, the data used to estimate the later models includes days that belong to the test period. This does not constitute information leakage, and the guarantee that replaces the conventional held-out block is stated in two parts. First, the model that produces the forecasts for a given day is estimated only on observations strictly earlier than that day. Second, each individual forecast for hour t uses only observations up to hour $t - 1$, which corresponds to a one-hour-ahead horizon, so the advance of the window by one day is a retraining cadence and

not a forecast horizon. Together the two conditions reproduce the circumstances of operational forecasting, in which the load observed up to the present instant is always known.

The seasonal mismatch between a training period confined to the colder months and a test period covering the entire summer is imposed by the eleven months of data available, and its consequences are far more severe for a probabilistic forecast than for a point forecast. A conditional mean is supported by the bulk of the observations and can be interpolated from a training period that resembles the test period only in part, so a mismatch of this kind leaves it with a bounded error. The extreme quantiles, in contrast, are determined by the infrequent observations in the tails and cannot be recovered from peaks that the training period never contained, so estimating them across a change of regime yields unstable and internally inconsistent estimates rather than a mere loss of accuracy. Estimating the models on a window that advances with time lets the summer observations enter the training set as they become available, an effect that should therefore be concentrated in the upper quantiles. The arrangement follows the one used in [1] for probabilistic forecasting of renewable energy, where the window advances by five months at a time, the length of the test period, while here it advances by one day. Whether the moving window repays its cost against a single static fit under otherwise identical conditions is settled empirically in Section 5.3.

The moving window was investigated in the probabilistic strand alone, since a daily re-estimation of the point-forecasting pipeline would require its cross-level chain of out-of-fold forecasts to be regenerated for each of the 122 test days. The point-forecast results therefore correspond to the static regime, and extending the window to that approach remains open.

Cluster representatives. Since tuning the hyperparameters per series is impractical where 464 and 69 series are forecast, the series are grouped and each group's medoid, the one most correlated on average with the others, carries out the search while the members inherit its configuration. The four-digit grouping of Section 3.3 at a Pearson threshold of 0.6 is subdivided into clusters of at most twenty series, giving 227 representatives, and the smoother two-digit series are clustered at a higher threshold of 0.7 and subdivided likewise, giving 26. The eight one-digit series are each their own representative. Only the hyperparameters are inherited, as every series is still fitted and forecast on its own data and selects its own predictors.

The predictive distribution. The predictive distribution is summarised by nine quantiles, at the levels 0.05, 0.10, 0.20, 0.35, 0.50, 0.65, 0.80, 0.90 and 0.95. The model estimates one quantile at a time, so each level is fitted independently of the others.

Searching for hyperparameters at all nine levels would multiply the cost of the tuning stage by three without a corresponding benefit, since adjacent quantile levels are estimated from overlapping regions of the distribution and their optimal configurations are close. The search is therefore restricted to three levels, 0.10, 0.50 and 0.90, referred to hereafter as the *pivot quantiles*, and each of the remaining six levels inherits the configuration of the pivot nearest to it, as shown in Figure 4.6. The three pivots are chosen so that each one anchors a group of three levels covering the lower tail, the centre and the upper tail of the distribution.

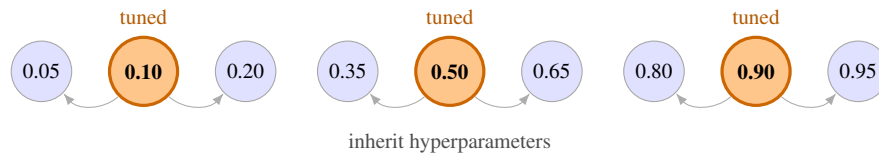


Figure 4.6: Quantile levels of the predictive distribution and inheritance of hyperparameters from the pivot quantiles.

Selection of the predictors. The predictors are selected once over the training period, and the resulting set is then held fixed for every model estimated afterwards. The procedure concerns only the multivariate scenario, whose candidates are the lags of all the series at the hierarchical level, and is carried out independently for each series. The selector is a median quantile regressor in the default configuration of the library, deliberately left untuned so that the ranking is not driven by an over-fitted model, and the relevance of each candidate is measured by permutation importance scored with the pinball loss. Candidates whose relative importance reaches at least 1% are retained, a threshold that discards the many cross-series lags carrying no usable signal for a given series.

Validation at this stage uses a single chronological block covering the last 20% of the training period, shown in Figure 4.7, which differs deliberately from the LOMO arrangement used for the hyperparameters. The single block was retained because the two schemes were compared under otherwise identical conditions and the difference between them did not justify the additional cost of the LOMO alternative, as reported in Section 5.3.

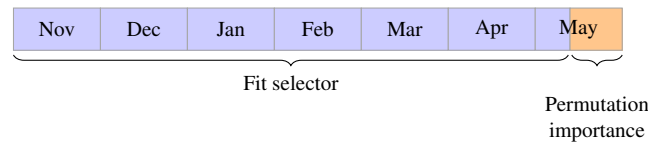


Figure 4.7: Validation block used for predictor selection within the training period.

Optimisation of the hyperparameters. The hyperparameters are tuned by a LOMO search over the seven training months, shown in Figure 4.8. In each fold one month is held out, the model is fitted on the remaining six and the pinball loss at the quantile being tuned is measured on the held-out month, and a candidate configuration is scored by the mean of the seven resulting losses.

Unlike the point-forecasting pipeline, this loop serves a single purpose, namely the selection of the hyperparameters, and produces no out-of-fold predictions. It is run once for each pivot quantile and only for the representatives defined above.

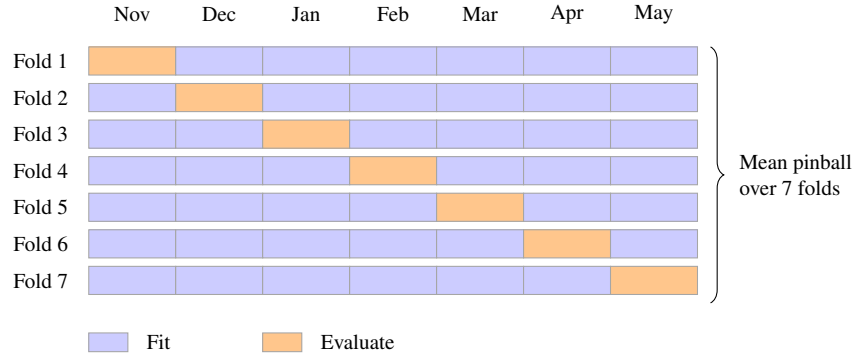


Figure 4.8: Leave-one-month-out scheme used for the hyperparameter search.

4.3.2.2 Forecast Generation

The probabilistic forecasts are produced with LightGBM, a gradient-boosted decision tree model of the family reviewed in Section 2.2.3, configured for quantile regression. The choice rests on properties established in the review of the state of the art. Tree ensembles are natively probabilistic, since a conditional quantile is obtained by replacing the training loss rather than by altering the architecture (Section 2.2.4), and they are repeated benchmark winners in energy forecasting competitions while remaining far cheaper to train than the deep architectures that rival them (Section 2.2.3). That last property is decisive here rather than incidental, because the sliding regime re-estimates every model once per forecast day, so the training cost is multiplied by the number of days and constrains the choice of algorithm far more than it would under a static regime.

Each quantile level is estimated by a separate model that minimises the pinball loss of Section 2.3 at that level,

$$\hat{f}_\tau = \arg \min_f \sum_t L^\tau(y_t, f(\mathbf{x}_t)), \quad (4.9)$$

so that the nine quantile levels of Section 4.3.2.1 require nine independent fits for each series and each training window.

This organisation differs from the point-forecasting pipeline in one fundamental respect. There a single LASSO-VAR-X model was fitted jointly to all the series of a level, so that the multivariate information was exploited inside the estimation itself. Here each series is modelled on its own, by its own nine quantile regressors, and the information from the other series enters only through the cross-series lags retained as its predictors, never through a joint fit. The target is therefore the consumption of a single series rather than a matrix of all of them, and the whole procedure is repeated independently for every series of the level.

The two stages shared by all the probabilistic scripts are summarised in Figure 4.9. The first is carried out once over the training period and fixes the predictors and the hyperparameters, while the second produces the forecasts by advancing through the test period in the manner established in Section 4.3.2.1. The blocks are detailed below, taking the four-digit level under the multivariate scenario as the running example.

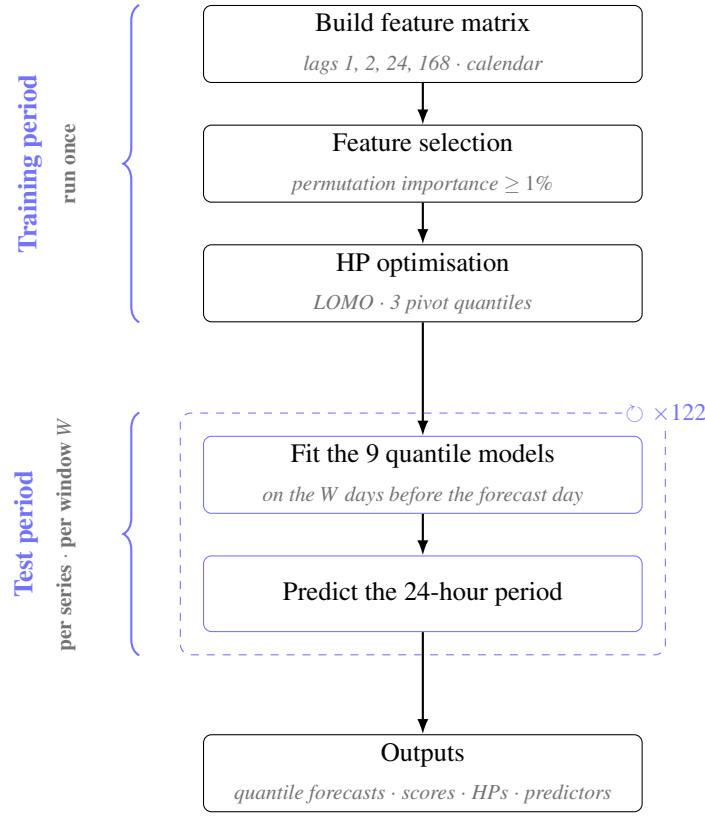


Figure 4.9: Two-stage pipeline of the probabilistic forecasts at a given hierarchical level.

Predictors and target. Let $y_i(t)$ denote the consumption of series i at hour t . For a given series the target is the scalar $y_i(t)$, and its predictors are drawn from a matrix shared by all the series of the level, formed by the four lag blocks at $t - 1$, $t - 2$, $t - 24$ and $t - 168$ of every series, followed by the cyclic hour and the cyclic day of the week. At the four-digit level, with 464 series, this shared matrix holds $4 \times 464 + 4 = 1860$ candidate predictors. The selection of Section 4.3.2.1, shown in Figure 4.7, reduces this set to the small subset that carries usable signal for each series, independently from one series to the next. In the univariate scenario the matrix is not shared and reduces to the eight own features, the four own lags together with the cyclic hour and the cyclic day of the week, which are used without selection. The predictors are not normalised, since a gradient-boosted tree splits each feature by rank rather than by magnitude and is therefore invariant to their scale.

Training period. This first stage runs once and fixes the two ingredients that every later model reuses, namely the predictor set of each series and the hyperparameters of each representative, both obtained as defined in Section 4.3.2.1. Its cost is dominated by the tuning. Each pivot is scored over 20 trials and each trial fits the model on 6 of the 7 training months, so a representative is tuned with $3 \times 20 \times 7 = 420$ fits, while the members of its cluster add nothing to this stage beyond their own selection.

Test period. With those fixed, the sliding window of Section 4.3.2.1 drives the second stage, and the cost moves from the representatives to every series individually. Nine quantile models are refitted for each of the 122 forecast days, so a single series requires $122 \times 9 = 1098$ fits for one window length, and the whole stage is repeated for every length in the swept range from 90 to 210 days. This is where the efficiency of the engine becomes binding rather than convenient, since the same sweep with a model an order of magnitude slower to fit would not have been tractable.

Hyperparameter search space. The hyperparameter search is carried out by the tree-structured Parzen estimator of the Optuna library over the space of Table 4.3, which governs the learning rate, the size and number of the trees and the ℓ_2 regularisation of the leaves.

Table 4.3: Search space of the LightGBM hyperparameters tuned by the tree-structured Parzen estimator.

Hyperparameter	Range	Scale
Learning rate	[0.01, 0.1]	log-uniform
Number of trees	[50, 300]	integer
Number of leaves	[15, 50]	integer
Minimum samples per leaf	[10, 50]	integer
ℓ_2 regularisation	$[10^{-3}, 5]$	log-uniform

Parallelisation. The two stages are parallelised at their natural unit. On the training period the clusters are processed concurrently, each representative running its own sequential search over the pivots. On the test period the series are distributed across the available cores. In both cases the LightGBM fit is held to a single thread, so that the parallelism operates across series rather than inside each model and the cores are not oversubscribed.

Outputs. On completion, and separately for each window length, each script writes a fixed set of files.

- The quantile forecasts over the test period, with one column per quantile level alongside the realised consumption, in real units.
- A metrics file recording, for each series, the pinball loss, the CRPS, the calibration error, the coverage of the 80% and 90% intervals and their average width, together with the cluster of the series and the number of predictors it retained.
- The tuned hyperparameters and the selected predictors of each representative.
- A record of the execution time of the window.

The pipeline described here applies unchanged to every hierarchical level and to both the univariate and the multivariate scenarios. The levels differ only in the number of series and in the clustering that supplies the representatives, and the two scenarios differ only in the predictor set, the shared cross-series matrix with selection for the multivariate case and the eight own features without selection for the univariate one. The model, the tuning and the sliding forecast are identical throughout.

4.4 Evaluation Metrics and Comparison Criteria

The forecasting approaches are assessed with the metrics introduced in Section 2.3, applied according to the nature of each forecast.

For point forecasting, performance is measured with the mean absolute error (MAE), the root mean square error (RMSE), the mean absolute percentage error (MAPE) and the mean absolute scaled error (MASE). The four metrics are complementary. MAE reports the average error magnitude in the units of the load, RMSE emphasises the largest deviations, MAPE expresses the error in relative terms as a percentage of the observed value, and MASE scales the model error by that of a seasonal naive benchmark, which shows whether the model improves on a naive forecast and keeps the comparison scale-independent. These metrics serve three purposes in this work. They quantify the accuracy of each scenario on its own, they allow the scenarios to be ranked against one another, and, most importantly, they support the central comparison of the study, namely the baseline scenarios defined in Section 4.2 against the new spatio-temporal and reconciliation-based approaches developed in this project. This comparison is carried out separately at each hierarchical level, so that any improvement can be attributed to a specific aggregation level.

For probabilistic forecasting, the predictive distribution is summarised by a set of conditional quantiles, and its quality is assessed with two metrics. The pinball loss, also known as the quantile score, measures how well each individual quantile is estimated, applying an asymmetric penalty according to the quantile level. The continuous ranked probability score (CRPS) summarises the quality of the whole predictive distribution in a single value and can be read as a probabilistic counterpart of the MAE. These two metrics are used both to evaluate the probabilistic forecasts internally, across quantiles and series, and to compare them against the corresponding baseline model.

