TU Dortmund

Master Thesis

Analysis of the Correlation Structure in the Simbench data set

Author: Amritha Sukhdev Singh Agarwal

Matriculation Id: 230854

Supervisor: Prof. Dr. Christine Müller

Second Supervisor: Prof. Dr. Roland Fried

April 1, 2026

Contents

1 Introduction

Contemporary electrical grids have deviated from a simple, centralised paradigm. They now incorporate distributed energy resources, including rooftop solar, wind turbines, and battery storage, alongside flexible loads. These resources exhibit variable, often unpredictable production and consumption patterns. Consequently, sudden fluctuations in power flows arise, complicating conventional grid planning.

To manage these complexities, operators need accurate forecasting and analysis tools. Without them, grid instability and supply issues may arise. Conventional planning approaches rely on conservative assumptions and often result in overbuilt infrastructure. Data-driven alternatives extract patterns from time-series measurements to capture real operating behaviour. However, they require high-quality data and rigorous statistical methods for effective analysis. (Nowotarski and Weron, 2016)

1.1 The Role of Load Profiles in Grid Analysis

Load profiles document power consumption and generation over time at specific locations and are crucial for comprehensive grid analysis. Utilities commonly employ standardised load profiles based on customer categories; however, such aggregate profiles obscure both inter- and intra-customer variability, potentially leading to undetected local overloads within distribution networks. (Chicco, 2012)

In practice, actual power consumption varies with season, day of the week, weather, and user behaviour. Although individual changes are typically small, they accumulate in local networks, contributing to overloads or voltage fluctuations. Modern metering infrastructure now enables high-resolution measurements, supporting the creation of empirically grounded load datasets that capture these dynamics in greater detail. As a result, grid operators can more accurately assess and address local variations, improving both grid reliability and operational decision-making. Understanding the covariance structure of these load profiles, how variability and correlation patterns differ across seasons, weekdays, and network locations, is therefore essential for effective network planning and operation. This thesis focuses on analysing the covariance structure of power profiles across a distribution network.

1.2 The SimBench Dataset

This study uses the SimBench dataset, a publicly available, open-source benchmark for power system research. It provides time-series measurements for a complete electrical distribution grid, integrating measured consumption profiles from real commercial and household customers with realistically simulated data for distributed renewable generation, battery storage, heat pumps, and electric vehicle charging, all grounded in actual weather and physically validated models.

While SimBench seeks to mirror real-world conditions, the dataset mixes measured and simulated data, which may introduce certain limitations or biases, particularly in representing the variability of distributed generation and novel load types. Results should therefore be interpreted in light of these constraints.

SimBench was chosen because real-world grid measurement data are typically proprietary and access-restricted. Critically, it provides both detailed load patterns and complete grid topology, capabilities necessary for analysing inter-node correlations and network-level behaviour. Most public datasets offer only isolated time series without network context, limiting their suitability for grid-focused research. (Meinecke et al., 2020)

1.3 Research Questions and Objectives

This thesis aims to analyse the correlation and covariance structures of power profiles within a distribution grid. This objective directly informs the formulation of the primary research questions. The analysis is expected to contribute new insights into how temporal and spatial variations in load profiles affect grid reliability and to identify patterns that could guide more efficient infrastructure planning. By providing a detailed empirical characterisation of variability and correlation in real-world and realistically simulated load data, this work seeks to improve modelling approaches for distribution networks and inform practical decision-making in grid operation. The specific questions guiding this study are therefore:

- How do seasonal and weekday effects influence power load profiles across nodes?
- Does a sinusoidal model provide a more efficient representation of seasonal variation than categorical approaches?

- Are covariance structures stable across seasons, weekdays, and nodes, or do they exhibit significant variation?
- To what extent can model complexity be reduced without compromising the representation of variability?

These questions are examined using descriptive analysis, hypothesis testing, and covariance estimation. This helps to understand the data intuitively and supports the findings with statistical evidence.

1.4 Thesis Structure

The thesis is organised as follows. Section 2 describes the problem statement and dataset, outlining data sources and preparation steps. Section 3 details statistical methodologies, including seasonal modelling, multivariate hypothesis testing, and changepoint detection. Section 4 then covers initial data exploration, including redundancy analysis and strange days identification. Section 5 presents the main results on temporal variation in power consumption across seasons and weekdays and examines differences in the covariance structure across temporal and network factors. Finally, Section 6 summarises the findings and discusses implications for grid analysis and future research.

2 Data Description & Notation

This section introduces the dataset used throughout the analysis and establishes the notation required for subsequent modelling and interpretation. A clear understanding of both is essential, as all methodological developments and empirical results are based on these definitions.

2.1 Data Description

The following analysis utilises SimBench data for power distribution grids, specifically the `1-LV-rural1-1-no_sw` configuration (SimBench Project, 2024). The dataset comprises per-unit, normalised active and reactive power, measured at 15-minute intervals during 2016 across 13 grid nodes, yielding $366 \times 96 = 35{,}136$ observations per node per power type for both residential and agricultural consumers. Two dataset versions were available: one with all 366 days, and another excluding 26 March and 30 October due to clock adjustments. This analysis employs the complete 366-day dataset.

For each node, the power measurements are available as complete time series, $P_n(t)$ and $Q_n(t)$, with t denoting 15-minute intervals spanning one year. The time series is divided into 366 days and 96 quarter-hour intervals, resulting in discrete variables $P_{n,d,h}$ and $Q_{n,d,h}$ for subsequent analyses. The distinction between h and H is important, as some analyses, such as Section 4's MANOVA, are conducted at the original 15-minute resolution, while Section 5 relies on hourly-aggregated data.

2.2 Notation & Indexing

The following notations are consistently used throughout the thesis:

- $n \in \{1, \dots, 13\}$ denotes the node index.
- $d \in \{1, \dots, 366\}$ denotes the day.
- $h \in \{1, \dots, 96\}$ denotes the quarter-hour index within a day.
- $H \in \{0, 1, \dots, 23\}$ denotes the hourly index.
- $\omega \in \{1, \dots, 52\}$ denotes the week.
- $w \in \{1, \dots, 7\}$ denotes the weekdays with $w = 1$ for Monday and $w = 7$ for Sunday.
- $P_{n,d,h}$ = active power.
- $Q_{n,d,h}$ = reactive power.
- $f_w : \{1, \dots, 366\} \ni d \to f_w(d) \in \{1, 2, .., 7\}$.
- $f_s : \{1, \dots, 366\} \ni d \to f_s(d) \in \{1, 2, 3\},\ f_s(d) = \begin{cases} 1, \text{if } d \in \text{Summer (15 May – 14 Sept)}, \\ 2, \text{if } d \in \text{Winter (1 Nov – 20 March)}, \\ 3,\ \text{Rest of the year.} \end{cases}$

The grid structure and the node layout are shown in Figure 1. The nodes were renumbered from the original grid labelling for consistency, as given in Table 1.

Old Node	1	2	3	4	5	6	7	8	9	10	11	12	13	14
New Node	1	2	13	14	9	8	5	10	3	12	11	6	4	7

Table 1: Node renumbering from the old grid structure to the new grid structure.

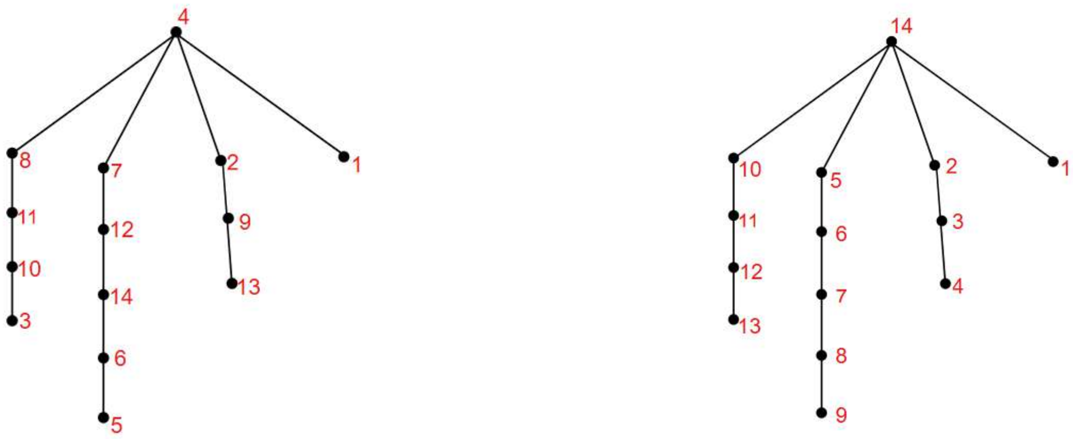

Figure 1: Original (left) and renamed (right) grid structures.

3 Methods

In this section, the statistical methods used throughout the analysis are explained. R (R Core Team, 2021) version 4.1.2 with packages `changepoint` (Killick and Eckley, 2023), `corrplot` (Wei and Simko, 2024), `ggplot2` (Wickham, 2016), `lubridate` (Grolemund and Wickham, 2024), `RColorBrewer` (Neuwirth, 2022), and `tidyverse` (Wickham et al., 2019) is used for all calculations and visualisations. RStudio (Posit team, 2024) is used to run the utilities. The analysis employs changepoint detection, MANOVA with Pillai's trace to assess seasonal and weekday effects on load profiles, and permutation-based Frobenius distance testing to compare covariance structures across groups. Pearson correlation and sample covariance matrices are used for descriptive analysis of residual and inter-weekday dependencies. A random seed of 12345 ensures reproducibility of all results.

3.1 Smoothing

In many real-world time series, observed values are affected by random fluctuations that hide the underlying structure of interest. Smoothing methods reduce this noise by decomposing the observed series into two components,

$$\text{Data} = \text{Signal} + \text{Rough}.$$

A symmetric distribution with mean zero and impulsive outliers makes up the rough, while the signal captures the systematic structure. Smoothing methods are flexible and data-driven, imposing no rigid functional form on the data. The primary objective is to suppress short-term irregularities so that systematic structures, such as seasonal and weekday patterns, can be identified.

The running median filter is a common smoother that preserves abrupt shifts and is robust to outliers, as the median ignores extreme values within the window. Let Z_t^n denote observations at time t for node n. With a half-window bandwidth $k_1 \in \mathbb{N}_0$, the running median smooths using

$$\tilde{Z}_t^n := \text{median}\left\{Z_\tau^n : \tau = t - k_1, \ldots, t + k_1\right\}, \quad t = 1 + k_1, \ldots, T - k_1.$$

Here, $2k_1 + 1$ is the total window width. The parameter k_1 controls the degree of smoothing. Larger values produce smoother estimates but may suppress genuine short-term patterns, while smaller values retain more local variation. (Tukey, 1977, p. 205–235)

3.2 Covariance & Correlation Estimation

Covariance measures the linear dependence between random variables X and Y as follows:

$$\sigma_{XY} = \text{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y),$$

where $\mathbb{E}$ is the expected value of the random variables. In the multivariate setting, let $X = (X_1, \ldots, X_p)^\top \in \mathbb{R}^p$ be a p-dimensional random vector. The covariance matrix

$\Sigma \in \mathbb{R}^{p \times p}$ collects all pairwise covariances:

$$\Sigma = \begin{pmatrix} \sigma_{X_1X_1} & \cdots & \sigma_{X_1X_p} \\ \vdots & \ddots & \vdots \\ \sigma_{X_pX_1} & \cdots & \sigma_{X_pX_p} \end{pmatrix}.$$

The diagonal entries $\sigma_{ii} = \mathrm{Var}(X_i)$ are the marginal variances, and the off-diagonal entries $\sigma_{ij} = \mathrm{Cov}(X_i, X_j)$ capture pairwise linear dependence.

Estimating covariance in practice typically involves data, since the true covariance matrix is usually unknown. For n observations $x_1, \ldots, x_n \in \mathbb{R}^p$, the sample covariance matrix is calculated as follows:

$$\hat{\Sigma} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^\top,$$

with $\bar{x}$ being the mean. This $n-1$ factor ensures an unbiased estimate. Note that $\hat{\Sigma}$ is symmetric and positive semi-definite. When data are multivariate normal and $n > p$, $\hat{\Sigma}$ is positive definite with probability 1.

The correlation can be defined by standardising the covariance matrix and is given as

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X) \cdot \mathrm{Var}(Y)}}.$$

The correlation value can range from -1 to 1. The sign indicates the relationship: a positive correlation indicates an upward slope, and a negative correlation indicates a downward slope. If the value is zero, it means there is no correlation.