4.5 Chapter Summary

This chapter has formulated the forecasting problem and defined the methodology used to answer it. The problem was stated as a one-step-ahead forecast produced for every spatial unit of every aggregation level, with the predictor vector rather than the model as the object of comparison. Eight scenarios were then defined over three groups of predictors, namely the calendar features in cyclic encoding, the lagged loads at one, two, twenty-four and one hundred and sixty-eight hours, and

the temperature forecasts, giving two univariate baselines, three spatio-temporal configurations and three hierarchical ones.

The point-forecasting pipeline was built around a LASSO-VAR-X model estimated by the alternating direction method of multipliers, with its two parameters selected by Bayesian optimisation under an inner leave-one-month-out validation. The same scheme produces the out-of-fold predictions that allow a forecast generated at one level to feed the model at the level above without leakage. Those forecasts are afterwards made coherent through the projection of Section 4.3.1.3, with ordinary least squares excluded a priori and the weighting matrix estimated from out-of-fold residuals.

The probabilistic pipeline reuses the same temporal split but assigns a different role to each partition, since the models are re-estimated on a window that advances through the test period one day at a time, so that the observations of the season being forecast enter the training set as they become available. The predictive distribution is summarised by nine quantiles estimated independently with gradient-boosted trees, with the tuning restricted to three pivot levels and to the cluster representatives obtained in Chapter 3.

Two properties of this design govern how the following chapter should be read. Every scenario is estimated by the same model and evaluated over the same held-out months, so that any difference in accuracy is attributable to the information supplied rather than to the model itself, and every quantity derived from a forecast and reused elsewhere is generated out of fold. Chapter 5 applies this methodology to the four test months, comparing the scenarios at every level, reconciling the resulting forecasts and repeating the whole comparison in the probabilistic setting.

Chapter 5

Results and Discussion

This chapter presents and discusses the results obtained using the methodology introduced in Chapter 4. Its first objective is to assess whether, and to what extent, the spatial and hierarchical dependencies identified in Chapter 3 improve short-term load forecasting across the hierarchy. Its second objective is to explain the differences observed between aggregation levels by relating the aggregate metrics to the estimated coefficients, the behaviour of individual series, and changes in performance over the test period.

All results refer to the four held-out test period from June to September 2023, which was not used for model fitting or configuration selection. The evaluation covers the four levels of the hierarchy, namely the four-digit, two-digit and one-digit postal codes and the national aggregate. The evaluation follows the two forecasting approaches in turn, so that the point forecasts are assessed before and after hierarchical reconciliation, followed by the probabilistic forecasts. The scenario codes introduced in Section 4.2 are used throughout: **B1** and **B2** denote the univariate baselines, **S1** to **S3** the spatio-temporal scenarios and **H1** to **H3** the hierarchical scenarios.

This chapter is organised as follows. Section 5.1 presents the base point-forecast results produced by the LASSO-VAR-X models and examines their variation across levels and individual series. Section 5.2 evaluates hierarchical reconciliation under weighted least squares and minimum trace. Section 5.3 reports the probabilistic results, focusing on the training regime, scenario comparison, calibration, and temporal behaviour. Finally, Section 5.4 summarises the main findings.

5.1 Base Point Forecast Results

This section evaluates the point forecasts produced by the LASSO-VAR-X models before hierarchical reconciliation, thereby separating the contribution of each forecasting scenario from the effect of the projection applied in Section 5.2. The eight forecasting scenarios defined in Table 4.1 are compared using MAE, RMSE, MAPE and MASE. Since the hierarchical scenarios rely on information propagated from the level below, they are not applicable at the four-digit level.

Rather than proceeding one level at a time, the exposition is organised around three questions asked of the whole hierarchy at once. Section 5.1.1 establishes which scenario is the most accurate at each level, and how far that accuracy is representative of the series it summarises. Section 5.1.2 turns to the estimated coefficients, in order to determine which sources of information each model actually relies upon and whether the spatial dependence it retains matches the one measured in the exploratory analysis. Section 5.1.3 then examines how the advantage of the multivariate scenarios is distributed across the individual series, and identifies the mechanism that makes the cross-series information harmful at the bottom of the hierarchy and useful above it.

Complete validation errors and coefficient inventories for the case-study series are reported in Appendix A.

5.1.1 Overview

Table 5.1 summarises the accuracy of the eight forecasting scenarios across the four levels of the hierarchy. At the four-digit, two-digit and one-digit levels, each entry reports the mean of the metric across the series of that level, followed by the corresponding standard deviation. The national level contains a single series and therefore has no cross-series dispersion. At each level, shading identifies the model family in which the best-performing scenarios are concentrated. It does not imply that one scenario within that family attains the best value for every metric.

At the four-digit level the two baselines perform similarly. Scenario **B2**, which includes calendar information, has a marginal advantage over the purely autoregressive scenario **B1**. By contrast, the three spatio-temporal scenarios are less accurate than both baselines. Relative to **B2**, the best spatio-temporal scenario increases the mean MAE from 365 to 470 kWh, corresponding to a degradation of approximately 29%. The differences among **S1**, **S2** and **S3** are small, indicating that the day-of-week and temperature features provide little additional value once the cross-series lags have been introduced.

The hierarchical scenarios require forecasts from the level immediately below, which is why they are not available at the four-digit level. Among the scenarios eligible to supply them, and for the reason set out in Section 4.3.1.1, the leave-one-month-out validation error over the training period selects scenario **S1** at every level, ahead of scenarios **S2** and **S3** by less than two percent of validation error, so it is scenario **S1** that is propagated upward throughout the hierarchy. The complete validation table is reported in Appendix A.

The ordering changes at the two-digit level. Scenario **B2** remains the better of the two baselines, but all three spatio-temporal scenarios improve upon it. Scenario **S2** attains the lowest MAE, RMSE, MAPE and MASE, reducing the mean MAE from 1871 to 1772 kWh, or by approximately 5%. The day-of-week features therefore provide a small additional gain over **S1**, whereas the inclusion of temperature in **S3** does not improve the result.

None of the hierarchical scenarios outperforms the direct modelling of the two-digit series. The aggregate four-digit forecast used by scenario **H2** is the least harmful, the individual four-digit forecasts of scenario **H1** degrade the error further, and the observed four-digit lags of scenario **H3** give the worst result of the level, more than sixty percent above scenario **S2**. This ordering follows

Table 5.1: Point forecast accuracy across aggregation levels and forecasting scenarios. The model type that attains the best results at each level is shaded. The hierarchical scenarios (**H1** to **H3**) do not apply at the four-digit level.

Level	Model type	Scenario	MAE (kWh)	RMSE (kWh)	MAPE (%)	MASE
Four-digit	Baseline	B1	372 ± 773	520 ± 1150	4.06 ± 2.43	0.45 ± 0.20
		B2	365 ± 781	511 ± 1158	3.97 ± 2.45	0.45 ± 0.21
	Spatio-temporal	S1	478 ± 2191	610 ± 2545	4.88 ± 5.17	0.53 ± 0.29
		S2	471 ± 2036	603 ± 2387	4.87 ± 4.90	0.53 ± 0.29
		S3	470 ± 2170	602 ± 2524	4.72 ± 5.22	0.50 ± 0.29
Two-digit	Baseline	B1	1934 ± 2687	2607 ± 3685	2.88 ± 1.18	0.41 ± 0.20
		B2	1871 ± 2555	2541 ± 3541	2.82 ± 1.22	0.41 ± 0.23
	Spatio-temporal	S1	1778 ± 2660	2326 ± 3510	2.70 ± 1.47	0.38 ± 0.22
		S2	1772 ± 2656	2320 ± 3504	2.69 ± 1.47	0.37 ± 0.22
		S3	1799 ± 2574	2342 ± 3407	2.81 ± 1.36	0.39 ± 0.22
	Hierarchical	H1	2399 ± 3323	3035 ± 4238	3.62 ± 1.48	0.49 ± 0.23
		H2	2199 ± 3365	2796 ± 4278	3.16 ± 1.93	0.42 ± 0.25
		H3	2841 ± 5118	3534 ± 6075	4.18 ± 2.62	0.57 ± 0.33
One-digit	Baseline	B1	13320 ± 12069	17376 ± 15712	2.19 ± 0.37	0.36 ± 0.16
		B2	13340 ± 12068	17400 ± 15705	2.20 ± 0.37	0.36 ± 0.17
	Spatio-temporal	S1	22226 ± 19335	27579 ± 23742	3.86 ± 0.91	0.61 ± 0.21
		S2	21666 ± 18634	26908 ± 22895	3.80 ± 0.90	0.61 ± 0.22
		S3	24548 ± 20875	30009 ± 25304	4.36 ± 1.04	0.69 ± 0.25
	Hierarchical	H1	13048 ± 13084	16465 ± 16098	2.01 ± 0.42	0.32 ± 0.13
		H2	10788 ± 9418	14046 ± 12279	1.90 ± 0.36	0.31 ± 0.16
		H3	11276 ± 10090	14635 ± 13092	1.87 ± 0.19	0.30 ± 0.13
National	Baseline	B1	90834	118712	1.92	0.26
		B2	84902	113157	1.79	0.24
	Hierarchical	H1	79262	101885	1.71	0.23
		H2	79634	94531	1.65	0.23
		H3	170994	211382	3.65	0.49

the nature of the level below, the most granular and volatile of the hierarchy. The forecasts of the four-digit level import that level's own error, its raw lags import its volatility directly, and each two-digit series is in any case already a smooth aggregation of its four-digit components, so the lower-level information carries no signal that the two-digit lags do not already provide in a cleaner form. The aggregation model behind scenario **H2** smooths part of that volatility, which is why it is the least damaging of the three.

At the one-digit level, the two baselines become nearly indistinguishable, suggesting that the own lags already represent most of the daily and weekly structure captured by the calendar variables. The spatio-temporal scenarios perform substantially worse than the baselines, with mean MAE values more than 60% higher. Since this level contains only eight large and heterogeneous regional aggregates, the lagged loads of the other regions appear to add little information beyond each target series' own history.

The hierarchical scenarios produce the opposite result. All three improve upon the baselines, demonstrating that information from the two-digit level is useful for forecasting the one-digit aggregates. Scenario **H2**, which uses the aggregate two-digit forecast associated with each one-digit region, attains the lowest MAE and RMSE, reducing the mean MAE from 13 320 kWh for **B1** to 10 788 kWh, a decrease of approximately 19%. Scenario **H3**, which uses observed two-digit lags, attains the lowest MAPE and MASE. Although **H1** is the weakest hierarchical configuration, it also remains more accurate than both baselines. The explanation mirrors in reverse the one given for the level below, since the level immediately under the one-digit series is now the two-digit level, a smooth and reliable aggregation whose forecasts and lagged loads carry genuine signal instead of the volatility that the four-digit level imported into the two-digit forecasts. Where the cross-series lags among the eight one-digit series fail, the structured information propagated from the well-behaved level below succeeds, and it does so more effectively than any direct model of the one-digit series on its own. The shading of the hierarchical block at this level is a case where the caution stated earlier applies, because the four best metrics are split evenly between scenarios **H2** and **H3**.

At the national level, the set of applicable scenarios changes because the entire hierarchy reduces to a single target series. Same-level spatio-temporal scenarios are therefore not defined, and the hierarchical scenarios add lower-level information to a univariate national model rather than to a same-level multivariate specification. Scenario **B2** improves slightly upon **B1**, indicating that the calendar variables retain some predictive value at the national scale.

Among the scenarios that add lower-level information the outcome follows the accuracy of the level below. The forecasts propagated to the national level are those of the one-digit scenario selected on validation, and the two ways of supplying them end up close together. Taken individually in scenario **H1** they give the best result of the level, roughly seven percent below the best baseline, and combined into a single aggregate in scenario **H2** they come within half a percent of it, since both draw on the same one-digit forecasts. The raw one-digit lags of scenario **H3** carry too much high-frequency noise for a series as smooth as the national total, and at twice the error of its own baseline this is the largest relative degradation of the whole study. What the comparison establishes is that accurate forecasts of the level below carry genuine predictive value, which holds at the one-digit and the national levels and fails at the two-digit level, where the four-digit forecasts that would have to be propagated are the least accurate of the hierarchy.

The comparisons above are based on the mean value of each metric. However, the standard deviations in Table 5.1 indicate that the distribution of errors differs considerably among scenarios. At the four-digit level the dispersion of the spatio-temporal scenarios exceeds four times their own mean, against roughly twice for the baselines, so a comparison of means alone leaves open whether a difference reflects the behaviour of the level or that of a few series within it. The distribution of the per-series relative MAE is therefore examined in Figure 5.1, for the three levels that contain more than one series, with the mean marked by a diamond beside the median so that its displacement by the upper tail remains visible.

At the four-digit level, the five distributions are almost superimposed over their central half,

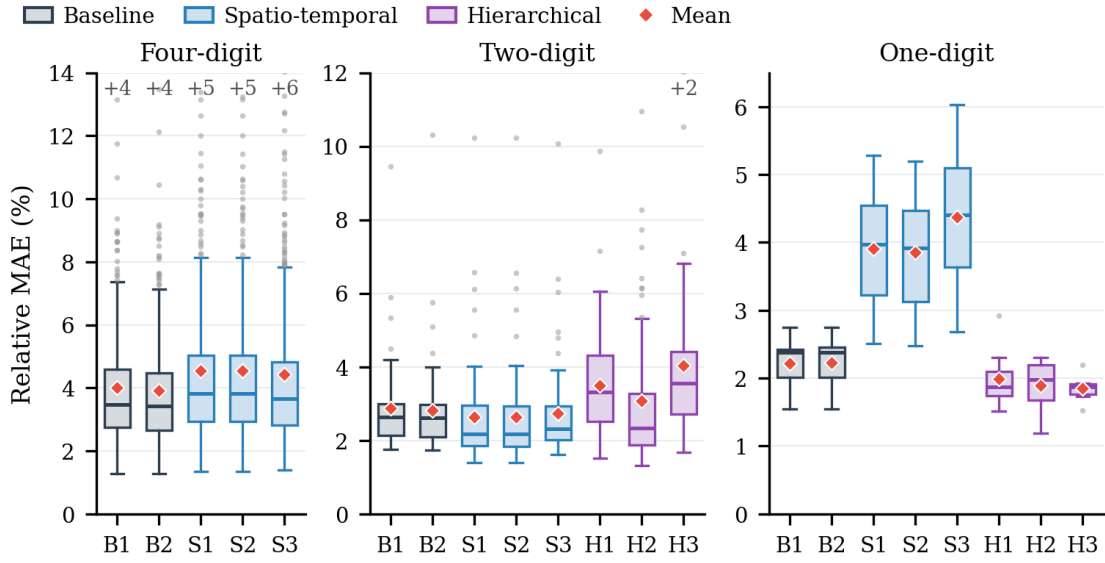


Figure 5.1: Distribution of the per-series relative MAE by scenario, at the three levels with more than one series.

with medians between 3.41% and 3.83% and interquartile ranges that overlap almost entirely, so on the typical postal code the choice of scenario is close to irrelevant. The families separate above the third quartile, where the multivariate whiskers and outliers extend to 54% of relative error against 26% for the baselines. The mean of every scenario sits above its own median, and the distance grows from 0.50 percentage points in the baselines to 0.77 in the spatio-temporal scenarios, which is what places the multivariate means nearly thirty percent above the baselines in Table 5.1 while their medians differ by only 0.4 points, a concentration whose cause is taken up in Section 5.1.3.

At the two-digit level, the spatio-temporal scenarios are shifted downward as a whole rather than only in their tail, with a median of 2.17% for scenario **S2** against 2.62% for scenario **B2** and a first quartile that is lower by a comparable amount, so the advantage reported in the table is spread across the dataset. Among the hierarchical scenarios the figure separates two behaviours that the mean alone merges. Scenarios **H1** and **H3** are worse than the baselines in median as well as in mean, and their boxes sit above every other scenario of the level, so their degradation is general. Scenario **H2** instead reaches a median of 2.34%, below both baselines and second only to the spatio-temporal group, together with a mean of 3.08% that lies close to its own third quartile, the largest gap between the two statistics anywhere in the study. The aggregated four-digit forecast is therefore useful on the majority of the two-digit series and severely harmful on a small group of them, a distinction that its mean of 2199 kWh presents as a uniform penalty.

At the one-digit level, where each box summarises only eight series, the three families separate almost completely. The spatio-temporal boxes lie entirely above the baselines and the hierarchical boxes lie below them, overlapping the baseline range only at their upper edge, which confirms that the penalty and the gain reported in the table apply to every region rather than to a subset. Within the hierarchical group the figure explains why the best metrics of the level are split between two

scenarios. Scenario **H2** reaches the lowest values of the level but spreads over an interquartile range of 1.68% to 2.19%, whereas scenario **H3** occupies a range of 1.77% to 1.91%, almost four times narrower, and even its worst region stays below the median of the baselines. Scenario **H2** is therefore the more accurate on average and scenario **H3** the more consistent across regions, which is why the former holds the MAE and the RMSE of the level and the latter the MAPE and the MASE.

5.1.2 Impact of Intra- and Cross-Level Dynamics

The previous subsection compared the forecasting scenarios in terms of accuracy. This subsection turns to the estimated LASSO-VAR-X coefficients, to determine which sources of information contribute most strongly to the forecasts and whether the retained spatial dependences are consistent with the dependences identified in the exploratory analysis of Chapter 3. A direct comparison of the coefficients would be misleading, since the number of candidate features of each type differs by several orders of magnitude and a numerous but individually weak group can accumulate a large total weight without moving the forecast. The models are therefore compared by how much each group of features actually shifts the prediction, obtained by adding up the contributions of the features of a group at every hour and averaging the size of that combined effect over the test period. This quantity reflects both the magnitude of the coefficients and the scale of the features they multiply, and it lets the terms of opposite sign within a group offset one another, so it measures the effect of a group on the forecast rather than the number of features it contains.

Figure 5.2 reports these contributions for the strongest spatio-temporal and hierarchical scenarios at each level, identified retrospectively from the test results in Table 5.1.

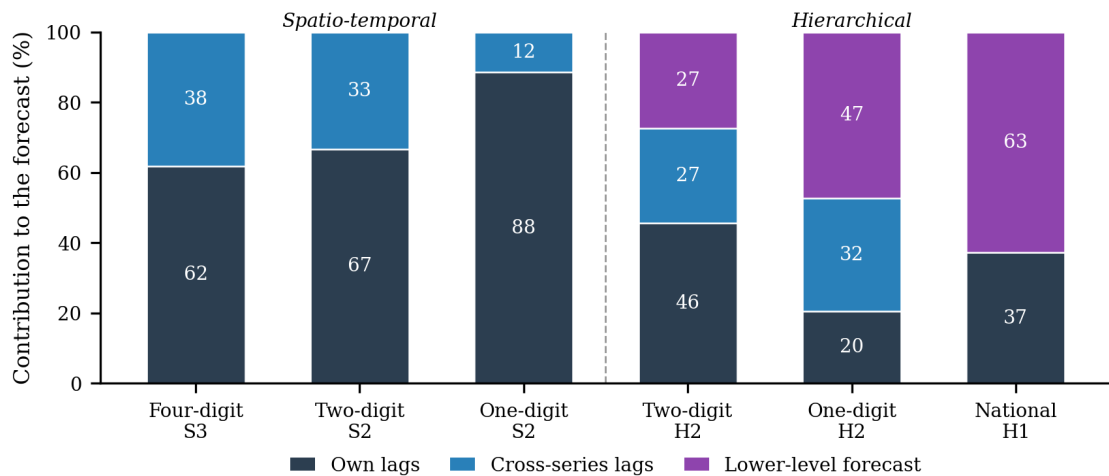


Figure 5.2: Effective contribution of the load-based information sources to the forecast, for the best spatio-temporal and the best hierarchical scenario of each level.

The four-digit level has no hierarchical scenario and the national level no spatio-temporal one, and a single bar is shown in each case. The three groups compared are the load-based sources of

information, namely the own lags of the series, the cross-series lags and the forecast propagated from the level below, all measured on the same scale.

Across the three spatio-temporal scenarios the own lags of each series are the dominant source of the forecast, and their contribution rises with aggregation from 62% at the four-digit level to 67% at the two-digit level and 88% at the one-digit level, the cross-series information supplying the complementary 38%, 33% and 12%. Because these weights were fixed on the training period, they show how much cross-series structure each model found useful in sample, and the test accuracy shows where that structure survived out of sample. The four-digit model was led to draw the largest of the three shares from the cross-series lags, and it is nonetheless less accurate than its univariate baseline, so the cross-series relationships learned during training did not hold through the test months. Only at the two-digit level, where the series are numerous and share a clean dynamic, does the in-sample reliance carry over into a test gain. At the one-digit level, where the cross-series share is already small, the limited cross-series contribution does not translate into improved accuracy and appears to act primarily as noise, in agreement with the accuracy ordering of Section 5.1.1. The mechanism behind the four-digit case is examined in Section 5.1.3.

In the hierarchical scenarios the forecast propagated from the level below becomes progressively more important as the hierarchy is climbed, contributing 27% at the two-digit level, 47% at the one-digit level and 63% at the national level. The two-digit model was fitted to draw 27% of its forecast from the four-digit forecasts and is nonetheless less accurate than the direct two-digit model, for the reason given in Section 5.1.1. The one-digit and the national models draw 47% and 63% from the smooth level immediately below and are the most accurate models of their respective levels, so here the reliance learned in training is rewarded in the test period. At the one-digit level the propagated forecast already contributes more than the own lags and the cross-series lags taken separately, which is the coefficient-level counterpart of scenario **H2** being the best model of that level, and within that contribution the forecast of a region own aggregate carries roughly three times the weight of those of the other regions, so each forecast is anchored to its own area and draws on the remaining regions as an adjustment.

The temporal encoders of the hour and the day of the week are absent from the figure, even though they are retained for most series, at rates above 90% at the two-digit level. Relative to the load-based predictors, their effective contribution is negligible, and the daily and weekly cycles they represent are in any case already carried by the own lags at 24 and 168 hours. The temperature is retained in only 0.03% of its candidate coefficients at the four-digit level, which is consistent with the nearly identical accuracy of scenarios **S2** and **S3**. This does not imply that temperature is unrelated to load, only that it provides little incremental information once recent load and calendar variables are included.

The effective contribution establishes how much the cross-series information matters, but not whether it is spatially meaningful. To examine whether the retained cross-series terms are consistent with the dependence measured in Chapter 3, each retained cross-series coefficient of the four-digit scenario **S3** model is compared with the partial cross-correlation between the same pair of series at the same lag, in Figure 5.3, for the one and two hour lags. Agreement between the two

quantities corresponds to the two quadrants where they share their sign, the upper right and the lower left.

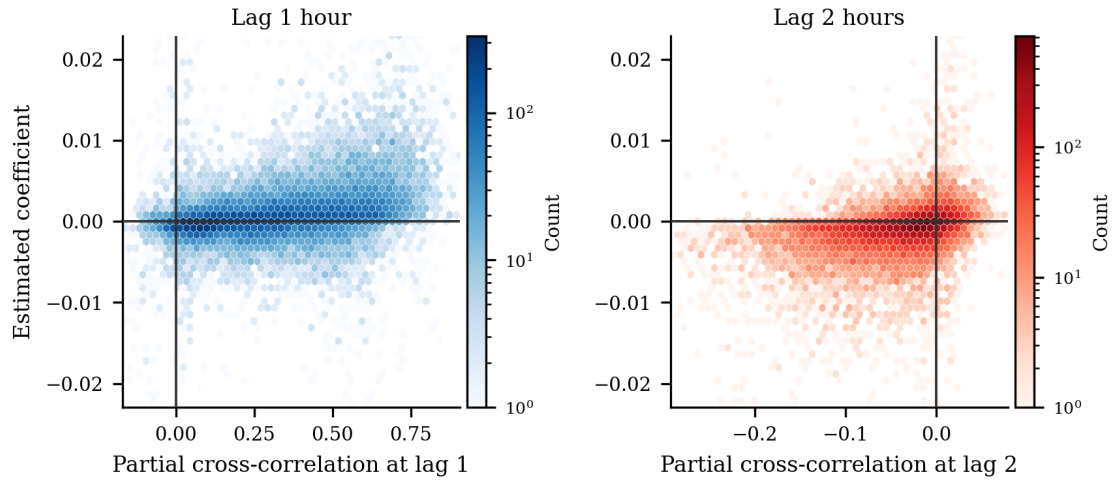


Figure 5.3: Retained cross-series coefficients of the four-digit scenario **S3** model against the partial cross-correlation of the same pair of series at the same lag, at the one and two hour lags. Colour encodes the number of pairs per cell.

At the one hour lag the densest cells form a horizontal band along the zero coefficient line, which is the large population of retained cross-series terms whose weight is small. Above this band the distribution opens into a wedge that climbs to the right, since the positive coefficients grow and spread further from zero as the partial cross-correlation increases, reaching values above 0.015 for the pairs whose correlation approaches 0.8. The bulk of the mass with a non-negligible weight therefore sits in the upper right quadrant, where the two quantities are both positive, and the sign matches in 63% of the 25 700 retained pairs. Almost no mass falls to the left of the vertical axis, in accordance with the exploratory analysis, where the one hour dependence was positive for nearly every pair of series, and the linear correlation between the two quantities is 0.31.

At the two hour lag the same picture is reflected into the opposite corner, with the mass fanning out from a compact core near the origin towards the lower left quadrant, where the partial cross-correlation and the coefficient are both negative, and the sign matches in 76% of the pairs. Agreement is higher than at the one hour lag because the sign is now shared on both branches of the distribution rather than concentrated on the positive side. The joint pattern across the two lags, a positive co-movement at one hour followed by a negative adjustment at two hours, reproduces both the sign and the time structure of the spatial dependence identified in Section 3.3.2, now over more than twenty thousand series pairs at each lag rather than a handful of examples.

The same correspondence holds at the level of the individual series. Among the ten largest coefficients of three postal codes from contrasting environments, an urban centre in Lisbon, a coastal area in the Algarve and an inland town near Évora, each series is led by its own one hour lag, and its strongest cross-series terms are the neighbouring codes whose partial cross-correlation is highest, between 0.62 and 0.94, which places them in the dense upper right region of Figure 5.3.

The two hour lag contributes with the opposite sign, as in the Évora series, whose second largest coefficient is a negative term at that lag matching a negative partial cross-correlation, while the terms at 24 and 168 hours carry appreciable coefficients against correlations close to zero and therefore follow the shared daily and weekly seasonality rather than a spatial link. The three coefficient inventories are reported in Appendix A.

5.1.3 Precision and Dispersion

The aggregate metrics and boxplots of Section 5.1.1 showed that the advantage of the spatio-temporal and hierarchical scenarios is unevenly distributed across the series of each level. This subsection examines that distribution directly and identifies the mechanism behind it, in particular the reason the cross-series information degrades the four-digit forecasts even though it lowered their training error, which was left open in Section 5.1.2.

A first sense of that distribution comes from following one series at each level over the whole test period. The comparison in Figure 5.4 places the observed load, the best model of the level and the univariate baseline side by side for four such series, in daily means so that the four months remain legible. The vertical scale contracts from a range of a few thousand kWh at the four-digit level to a band of several million kWh at the national total, and the load grows more regular as the aggregation rises, to the point where the one-digit series of Porto follows almost the same shape as the national total. Near the top of the hierarchy the regional aggregates take on the characteristics of the national series itself, and the abrupt holiday movements that dominate an individual postal code survive only as the broad August depression visible in the lower panels.

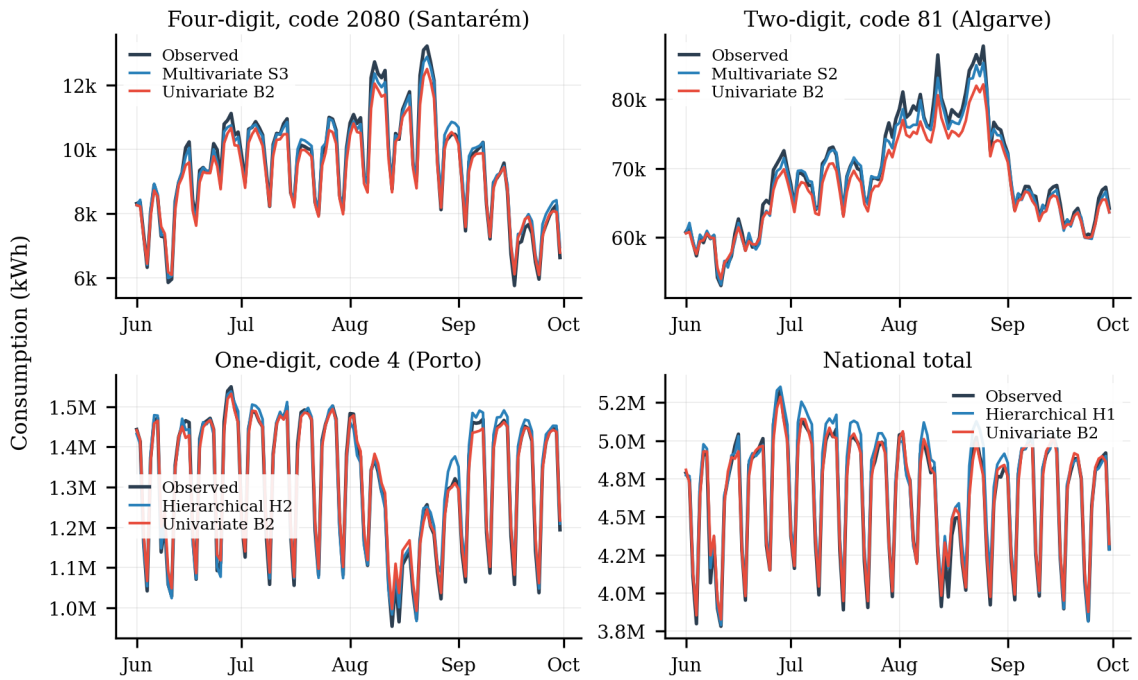


Figure 5.4: One representative series per level over the test period, in daily means, with the observed load, the best model of each level, and the univariate baseline **B2**.