In the multivariate setting, the correlation matrix is given as

$$\mathcal{P} = \begin{pmatrix} \rho_{X_1X_1} & \cdots & \rho_{X_1X_p} \\ \vdots & \ddots & \vdots \\ \rho_{X_pX_1} & \cdots & \rho_{X_pX_p} \end{pmatrix}.$$

(Härdle et al., 2024, p. 72–80)

3.3 Permutation-based Frobenius Distance Testing

To test whether covariance matrices differ across groups, a permutation-based test using the Frobenius norm is used. The Frobenius norm of a matrix A is defined as $\|A\|_F = \sqrt{\sum_{i,j} A_{ij}^2}$, and the squared Frobenius distance between two symmetric matrices A and B of the same dimension is

$$T_F = \|A - B\|_F^2 = \sum_{i,j} (A_{ij} - B_{ij})^2 .$$

Here, i and j index the rows and columns of the matrix. This measures the overall discrepancy between two covariance matrices, summing the squared differences across all entries. When comparing more than two groups $g = 1, \ldots, G$, the test statistic is the total pairwise Frobenius distance, summing over all $\binom{G}{2}$ pairs:

$$T_{\text{obs}} = \sum_{g<g'} \left\| \hat{\Sigma}_g - \hat{\Sigma}_{g'} \right\|_F .$$

The null hypothesis states that each group has an identical covariance matrix, $H_0 : \Sigma_1 = \cdots = \Sigma_G$. To obtain a p-value without distributional assumptions, a permutation test is used:

1. Compute T_{obs} from the data with the original group labels.
2. Randomly permute the group labels $B = 1000$ times, keeping the data fixed.
3. For each permutation b, recompute the sample covariance matrix per group and compute the total Frobenius distance $T^{(b)}$.
4. The p-value is estimated as

$$\hat{p} = \frac{1}{B} \# \left\{ T^{(b)} \geq T_{\text{obs}} \right\} .$$

$\hat{p} < 0.05$ indicates that the covariance structures of the groups differ significantly, thus rejecting the null hypothesis. When the global test is significant, pairwise comparisons between individual groups are carried out using the same permutation procedure applied to each pair separately with B permutations. Following the closure principle, pairwise tests are conducted only after the global null hypothesis is rejected, thereby controlling the familywise error rate without requiring a Bonferroni correction. All permutation tests use a fixed random seed of 12345 for reproducibility. The test is implemented in R using custom functions without additional packages. (Cabrieto et al., 2018)

3.4 MANOVA

Multivariate analysis of variance (MANOVA) treats each observation as a vector of responses to calculate the mean vector across multiple dependent variables. This method simultaneously tests all dependent variables, effectively controlling the risk of Type I errors across the entire set of responses. MANOVA is widely used in both experimental and observational research, as it evaluates the effects of categorical independent variables on multiple dependent variables.

Let $\mathbf{y}_{\ell j} = (y_{\ell j1}, \ldots, y_{\ell jp})^\top \in \mathbb{R}^p$ represent the p-dimensional observation vector for the j-th observation in group ℓ. Here, ℓ ranges from 1 to g, indicating the different groups, while j ranges from 1 to n_ℓ, representing the observations within each group.The one-way MANOVA model is given by

$$\mathbf{y}_{\ell j} = \boldsymbol{\mu} + \boldsymbol{\tau}_\ell + \mathbf{e}_{\ell j}, \quad j = 1, \ldots, n_\ell, \quad \ell = 1, \ldots, g,$$

where $\boldsymbol{\mu} \in \mathbb{R}^p$ is the overall mean vector, $\boldsymbol{\tau}_\ell$ is the effect of group ℓ subject to $\sum_{\ell=1}^{g} n_\ell \boldsymbol{\tau}_\ell = \mathbf{0}$, and $\mathbf{e}_{\ell j}$ is the random error vector with $\mathbf{e}_{\ell j} \overset{\text{iid}}{\sim} N_p(\mathbf{0}, \boldsymbol{\Sigma})$. The errors for the components of $\mathbf{y}_{\ell j}$ are correlated, but the covariance matrix $\boldsymbol{\Sigma}$ is assumed to be the same for all groups.

When observations are classified by two factors simultaneously, the model extends to the two-way case. Let factor A have a levels indexed by $i = 1, \ldots, a$, and factor B have b levels indexed by $k = 1, \ldots, b$, with $j = 1, \ldots, n_{ik}$ observations in cell (i, k). The two-way fixed-effects MANOVA model with interaction is:

$$\mathbf{y}_{ikj} = \boldsymbol{\mu} + \boldsymbol{\alpha}_i + \boldsymbol{\beta}_k + \boldsymbol{\gamma}_{ik} + \mathbf{e}_{ikj}, \qquad \mathbf{e}_{ikj} \overset{\text{iid}}{\sim} N_p(\mathbf{0}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\mu} \in \mathbb{R}^p$ is the overall mean vector, $\boldsymbol{\alpha}_i$ represents the main effect at level i of factor A, while $\boldsymbol{\beta}_k$ represents the main effect at level k of factor B, $\boldsymbol{\gamma}_{ik}$ is the interaction effect of the combination (i, k), and $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$ is the common within-cell covariance matrix. The identifiability constraints are:

$$\sum_{i=1}^{a} \boldsymbol{\alpha}_i = \mathbf{0}, \quad \sum_{k=1}^{b} \boldsymbol{\beta}_k = \mathbf{0}, \quad \sum_{i=1}^{a} \boldsymbol{\gamma}_{ik} = \mathbf{0}, \quad \sum_{k=1}^{b} \boldsymbol{\gamma}_{ik} = \mathbf{0}.$$

When the interaction term is not significant, the additive model with $\boldsymbol{\gamma}_{ik} = \mathbf{0}$ is used instead. (Johnson and Wichern, 2007, p. 301–315)

3.4.1 Multivariate Regression Extension

MANOVA extends naturally to settings where the multivariate response vectors are not raw observations but estimated quantities, such as regression coefficients. Given n observations with p-dimensional response vectors $\mathbf{y}_1, \ldots, \mathbf{y}_n \in \mathbb{R}^p$ and q-dimensional covariate vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^q$, the multivariate linear regression model is:

$$\mathbf{y}_i = \mathbf{B}^\top \mathbf{x}_i + \mathbf{e}_i, \qquad \mathbf{e}_i \overset{\text{iid}}{\sim} N_p(\mathbf{0}, \boldsymbol{\Sigma}),$$

where $\mathbf{B} \in \mathbb{R}^{q \times p}$ is the unknown coefficient matrix.

When observations belong to g subgroups, one can test whether the coefficient vectors are equal across groups. This reduces to the standard MANOVA hypothesis:

$$H_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_g,$$

where $\boldsymbol{\mu}_\ell = \mathbf{B}_\ell^\top \bar{\mathbf{x}}_\ell$ is the mean response vector in group ℓ, tested using the $\mathbf{H}$ and $\mathbf{E}$ matrices and Pillai's trace as described subsequently. (Johnson and Wichern, 2007, p. 387–388)

3.4.2 Assumptions

MANOVA, like any other statistical test, is based on some assumptions, and they are:

1. Independence: the observations $\mathbf{y}_{\ell j}$ are independent.
2. Multivariate Normality: all the dependent variables are multivariately normally distributed.
3. Homogeneity of Covariance Matrix: the covariance matrix $\boldsymbol{\Sigma}$ across the groups is equal.

(Johnson and Wichern, 2007, p. 296–297)

3.4.3 Hypothesis Test

The null hypothesis for the one-way case states that all group mean vectors are equal and is given as

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_g,$$

$$H_1 : \boldsymbol{\mu}_\ell \neq \boldsymbol{\mu}_{\ell'} \text{ for at least one pair } \ell \neq \ell',$$

where $\boldsymbol{\mu}_\ell = \mathbb{E}[\mathbf{y}_{\ell j}] \in \mathbb{R}^p$ is the mean vector of group ℓ.

In the case of the two-way case, three separate hypotheses are tested:

$$H_0^{(1)} : \boldsymbol{\alpha}_1 = \cdots = \boldsymbol{\alpha}_a = \mathbf{0} \quad \text{(no main effect of factor A)}, \tag{1}$$

$$H_0^{(2)} : \boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_b = \mathbf{0} \quad \text{(no main effect of factor B)}, \tag{2}$$

$$H_0^{(3)} : \boldsymbol{\gamma}_{ik} = \mathbf{0}\ \forall\, i, k \quad \text{(no interaction)}. \tag{3}$$

(Johnson and Wichern, 2007, p. 316)

3.4.4 Statistical Tests

Evaluation of the null hypothesis involves partitioning the total variation into between-group and within-group components, expressed as matrices of sums of squares and cross-products (SSCP).

3.4.4.1 One-way case. The total SSCP matrix decomposes as $\mathbf{T} = \mathbf{H} + \mathbf{E}$, where:

$$\mathbf{H} = \sum_{\ell=1}^{g} n_\ell (\bar{\mathbf{y}}_\ell - \bar{\mathbf{y}})(\bar{\mathbf{y}}_\ell - \bar{\mathbf{y}})^\top,$$

$$\mathbf{E} = \sum_{\ell=1}^{g} \sum_{j=1}^{n_\ell} (\mathbf{y}_{\ell j} - \bar{\mathbf{y}}_\ell)(\mathbf{y}_{\ell j} - \bar{\mathbf{y}}_\ell)^\top,$$

with $\bar{\mathbf{y}}_\ell = \frac{1}{n_\ell} \sum_j \mathbf{y}_{\ell j}$ and $\bar{\mathbf{y}} = \frac{1}{n} \sum_\ell \sum_j \mathbf{y}_{\ell j}$.

3.4.4.2 Two-way case. The total SSCP matrix decomposes into four components:

$$\mathbf{T} = \mathbf{H}_A + \mathbf{H}_B + \mathbf{H}_{AB} + \mathbf{E},$$

where the components are defined as follows. Let

- $\bar{\mathbf{y}}_{i\cdot} = \frac{1}{bn}\sum_{k=1}^{b}\sum_{j=1}^{n}\mathbf{y}_{ikj}$ be the marginal mean vector for level i of factor A (averaging over all bn observations at that level),
- $\bar{\mathbf{y}}_{\cdot k} = \frac{1}{an}\sum_{i=1}^{a}\sum_{j=1}^{n}\mathbf{y}_{ikj}$ be the marginal mean vector for level k of factor B (averaging over all an observations at that level),
- $\bar{\mathbf{y}}_{ik} = \frac{1}{n}\sum_{j=1}^{n}\mathbf{y}_{ikj}$ be the cell mean vector for cell (i, k),
- $\bar{\mathbf{y}} = \frac{1}{abn}\sum_{i=1}^{a}\sum_{k=1}^{b}\sum_{j=1}^{n}\mathbf{y}_{ikj}$ be the overall grand mean vector.

The four SSCP matrices are then:

$$\mathbf{H}_A = bn\sum_{i=1}^{a}(\bar{\mathbf{y}}_{i\cdot} - \bar{\mathbf{y}})(\bar{\mathbf{y}}_{i\cdot} - \bar{\mathbf{y}})^\top, \quad \text{d.f.} = a - 1,$$

$$\mathbf{H}_B = an\sum_{k=1}^{b}(\bar{\mathbf{y}}_{\cdot k} - \bar{\mathbf{y}})(\bar{\mathbf{y}}_{\cdot k} - \bar{\mathbf{y}})^\top, \quad \text{d.f.} = b - 1,$$

$$\mathbf{H}_{AB} = n\sum_{i=1}^{a}\sum_{k=1}^{b}\left(\bar{\mathbf{y}}_{ik} - \bar{\mathbf{y}}_{i\cdot} - \bar{\mathbf{y}}_{\cdot k} + \bar{\mathbf{y}}\right)\left(\bar{\mathbf{y}}_{ik} - \bar{\mathbf{y}}_{i\cdot} - \bar{\mathbf{y}}_{\cdot k} + \bar{\mathbf{y}}\right)^\top, \quad \text{d.f.} = (a-1)(b-1),$$

$$\mathbf{E} = \sum_{i=1}^{a}\sum_{k=1}^{b}\sum_{j=1}^{n}(\mathbf{y}_{ikj} - \bar{\mathbf{y}}_{ik})(\mathbf{y}_{ikj} - \bar{\mathbf{y}}_{ik})^\top, \quad \text{d.f.} = ab(n-1).$$

The quantity $\hat{\boldsymbol{\gamma}}_{ik} = \bar{\mathbf{y}}_{ik} - \bar{\mathbf{y}}_{i\cdot} - \bar{\mathbf{y}}_{\cdot k} + \bar{\mathbf{y}}$ in $\mathbf{H}_{AB}$ is the estimated interaction effect for cell (i, k): it measures the deviation of the cell mean from what would be expected under the additive model (no interaction). Each hypothesis $H_0^{(1)}$–$H_0^{(3)}$ from equations (1)–(3) is tested by replacing $\mathbf{H}$ with $\mathbf{H}_A$, $\mathbf{H}_B$, or $\mathbf{H}_{AB}$ respectively in Pillai's trace statistic, using $\mathbf{E}$ as the error matrix in all three cases.