Each panel has the best model of its level ahead of the baseline, and the way it gets there changes with the aggregation. In the Santarém district, the four-digit series of Almeirim confirms that individual series at the base of the hierarchy do benefit from the cross-series information even where the level as a whole does not, with scenario **S3** reaching weekday peaks that the baseline leaves short. In the Algarve, the two-digit series 81 makes the same point for a series whose annual behaviour is unlike that of the rest of the country, since its tourism cycle lifts the summer load well above the national pattern and scenario **S2** still follows the August peaks that the baseline misses. At the one-digit level the panel of region 4 in Porto is the clearest of the four, with scenario **H2** tracking the deep August trough that the baseline places too high. The national panel is the least convincing, since scenario **H1** improves on the baseline over most of the period while overshooting the peaks of June and July.

A trait common to the four panels is the way the univariate baseline compresses the amplitude of the load at every level, sitting above the observed series on the days of lowest consumption and below it on the days of highest, while the best model of each level recovers roughly half of the range that the baseline leaves out. The compression is strongest at the four-digit and two-digit levels and weakens towards the national total, which is consistent with the smoothing seen across the panels, and the correction acts in the same direction at every level.

Those four series show the favourable side of the comparison, and the one in Figure 5.5 extends it to every series of the three levels that hold more than one, placing the relative error of the best model of the level against that of the univariate baseline. Each point is a series and the dashed line is parity, so a series below the line is one where the best model of the level is the more accurate. The range of the axes alone confirms the smoothing described above, since the errors reach 32% at the four-digit level, 13% at the two-digit level and only 3.6% at the one-digit level.

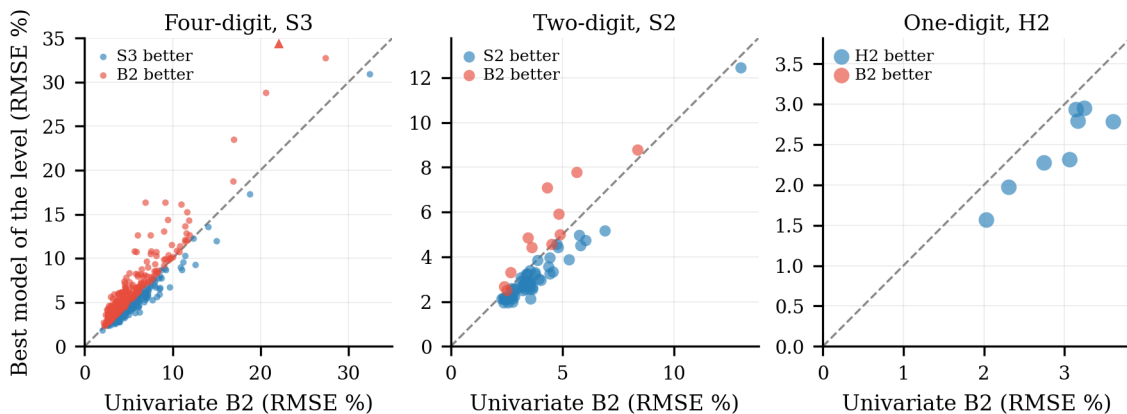


Figure 5.5: Per-series relative error (RMSE%) of the best model of each level against the univariate baseline (**B2**). The triangle marks one four-digit series beyond the axis range.

The three panels differ in shape as much as in scale, since at the one-digit level every one of the eight series lies below parity, with a mean advantage of 0.46 percentage points (pp) for the hierarchical scenario **H2** and no exception in either direction. At the two-digit level 58 of the

69 series lie below parity, and the eleven that do not are concentrated in the middle of the range, while the most difficult series of the level, at 13% baseline error, still gains from the cross-series information. The four-digit panel is the one that departs from this picture, and it does so through its shape rather than through a simple count, since the series below parity are 217 against 247 above it, close to an even split. The difference between the two sides lies in the magnitude rather than in the number, with the points below parity hugging the diagonal and gaining 0.64 pp on average, whereas the points above it fan upward and away from the diagonal, losing 1.26 pp on average and reaching 39 pp in the worst case. The cross-series information at the four-digit level therefore buys small gains on roughly half the series, of the kind seen in the Almeirim panel, and pays for them with much larger losses on the other half, and those losses are what the rest of this subsection sets out to explain.

The asymmetry visible in that panel is quantified by the summary in Figure 5.6, which reports the difference in mean relative error between the best multivariate scenario of the level and the univariate baseline, as the series worst predicted by the multivariate model are removed one by one, worst first. A value above zero means the multivariate model is less accurate than the baseline and a value below zero the opposite, with the difference expressed in pp.

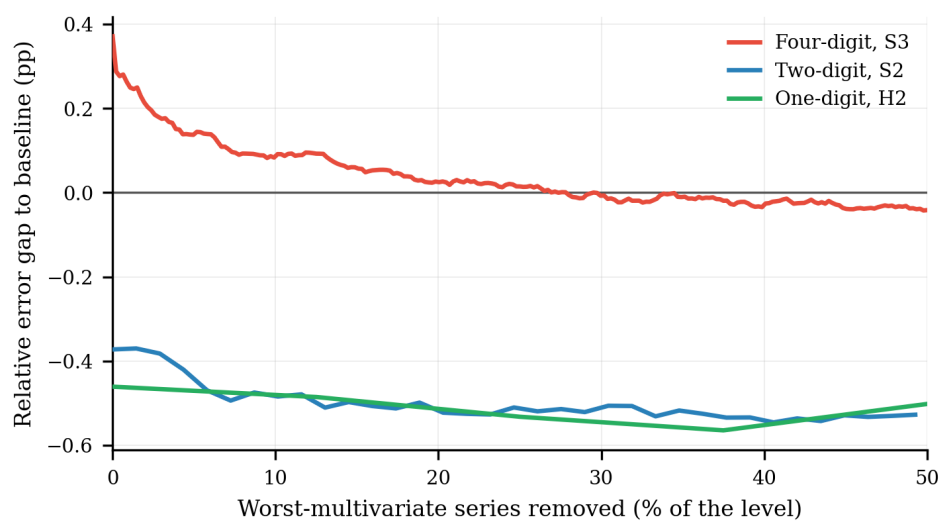


Figure 5.6: Difference in mean relative error between the best multivariate scenario of each level and the baseline **B2**, as the worst-predicted multivariate series are removed, worst first.

At the four-digit level the gap starts at nearly 0.4 pp and falls steeply as the worst multivariate series are removed, losing four fifths of its value once a tenth of the series are set aside and reaching parity at about a quarter, which places the penalty of the multivariate model in the magnitude of a few extreme cases rather than in the general behaviour of the level. The two-digit curve lies below zero throughout, and its early descent is the work of two interior aggregates of the Baixo Alentejo, around Beja and Ferreira do Alentejo, on which the multivariate model loses more than 2 pp for reasons examined further below, and once these are set aside the advantage over the remaining series is larger and roughly constant. The one-digit curve is almost flat, since with only eight large

regional aggregates there are no outlying series to remove, the local irregularities that produce the heavy tails of the four-digit level having been averaged out by the aggregation. The influence of a few extreme series therefore diminishes as the level rises, which is why the concentrated penalty of the four-digit level has no counterpart at the levels above it.

The difficulty of those four-digit series is seasonal, and the mean relative error broken down by month, in Figure 5.7, shows that the multivariate model is the more accurate only in June, at 4.9% against 5.6% for the baseline, and that it falls behind from July onward, reaching its widest margin in August, at 6.8% against 5.3%. The degradation concentrates in the height of the summer, when consumption departs most from the autumn, winter and spring patterns on which the cross-series relationships were estimated.

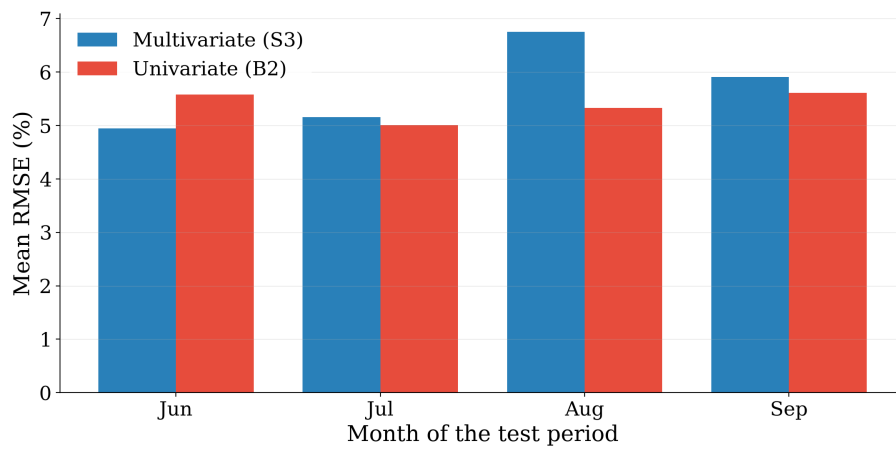


Figure 5.7: Mean relative error of the four-digit multivariate (**S3**) and univariate (**B2**) forecasts by month of the test period.

The origin of the summer degradation is a forecast that is built in part from the recent load of other series. When a target series and the series that predict it do not reduce their activity at the same time, which is what the summer holidays cause, the cross-series terms carry the behaviour of the wrong series into the forecast. The sign of the resulting error depends on which side keeps its usual level, and one case of each is examined below at the level of the individual series.

An example of overprediction is the four-digit series 4835, shown in Figure 5.8 together with two of its predictor series. From the middle of August the series falls to about two thirds of its usual level for roughly three weeks, in a pattern consistent with a holiday shutdown. The univariate baseline, which uses only the recent load of the series itself, follows the drop closely and remains within 3% of the observed load, whereas the multivariate model stays too high and overpredicts by 13%. The two predictor series in the lower panels keep their usual level throughout August, and since the multivariate forecast draws on their recent load, it is held up while the target itself falls. Over the whole test period the relative error of the series rises from 5.4% for the univariate model to 7.0% for the multivariate one.

The opposite case is the four-digit series 5225, in Figure 5.9. Here the observed load keeps its usual level of around 1 000 kWh throughout August, and the univariate baseline again follows it

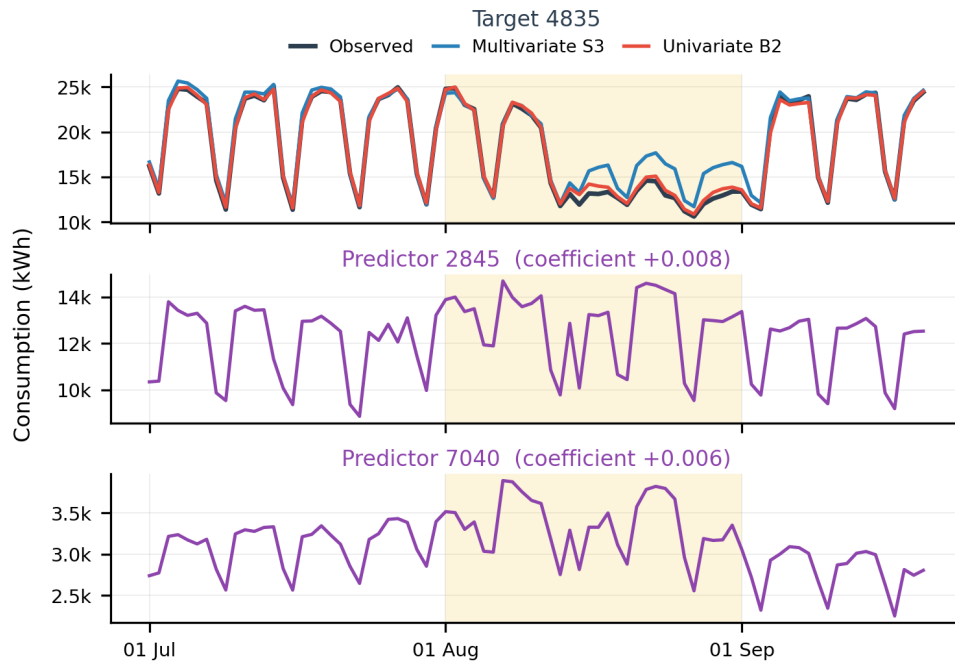


Figure 5.8: The four-digit series 4835 and its forecasts (top), with two of its predictor series below. August is shaded.

closely, remaining within 2% of the observed load. The multivariate model instead settles below the observed series and underpredicts by 9%, because the two predictor series in the lower panels reduce their activity sharply during August in a clear holiday pattern. Since the multivariate forecast depends on the recent load of these predictors, their drop pulls the prediction of the 5225 series down even though the series itself does not change.

Both cases therefore leave the univariate baseline as the more robust option on the difficult four-digit series, and the same mechanism appears in more isolated cases in July and September, with the error simply displaced to a different point of the holiday period, as anticipated by the monthly breakdown of Figure 5.7.

The mechanism also accounts for the behaviour of the levels above, although it survives the aggregation only where a region is sparse enough for the local shutdowns not to cancel one another out. At the two-digit level the losses of any size fall on the interior aggregates of the Baixo Alentejo already noted, smaller and more irregular series whose load does not follow that of their predictors, and their monthly profile repeats that of the four-digit cases, since the multivariate model is the more accurate on them in June and then overpredicts them from July onward, by as much as 8% in August. The remaining series below parity in Figure 5.5 lose by margins too small to matter, and on the large regional aggregates that dominate the level the multivariate model stays the more accurate throughout. At the one-digit level the mechanism disappears, because the eight regional series draw their advantage not from each other but from the smooth two-digit forecasts propagated from the level below, which carry genuine signal rather than the volatility that breaks in summer, so the hierarchical model is the more accurate on every one of the eight series without exception,

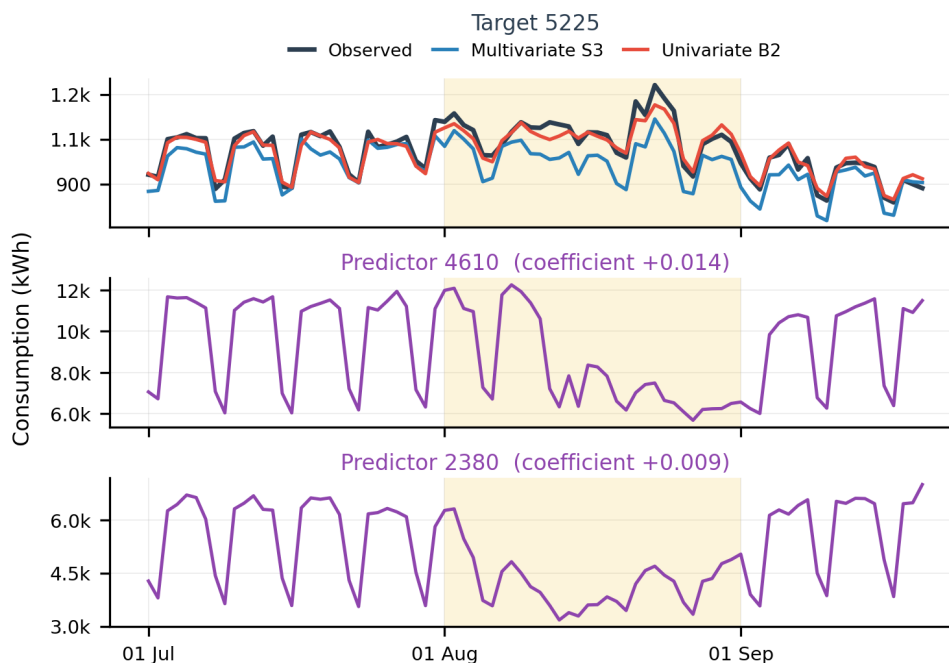


Figure 5.9: The four-digit series 5225 and its forecasts (top), with two of its predictor series below. August is shaded.

as the third panel of Figure 5.5 shows.