Pillai's test statistic is then based on the eigenvalues $\lambda_1, \ldots, \lambda_s$ of $\mathbf{E}^{-1}\mathbf{H}$ and is given by

$$V^{(s)} = \operatorname{tr}\Big[(\mathbf{E} + \mathbf{H})^{-1}\mathbf{H}\Big] = \sum_{i=1}^{s}\frac{\lambda_i}{1 + \lambda_i},$$

where $s = \min(p, g-1)$, with p representing the dimension of the response vector and g indicating the number of groups being compared. The eigenvalues $\lambda_1, \ldots, \lambda_s$ of $\mathbf{E}^{-1}\mathbf{H}$ summarise the ratio of between-group to within-group variation. The hypothesis is rejected for larger values of $V^{(s)}$. Pillai's trace is the most robust to violations of MANOVA's assumptions and is frequently used, especially with small or unequal sample sizes. The p-value is obtained from an F-approximation to the distribution of $V^{(s)}$ and is given as

$$F \approx \frac{2(N-s-1)}{2m+s+1} \cdot \frac{V^{(s)}}{m-V^{(s)}},$$

where $m = \max(a-1,\, b-1,\, p)$ and $N = \sum n_{ij}$. (Johnson and Wichern, 2007, p. 302–320, 336)

3.5 Change point detection

A changepoint is a time point where the statistical properties of a time series, such as mean or variance, change abruptly. Given a sequence $y_1, y_2, \ldots, y_T$, the goal is to identify m changepoints at locations $\tau_1 < \tau_2 < \cdots < \tau_m$. This partitions the data into $m+1$ segments. Each segment contains observations from a common distribution. The segments are $\{y_1, \ldots, y_{\tau_1}\}$, $\{y_{\tau_1+1}, \ldots, y_{\tau_2}\}$, ..., $\{y_{\tau_m+1}, \ldots, y_T\}$.

A common approach to detecting multiple changepoints is to minimise a penalised cost function. This method considers all possible numbers and locations of changepoints. Let $\mathcal{C}(y_{(s+1):t})$ denote the cost of fitting a segment to the observations from time $s+1$ to t. The optimal set of changepoints $\{\hat{\tau}_1, \ldots, \hat{\tau}_m\}$ minimises the total penalised cost:

$$\sum_{i=1}^{m+1} \mathcal{C}\Big(y_{(\tau_{i-1}+1):\tau_i}\Big) + m\beta,$$

with $\tau_0 = 0$, $\tau_{m+1} = T$, and $\beta > 0$ as a penalty balancing model fit and the number of detected changepoints. Increasing β results in fewer changepoints. This thesis uses two algorithms to minimise this criterion, described below. (Killick et al., 2011, p.1590-1591)

3.5.1 PELT

The Pruned Exact Linear Time (PELT) algorithm exactly minimises the penalised cost function. It uses dynamic programming with a pruning step. Pruning eliminates changepoint candidates that cannot be optimal. It reduces computational complexity while retaining solution accuracy.

Define $F(t)$ as the minimum penalised cost for the data $y_{1:t}$, with $F(0) = -\beta$ as initialisation. The dynamic programming recursion is:

$$F(t) = \min_{\tau \in \{0,1,\ldots,t-1\}} \Big[F(\tau) + \mathcal{C}(y_{(\tau+1):t}) + \beta\Big].$$

The optimal solution is found by back-tracking through the τ values that achieved each minimum.

At each step of the recursion, the set of candidate changepoint locations R_t is updated using the pruning rule. A candidate location τ is retained in R_{t+1} only if

$$F(\tau) + \mathcal{C}(y_{(\tau+1):t}) + K < F(t),$$

and is permanently discarded otherwise. Here K is a constant satisfying

$$\mathcal{C}(y_{(\tau+1):t}) + \mathcal{C}(y_{(t+1):T}) + K \leq \mathcal{C}(y_{(\tau+1):T}),$$

for all $\tau < t < T$. Any candidate τ failing this condition can never be the optimal last changepoint prior to any future time $T > t$, and is therefore safely removed from consideration. Formally, the candidate set is updated as

$$R_{t+1} = \left\{ \tau \in R_t : F(\tau) + \mathcal{C}(y_{(\tau+1):t}) + K < F(t) \right\}.$$

If the number of changepoints scales linearly with T, pruning reduces the expected cost from $\mathcal{O}(T^2)$ for standard dynamic programming down to $\mathcal{O}(T)$ for PELT. In R, PELT is implemented via `cpt.mean()` in the `changepoint` package, with `method = "PELT"`. (Killick et al., 2011, p. 1592)

3.5.2 Binary Segmentation

Binary Segmentation (BinSeg) is one of the most established methods for multiple changepoint detection. The algorithm begins by applying a single changepoint method to the entire sequence $y_{1:T}$. It tests whether a τ exists such that

$$\mathcal{C}(y_{1:\tau}) + \mathcal{C}(y_{(\tau+1):T}) + \beta < \mathcal{C}(y_{1:T}).$$

If this condition is false, no changepoint is detected. The method then stops. If true, the data are split at the identified changepoint τ_a. The single changepoint method is then applied to each new segment independently. This process repeats until no further changepoints are detected in any segment.

BinSeg minimises the penalised cost function by introducing a changepoint at each step only if it reduces the total cost. It has a computational cost of $\mathcal{O}(T \log T)$, which is efficient. However, it does not guarantee the global minimum. Each split depends on previously detected changepoints, so earlier errors cannot be corrected later. In R, Binary Segmentation is implemented via `cpt.mean()` with `method = "BinSeg"`. (Killick et al., 2011, p. 1591)

3.6 Goodness of Fit

Goodness-of-fit measures how well a statistical model matches the data. R-squared is a common measure of goodness-of-fit. It shows how much variance in the response variable is explained by the predictor variable, based on the response variable, and is defined as

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}.$$

Here y_i are the observed values, $\hat{y}_i$ are the fitted values, $\bar{y}$ is the sample mean, and $\hat{\varepsilon}_i = y_i - \hat{y}_i$ are the residuals. Higher R-squared values mean a better fit, with values ranging from 0 to 1. A value of 1 means all residuals are zero and the fit is perfect. (Fahrmeir et al., 2013, p.112-115)

3.7 F-test

When two linear models are nested, the reduced model is a special case of the full model with some parameters set to zero. An F-test determines if the extra parameters in the full model significantly improve the fit by comparing how much additional variance they explain relative to the unexplained variance. Let SSE be the residual sum of squares for the full model and SSE_{H_0} for the reduced model under the null hypothesis. The test statistic is

$$F = \frac{n-p}{r} \cdot \frac{\Delta \mathrm{SSE}}{\mathrm{SSE}} = \frac{n-p}{r} \cdot \frac{\mathrm{SSE}_{H_0} - \mathrm{SSE}}{\mathrm{SSE}},$$

Here, n is the number of observations, p is the number of full model parameters, and r is the number of extra parameters being tested. Under the null hypothesis with Gaussian errors, F follows an F-distribution with $(r, n-p)$ degrees of freedom. A large F indicates that removing parameters increases the residual sum of squares, contradicting the null hypothesis. (Fahrmeir et al., 2013, p. 128–130)

3.8 Multiple Testing Correction: Benjamini–Hochberg FDR

When m hypothesis tests are performed simultaneously, the probability of at least one false positive increases with m. The Benjamini–Hochberg (BH) procedure is used to control the false discovery rate (FDR), reflecting the average tendency to incorrectly reject null hypotheses among all rejections. Given ordered p-values $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$ and significance level α, the procedure finds the largest index k such that

$$p_{(k)} \leq \frac{k}{m}\,\alpha,$$

and rejects all hypotheses $H_{(1)}, \ldots, H_{(k)}$. The BH procedure is less conservative than the Bonferroni correction and has greater power when many tested hypotheses are truly non-null. (Benjamini and Hochberg, 1995, p. 289–300)

4 Cluster and Correlation Analysis

Before analysis, three preparatory steps ensured data quality and relevance. First, raw 15-minute power series were smoothed with a running median to remove noise and spikes. Next, 13 nodes were checked for redundancy and grouped by identical patterns, selecting a smaller set of representative nodes. Finally, changepoint analysis identified changes in these nodes and removed atypical days. Each step is outlined below with transitions for clarity.

4.1 Data Smoothing and Preprocessing

Before statistical analyses, the raw power time series were preprocessed to remove short-term fluctuations from measurement noise and transient demand. A running median filter with window size k = 97 (about one day at 15-minute intervals) was applied independently to each node's active and reactive power series. The window size of 97 balances two goals: smaller windows ($k < 50$) retain too much high-frequency noise, while larger windows ($k > 150$) over-smooth and obscure intra-day structure.

The running median is robust to outliers and extreme values, which is crucial for changepoint detection analysis (Section 4.3) and mean-based load profile estimation. All subsequent analyses use the smoothed series $\tilde{P}_{n,d,h}$ and $\tilde{Q}_{n,d,h}$.

4.2 Cluster Identification

With 13 nodes in the grid, analysing each node separately would be computationally demanding and difficult to interpret. After smoothing, pairwise correlations were computed for both active and reactive power to identify groups of nodes with structurally identical consumption patterns. Nodes with pairwise correlation equal to 1 were grouped into clusters, as can be seen in Figures 2 and 3. Two clusters were identified for each power type:

- Active power clusters: C1 $= \{1, 3, 5, 9, 12\}$, C2 $= \{2, 6, 7, 8, 13\}$ (agricultural load profiles).
- Reactive power clusters: C1 $= \{1, 3, 5, 9, 12, 13\}$, C2 $= \{2, 6, 7, 8\}$.
- Remaining nodes: 4, 10, and 11, do not belong to either cluster and are characterised by residential load profiles.

To formally confirm this, pairwise linear regressions of the form

$$P_{n_2,t} = \beta_1 \cdot P_{n_1,t} + \varepsilon_t, \quad n_1, n_2 \in C_k,$$

were fitted between all node pairs within each candidate cluster. Here, $P_{n_1,t}$ and $P_{n_2,t}$ denote the active power at nodes n_1 and n_2 at time point t. β_1 is the slope representing the multiplicative scaling factor between the two nodes. ε_t is the residual error term, and C_k denotes the k-th cluster. The estimated intercepts were negligible in all cases (of order 10^{-10} or smaller). This confirms that the relationship between nodes within a cluster is purely multiplicative with no constant offset. All regressions yielded $R^2 = 1$. Differences appeared only in the slope β_1, as seen in Tables 2 and 3. The same analysis was repeated for reactive power, replacing $P_{n_1,t}$ and $P_{n_2,t}$ with $Q_{n_1,t}$ and $Q_{n_2,t}$ throughout.

Table 2: Regression results between P-node pairs by cluster.