Taken together, these results complete the reading of the point-forecasting study by supplying what the aggregate comparison could not. The value of the cross-series information depends not on how much of it a model uses but on whether the relationships it encodes survive from the training months into the test months, and the summer holidays are precisely where that stability fails, on a minority of series large enough to reverse the verdict of an entire level.

5.2 Reconciliation Results

The point forecasts of the previous section were produced one level at a time, so they do not add up across the hierarchy, and reconciliation is applied to them as defined in Section 4.3.1.3. Of the weighting schemes considered there, WLS and MinT(Shrink) are the two evaluated here, and the metrics follow the same per-series convention used in the previous section.

To isolate what the projection contributes, it is applied to two base-forecast sets that bracket the range of information a base can carry. On one side stands the univariate baseline, scenario **B2**, whose levels are forecast in isolation from one another, and on the other scenario **S1**, which adds the cross-series lags wherever they exist so that the spatial structure of the load is already exploited before any correction is made. Among the three spatio-temporal scenarios, **S1** is the one used here because the leave-one-month-out validation selected it at every level, as reported in Appendix A, which keeps the choice free of any contact with the test period, and because the comparison over

eight base sets at the end of this section confirms that the three of them reconcile to almost the same accuracy.

Since the incoherence of the base is both what reconciliation removes and what bounds how far the forecasts can move, it is the natural place to start. As Table 5.2 reports, the mean absolute difference between the forecast of each aggregate and the sum of the four-digit forecasts it contains, relative to the load of that aggregate, separates the two bases by a wide margin. Having applied the same model type at every level, so that its levels move together, the univariate base is already close to coherent, within 1.10% of the aggregate load at worst. Much further from coherence lies the spatio-temporal base, reaching 5.57% at the one-digit level, since its cross-series regressors differ from level to level and drive the forecast of each aggregate away from the sum of its parts. Coherence is therefore a property of the base rather than of the hierarchy. In both cases the projection removes the discrepancy exactly, since the corrected forecasts are coherent by construction and the residual gap falls to the level of numerical error.

Table 5.2: Mean absolute aggregation gap of the base forecasts relative to aggregate load.

Aggregation level	Base B2	Base S1*	Reconciled
Two-digit	0.65%	2.73%	≈ 0
One-digit	1.10%	5.57%	≈ 0
National	0.51%	2.54%	≈ 0

* At the national level scenario **S1** reduces to own lags with the cyclic hour.

The reading of Table 5.3 proceeds level by level, and each level occupies five rows. The first of them is not a reconciled forecast at all, but the scenario that attained the lowest MAE of that level in Table 5.1, which serves as the reference against which everything below it is measured. That reference is not the same scenario at every level, being the univariate baseline at the four-digit level, the spatio-temporal scenario **S2** at the two-digit level and the hierarchical scenarios **H2** and **H1** above it. The four rows that follow apply the two weighting schemes to the two base-forecast sets under study, with the shading marking, within each base, the scheme that attains the better results at that level, and the bold values marking the best result of the level across both bases. The last column reports by how much each combination improves on the reference of its level in MAE, so that a positive value is an improvement and a negative one a degradation.

At the national level scenario **S1** has no cross-series lags available, since the whole demand reduces to a single series, and it therefore degenerates into own lags with the cyclic hour, which is scenario **B2** without its day-of-week terms. In its place a hierarchical scenario is deliberately not substituted, because this base set exists to apply one specification uniformly across the hierarchy, and a base assembled from different specifications at different levels is a distinct case, examined at the end of this section.

Table 5.3: Accuracy of the reconciled forecasts at every hierarchical level, set against the most accurate scenario obtained at that level without reconciliation. Within each base, the better weighting scheme is shaded.

Level	Base	Method	MAE (kWh)	RMSE (kWh)	MAPE (%)	MASE	MAE gain (%)
Four-digit	B2	—	365 ± 781	511 ± 1158	3.97 ± 2.45	0.45 ± 0.21	—
	B2	WLS	366 ± 785	511 ± 1161	3.98 ± 2.44	0.45 ± 0.21	−0.3
		MinT	342 ± 811	471 ± 1168	3.78 ± 2.55	0.42 ± 0.22	+6.3
	S1	WLS	407 ± 1126	534 ± 1465	4.60 ± 3.35	0.51 ± 0.29	−11.5
		MinT	429 ± 1054	560 ± 1400	4.97 ± 3.81	0.54 ± 0.28	−17.5
Two-digit	S2	—	1772 ± 2656	2320 ± 3504	2.69 ± 1.47	0.37 ± 0.22	—
	B2	WLS	1796 ± 2448	2449 ± 3410	2.72 ± 1.21	0.39 ± 0.23	−1.4
		MinT	1608 ± 2233	2158 ± 3130	2.53 ± 1.23	0.37 ± 0.24	+9.3
	S1	WLS	1769 ± 2997	2317 ± 3855	2.61 ± 1.55	0.36 ± 0.22	+0.2
		MinT	2018 ± 3313	2602 ± 4254	2.97 ± 1.46	0.40 ± 0.20	−13.9
One-digit	H2	—	10788 ± 9418	14046 ± 12279	1.90 ± 0.36	0.31 ± 0.16	—
	B2	WLS	11954 ± 10368	15871 ± 13809	2.02 ± 0.35	0.34 ± 0.18	−10.8
		MinT	10139 ± 8179	13250 ± 10756	1.83 ± 0.30	0.31 ± 0.19	+6.0
	S1	WLS	10756 ± 10665	13841 ± 13295	1.71 ± 0.30	0.28 ± 0.11	+0.3
		MinT	13642 ± 13419	17161 ± 16813	2.21 ± 0.54	0.34 ± 0.10	−26.5
National	H1	—	79262	101885	1.71	0.230	—
	B2	WLS	80824	108249	1.71	0.232	−2.0
		MinT	66937	86098	1.44	0.192	+15.5
	S1	WLS	65047	85029	1.42	0.187	+17.9
		MinT	98376	120896	2.14	0.282	−24.1

MinT denotes the minimum trace estimator with Schäfer–Strimmer shrinkage.

Falling in a regular pattern, the shaded cells select MinT for the univariate base and WLS for the spatio-temporal one, at every one of the four levels and on every metric. Neither scheme is therefore uniformly superior, and which of the two to use is decided by the base forecasts it is applied to rather than by the hierarchy.

A different question is answered by the bold values, namely which combination of base and scheme to prefer at each level, and the answer follows the pattern of the previous section. Recon-

ciled with MinT, the univariate base is the most accurate option at the four-digit level on all four metrics and at the two-digit level on three of them, whereas the spatio-temporal base reconciled with WLS takes the national level outright and leads the scale-free metrics from the one-digit level upward, with the two approaches splitting the one-digit level between them. As before the projection was applied, the cross-series information remains a liability at the bottom of the hierarchy and an asset at the top, and reconciliation does not overturn that division.

The last column answers a stricter question than whether each base improves on itself, since the reference of a level is often not the scenario being reconciled. Provided the base is chosen appropriately, that test is cleared at every level. Reconciled with MinT, the univariate base improves on the reference throughout, by between 6.0% and 15.5% depending on the level, and it is the better of the two options at the three lower ones, whereas the spatio-temporal base reconciled with WLS falls 11.5% behind at the four-digit level and takes the lead only at the national total. Across the column the gain grows with aggregation, from around 6% at the bottom of the hierarchy to nearly 18% at the top, since an aggregate is corrected by the whole bottom level summed underneath it while an individual postal code is corrected only by its own share of the discrepancy.

Reading that column downwards within each level shows how much the pairing of base and scheme matters, since the same base reconciled by the wrong one loses more than the right one gains. At the one-digit level the spatio-temporal base moves from a gain of 0.3% under WLS to a loss of 26.5% under MinT, and at the national total from a gain of 17.9% to a loss of 24.1%. More consequential still, the univariate base, weaker than the spatio-temporal one before reconciliation at three of the four levels, becomes the stronger of the two at those same levels once the projection is applied, so that reconciliation does not merely improve each base but changes which base is worth starting from.

Where the reversal comes from is what each base leaves in its residuals, since the weighting matrix is estimated from them and the full covariance of MinT(Shrink) can improve on the WLS diagonal only where the errors of different nodes are correlated strongly enough to be worth estimating. Measured over the nodes of the hierarchy, the mean absolute correlation is 0.42 for the univariate base and about 0.28 for every base that uses cross-series or cross-level information, and what sets it apart is not the degree of spatial information it uses but the fact that it uses none, so the structure of the load stays in its residuals for the off-diagonal terms to exploit. Its covariance is also the only one that the shrinkage rescues rather than merely refines, since the unregularised estimator built on it is numerically unusable and returns MASE values above 7 at every level.

Whether the pattern holds beyond these two bases was tested on seven of them, each applying one scenario of the previous section at every level where that scenario is defined and falling back elsewhere, since not all of them are available throughout the hierarchy. At the four-digit level the hierarchical sets use scenario **S1**, because there is no level below it to draw forecasts or lags from, and at the national level the spatio-temporal sets reduce to the univariate model, because a single series admits no cross-series regressors. Included as an eighth case is a mixed set that takes the best model of each level. Reported in Table 5.4 is the mean MASE over the four levels, each weighted equally so that a method is judged on whether it helps throughout the hierarchy rather

than only where most series lie.

Table 5.4: Mean MASE over the four hierarchical levels, each level weighted equally, for every base-forecast set and reconciliation method.

Set	Base forecasts	Base	WLS	MinT(Shrink)	Best
B2	Univariate	0.364	0.352	0.324	MinT(Shrink)
S1	Cross-series + hour	0.441	0.335	0.389	WLS
S2	Cross-series + day of week	0.438	0.335	0.380	WLS
S3	Cross-series + temperature	0.471	0.331	0.332	WLS
H1	Lower-level forecasts	0.392	0.393	0.394	Base
H2	Aggregate lower-level forecast	0.367	0.362	0.453	WLS
H3	Observed lower-level lags	0.472	0.415	0.601	WLS
–	Best model per level	0.359	0.353	0.384	WLS

Across the eight sets WLS is the best method in six and MinT(Shrink) leads only for the univariate base, which confirms that the reversal is a property of the base forecasts and not an accident of the two sets examined in detail. Among the three spatio-temporal sets the reconciled values are almost identical, so the choice among them is immaterial to this study. Largest of all is the gain where the base forecasts are modelled independently at each level, as in those three sets, whose mean MASE falls from about 0.44 to 0.33, and it narrows as the base forecasts begin to borrow across levels.

In the remaining set reconciliation fails outright, since for scenario **H1**, whose aggregates are predicted from the forecasts of the level below, the unreconciled base is already the best of the three options and neither scheme recovers anything. Affecting the mixed set as well, the same difficulty explains an expectation that the results do not meet, because reconciling the most accurate model of each level starts from the lowest base error of the nine and still ends above the value reached by applying a single cross-series scenario throughout. Behind both cases lies one cause, since a base that already imports information from adjacent levels is closer to coherent to begin with and leaves less for the projection to correct, while its error structure is no longer consistent from level to level and a single covariance describes it poorly. For this reason the reconciliation is carried out on one scenario applied throughout rather than on the best model of each level.

Paralleling the central finding of Section 5.1.3, these results turn on the same distinction. There, the value of the cross-series and cross-level information depended on whether the forecast had already captured it. Here, the value of modelling the error correlations depends on whether the base forecasts left that correlation in their residuals, and the value of reconciliation itself depends on whether the base was incoherent enough for the projection to have work to do. Of the two schemes, WLS is therefore the robust recommendation for this hierarchy, and the full error

covariance pays off only where the base forecasts deliberately ignore the spatial structure of the load.

5.3 Probabilistic Forecast Results

This section evaluates the probabilistic forecasts produced by the pipeline of Section 4.3.2.1. The univariate scenario **B2** under the sliding window constitutes the reference configuration of the study, against which the multivariate scenario **S2** is compared. The analysis covers the four levels of the hierarchy over the 122-day test period, with training windows of 90, 180 and 210 days. Unless stated otherwise, results refer to the longest window. Forecast quality is measured using the pinball loss and CRPS introduced in Section 4.4. These scores are complemented by three interval diagnostics: the calibration error, defined as the mean absolute deviation between nominal and empirical coverage over the nine quantile levels, the empirical coverage of the 80% and 90% intervals, and their average width. All per-level values are averages over the series of that level.

The exposition is organised by question rather than by hierarchical level. Aggregate tables and figures report results for all four levels at once, while per-series and intra-day detail is examined at a single representative level, namely the two-digit one, whose 69 series remain individually legible, while the national series illustrates the window analysis, where the seasonal mismatch is most visible.

5.3.1 Effect of the Training Window

The window length W is the dominant lever of the whole probabilistic study. Table 5.5 reports the improvement obtained by the 180- and 210- day windows relative to the 90-day window for scenario **B2**. The absolute values of every metric, level, scenario and window length, are gathered in Appendix B. Extending the window improves every metric at every level, and the magnitude of the improvement grows monotonically with aggregation, from 9.0% of pinball loss at the four-digit level to 20.6% at the national one. The aggregate series are the smoothest and therefore the most dependent on the training window matching the regime of the forecast day, whereas the noisy four-digit series carry an irreducible component that no window length removes.

The gain is essentially exhausted at 210 days, since the step from 180 to 210 days moves the pinball loss by less than 1% at every level, and longer windows are precluded by the extent of the dataset itself.

Table 5.5: Improvement of scenario **B2** over the 90-day window, per hierarchical level.

Level	Pinball		CRPS	
	$W = 180$	$W = 210$	$W = 180$	$W = 210$
National	20.4%	20.6%	18.0%	18.2%
One-digit	15.5%	15.8%	13.6%	13.8%
Two-digit	11.1%	11.9%	9.6%	10.4%
Four-digit	8.4%	9.0%	7.2%	7.7%

At the national level, where shorter windows were also evaluated, the full sweep from 30 to 210 days makes the dependence explicit. The pinball loss falls from 31 739 kWh at 30 days to 15 251 kWh at 210 days, a reduction of 52%, while the calibration error improves from 0.123 to 0.049 and the coverage of the 90% interval rises from 0.613 to 0.830. A window of one month is simply too short to represent the conditions of the forecast day, and every additional month of history pays until the saturation described above.

The monthly decomposition in Figure 5.10 shows that this average conceals two distinct situations. In June and in August the long windows dominate, with a CRPS reduction above 20% and 30% respectively, because the short window either contains no summer at all, as in June, when its 90 days reach back only to spring, or is dominated by ordinary summer days that do not represent the anomalous holiday period of mid August.

In July the ordering inverts slightly, and the 90-day window becomes marginally the best, since by then its content is almost entirely recent summer data, which matches the forecast days better than a window still carrying the winter months. The long window is therefore best understood as insurance, indispensable when the recent past is unrepresentative and almost costless when it is not.

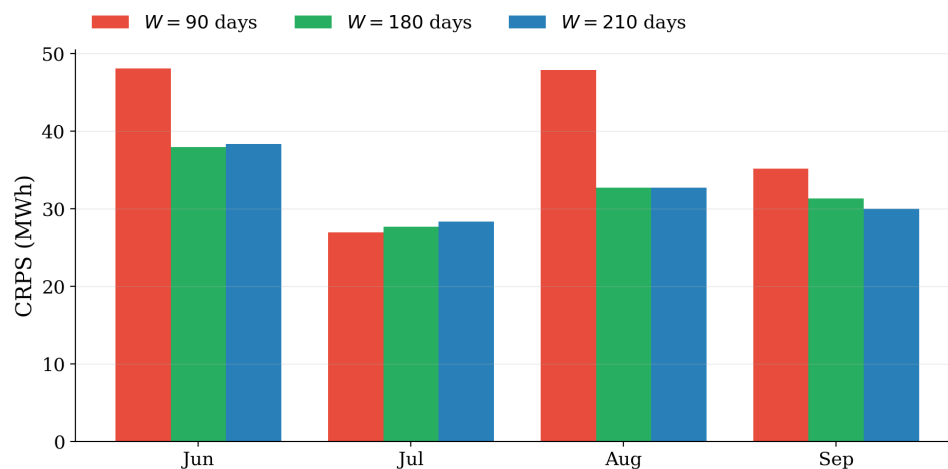


Figure 5.10: Monthly CRPS of the national forecast for the three window lengths.

Window length establishes how much history each fit should see, but not whether refitting matters at all. This is assessed by comparing the sliding regime against a single fit on the whole training period under an identical engine, scenario and feature set. Table 5.6 reports the relative improvement of the sliding regime for scenario **B2** at the longest window. Daily refitting reduces the pinball loss by between 18.9% and 27.0%, depending on the level and lowers the calibration error by roughly a third. It does so while narrowing the 90% interval rather than widening it, so the gain in coverage is not bought with a loss of sharpness.

Table 5.6: Improvement of the sliding regime over a single static fit, for scenario **B2** at the longest window.

Level	Pinball	Calibration error	90% interval width
National	18.9%	30.9%	8.7%
One-digit	21.9%	33.5%	7.5%
Two-digit	27.0%	38.9%	4.6%
Four-digit	26.9%	32.3%	4.5%

The two design choices have opposite relationships with aggregation. The gain from refitting grows towards disaggregation, from 18.9% at the national total to 27.0% at the two-digit level, whereas the gain from lengthening the window moves in the opposite direction, from 9.0% at the four-digit level to 20.6% at the national total. What a smooth aggregate needs is a window that contains the right season, while what a noisy postal code needs is a fit that is never far behind the present, and the two requirements are met by different halves of the same design. Applied to the multivariate scenario the effect is larger still, between 30.6% and 36.1% of pinball loss and about half of the calibration error, so the approach that depends most on the sliding regime is also the one that remains behind the univariate baseline once both have it.

Regarding computational cost, the sliding regime multiplies the number of fits by the number of forecast days, as discussed in Section 4.3, and the efficiency of the engine keeps this tractable, with the complete national sweep running in minutes and the heaviest configuration, the multivariate scenario at the four-digit level, in hours per window length.

5.3.2 Univariate and Multivariate Forecasts

In the point-forecasting study, cross-series information was beneficial at the upper levels of the hierarchy. This advantage does not carry over to the probabilistic setting. As shown in Table 5.7, the multivariate scenario **S2** is outperformed by the univariate scenario **B2** in CRPS by 5.8% at the one-digit level, by 13.6% at the two-digit level and by 18.0% at the four-digit level, with the deficit growing towards disaggregation, precisely where the series are noisiest and the cross-series

lags carry the least usable signal. The national level admits no multivariate scenario, since a single series offers no cross-series lags.

Beyond the centre of the distribution, the multivariate intervals are systematically wider, by 8% to 19% at the 90% level, and yet no better calibrated except at the one-digit level, so the additional width is not matched by better coverage.

Table 5.7: Univariate and multivariate forecasts at the longest window, with the best value of each metric in bold.

Level	Pinball (kWh)		CRPS (kWh)		Calibration error	
	B2	S2	B2	S2	B2	S2
One-digit	2581	2750	5443	5759	0.059	0.058
Two-digit	411.1	469.3	865.9	983.6	0.052	0.056
Four-digit	85.3	100.9	179.6	212.0	0.047	0.054

Figure 5.11 examines whether this deficit is concentrated in a small number of poorly predicted series. The series with the largest multivariate disadvantage are removed one at a time, and the aggregate CRPS gap is recomputed over the remaining set. The curves fall steadily but never reach zero, and after removing half of the series of each level the multivariate forecasts remain 2.5% to 7% worse, so the deficit is pervasive rather than concentrated, with scenario **S2** beating scenario **B2** in only 17 of the 464 four-digit series, 6 of the 69 two-digit series and 1 of the 8 one-digit series.

This is a sharp contrast with the point-forecasting results, where the multivariate disadvantage at the four-digit level was driven by a minority of series and vanished once they were removed. Under the sliding regime, with at most 210 days of training data per fit, the cross-series structure appears to cost more in estimation variance than it returns in signal, uniformly across space. The monthly decomposition at the two-digit level supports the same reading, with the multivariate gap varying only between 12.0% and 16.2% across the four test months, peaking in August.

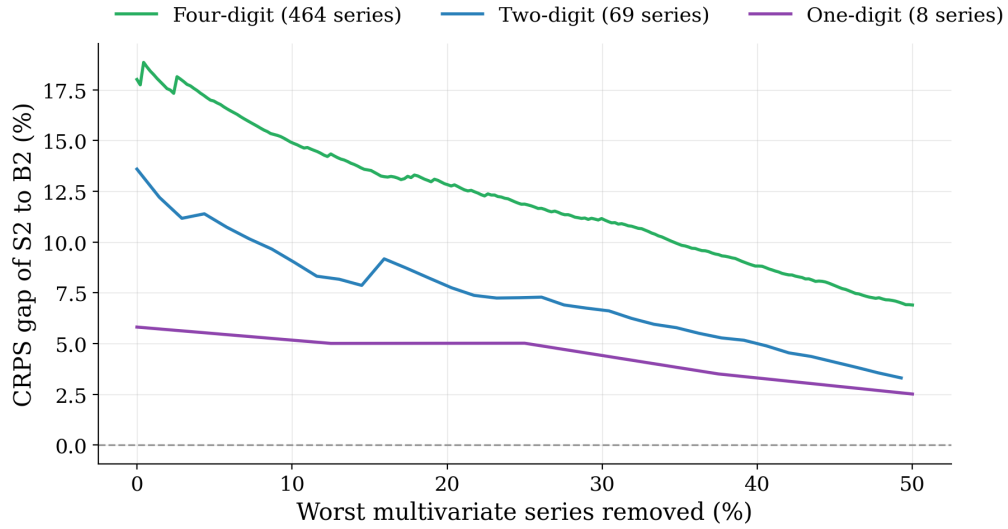


Figure 5.11: CRPS gap of scenario **S2** to scenario **B2** as the worst multivariate series are removed.

Several refinements to predictor selection and hyperparameter tuning were tested, but none changed the ranking. Raising the retention threshold of the predictor selection degraded the pinball loss by close to 3% at the two longest windows. Replacing the single validation block of Section 4.3.2.1 by a leave-one-month-out selection retains essentially the same predictors and moves the pinball loss by less than 1% at both levels, at seven times the computational cost, which is what justifies the single block adopted there, and re-running the selection inside each training window or widening the hyperparameter search space likewise produced variations within a few percent. The configuration of the machinery matters far less than the training regime it operates under.

5.3.3 Calibration Across the Hierarchy

Accuracy scores alone do not establish whether a predictive distribution reports its own uncertainty correctly, since a model may achieve a competitive CRPS with intervals that are systematically too narrow. The calibration of the multivariate forecasts is therefore examined directly, first in aggregate and then across space, since the spatio-temporal approach is the object of this study even where the univariate baseline outperforms it. The national level is absent from this analysis, since it admits no multivariate scenario.

The reliability diagram in Figure 5.12 shows that the three levels share a single calibration signature. The lower quantiles over-cover, with the 0.05 level observed between 0.082 and 0.105, the curves cross the diagonal near the 0.35 level, and the upper half of the distribution under-covers, most severely at the 0.80 level, where the empirical coverage falls short of the nominal one by 0.06 at the four-digit level and by 0.10 at the one-digit level.

The predictive distribution is therefore shifted slightly downward relative to the observations, which is the residual trace of the seasonal mismatch discussed in Section 4.3.2.1. The models keep underestimating the upper region of a summer load that exceeds most of their training window, and

the effect persists, attenuated, even at the longest window. The deviation shrinks as aggregation increases, so the smoother upper-level series are also the better calibrated ones, and the univariate forecasts follow the same shape, differing by less than 0.02 at every quantile level, so the diagnosis below applies to the recommended configuration as well.

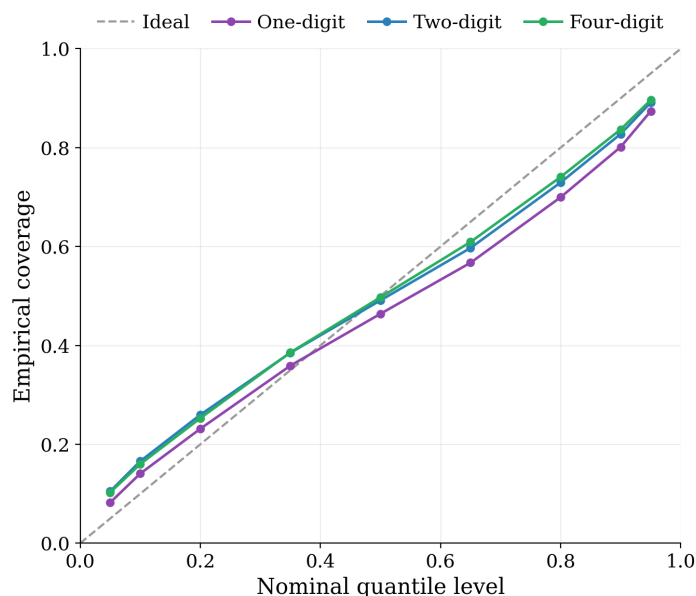


Figure 5.12: Empirical against nominal coverage of the scenario **S2** forecasts.

Figure 5.13 examines the spatial variation behind these aggregate results at the two-digit level. Calibration deviation is reported by series and quantile level, with the series ordered by relative CRPS. The most immediate reading is that the larger deviations are confined to a region of the distribution rather than spread over it, since almost every row is blue on the left and red from the 0.65 level onwards. The intervals are consistently too low at both ends, which is the spatial confirmation of the aggregate signature above.

Nor are the larger deviations spatially uniform, since in the eastern Algarve code 89 combines the worst relative CRPS of the level, at 5.8%, with a pronounced under-coverage across the whole upper half, consistent with a strongly seasonal tourist region whose summer regime departs furthest from the training window. At the opposite extreme, code 30 shows an almost white row together with one of the lowest relative CRPS values, meaning that both its central tendency and its uncertainty are well captured.

CRPS and calibration are not aligned across the map, and if anything they oppose each other. The five best-predicted series, with a relative CRPS around 1%, average a deviation of -0.07 at the 0.80 level, while the five worst, with a relative CRPS between 2.3% and 5.8%, average -0.05 . The clearest case is code 28, which contains the irregular industrial series of Seixal identified in Section 3.3, and whose upper-quantile deviations are among the mildest of the level while its relative CRPS is the second worst.

The explanation lies in what dominates the error budget of each series. For smooth series the residual noise is small, so the seasonal drift is the main source of error and it compresses the intervals below their nominal level. For genuinely volatile series the models learn wide distributions whose width matches the dispersion they must accommodate, since that dispersion is present in the training window as well. Uncertainty is therefore easier to quantify correctly precisely where it is largest, which is a caution against reading the CRPS ranking as a ranking of trustworthiness.

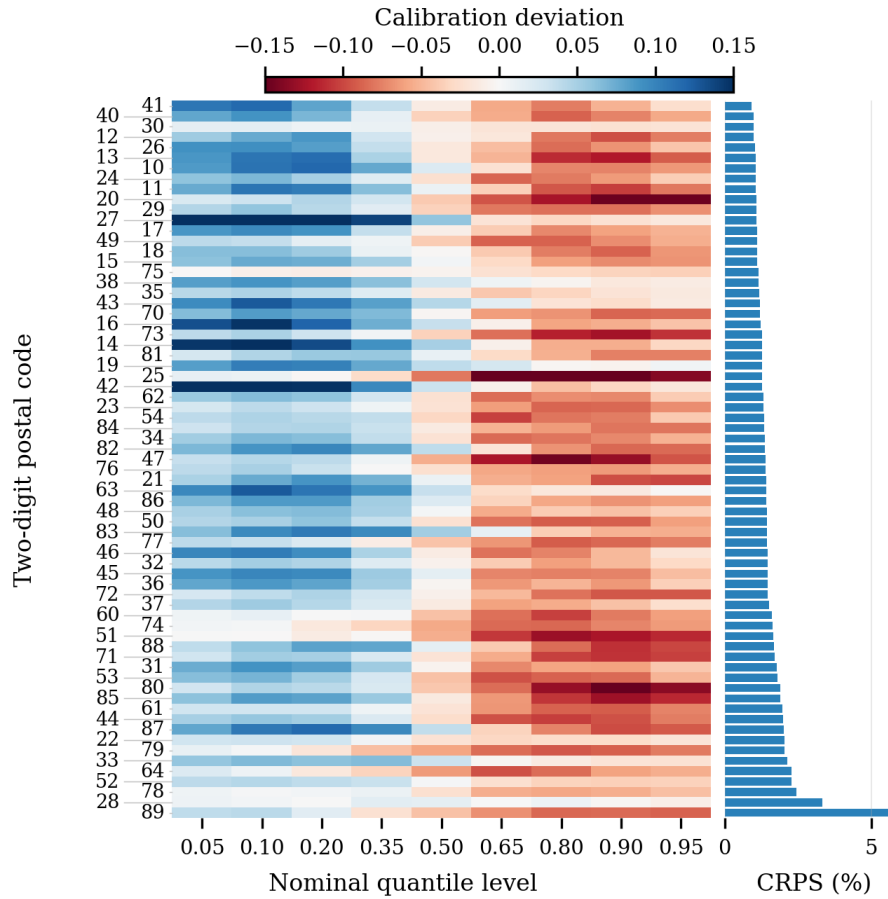


Figure 5.13: Calibration deviation per quantile level and relative CRPS of the two-digit series, for scenario S2.

5.3.4 Temporal Behaviour of the Predictive Intervals

The aggregate scores and calibration maps describe performance over the whole test period, but not how the predictive intervals evolve from hour to hour. This last analysis therefore examines the multivariate forecasts from scenario S2 in the time domain, for postal code 47, whose relative CRPS of 1.4% is the median of the two-digit level. Two weeks of the test period are compared in Figure 5.14, an ordinary week of June and the mid-August week that contains the usual holidays.

In June the intervals are narrow, with the 80% band width average of 18.8 MWh, and the

observed load remains inside them almost permanently. The exceedances that do occur are concentrated at the daily peaks, which is the intra-day expression of the upper-tail under-coverage quantified in Figure 5.12. particular hour.

In August, the same model, re-estimated daily on a window that by then contains the full summer, responds to the change of regime by widening the intervals markedly. Over the three days around the holiday the 80% band averages 39.3 MWh, more than double its June width, and the observed load, which departs from its usual shape, remains inside it. The widening then recedes in the second half of the week as the regime normalises, so the model does not simply become permanently more uncertain.