Cluster 1				Cluster 2			
Node X	Node Y	Slope ($\hat{\beta}_1$)	R^2	Node X	Node Y	Slope ($\hat{\beta}_1$)	R^2
P_Node1	P_Node3	0.29	1.00	P_Node2	P_Node6	3.00	1.00
P_Node1	P_Node5	0.21	1.00	P_Node2	P_Node7	1.25	1.00
P_Node1	P_Node9	1.00	1.00	P_Node2	P_Node8	0.75	1.00
P_Node1	P_Node12	0.43	1.00	P_Node2	P_Node13	2.00	1.00
P_Node3	P_Node5	0.75	1.00	P_Node6	P_Node7	0.42	1.00
P_Node3	P_Node9	3.50	1.00	P_Node6	P_Node8	0.25	1.00
P_Node3	P_Node12	1.50	1.00	P_Node6	P_Node13	0.67	1.00
P_Node5	P_Node9	4.67	1.00	P_Node7	P_Node8	0.60	1.00
P_Node5	P_Node12	2.00	1.00	P_Node7	P_Node13	1.60	1.00
P_Node9	P_Node12	0.43	1.00	P_Node8	P_Node13	2.67	1.00

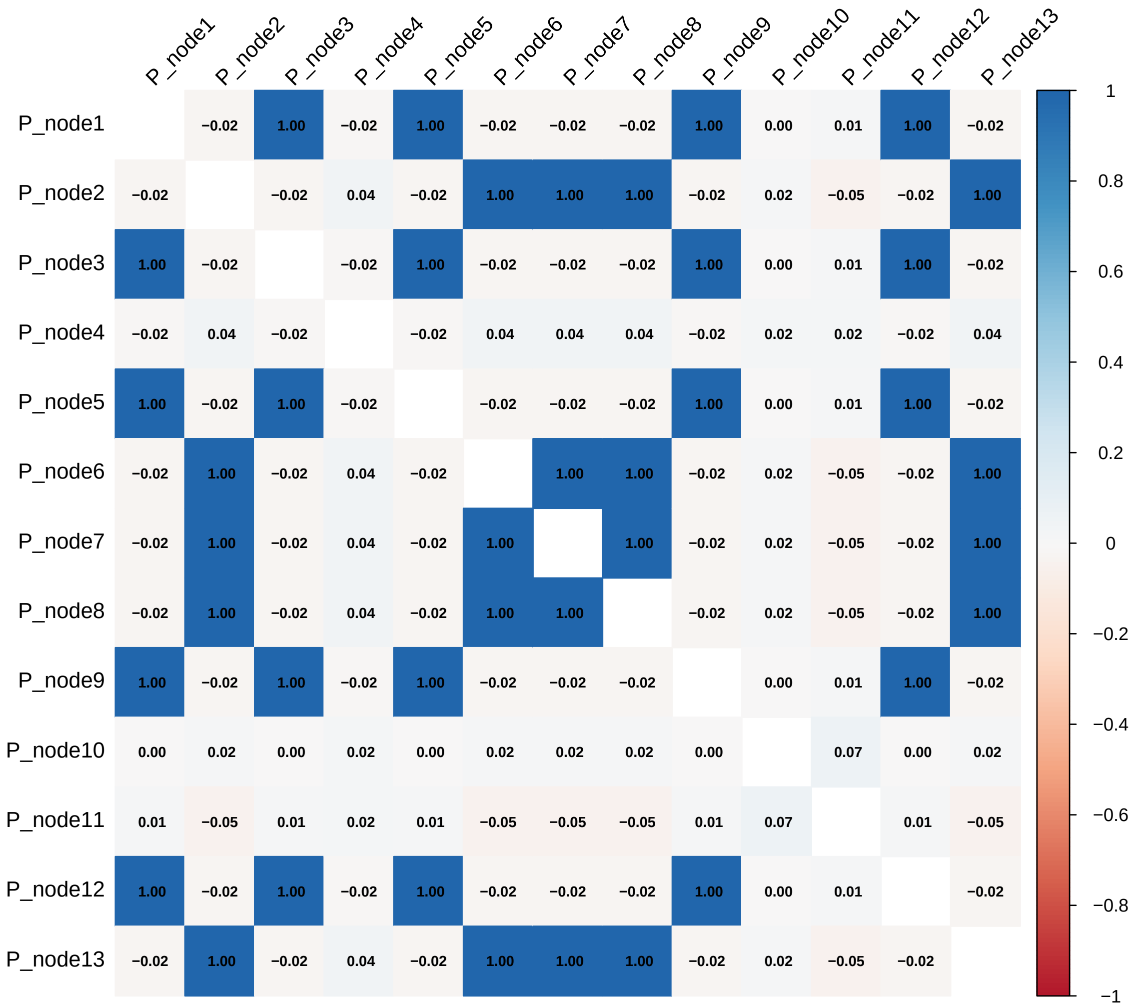

Figure 2: Correlation matrix for active (P) power.

Table 3: Regression results between Q-node pairs by cluster

Cluster 1 (Q nodes)

Node X	Node Y	Slope ($\hat{\beta}_1$)	R^2
Q_node1	Q_node3	0.29	1.00
Q_node1	Q_node5	0.21	1.00
Q_node1	Q_node9	2.72	1.00
Q_node1	Q_node12	0.43	1.00
Q_node1	Q_node13	0.57	1.00
Q_node3	Q_node5	0.75	1.00
Q_node3	Q_node9	9.52	1.00
Q_node3	Q_node12	1.50	1.00
Q_node3	Q_node13	2.00	1.00
Q_node5	Q_node9	12.7	1.00
Q_node5	Q_node12	2.00	1.00
Q_node5	Q_node13	2.67	1.00
Q_node9	Q_node12	0.16	1.00
Q_node9	Q_node13	0.21	1.00
Q_node12	Q_node13	1.33	1.00

Cluster 2 (Q nodes)

Node X	Node Y	Slope ($\hat{\beta}_1$)	R^2
Q_node2	Q_node6	3.00	1.00
Q_node2	Q_node7	1.25	1.00
Q_node2	Q_node8	0.75	1.00
Q_node6	Q_node7	0.42	1.00
Q_node6	Q_node8	0.25	1.00
Q_node7	Q_node8	0.60	1.00

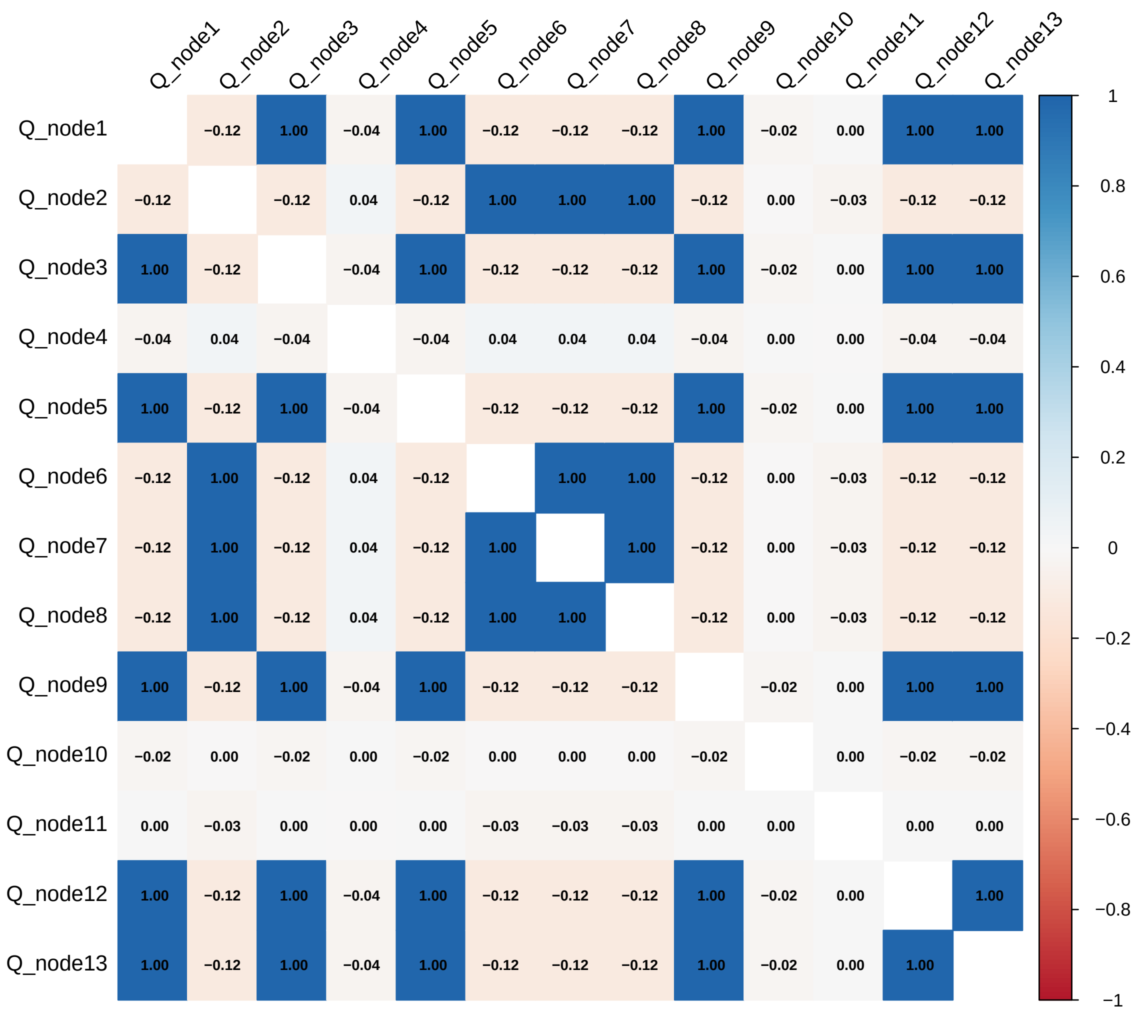

Figure 3: Correlation matrix for reactive (Q) power.

This finding confirms that nodes within the same cluster have nearly identical consumption patterns. They differ only in the overall magnitude of power consumed. A particularly notable observation was that nodes 1 and 9 in Cluster 1 had identical power values, corresponding to a regression slope of exactly 1. Since all nodes within a cluster carry identical structural information, a single representative node per cluster suffices for subsequent analyses. In addition to the three residential nodes, five representative nodes were selected: P_node1 (agricultural), P_node2 (agricultural), P_node4 (residential), P_node10 (residential), and P_node11 (residential), along with their reactive power counterparts.

4.3 Change-point analysis

The cluster analysis described above used the full 366-day dataset. Even after smoothing, the time series may include isolated days when power consumption deviates from the expected pattern because of unusual events. If kept, these atypical days could distort the mean profiles and covariance estimates that support later analyses. Changepoint analysis was performed on the five representative nodes to detect and remove these days before proceeding.

Changepoint detection was applied to identify time points at which the mean of the smoothed signal changes significantly. Two algorithms from the `changepoint` package (Killick and Eckley, 2023) were considered:

1. **Pruned Exact Linear Time (PELT)**: minimises a penalised cost function over the mean to efficiently detect multiple changepoints. Automatic penalties (AIC, BIC, MBIC) were tested with different smoothing windows, but were too conservative for this dataset, failing to detect changepoints for smoothing windows of 3 days or more. A manual penalty value of 0.1 was chosen, targeting 3 to 6 changepoints per signal.
2. **Binary Segmentation (BinSeg)**: recursively partitions the time series by selecting the split point that minimises the total segment cost at each step. A maximum of Q=5 changepoints was specified with no penalty.

Both methods were applied independently to the active and reactive power series of each representative node. The number of changepoints detected by each method is summarised in Table 4. Since BinSeg always returns exactly 5 changepoints by construction,

Method	P_node1	P_node2	P_node4	P_node10	P_node11
PELT	6	1	0	0	2
BinSeg	5	5	5	5	5
Method	**Q_node1**	**Q_node2**	**Q_node4**	**Q_node10**	**Q_node11**
PELT	1	1	0	0	0
BinSeg	5	5	5	5	5

Table 4: Number of changepoints detected per method and node. Nodes 1 and 2 are agricultural, while 4, 10 and 11 are residential.

regardless of the underlying signal, PELT is the more informative method for identifying genuine structural shifts. Days with at least one PELT-detected changepoint were flagged as atypical. Taking the union of flagged dates across both active and reactive power yielded 11 unique atypical days out of 366, listed in Table 5.

Date	Weekday	Public Holiday
2016-03-17	Thursday	No
2016-05-29	Sunday	No
2016-08-07	Sunday	No
2016-08-28	Sunday	No
2016-08-29	Monday	No
2016-09-16	Friday	No
2016-09-25	Sunday	No
2016-10-13	Thursday	No
2016-11-07	Monday	No
2016-11-08	Tuesday	No
2016-11-22	Tuesday	No

Table 5: Atypical days removed from the dataset. None coincide with German public holidays in 2016.

To assess whether these days corresponded to known events, the flagged dates were compared against the nine official German public holidays in 2016. None of the 11 flagged days matched any public holiday, confirming that the detected shifts reflect unusual consumption events unrelated to the standard holiday calendar. After removing these 11 days, the clean dataset contains 355 days and is used for all subsequent analyses.

4.4 Seasonal Modelling and Load Profile Estimation

This section aims to identify systematic patterns in the smoothed daily power series by modelling seasonal and weekday effects. This approach yields seasonal–weekday load profiles and characterises residual variation.

Two complementary modelling approaches are employed:

1. **Categorical seasonal model** : Mean load profiles are estimated independently for each season category ($s \in 1, 2, 3$) and weekday ($w \in 1, \ldots, 7$) combination. Model selection utilises multivariate analysis of variance (MANOVA), resulting in a maximum of $3 \times 7 = 21$ unique daily profiles per node.
2. **Continuous seasonal model via sine function**: A parametric regression model with a sine-based seasonal component is fitted per weekday and hour, allowing the seasonal trend to vary smoothly throughout the year. After fitting, residuals from this model are analysed to test whether residual variation differs by season. If not, the sine model has successfully removed seasonal effects.