This behaviour is consistent with the intended role of the sliding regime described in Section 4.3.2.1, since the sliding window feeds the recent anomalous observations into the training set within days, letting the predictive distribution inflate exactly when the uncertainty is genuine instead of retaining intervals calibrated on the colder months. The remaining levels display the same contrast, which at the aggregate scale is what drives the August behaviour of Figure 5.10, while the four-digit series superimpose their local noise on it.

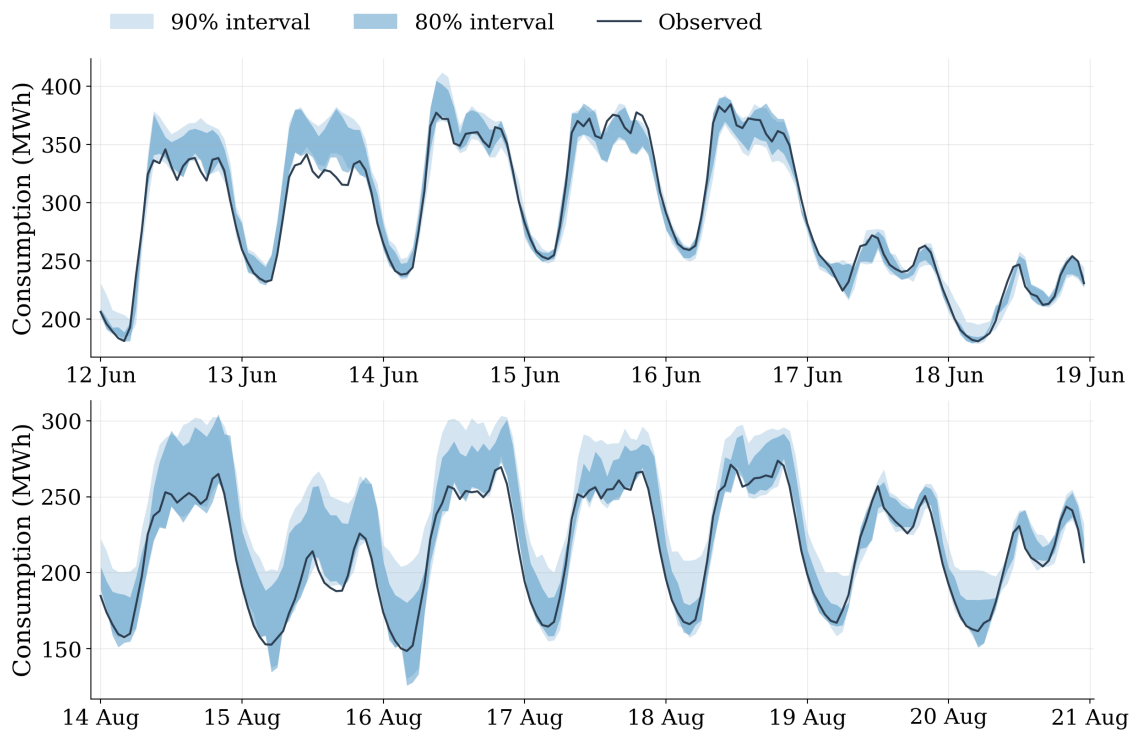


Figure 5.14: Predictive intervals of scenario **S2** for postal code 47 over a June and an August week.

Taken together, the probabilistic results show that the conclusions of the point-forecasting study do not transfer directly to the distributional setting. There the choice of features and of cross-level information was the decisive margin, while here every variation of features, selection and tuning is secondary to a single decision, the training regime, which alone accounts for a fifth to a quarter of the error at every level, against the few percent separating one configuration of

the machinery from another. A univariate model with eight fixed features, re-estimated daily on a window of 210 days, delivers the best predictive distributions at all four levels of the hierarchy, with a calibration error of 0.047 to 0.059 and a coverage of the 90% interval between 0.79 and 0.83.

5.4 Chapter Summary

This chapter has evaluated the methodology of Chapter 4 over the four held-out months of June to September 2023. The eight forecasting scenarios were compared at every level of the hierarchy, the resulting forecasts were reconciled under both weighting schemes and over several sets of base forecasts, and the same hierarchy was then evaluated in the probabilistic setting.

The point-forecast results show that the value of cross-location information depends on the level at which the forecast is produced. At the four-digit level the univariate baseline remains the most accurate option, whereas at the two-digit level the cross-series lags reduce the error by roughly 5%, and among the eight one-digit regions the same information becomes markedly counterproductive. Cross-level information behaves in the opposite manner and becomes valuable precisely where cross-location information fails, reducing the error by roughly 19% at the one-digit level and by 7% for the national total.

Reconciliation was found to do considerably more than restore coherence, improving accuracy at all four levels by between 6% and 18% when measured against the best unreconciled scenario of each level. The preferred weighting scheme proved to be a property of the base forecasts rather than of the hierarchy, and the projection also changes which base is worth starting from, since the univariate base becomes the stronger of the two at most levels once it is applied.

In the probabilistic setting the ranking established for point forecasts did not survive, since the multivariate scenario is outperformed by the univariate one at every level and the deficit is pervasive rather than concentrated in a few series. What dominates these results is instead the training regime, because re-estimating the models daily on a moving window lowers the error by roughly a quarter, against the few percent separating one configuration of the machinery from another.

Read together, the three strands turn on the same distinction, namely whether the information being added had already been captured by other means. Chapter 6 draws these findings together, relates each of them back to the research questions stated in the introduction, states the limitations that bound their scope, and outlines the directions in which the work points next.

Chapter 6

Conclusions

This dissertation investigated whether the spatial and hierarchical organisation of distribution-level electrical load carries predictive information beyond that already captured by purely temporal models. Using eleven months of hourly consumption from the E-REDES open data portal, a four-level hierarchy was constructed from four-digit postal-code series up to the national total. Eight forecasting scenarios were then compared under a common LASSO-VAR-X estimator at every level, their point forecasts were reconciled across the hierarchy, and the analysis was extended to probabilistic forecasts obtained by quantile regression. All conclusions are based on four held-out months that were excluded from both model fitting and configuration selection.

6.1 Main Findings

The exploratory analysis first confirmed that the spatial dependence between postal-code areas remains present after the dominant temporal components are removed. Following multi-seasonal decomposition, the residual series remain broadly and positively correlated, with visible blocks that group the denser urban regions of the country. That dependence is, however, sharply confined in time, since the partial cross-correlation between areas is concentrated at delays of one and two hours and becomes indistinguishable from zero beyond them. The clustering analysis of the residuals further showed that the dataset is far from homogeneous, since fewer than half of the postal-code areas group with any other at a moderate correlation threshold, while the remaining series have residual dynamics that resemble no other area closely enough.

Turning to the first research question, the predictive value of cross-location information proved to depend on the hierarchical level at which the forecast is produced, and not in the direction that increasing aggregation would suggest. At the four-digit level, the univariate baseline remains the most accurate option, while the spatio-temporal scenarios trail it by close to a third. At the two-digit level, the ordering reverses and cross-series lags reduce the error by roughly 5%. Among the eight one-digit regions, the same information becomes markedly counterproductive. This result is consistent with the small number and heterogeneity of the regional series, which limit the shared

structure available through cross-series lags. What governs the usefulness of cross-location information is therefore the number of series sharing a level and their mutual homogeneity, rather than the degree of aggregation itself.

The second research question concerned information transferred across adjacent levels of the hierarchy. Cross-level information behaves in the opposite manner and becomes valuable precisely where cross-location information fails, reducing the error by roughly 19% at the one-digit level and by 7% for the national total. Both gains depend on the quality of the level supplying the information, and the same mechanism explains why these scenarios fail at the two-digit level, where the level below is the most volatile of the hierarchy and imports its own error rather than usable signal. What the comparison establishes is that accurate forecasts of the level below carry genuine predictive value. This is confirmed at the one-digit and the national levels, where the supplying level is regular and its forecasts are accurate, and it fails at the two-digit level, where the four-digit forecasts that would have to be propagated are the least accurate of the hierarchy.

On the third question, hierarchical reconciliation was found to do considerably more than restore coherence. Relative to the best unreconciled scenario at each level, the reconciled forecasts improve accuracy at all four levels, by between 6% and 18%, with the gain growing towards the top of the hierarchy because an aggregate is corrected by the whole bottom level summed underneath it while an individual postal code is corrected only by its own share of the discrepancy. Two further findings emerge from this analysis. First, the preferred weighting scheme is a property of the base forecasts and not of the hierarchy, since the minimum trace estimator is superior for the univariate base at every level while weighted least squares is superior for every base that already uses spatial or cross-level information. Because the univariate base is the only one that leaves the spatial structure of the load in its residuals, it is also the only one for which modelling the off-diagonal terms of the error covariance repays their estimation. Second, reconciliation changes which base is worth starting from, since the univariate base, weaker than the spatio-temporal one before the projection at most levels, becomes the stronger of the two at those same levels once it is applied.

The fourth research question asked whether the point-forecast conclusions extend to the full predictive distribution. They do not transfer directly. The multivariate probabilistic scenario is outperformed by the univariate reference at every level and by a margin that grows towards disaggregation, and the deficit is pervasive rather than concentrated in a few pathological series. What dominates the probabilistic results is instead the training regime, since re-estimating the models daily on a moving window rather than fitting them once lowers the error by roughly a quarter and the calibration error by about a third, while narrowing the predictive intervals rather than widening them. Against these margins, every variation of feature selection and hyperparameter tuning attempted produced differences of a few percent, so the recommended configuration is a univariate model with a small fixed feature set re-estimated daily on the longest window the dataset permits.

Taken together, these results qualify rather than confirm the expectation that motivated the work. Spatial and hierarchical information does carry predictive value at distribution level, but that value is conditional on the number and homogeneity of the series involved, on the form in

which the information is supplied, on whether the forecast has already captured it by other means, and on whether the target is the centre of the predictive distribution or its full shape. A univariate model combined with an appropriate reconciliation scheme proved competitive with, and frequently superior to, the multivariate alternatives, which is a result of practical relevance for network operators, since it is also the cheaper option to estimate and to maintain.

6.2 Limitations

Four limitations bound the scope of these conclusions and should be weighed when interpreting them.

The first and most consequential is the temporal extent of the dataset. Eleven months of history do not contain a complete annual cycle, so the test months were necessarily forecast by models whose training period never observed the corresponding season. Because those months are the summer ones, with school breaks and the August holiday period, the test set coincides with the segment of the year identified in Section 3.5 as the least regular, and the error levels reported throughout should be read as an upper bound on what the same methodology would attain over a full year. The effect is visible in the probabilistic results, where the residual under-coverage of the upper quantiles is the direct trace of a training window that only gradually acquires the summer regime.

The second limitation concerns the forecasting horizon. All results refer to a one-step-ahead horizon, which is the setting in which the recent load of neighbouring areas is most informative. Whether the spatial dependence identified here survives at the horizons of several hours or one day that most operational applications require was not tested, and there is reason to expect it to weaken, given that the measured cross-correlations vanish beyond two hours.

The third limitation is the coverage of the comparison. The point forecasts were produced under a static regime, since re-estimating the cross-level chain of out-of-fold forecasts for every test day was not tractable within this work, so the two approaches were not evaluated under identical training conditions. Given how large the effect of the sliding regime proved to be in the probabilistic strand, that asymmetry is likely to understate what the point forecasts could achieve.

The fourth limitation concerns the hierarchy itself, which follows the postal-code system rather than the physical topology of the distribution network. Since the published dataset is released at that resolution, the choice was imposed rather than made, but its consequences reach beyond the labels on the levels. The aggregates being forecast are not the quantities a network operator manages, because congestion, asset loading and flexibility procurement are handled at substation level, and the coherence that reconciliation enforces therefore holds along administrative boundaries rather than along the paths through which power actually flows. More consequential for the reading of the results, the usefulness of cross-location information was found to be governed by the number and homogeneity of the series sharing a level, and a partition drawn for postal purposes is arbitrary with respect to exactly that criterion. What is reported here is therefore what an administrative hierarchy delivers, which may understate what a physically grounded one would.

6.3 Future Work

Four directions follow naturally from the limitations above and from the results themselves.

The most promising follows directly from the last of those limitations. Rebuilding the hierarchy on the topology of the grid, aggregating consumption by secondary substation and then by primary substation and by HV/MV node, would group together the areas that are electrically rather than administratively related, and would align the levels with those over which operational decisions are actually taken. Reconciliation would then enforce coherence along the paths through which power genuinely flows, and the experiment would establish whether the moderate gains obtained here grow once the partition is drawn on a physically meaningful criterion, which is, in the view of this work, the single most informative question left open. A weaker version of the same idea, available without any additional data, is to replace the administrative levels by the clusters identified in Chapter 3, which would test whether the relevant neighbourhood is statistical rather than geographical.

A second direction is to repeat the study over a longer historical period. A dataset spanning two or more complete years would allow each test month to be forecast by models that have observed that month before, would permit the seasonal effect to be separated from the model effect, and would remove the ceiling that the eleven-month window imposes on the training length in the probabilistic strand, where the calibration had not yet saturated at the longest window available.

A third is to extend the moving-window regime to the point-forecasting pipeline. Since that regime accounted for the largest single share of the probabilistic error at every level, applying it to the LASSO-VAR-X forecasts would establish whether the ranking between the univariate and the multivariate scenarios is altered once both are re-estimated as the forecast origin advances, and whether the reconciliation gains reported here persist under a covariance estimated on a moving basis rather than fixed on the training period.

Finally, the probabilistic strand would benefit from coherent reconciliation across the hierarchy. The point forecasts were made coherent while the predictive quantiles were not, so the intervals reported at adjacent levels are mutually inconsistent, and extending the projection to the full predictive distribution would allow uncertainty to be propagated consistently from an individual area up to the national total.

References

- [1] José R. Andrade and Ricardo J. Bessa. “Improving Renewable Energy Forecasting with a Grid of Numerical Weather Predictions”. In: *IEEE Transactions on Sustainable Energy* 8.4 (Oct. 2017), pp. 1571–1580. DOI: [10.1109/TSTE.2017.2694340](https://doi.org/10.1109/TSTE.2017.2694340).
- [2] Ayoub Atanane, Loubna Benabbou, and Abderrazak El Ouafi. “Electricity demand Forecasting: A systematic literature review”. en. In: *2023 14th International Conference on Intelligent Systems: Theories and Applications (SITA)*. Casablanca, Morocco: IEEE, Nov. 2023, pp. 1–8. ISBN: 979-8-3503-0821-1. DOI: [10.1109/SITA60746.2023.10373741](https://doi.org/10.1109/SITA60746.2023.10373741). URL: <https://ieeexplore.ieee.org/document/10373741/>.
- [3] K. Bandara, R. J. Hyndman, and C. Bergmeir. “MSTL: A Seasonal-Trend Decomposition Algorithm for Time Series with Multiple Seasonal Patterns”. In: *International Journal of Operational Research* 52 (2025), p. 79. DOI: [10.1504/IJOR.2025.143957](https://doi.org/10.1504/IJOR.2025.143957).
- [4] Souhaib Ben Taieb, James W. Taylor, and Rob J. Hyndman. “Hierarchical Probabilistic Forecasting of Electricity Demand with Smart Meter Data”. In: *Journal of the American Statistical Association* 116.533 (2021), pp. 27–43. DOI: [10.1080/01621459.2020.1736081](https://doi.org/10.1080/01621459.2020.1736081).
- [5] Leonard Burg et al. “Comparative Analysis of Load Forecasting Models for Varying Time Horizons and Load Aggregation Levels”. In: *Energies* 14.21 (2021), p. 7128. DOI: [10.3390/en14217128](https://doi.org/10.3390/en14217128).
- [6] R. Chandrasekaran and S. K. Paramasivan. “Advances in Deep Learning Techniques for Short-term Energy Load Forecasting Applications: A Review”. In: *Archives of Computational Methods in Engineering* 32 (2025), pp. 663–692. DOI: [10.1007/s11831-024-10155-x](https://doi.org/10.1007/s11831-024-10155-x).
- [7] Tie Chen et al. “A regional short-term load forecasting model based on a cross-attention causal spatio-temporal convolutional network”. In: *International Journal of Electrical Power & Energy Systems* 172 (2025), p. 111275. DOI: [10.1016/j.ijepes.2025.111275](https://doi.org/10.1016/j.ijepes.2025.111275).
- [8] Robert B. Cleveland et al. “STL: A Seasonal-Trend Decomposition Procedure Based on Loess”. In: *Journal of Official Statistics* 6.1 (1990), pp. 3–73.
- [9] E-REDES – Distribuição de Eletricidade. *E-REDES Open Data Portal*. Accessed: 2026-05-21. 2024. URL: <https://e-redes.opendatasoft.com/pages/homepage/>.
- [10] E-REDES – Distribuição de Eletricidade. *Hourly Consumption by Postal Code - 4 Digits*. E-REDES Open Data Portal. Dataset Identifier: consumos_horario_codigo_postal, Accessed: 2026-05-21. 2024. URL: https://e-redes.opendatasoft.com/explore/dataset/consumos_horario_codigo_postal/.
- [11] Ciaran Gilbert, Jethro Browell, and David McMillan. “Leveraging Turbine-Level Data for Improved Probabilistic Wind Power Forecasting”. In: *IEEE Transactions on Sustainable Energy* 11.3 (July 2020), pp. 1152–1160. DOI: [10.1109/TSTE.2019.2920085](https://doi.org/10.1109/TSTE.2019.2920085).

- [12] Stephen Haben et al. “Spatio-temporal load forecasting at different aggregation levels”. In: *International Journal of Forecasting* 37.2 (2021), pp. 667–682. DOI: [10.1016/j.ijforecast.2020.09.005](https://doi.org/10.1016/j.ijforecast.2020.09.005).
- [13] Tao Hong and Shu Fan. “Probabilistic electric load forecasting: A tutorial review”. In: *International Journal of Forecasting* 32.3 (2016), pp. 914–938. DOI: [10.1016/j.ijforecast.2015.11.011](https://doi.org/10.1016/j.ijforecast.2015.11.011).
- [14] He Jiang et al. “Power load forecasting based on spatial–temporal fusion graph convolution network”. In: *Technological Forecasting and Social Change* 204 (2024), p. 123435. DOI: [10.1016/j.techfore.2024.123435](https://doi.org/10.1016/j.techfore.2024.123435).
- [15] Taesik Kim et al. “Structured stochastic models based on multi-source heterogeneous data for probabilistic prediction of building electric load”. In: *Journal of Building Engineering* 84 (2024), p. 108411. DOI: [10.1016/j.jobe.2023.108411](https://doi.org/10.1016/j.jobe.2023.108411).
- [16] Eryk Lewinson. *Introduction to Hierarchical Time Series Forecasting, Part I*. Accessed: 2026-05-20. 2021. URL: <https://medium.com/data-science/introduction-to-hierarchical-time-series-forecasting-part-i-88a116f2e2>.
- [17] Shuo Liu et al. “Performance enhancement of building electricity load forecasting based on nonparametric time series decomposition algorithm”. In: *Journal of Building Engineering* 114 (2025), p. 114246. DOI: [10.1016/j.jobe.2025.114246](https://doi.org/10.1016/j.jobe.2025.114246).
- [18] Shuo Liu et al. “Performance enhancement of building electricity load forecasting based on nonparametric time series decomposition algorithm”. In: *Journal of Building Engineering* 114 (2025), p. 114246. DOI: [10.1016/j.jobe.2025.114246](https://doi.org/10.1016/j.jobe.2025.114246).
- [19] Peter Nystrup et al. “Temporal hierarchies with autocorrelation for load forecasting”. In: *European Journal of Operational Research* 280 (2020), pp. 876–888. DOI: [10.1016/j.ejor.2019.07.061](https://doi.org/10.1016/j.ejor.2019.07.061).
- [20] Fotios Petropoulos et al. “Forecasting: theory and practice”. In: *International Journal of Forecasting* 38 (2022), pp. 705–871. DOI: [10.1016/j.ijforecast.2021.11.001](https://doi.org/10.1016/j.ijforecast.2021.11.001).
- [21] Marco G. Pinheiro, Sara C. Madeira, and Alexandre P. Francisco. “Short-term electricity load forecasting—A systematic approach from system level to secondary substations”. In: *Applied Energy* 332 (2023), p. 120493. DOI: [10.1016/j.apenergy.2022.120493](https://doi.org/10.1016/j.apenergy.2022.120493).
- [22] Plain English. *Unlocking Time Series Insights: Mastering Seasonal-Trend Decomposition Using LOESS (STL)*. <https://ai.plainenglish.io/unlocking-time-series-insights-mastering-seasonal-trend-decomposition-using-loess-stl-c205e7439fab>. 2023.
- [23] Han Qiu et al. “Short-Term Electricity Load Forecasting Based on Complete Ensemble Empirical Mode Decomposition with Adaptive Noise and Improved Sparrow Search Algorithm–Convolutional Neural Network–Bidirectional Long Short-Term Memory Model”. In: *Mathematics* 13 (2025), p. 813. DOI: [10.3390/math13050813](https://doi.org/10.3390/math13050813).
- [24] Nicolò Rubattu, Gabriele Maroni, and Giorgio Corani. “Electricity Load and Peak Forecasting: Feature Engineering, Probabilistic LightGBM and Temporal Hierarchies”. In: *Proceedings of the BigDeal Challenge 2022*. Lugano, Switzerland, 2022.
- [25] B. Selvi and S. Mishra. “Investigation of Performance of Electric Load Power Forecasting in Multiple Time Horizons With New Architecture Realized in Multivariate Linear Regression and Feed-Forward Neural Network Techniques”. In: *IEEE Transactions on Industry Applications* 56.5 (2020), pp. 5051–5062. DOI: [10.1109/TIA.2020.3009313](https://doi.org/10.1109/TIA.2020.3009313).

- [26] Navneet Sharma, Rohit Bhakar, and Prerna Jain. “Optimal reconciliation of hierarchical wind power forecasts of correlated wind farms”. In: *Sustainable Energy, Grids and Networks* 35 (2023), p. 101091. DOI: [10.1016/j.segan.2023.101091](https://doi.org/10.1016/j.segan.2023.101091).
- [27] P. Silva and R. J. Bessa. “Fuzzy-Enhanced Smart Building Energy Forecasting: A Case Study from the 2025 GECAD Competition”. In: *Energy Reports* 11 (2025), pp. 1120–1135. DOI: [10.1016/j.egyr.2025.01.042](https://doi.org/10.1016/j.egyr.2025.01.042).
- [28] James W. Taylor. “Triple seasonal methods for short-term electricity demand forecasting”. In: *European Journal of Operational Research* 204.1 (2010), pp. 139–152. DOI: [10.1016/j.ejor.2009.10.003](https://doi.org/10.1016/j.ejor.2009.10.003).
- [29] Hao Tong and Jun Liu. “MFformer: An improved transformer-based multi-frequency feature aggregation model for electricity load forecasting”. In: *Electric Power Systems Research* 243 (2025), p. 111492. DOI: [10.1016/j.epsr.2025.111492](https://doi.org/10.1016/j.epsr.2025.111492).
- [30] United Nations. *The Sustainable Development Goals*. 2025. URL: <https://sdgs.un.org/goals> (visited on 06/20/2026).
- [31] Yi Wang et al. “Review of Smart Meter Data Analytics: Applications, Methodologies, and Challenges”. In: *IEEE Transactions on Smart Grid* 9.6 (2018), pp. 6531–6547. DOI: [10.1109/TSG.2018.2818167](https://doi.org/10.1109/TSG.2018.2818167).
- [32] Raniyah Wazirali et al. “State-of-the-art review on energy and load forecasting in micro-grids using artificial neural networks, machine learning, and deep learning techniques”. en. In: *Electric Power Systems Research* 225 (Dec. 2023), p. 109792. ISSN: 03787796. DOI: [10.1016/j.epsr.2023.109792](https://doi.org/10.1016/j.epsr.2023.109792). URL: <https://linkinghub.elsevier.com/retrieve/pii/S0378779623006818>.
- [33] A. Yang, W. Li, and X. Yang. “Short-term electricity load forecasting based on feature selection and least squares support vector machines”. In: *Knowledge-Based Systems* 163 (2019), pp. 159–173. DOI: [10.1016/j.knosys.2018.08.027](https://doi.org/10.1016/j.knosys.2018.08.027).
- [34] Suyang Zhou et al. “Robust Energy Management in Active Distribution Systems Considering Temporal and Spatial Correlation”. In: *IEEE Access* 7 (2019), pp. 153635–153649. DOI: [10.1109/ACCESS.2019.2947415](https://doi.org/10.1109/ACCESS.2019.2947415).
- [35] Florian Ziel. “Modelling and Forecasting Electricity Load Using Lasso Methods”. In: *2015 Modern Electric Power Systems (MEPS)*. 2015. DOI: [10.1109/MEPS.2015.7477217](https://doi.org/10.1109/MEPS.2015.7477217).

Appendix A

Detailed Point Forecast Results

A.1 Validation Errors

Table A.1: Leave-one-month-out validation error of the spatio-temporal and hierarchical scenarios, by training month and hierarchical level. The lowest mean of each level is in bold.

Level	Scen.	Nov	Dec	Jan	Feb	Mar	Apr	May	Mean
Four-digit	S1	0.1164	0.1166	0.1126	0.1133	0.1130	0.1202	0.1125	0.1150
	S2	0.1165	0.1166	0.1128	0.1134	0.1131	0.1203	0.1126	0.1150
	S3	0.1178	0.1182	0.1140	0.1129	0.1135	0.1237	0.1136	0.1162
Two-digit	S1	0.0924	0.0920	0.0876	0.0873	0.0890	0.0895	0.0881	0.0894
	S2	0.0924	0.0919	0.0878	0.0874	0.0891	0.0896	0.0883	0.0895
	S3	0.0949	0.0941	0.0893	0.0880	0.0917	0.0918	0.0898	0.0914
	H1	0.0988	0.1006	0.0920	0.0940	0.0999	0.1035	0.0959	0.0978
	H2	0.1035	0.1053	0.0972	0.0957	0.1008	0.1033	0.1002	0.1009
	H3	0.0921	0.0920	0.0870	0.0889	0.0874	0.0928	0.0867	0.0896
One-digit	S1	0.0790	0.0800	0.0759	0.0773	0.0761	0.0806	0.0794	0.0783
	S2	0.0792	0.0800	0.0760	0.0776	0.0762	0.0809	0.0796	0.0785
	S3	0.0802	0.0798	0.0769	0.0785	0.0775	0.0856	0.0792	0.0797
	H1	0.0812	0.0799	0.0739	0.0724	0.0796	0.0777	0.0726	0.0767
	H2	0.0846	0.0861	0.0778	0.0830	0.0865	0.0845	0.0762	0.0827
	H3	0.0702	0.0683	0.0635	0.0651	0.0668	0.0673	0.0664	0.0668

A.2 Coefficient Inventories

Table A.2: Ten largest coefficients of three four-digit series in different regions, with the partial cross-correlation of each pair at the matching lag.

Rank	1200 (Lisbon)				8125 (Faro)				7000 (Évora)			
	Source	Lag	Coef.	PCCF	Source	Lag	Coef.	PCCF	Source	Lag	Coef.	PCCF
1	Own	1	+0.0942	–	Own	1	+0.0453	–	Own	1	+0.1050	–
2	2740	1	+0.0125	0.81	8135	1	+0.0151	0.86	5335	2	-0.0107	-0.13
3	1250	1	+0.0116	0.94	Own	24	+0.0141	–	2740	1	+0.0103	0.75
4	4930	168	+0.0082	0.02	5335	168	+0.0130	0.00	1050	1	+0.0075	0.80
5	4930	1	+0.0072	0.64	8400	24	+0.0097	0.01	2735	2	-0.0060	-0.19
6	4580	2	-0.0067	-0.20	2845	168	+0.0097	0.04	5335	168	+0.0058	0.02
7	Hour	–	-0.0067	–	5335	2	-0.0089	-0.12	Own	24	+0.0057	–
8	2845	168	+0.0063	0.04	8200	1	+0.0071	0.89	2845	1	+0.0043	0.56
9	5335	2	-0.0060	-0.08	2845	1	+0.0057	0.44	3680	1	+0.0041	0.64
10	1050	1	+0.0059	0.90	8400	1	+0.0056	0.84	4605	168	+0.0041	0.00

Appendix B

Complete Probabilistic Forecast Metrics

Table B.1: Probabilistic forecast metrics for every level, scenario and training window length W , averaged over the series of each level. The pinball loss, the CRPS and the interval width are expressed in kWh. The national level admits no multivariate scenario.

Level	Scen.	W	Pinball	CRPS	Calib.	Cov. 80 %	Cov. 90 %	Width 90 %
National	B2	90	19 201	39 469	0.089	0.581	0.720	196 845
	B2	180	15 291	32 354	0.054	0.689	0.825	195 154
	B2	210	15 251	32 299	0.049	0.708	0.830	197 158
One-digit	B2	90	3 064	6 312	0.091	0.563	0.698	30 574
	B2	180	2 590	5 451	0.065	0.645	0.781	30 154
	B2	210	2 581	5 443	0.059	0.661	0.792	30 733
	S2	90	3 256	6 709	0.085	0.578	0.710	33 559
	S2	180	2 774	5 792	0.064	0.648	0.773	32 663
	S2	210	2 750	5 759	0.058	0.661	0.791	33 214
Two-digit	B2	90	466.7	966.0	0.078	0.595	0.718	4 839
	B2	180	415.1	872.8	0.056	0.666	0.790	4 890
	B2	210	411.1	865.9	0.052	0.677	0.800	4 961
	S2	90	517.4	1 072.3	0.079	0.598	0.718	5 200
	S2	180	473.5	991.1	0.060	0.652	0.775	5 451
	S2	210	469.3	983.6	0.056	0.662	0.786	5 540
Four-digit	B2	90	93.7	194.6	0.069	0.620	0.740	967
	B2	180	85.9	180.6	0.051	0.679	0.797	989
	B2	210	85.3	179.6	0.047	0.688	0.807	999
	S2	90	106.7	222.2	0.070	0.623	0.742	1 131
	S2	180	101.4	212.9	0.057	0.668	0.785	1 175
	S2	210	100.9	212.0	0.054	0.677	0.793	1 191