Both models use the smoothed series $\tilde{P}_{n,d,h}$ and $\tilde{Q}_{n,d,h}$ computed in the data preprocessing step (Section 4.1). The categorical model is simpler and easier to interpret for describing load profiles. The sine-based model is more parsimonious, requiring only 2 parameters per weekday and hour, and its residuals are used to test whether seasonal effects have been removed. Model choice balances interpretability, parsimony, and suitability for hypothesis testing.

Residuals are computed after each modelling phase, and their correlation structures are analysed to determine whether nodes exhibit statistical coupling or independence.

4.4.1 Vector Representations for Daily Analysis

The full smoothed time series $\tilde{P}_{n,d,h}$ provides $366 \times 96 = 35{,}136$ observations per node, where $d = 1, \ldots, 366$ indexes the day and $h = 1, \ldots, 96$ indexes the quarter-hour within each day. Rather than working with the raw time series directly, the analysis instead treats each day as a single multivariate observation. To facilitate this approach, two different per-day vector representations, both derived from the smoothed series $\tilde{P}_{n,d,h}$, are used, with the choice depending on the specific analysis context.

96-Dimensional Quarter-Hourly Vectors: For each node n and day d, the 96 smoothed quarter-hourly values are collected into a single vector:

$$\tilde{\mathbf{P}}_{n,d} = \left(\tilde{P}_{n,d,1},\, \tilde{P}_{n,d,2},\, \ldots,\, \tilde{P}_{n,d,96}\right)^{\top} \in \mathbb{R}^{96}.$$

This yields 96-dimensional observations per day per node, with each vector representing the daily load profile. Over the full year, there are 366 such vectors per node. To prepare for analysis, these vectors are grouped by node, season, and weekday, and used as the multivariate response in a two-way MANOVA. This analysis then tests whether the full daily load shape differs across seasons and weekdays.

24-Dimensional Hourly Vectors: For each node n and day d, the 96 quarter-hourly values are averaged within each clock hour $H = 0, 1, \ldots, 23$ to produce a 24-dimensional hourly mean vector:

$$\tilde{P}_{n,d,H} = \frac{1}{4} \sum_{h=4H+1}^{4(H+1)} \tilde{P}_{n,d,h}, \qquad H = 0, 1, \ldots, 23,$$

$$\tilde{\mathbf{P}}_{n,d} = \left(\tilde{P}_{n,d,0},\, \tilde{P}_{n,d,1},\, \ldots,\, \tilde{P}_{n,d,23}\right)^{\top} \in \mathbb{R}^{24}.$$

Reducing from 96 to 24 dimensions serves two purposes. First, it reduces the covariance matrix from 96×96 to 24×24, thus avoiding estimation problems when the matrix dimension exceeds group sample sizes. Second, averaging within hours yields a smoother, more reliable representation for capturing daily variation and testing covariance structure. Consequently, these 24-dimensional vectors are used throughout Section 5 for all permutation-based Frobenius distance tests of covariance equality.

In summary, employing both representations is essential: the 96-dimensional form retains the quarter-hourly detail necessary for MANOVA model selection, while the 24-dimensional form ensures numerical stability for covariance structure testing. The transition between these forms aligns with the needs of each analysis stage, enhancing both interpretability and reliability.

The same constructions apply to reactive power $\tilde{Q}_{n,d,h}$, replacing $\tilde{P}$ with $\tilde{Q}$ throughout.

4.4.2 Seasonal & Weekday Load Profiles

Load consumption varies strongly by season, with higher demand in winter for heating and lower in summer. It also shows weekly periodicity, with clear differences between weekdays and weekends. To better understand the contribution of seasonality and weekly cycles before conducting covariance analysis, the smoothed power series was modelled using season and weekday as categorical predictors.

For each node n, day d, and quarter-hour h, the smoothed active power is decomposed as

$$\tilde{P}_{n,d,h} = \mu_{n,s(d),w(d),h} + E_{n,d,h},$$

where $s(d) \in (1, 2, 3)$ denotes the season (Summer, Winter, Rest) category, $w(d) \in (1, \ldots, 7)$ denotes the weekdays (Monday - Sunday), $E_{n,d,h} \sim N(0, \sigma_h^2)$ is the random error term, assumed independent across days, and $\mu_{n,s(d),w(d),h}$ is the mean load profile for node n in season s on weekday w at quarter-hour h. The same model applies to reactive power $\tilde{Q}_{n,d,h}$ with corresponding mean profiles $\mu^Q_{n,s,w,h}$.

The mean profile for node n in season s, weekday w, and quarter-hour h is estimated as the sample average over all days sharing that combination:

$$\hat{\mu}_{n,s,w,h} = \frac{1}{|\mathcal{D}_{n,s,w}|} \sum_{d \in \mathcal{D}_{n,s,w}} \tilde{P}_{n,d,h},$$

Here, $\mathcal{D}_{n,s,w} = \{d : s(d) = s, w(d) = w\}$ defines the set of days matching a given season and weekday. This results in $3 \times 7 = 21$ distinct mean profiles per node and hour. The approach provides a parsimonious representation of seasonal and weekly structure. Not all nodes require this level of detail. Some nodes may show mainly seasonal variation, others may be dominated by weekday effects, and some may have interactions in which weekday patterns differ by season. This motivates formal testing with MANOVA.

In this analysis, daily independence is not fully satisfied, as load profiles from consecutive days may be temporally dependent. After removing seasonal and weekday effects, the observations are approximately independent. Consequently, MANOVA results should be interpreted with caution. They highlight meaningful differences but are not exact in a strict theoretical sense. MANOVA also assumes multivariate normality and homogeneity of covariance matrices. Formal assumption tests are not done due to sample size limitations, but MANOVA is generally robust to moderate violations. For more rigorous inference, temporal dependence could be explicitly modelled using mixed-effects or time-series approaches. Here, MANOVA is mainly a screening tool to guide parsimonious mean modelling.

4.4.2.1 Model Selection via MANOVA To determine which factors to retain for each node, each day is treated as a 96-dimensional observation vector. This is constructed from the 15-minute smoothed power measurements, $\tilde{\mathbf{P}}_{n,d} = (\tilde{P}_{n,d,1}, \ldots, \tilde{P}_{n,d,96}) \in \mathbb{R}^{96}$. The explanatory variables are season ($s = 1, 2, 3$) and weekday ($w = 1, \ldots, 7$). Both are modelled as categorical factors. MANOVA with Pillai's trace is applied separately for each node. This assesses the significance of these effects, as given by

$$\tilde{\mathbf{P}}_{n,d} = \boldsymbol{\mu} + \boldsymbol{\alpha}_{s(d)} + \boldsymbol{\beta}_{w(d)} + \boldsymbol{\gamma}_{s(d),\,w(d)} + \mathbf{e}_{n,d},$$

where

$$\boldsymbol{\alpha}_s, \boldsymbol{\beta}_w, \boldsymbol{\gamma}_{s,w} \in \mathbb{R}^{96}, \qquad \mathbf{e}_{n,d} \sim \mathcal{N}_{96}(\mathbf{0}, \boldsymbol{\Sigma}).$$

To evaluate the contribution of each factor, three nested hypotheses are tested:

$$\begin{aligned}
&H_1 : \boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2 = \boldsymbol{\alpha}_3 = \mathbf{0} \quad \text{(no season effect)},\\
&H_2 : \boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_7 = \mathbf{0} \quad \text{(no weekday effect)},\\
&H_3 : \boldsymbol{\gamma}_{s,w} = \mathbf{0} \text{ for all } s, w \quad \text{(no season–weekday interaction)}.
\end{aligned}$$

Rejecting any null hypothesis shows that at least one factor is significant. Either season, weekday, or their interaction contributes to differences in the mean daily load profile.

4.4.2.2 MANOVA Model Fitting and Selection For each node, daily load profiles were organized into a $N_{\text{days}} \times 96$ response matrix (rows = days, columns = quarter-hours). Season (3 categories: Summer, Winter, Rest) and weekday (7 categories: Monday–Sunday) were assigned to each day using the date.

The full interaction model was fitted in R:

```
fit <- manova(response ~ Season * Weekday, data = daily_data)
summary(fit, test = "Pillai")
```

From the summary output, three p-values were extracted:

$$\begin{aligned}
p_{\text{season}} &= \Pr(> F) \text{ for the Season effect},\\
p_{\text{weekday}} &= \Pr(> F) \text{ for the Weekday effect},\\
p_{\text{interaction}} &= \Pr(> F) \text{ for the Season:Weekday interaction}.
\end{aligned}$$

Model selection followed a decision tree: if $p_{\text{interaction}} < 0.05$, the full interaction model was retained. Otherwise, non-significant terms were dropped according to the hierarchy: first drop interaction, then drop weekday if $p_{\text{weekday}} \geq 0.05$, then drop season if $p_{\text{season}} \geq 0.05$.

Special handling was required for P_node1 and Q_node1 (both agricultural) due to their near-zero power and variance during nighttime hours (approximately 22:00–05:00). This caused a rank deficiency in the response matrix. To fix this, an iterative dimension-reduction procedure was applied. Columns (quarter-hours) with the lowest variance were removed one at a time until the response achieved full rank. This reduced the effective dimension to 37 quarter-hours for P_node1 and 29 for Q_node1, resolving the rank deficiency and enabling the additive model (Season + Weekday, no interaction) to be fitted reliably. However, the full interaction model (Season × Weekday) could not be fitted for these two nodes. This was due not only to the high dimensionality but also to the small number of days remaining after removing holidays (~ 355) across 21 season–weekday groups, yielding only about 6–8 observations per cell. Since this is smaller than the response dimension, within-cell covariance matrices cannot be reliably estimated. This results in instability or singularity and violates MANOVA assumptions. In contrast, the additive model pools information across groups, effectively increasing the sample size available for covariance estimation and enabling reliable model fitting. Mean profiles were still computed on the full 96 quarter-hours for all nodes to ensure consistency in downstream analyses.

Table 6 summarises the MANOVA results for active power nodes. P_node1 (agricultural) shows significant effects for both season and weekday under the additive model. P_node2 (agricultural) and P_node11 (residential) retain the full interaction model, indicating that consumption patterns vary across both dimensions. P_node10 (residential) uses the additive model, while P_node4 (residential) shows only a weekday effect.

Node	**Season p-val**	**Weekday p-val**	**Interaction p-val**	**Selected Model**
P_node1	6.3×10^{-14}	5.2×10^{-13}	NA	Season + Weekday
P_node2	2.4×10^{-12}	1.1×10^{-16}	0.039	Season × Weekday
P_node4	2.2×10^{-1}	1.2×10^{-3}	0.055	Weekday only
P_node10	1.1×10^{-6}	1.6×10^{-12}	0.086	Season + Weekday
P_node11	6.1×10^{-16}	5.9×10^{-3}	0.032	Season × Weekday

Table 6: MANOVA-based model selection results for P-nodes.

The same procedure is used for the reactive power nodes. The results for each node are given in Table 7. For Q_node1 (agricultural), season and weekday both have clear effects in the additive model. Q_node2 (agricultural) keeps the full interaction model. For residential nodes: Q_node10 keeps the interaction model due to a significant interaction ($p = 0.006$), even though the weekday effect is not significant. Q_node4 shows no significant effects, so a single mean profile is used. Q_node11 keeps only the season model, with no other effects.

Node	Season p-val	Weekday p-val	Interaction p-val	Selected Model
Q_node1	1.7×10^{-4}	4.7×10^{-3}	NA	Season + Weekday
Q_node2	7.7×10^{-14}	5.6×10^{-11}	0.043	Season × Weekday
Q_node4	6.1×10^{-1}	8.5×10^{-1}	0.089	Annual
Q_node10	1.4×10^{-4}	9.3×10^{-1}	0.006	Season × Weekday
Q_node11	2.4×10^{-15}	2.02×10^{-1}	0.126	Season only

Table 7: MANOVA-based model selection results for Q nodes.

4.4.2.3 Residual Computation After model selection via MANOVA, residuals are computed identically across all nodes, regardless of whether interaction terms were retained or dropped:

$$R_{n,d,h} = \tilde{P}_{n,d,h} - \hat{\mu}_{s(d),w(d),h}.$$

where $\hat{\mu}_{s(d),w(d),h}$ is the fitted mean from the selected model (Season × Weekday, Season + Weekday, Weekday-only, or pooled, depending on significance tests). The residual formula remains consistent across all model structures: it is the deviation from the predicted mean under the fitted model.

This consistency ensures residuals across all nodes are comparable in scale and reflect unexplained variability after removing only significant factors. Nodes with simpler models (e.g., Weekday-only) will have larger residuals if seasonal effects exist but were not modelled, while nodes with complex models (Season × Weekday) will have residuals reflecting only the remaining unmodelled variation.

For this reason, the residual covariance structure (tested in Section 5) is comparable across nodes despite different model structures, as each residual set represents variation unexplained by its node-specific optimal model.

4.4.2.4 Residual Correlation Matrix

Daily-level residuals were obtained by averaging the 96 quarter-hour residuals $R_{n,d,h}$ across all hours for each day:

$$\bar{R}_{n,d} = \frac{1}{96} \sum_{h=1}^{96} R_{n,d,h}.$$

These daily residuals $\bar{R}_{n,d}$ were arranged into a matrix with days as rows ($d = 1, \ldots, D$) and representative nodes as columns ($n \in$ {P_node1, P_node2, ...}). Pairwise Pearson correlations between nodes were computed using the complete-case method:

$$r_{ij} = \frac{\sum_d (\bar{R}_{i,d} - \bar{\bar{R}}_i)(\bar{R}_{j,d} - \bar{\bar{R}}_j)}{\sqrt{\sum_d (\bar{R}_{i,d} - \bar{\bar{R}}_i)^2 \sum_d (\bar{R}_{j,d} - \bar{\bar{R}}_j)^2}}.$$

Here i and j represent the five representative nodes, $\bar{\bar{R}}_i = \frac{1}{D} \sum_d \bar{R}_{i,d}$ is the mean of daily residuals for node i across all clean days.

Results are shown in Tables 8 and 9. After removing seasonal and weekday mean profiles, residual correlations are weak, ranging from -0.239 to 0.254 for active power and -0.433 to 0.217 for reactive power. The strongest correlations are negative, especially between Q_node1 (agricultural) and Q_node4 (residential), r = -0.433, possibly reflecting opposing reactive power behaviours between agricultural and residential clusters.

This weak inter-node residual correlation indicates that nodes behave largely independently once seasonal and weekday factors are removed. This suggests that much of the observed correlation structure in the original data is driven by shared seasonal and weekly patterns rather than direct physical coupling or network effects.

	P_node1	**P_node2**	**P_node4**	**P_node10**	**P_node11**
P_node1	1.000	-0.063	-0.189	-0.027	-0.004
P_node2	-0.063	1.000	0.130	-0.001	-0.239
P_node4	-0.189	0.130	1.000	0.254	0.198
P_node10	-0.027	-0.001	0.254	1.000	0.251
P_node11	-0.004	-0.239	0.198	0.251	1.000

Table 8: Correlation matrix between selected P-nodes.

4.4.3 Seasonal Modeling via the Sine Function

Power consumption in the SimBench grid follows a pronounced annual cycle. Consumption is higher in winter due to heating loads and lower in summer, returning to winter

	Q_node1	**Q_node2**	**Q_node4**	**Q_node10**	**Q_node11**
Q_node1	1.000	-0.229	-0.433	-0.164	-0.027
Q_node2	-0.229	1.000	0.217	-0.011	-0.194
Q_node4	-0.433	0.217	1.000	0.197	0.104
Q_node10	-0.164	-0.011	0.197	1.000	0.089
Q_node11	-0.027	-0.194	0.104	0.089	1.000

Table 9: Correlation matrix between selected Q-nodes.

levels the next year. This 365-day periodic pattern makes a sinusoidal function a natural choice for modelling. A sine function is the simplest smooth, symmetric periodic function with a single frequency. It requires only two parameters per node-hour-weekday combination: an intercept (β^1) for the annual mean level and a slope coefficient (β^2) for seasonal amplitude. The sine function guarantees periodicity by construction, so the fitted curve returns to the same value one year later. Compared to the three-category season model of Section 4.4.2, the sine approach models seasonal change as a continuous smooth curve. This better reflects the gradual nature of seasonal transitions and avoids arbitrary category boundaries. The 80-day phase shift aligns the model minimum with early summer, around late June, matching the observed consumption pattern.

For each quarter-hour $h = 1, \dots, 96$, node n, and weekday $w = 1, \dots, 7$ (where 1 = Monday, 7 = Sunday), the 2-dimensional covariate vector is defined as

$$x(d) := \left(1,\ \sin\left(\frac{2\pi(d-80)}{365}\right)\right)^\top \in \mathbb{R}^2,$$

where $d = 1$ corresponds to 1 January. The smoothed active and reactive power values $\tilde{P}_{n,d,h}$ and $\tilde{Q}_{n,d,h}$ are modelled as:

$$\begin{aligned} \tilde{P}_{n,d,h} &= x(d)^\top \beta^P_{n,w(d),h} + E^P_{n,d,h}, \\ \tilde{Q}_{n,d,h} &= x(d)^\top \beta^Q_{n,w(d),h} + E^Q_{n,d,h}, \end{aligned}$$

where $E^P_{n,d,h}$ and $E^Q_{n,d,h}$ are random error terms, and:

$$\beta^P_{n,w,h} = \left(\beta^{P,1}_{n,w,h},\ \beta^{P,2}_{n,w,h}\right)^\top \in \mathbb{R}^2, \qquad \beta^Q_{n,w,h} = \left(\beta^{Q,1}_{n,w,h},\ \beta^{Q,2}_{n,w,h}\right)^\top \in \mathbb{R}^2,$$

are the intercept and seasonal amplitude coefficients to be estimated. The model is fitted separately for each weekday w using only the days d with $w(d) = w$.

Let $W(w) = \#\{d : w(d) = w\}$ denote the number of days falling on weekday w, and let $d_1^w < d_2^w < \cdots < d_{W(w)}^w$ denote those days in order. The design matrix for weekday w is:

$$X_w = \begin{pmatrix} x(d_1^w)^\top \\ \vdots \\ x(d_{W(w)}^w)^\top \end{pmatrix} = \begin{pmatrix} 1 & \sin\left(\frac{2\pi(d_1^w - 80)}{365}\right) \\ \vdots & \vdots \\ 1 & \sin\left(\frac{2\pi(d_{W(w)}^w - 80)}{365}\right) \end{pmatrix} \in \mathbb{R}^{W(w)\times 2},$$

and the corresponding power vectors restricted to weekday w are:

$$\mathbf{P}_{n,h}^w = \left(\tilde{P}_{n,d_1^w,h}, \ldots, \tilde{P}_{n,d_{W(w)}^w,h}\right)^\top \in \mathbb{R}^{W(w)},$$

and analogously $\mathbf{Q}_{n,h}^w$. The ordinary least squares estimators are:

$$\widehat{\beta}_{n,w,h}^P = (X_w^\top X_w)^{-1} X_w^\top \mathbf{P}_{n,h}^w, \qquad \widehat{\beta}_{n,w,h}^Q = (X_w^\top X_w)^{-1} X_w^\top \mathbf{Q}_{n,h}^w.$$

This produces 14 estimated parameters per quarter-hour (h) and node (n) for each power type. In R, the model is fitted as follows:

```
sine_term <- function(doy) sin((doy - 80) / 365 * 2 * pi)
fit <- lm(Smoothed ~ 1 + sine_term(DayOfYear), data = hh_df)
```

where `hh_df` contains all observations for node n, quarter-hour h, and weekday w.

The model's fit varies across nodes (Tables 10 and 11). R^2 values range from 0.04 to 0.66 for active power and 0.052 to 0.568 for reactive power. Residuals for all nodes are centered around zero (mean residual $\approx 10^{-18}$), confirming model unbiasedness for both power types. P_node1 and P_node2 (agricultural, R^2=0.255 and 0.381) show moderate seasonality, while P_node11 (residential, R^2=0.660) is highly seasonal. For reactive power, Q_node4 (residential, R^2=0.052) shows little seasonality. Q_node1 (agricultural, R^2=0.108) and Q_node10 (residential, R^2=0.230) show moderate seasonality, and Q_node11 (residential, R^2=0.568) is strongly seasonal.

The sine model fits the data well for all five nodes and all weekdays, as seen in Figures 4 and 5. For active power (Figure 4), the observed curves (red) and the model predictions (dashed cyan) match closely across all nodes and weekdays, with P_node11 (residential) showing the most pronounced seasonal curve, consistent with its high R^2, while P_node4 (residential) remains comparatively flat. For reactive power (Figure 5), the sine model similarly tracks the observed series well, with Q_node11 (residential) and Q_node2

Node	R^2	Mean Residual
P_node1	0.255	-7.78×10^{-18}
P_node2	0.381	-1.74×10^{-18}
P_node4	0.042	-4.67×10^{-19}
P_node10	0.315	2.35×10^{-19}
P_node11	0.660	-4.36×10^{-19}

Table 10: Goodness of fit metrics for P-nodes.

Node	R^2	Mean Residual
Q_node1	0.108	9.99×10^{-19}
Q_node2	0.348	-1.05×10^{-18}
Q_node4	0.052	-2.32×10^{-20}
Q_node10	0.230	-6.22×10^{-19}
Q_node11	0.568	-8.33×10^{-20}

Table 11: Goodness of fit metrics for Q-nodes.

(agricultural) showing clear seasonal curvature and Q_node4 (residential) remaining largely flat throughout the year. Near identical patterns across weekdays for both active and reactive power support the conclusion that seasonal effects are uniform across all days of the week.

Having established the overall fit of the sine model, the next step is to consider whether any residual seasonal effects persist after correction. Although the sine model provides a good fit (Tables 10 and 11), it remains to assess whether discrete season-category differences in covariance remain.

Table 12 shows that season effects remain significant for most nodes ($p < 0.05$ for 4 of 5 P-nodes and 4 of 5 Q-nodes). The sine model removes the dominant seasonal trend, but the residual covariance structure still varies across seasons. Analyses in Sections 5.2–5.3 pool weeks and permute across seasons to ensure detected effects are robust to these residual seasonal structures.

4.4.3.1 Weekday Effect Test The sine regression coefficients $\widehat{\beta}^{P,1}_{n,w,h}$(intercept) and $\widehat{\beta}^{P,2}_{n,w,h}$ (seasonal amplitude) vary across weekdays for many nodes and quarter-hours. To evaluate whether these differences are meaningful from a statistical perspective, the estimated coefficient vectors are compared across all seven weekdays to determine if they are equivalent. For each node n and quarter-hour h, the 7 estimated coefficient vectors $\widehat{\beta}^{P}_{n,w,h} = \left(\widehat{\beta}^{P,1}_{n,w,h}, \widehat{\beta}^{P,2}_{n,w,h}\right)^{\top}$ (one per weekday $w = 1, \ldots, 7$) are compared using a MANOVA test. The null hypothesis is: these seven coefficient vectors are compared

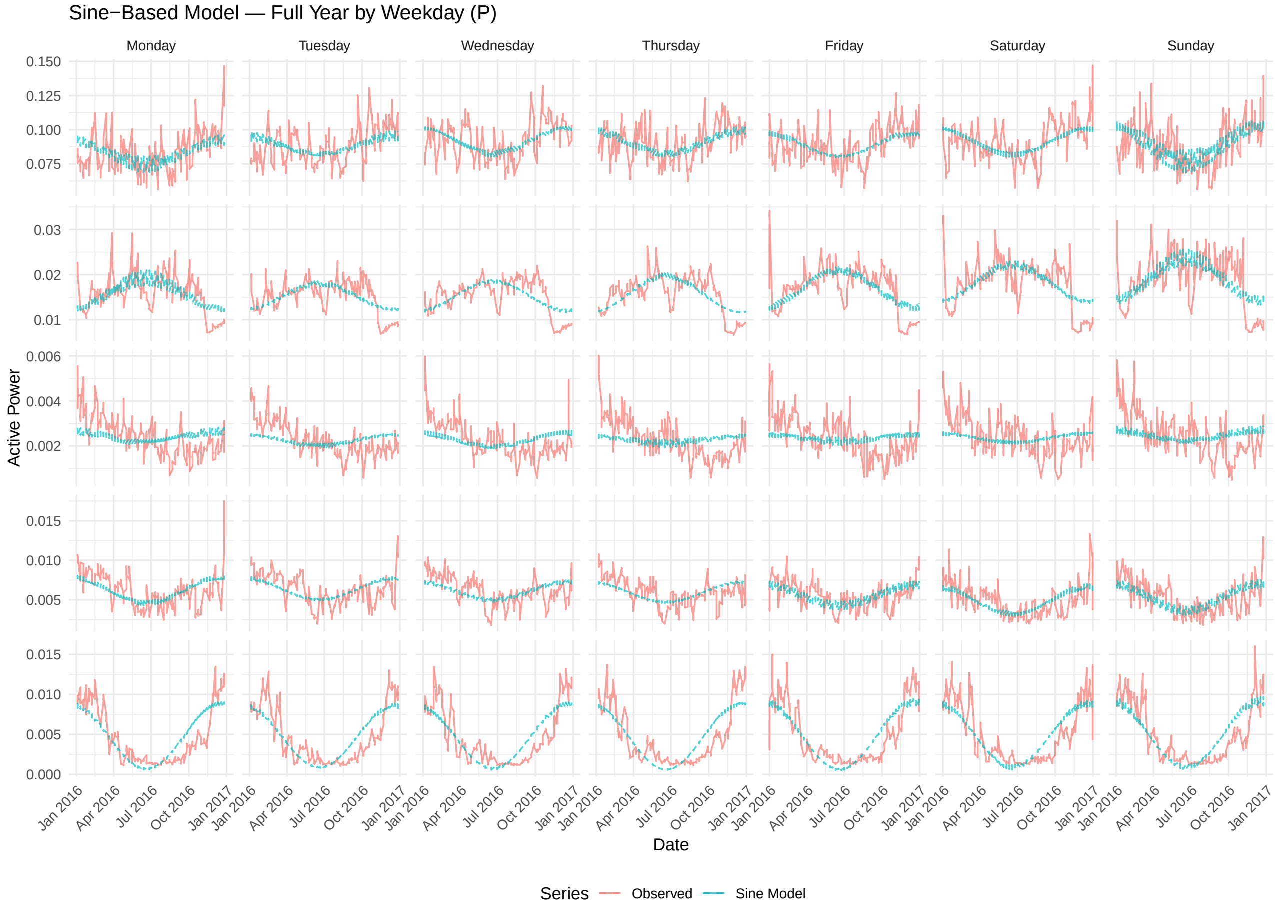

Figure 4: Observed (solid red) vs. Sine Model predicted (dashed cyan) power for full year across P-nodes.

using a MANOVA test. The null hypothesis is:

$$H_0 : \beta^P_{n,1,h} = \beta^P_{n,2,h} = \cdots = \beta^P_{n,7,h},$$

which states that the intercept and seasonal amplitude are identical across all weekdays. Rejection indicates that the sine regression parameters (and thus the seasonal pattern) differ significantly by day of the week.

The MANOVA model underlying this test is:

$$\widehat{\beta}^P_{n,w,h} = \nu_{n,h} + \delta_{n,w,h} + \epsilon_{n,w,h}, \quad \epsilon_{n,w,h} \overset{\text{iid}}{\sim} \mathcal{N}_2(0, \Sigma),$$

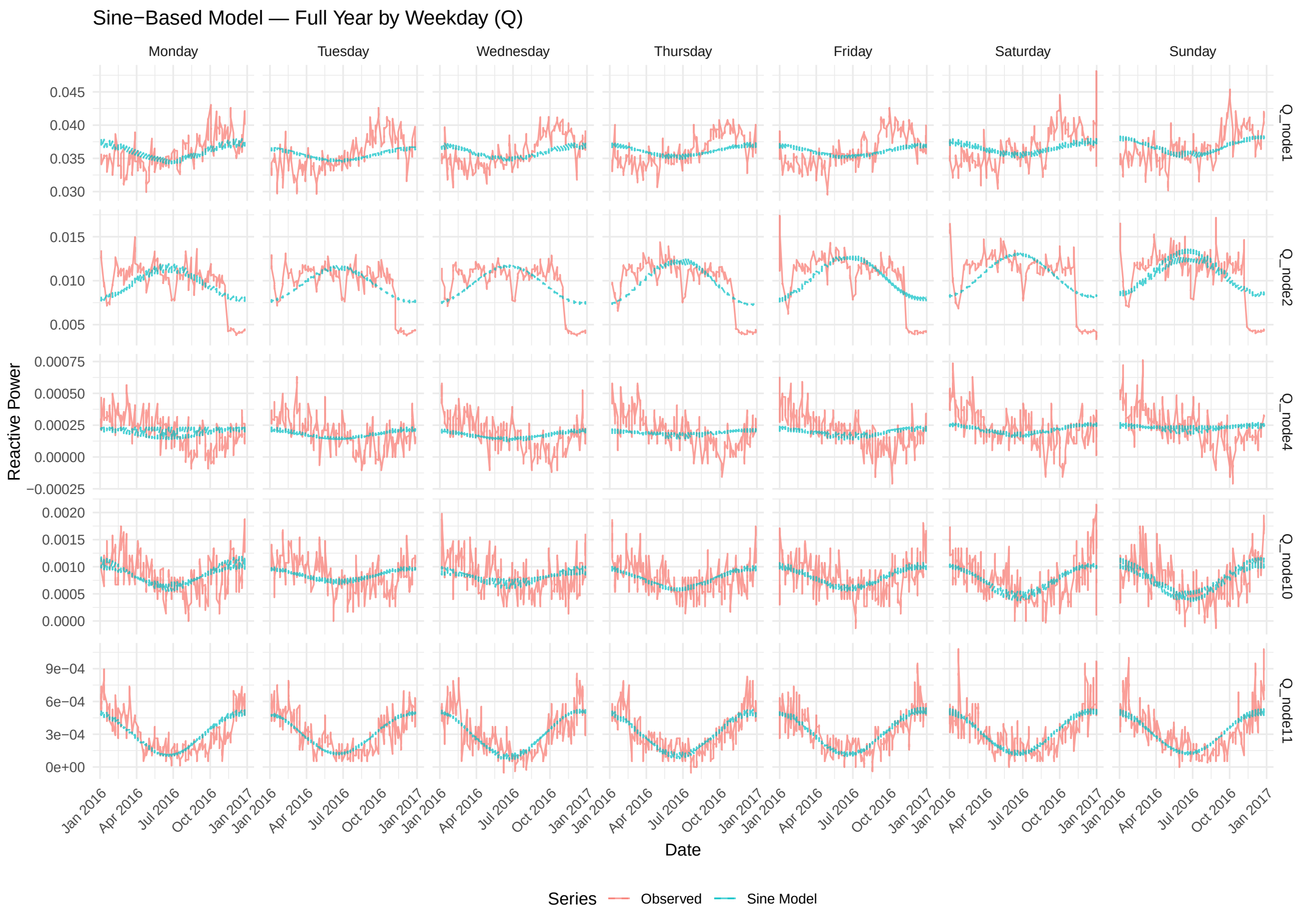

Figure 5: Observed (solid red) vs. Sine Model predicted (dashed cyan) power for full year across Q-nodes.

where $\nu_{n,h} \in \mathbb{R}^2$ is the overall mean coefficient vector, $\delta_{n,w,h} \in \mathbb{R}^2$ is the weekday-specific deviation (with $\sum_{w=1}^{7} \delta_{n,w,h} = 0$), and $\epsilon_{n,w,h}$ is the random error. The null hypothesis is equivalent to $H_0 : \delta_{n,h,1} = \cdots = \delta_{n,h,7} = 0$.

Since the response is two-dimensional (both intercept and slope), the MANOVA reduces to a standard F-test for nested linear models. The reduced model assumes a single set of sine parameters shared across all weekdays, while the full model allows weekday-specific parameters. The test evaluates whether weekday-specific parameters significantly improve the fit over a pooled model. As 96 separate tests are performed per node (one per quarter-hour), Benjamini–Hochberg false discovery rate (FDR) correction is applied at $\alpha = 0.05$ to control for multiple testing.

Additionally, a secondary hypothesis focusing on working days (Monday–Friday only) is tested:

$$\tilde{H}_0 : \beta^P_{n,1,h} = \beta^P_{n,2,h} = \cdots = \beta^P_{n,5,h},$$

to assess whether weekday variation is driven primarily by weekend differences.

In R:

```
# All 7 weekdays
fit <- manova(cbind(Beta_P1, Beta_P2) ~ factor(Weekday),
              data = df_h)
summary(fit, test = "Pillai")

# Monday-Friday only
fit_mf <- manova(cbind(Beta_P1, Beta_P2) ~ factor(Weekday),
                 data = df_h %>% filter(Weekday %in% 1:5))
summary(fit_mf, test = "Pillai")

# FDR correction across the 96 quarter-hour tests per node
p.adjust(p_values, method = "BH")
```

After Benjamini–Hochberg FDR correction across all 96 quarter-hours tested per node, the results differ across nodes. For active power, P_node1 (agricultural) and P_node10 (residential) show significant weekday effects across all 96 quarter-hours, and P_node2 (agricultural) similarly shows 96/96 significant quarter-hours, indicating that the sine regression relationship differs across days of the week for these nodes. P_node4 (residential) and P_node11 (residential) show no significant weekday effect (0/96). For reactive power, no node shows any significant weekday effect (0/96 for all nodes).

When the analysis is restricted to Monday–Friday, P_node1 (agricultural) still shows 74 of 96 significant quarter-hours. All other nodes show no significant Mon–Fri weekday effect. The R command used is:

```
fit_full    <- lm(Smoothed ~ SineTerm * Weekday_f, data = df_h)
fit_reduced <- lm(Smoothed ~ SineTerm, data = df_h)
anova(fit_reduced, fit_full)
p.adjust(p_values, method = "BH")
```

These results indicate that for P_node4 (residential), P_node11 (residential), and all reactive power nodes, a single unified sine correction is sufficient across all weekdays. For P_node1 (agricultural), P_node2 (agricultural), and P_node10 (residential), the sine model intercept or slope differs significantly across weekdays, suggesting that weekday-specific seasonal parameters improve the fit for these nodes. The week-specific correction applied in subsequent sections therefore uses the weekday-specific $\widehat{\beta}^{P}_{n,w,h}$ estimates, which already account for this variation.

As a further validation, it was tested whether the three-category season effect (winter/intermediate/summer, as in Section 4.4.2) remains detectable after subtracting the fitted sine model. The sine-corrected residuals for each day d and quarter-hour h are defined as

$$P^{*}_{n,d,h} = \tilde{P}_{n,d,h} - x(d)^{\top}\widehat{\beta}^{P}_{n,w(d),h},$$

and let $\mathbf{P}^{*}{}_{n,d,h} = (P^{*}_{n,d,1}, \ldots, P^{*}_{n,d,96})^{\top} \in \mathbb{R}^{96}$ be the corresponding residual day vector. Applying the MANOVA season test of Section 4.4.2 to $\mathbf{P}^{*}{}_{n,d,h}$ instead of $\tilde{\mathbf{P}}_{n,d,h}$ tests the null hypothesis

$$H_0^1 : \beta_n^1(s) = 0 \text{ for all } s = 1, 2, 3$$

on the residuals, i.e., whether any season-category mean difference remains after the sine correction. As a further check, the permutation-based covariance test from Section 3.3 was applied to the sine-corrected residuals $\mathbf{P}^{*}{}_{n,d,h}$ to assess whether any discrete season-category effect remains in the covariance structure after the sine correction. The results are shown in Table 12. For active power, season effects on the covariance structure remain significant for all five nodes after correction, with p-values ranging from 0.047 for P_node1 (agricultural) to below 0.001 for P_node4 (residential), P_node10 (residential), and P_node11 (residential). For reactive power, four of the five nodes remain significant, with only Q_node4 (residential) showing no residual season effect ($p = 0.146$). These results indicate that the sine model captures the dominant seasonal trend but does not fully eliminate discrete season-category differences in the covariance structure. This is consistent with the findings of Section 5.1, where season effects on covariance were significant for most nodes. In the cross-season analysis in Section 5.4, Q_node2 (agricultural) showed the largest residual cross-season correlation of -0.297.

Overall, the goodness-of-fit results confirm that the sine model captures the dominant seasonal trend for all nodes. However, the F-tests reveal that the seasonal parameters are not identical across weekdays for P_node1 (agricultural), P_node2 (agricultural), and P_node10 (residential). Since the model was already fitted separately per weekday

Node	Active Power p-value	Reactive Power p-value
Node 1	0.047*	<0.001*
Node 2	0.002*	<0.001*
Node 4	<0.001*	0.146
Node 10	<0.001*	0.003*
Node 11	<0.001*	<0.001*

Table 12: Permutation test p-values for season effects on the covariance structure of sine-corrected residuals. Asterisks denote $p < 0.05$.

(yielding weekday-specific $\widehat{\beta}^{P}_{n,w,h}$), this weekday variation is incorporated in all subsequent seasonal corrections. For the remaining nodes, a single unified sine correction is sufficient.

5 Results on Representative Nodes

The following sections analyse the covariance structure of the load profiles for the five representative nodes identified in Section 4.2. Each day is summarised by a 24-dimensional vector of hourly means. These means are computed by averaging the 96 quarter-hourly observations within each clock hour $H = 0, \ldots, 23$, from the 355 clean days retained after changepoint removal. Section 5.1 tests whether covariance matrices differ across seasons, weekdays, and nodes using discrete season categories. Section 5.2 repeats these tests using a continuous sine-based seasonal correction. Section 5.3 switches to a weekly representation to study the within-week dependence structure. Section 5.4 examines cross-season correlations to validate the seasonal correction applied in Sections 5.2 and 5.3.

5.1 Covariance Testing by Season Categories

This section analysed the covariance structure of the load profiles for the five representative nodes identified in Section 4.2, using the 355 clean days remaining after changepoint removal. Throughout Section 5.1, each day is represented by a 24-dimensional vector of hourly means, $\tilde{\mathbf{P}} = (P_{n,d,1}, \ldots, P_{n,d,24}) \in \mathbb{R}^{24}$, indexed by $H \in \{0, \ldots, 23\}$, obtained by averaging the smoothed quarter-hourly observations $\tilde{P}_{n,d,h}$ within each clock hour.

These daily vectors are the basis for all permutation-based tests of covariance equality using the Frobenius distance. For season-level covariance testing, days are grouped by season $s \in \{1, 2, 3\}$ within each weekday $w \in \{1, \dots, 7\}$, yielding up to 21 season–weekday groups. For each node n, weekday w, and season s, the corresponding sample covariance matrix is computed from the collection of daily vectors.

Ignoring dependencies between day units, a multivariate linear model is assumed:

$$\tilde{P}_{n,d} = \hat{\mu}_{n,s(d),w(d)} + E^P_{n,d}, \quad E^P_{n,d} \sim \mathcal{N}_{24}\Big(0_{24}, \, \Sigma^P_{n,s(d),w(d)}\Big),$$

where $\hat{\mu}_{n,s(d),w(d)}$ is the estimated mean hourly profile for node n in season $s(d)$ and weekday $w(d)$, from Section 4.4.2. The covariance matrix $\Sigma^P_{n,s,w}$ is estimated using the mean-corrected form. For each season-weekday combination, the sample covariance is computed as:

$$\hat{\Sigma}^P_{n,s,w} := \frac{1}{|\mathcal{D}_{s,w}| - 1} \sum_{d \in \mathcal{D}_{s,w}} \Big(\tilde{P}_{n,d} - \hat{\mu}_{n,s,w}\Big) \Big(\tilde{P}_{n,d} - \hat{\mu}_{n,s,w}\Big)^\top,$$

where $\mathcal{D}_{s,w} = \{d : s(d) = s, \, w(d) = w\}$ is the set of clean days with season s and weekday w.

When testing global season or weekday effects, the pooled estimators $\hat{\Sigma}^P_{n,w}$ and $\hat{\Sigma}^P_{n,s}$ are used, obtained by conditioning on $w(d) = w$ or $s(d) = s$ alone.

The same model and estimator apply to reactive power, with P replaced by Q throughout.

Season Effect: The first hypothesis tested is whether the covariance matrix of the daily load profiles differs across the three seasons, separately for each weekday. The null hypothesis is

$$H_0^s : \Sigma_{n,1,w} = \Sigma_{n,2,w} = \Sigma_{n,3,w}.$$

This tests whether the pattern of day-to-day variability within a given weekday changes across seasons. The season-specific model is $\tilde{P}_{n,d} \sim \mathcal{N}_{24}\Big(\hat{\mu}_{n,s(d),w(d)}, \, \Sigma^P_{n,s(d),w(d)}\Big)$. Rejecting this hypothesis means the pattern of day-to-day variability within a given weekday depends on the season.

Table 13 shows the full set of p-values for all node–weekday combinations for active power. The corresponding tables for reactive power are in Table 14.

For active power, P_node11 (residential) showed the strongest and most consistent seasonal covariance variation, with 6 out of 7 weekdays yielding significant results. This suggests that for this residential node, the daily profiles change substantially across

Node	Mon	Tue	Wed	Thu	Fri	Sat	Sun
P_node1	0.283	0.073	**0.039***	0.819	0.566	0.300	0.578
P_node2	0.354	0.056	0.053	0.368	0.078	**0.005***	0.230
P_node4	0.191	0.093	0.426	**0.045***	0.524	0.756	0.206
P_node10	0.071	0.404	0.869	0.514	0.598	**0.005***	**0.029***
P_node11	**0.036***	**0.013***	**0.039***	**0.013***	**0.019***	**0.034***	0.054

Table 13: Permutation test p-values for the season effect on covariance of daily active power profiles, tested separately for each node–weekday combination. Asterisks and bold values denote significant results.

seasons on almost every day of the week. In contrast, P_node1 (agricultural), P_node2 (agricultural), and P_node4 (residential) showed only one significant weekday each. P_node10 (residential) had two significant weekdays, Saturday and Sunday. Seasonal covariance effects are not uniform across weekdays. For most nodes, they appear on some days but not others, suggesting that season-by-weekday interactions exist in the covariance structure.

Node	Mon	Tue	Wed	Thu	Fri	Sat	Sun
Q_node1	**0.019***	**0.040***	**0.005***	**0.023***	**0.010***	0.106	0.136
Q_node2	**0.001***	**0.001***	**0.001***	**0.001***	**0.001***	**0.001***	**0.001***
Q_node4	0.696	0.899	0.970	0.195	0.269	0.971	0.540
Q_node10	0.329	0.260	0.456	**0.041***	0.319	0.386	0.204
Q_node11	**0.008***	0.111	**0.039***	**0.038***	0.189	0.209	**0.045***

Table 14: Permutation test p-values for the season effect on covariance of daily reactive power profiles, tested separately for each node–weekday combination. Asterisks (*) denote $p < 0.05$.

Similarly, for reactive power, Q_node2 (agricultural) stands out by showing significant results for all 7 weekdays ($p \approx 0$ throughout), reflecting the strongest and most consistent season-dependent covariance structure among all nodes compared. In contrast, Q_node1 (agricultural) has 5 significant weekdays, limited to Monday through Friday, indicating a more pronounced effect on weekdays. Q_node11 (residential) displays 4 significant weekdays spread across the week. Q_node10 (residential) is significant on only 1 day, Thursday, while Q_node4 (residential) shows no significant effect, suggesting a stable structure across seasons. This pattern highlights stronger seasonal variation in agricultural nodes than in residential ones, and a more persistent seasonal dependency in Q_node2 than in other nodes.

Global Season Effect (All weekdays combined): The global season test evaluates whether the covariance matrices for each node differ across the three seasons. It does this by pooling data from all weekdays within each season. The null hypothesis is

$$H_0^{s|w} : \Sigma_{n,Sum} = \Sigma_{n,Win} = \Sigma_{n,Rest}.$$

This is tested using the weekday-pooled model

$$\tilde{P}_{n,d} = \hat{\mu}_{n,s(d)} + E_{n,d}^P, \quad E_{n,d}^P \sim \mathcal{N}_{24}\left(0_{24}, \Sigma_{n,s(d)}^P\right),$$

with estimators

$$\hat{\mu}_{n,s} := \frac{1}{|\mathcal{D}_s|} \sum_{d:\, s(d)=s} \tilde{P}_{n,d},$$

$$\hat{\Sigma}_{n,s}^P := \frac{1}{|\mathcal{D}_s| - 1} \sum_{d:\, s(d)=s} \left(\tilde{P}_{n,d} - \hat{\mu}_{n,s}\right)\left(\tilde{P}_{n,d} - \hat{\mu}_{n,s}\right)^\top.$$

Results are shown in Table 15. For active power, P_node2 (agricultural), P_node4 (residential), P_node10 (residential), and P_node11 (residential) all show highly significant global season effects ($p < 0.001$), confirming that day-to-day variability for these nodes changes substantially across seasons. P_node1 (agricultural) is not significant at the global level ($p = 0.062$), consistent with its per-weekday results showing only one significant weekday. For reactive power, Q_node1 (agricultural), Q_node2 (agricultural), Q_node10 (residential), and Q_node11 (residential) all show significant global season effects. Q_node4 (residential) does not ($p = 0.242$), consistent with its per-weekday results, where no weekday was significant.

Node	**Active Power p-value**	**Reactive Power p-value**
Node 1	0.062	<0.001*
Node 2	<0.001*	<0.001*
Node 4	<0.001*	0.242
Node 10	<0.001*	0.005*
Node 11	<0.001*	<0.001*

Table 15: Global season effect p-values from permutation tests with 1000 permutations. Asterisks denote p<0.05.

Global Weekday Effect: The weekday test checks whether the covariance matrix differs across the seven days of the week, pooling all seasons. The null hypothesis is

$$H_0^w : \Sigma_{n,1} = \cdots = \Sigma_{n,7}.$$

This is tested using the model

$$\tilde{P}_{n,d} = \hat{\mu}_{n,w(d)} + E^P_{n,d}, \quad E^P_{n,d} \sim \mathcal{N}_{24}\left(0_{24}, \Sigma^P_{n,w(d)}\right).$$

All $D = 355$ days are pooled, without accounting for season or weekday. The estimators are

$$\hat{\mu}_{n,w} := \frac{1}{|\mathcal{D}_w|} \sum_{d:\, w(d)=w} \tilde{P}_{n,d},$$

$$\hat{\Sigma}^P_{n,w} := \frac{1}{|\mathcal{D}_w| - 1} \sum_{d:\, w(d)=w} \left(\tilde{P}_{n,d} - \hat{\mu}_{n,w}\right)\left(\tilde{P}_{n,d} - \hat{\mu}_{n,w}\right)^\top.$$

A significant result shows that day-to-day variability patterns differ across weekdays. Table 16 presents results for both active and reactive powers.

Node	**Active Power p-value**	**Reactive Power p-value**
Node1	0.958	0.538
Node2	<0.001*	0.167
Node4	0.783	0.765
Node10	0.287	0.081
Node11	0.534	0.481

Table 16: Global Weekday covariance effect for active and reactive power nodes.

Only P_node2 (agricultural) shows a significant weekday effect ($p < 0.001$). All other nodes show no significant weekday variation: P_node1 (agricultural) ($p = 0.958$), P_node4 (residential) ($p = 0.783$), P_node10 (residential) ($p = 0.287$), and P_node11 (residential) ($p = 0.534$). For reactive power, all nodes have non-significant weekday effects, with p-values from 0.167 to 0.765.

Non-Weekend Effect: This test examines whether weekday effects are driven solely by working days (Monday to Friday). It uses the same fully pooled model as the global weekday test, but is restricted to observations with $w(d) \in \{1, \ldots, 5\}$. The estimators $\hat{\mu}_{n,w}$ and $\hat{\Sigma}^P_{n,w}$ are therefore computed using weekday data only. The null hypothesis for the test is

$$\tilde{H}_0^w : \Sigma_{n,1} = \cdots = \Sigma_{n,5}.$$

Node	Active Power p-value	Reactive Power p-value
Node1	0.884	0.564
Node2	0.062	0.423
Node4	0.831	0.85
Node10	0.148	0.053
Node11	0.586	0.501

Table 17: Weekday covariance effect for active and reactive power nodes.

The results show no significant weekday effects for any active or reactive power nodes from Monday to Friday. Since P_node2 (residential) showed a significant global weekday effect, the Monday to Friday restricted test was applied to see if this effect persists within working days. For P_node2 (residential), this yields $p = 0.062$, which is no longer significant. This confirms that the weekday effect is driven by the weekend contrast rather than variation within the working week.

Node Effect: The node effect tests whether covariance matrices differ across nodes. In other words, it checks whether different nodes in the network exhibit structurally distinct patterns of day-to-day variability. The null hypothesis is

$$H_0^n : \Sigma_1 = \Sigma_2 = \ldots = \Sigma_5.$$

This is tested using the model

$$\tilde{P}_{n,d} = \hat{\mu}_n + E^P_{n,d}, \quad E^P_{n,d} \sim \mathcal{N}_{24}\left(0_{24}, \Sigma^P_n\right).$$

The analysis is pooled across all $D = 355$ days, regardless of season or weekday, with estimators defined as

$$\hat{\mu}_n := \frac{1}{D}\sum_{d=1}^{D} \tilde{P}_{n,d},$$

$$\hat{\Sigma}^P_n := \frac{1}{|\mathcal{D}|-1}\sum_{d=1}^{D}\left(\tilde{P}_{n,d} - \hat{\mu}_n\right)\left(\tilde{P}_{n,d} - \hat{\mu}_n\right)^\top.$$

This is tested under the fully pooled model across all nodes. For each node n, $\hat{\Sigma}^P_n$ is estimated as above. If this hypothesis cannot be rejected, a single covariance model could describe the whole network. The global test confirms that covariance matrices differ significantly between nodes for both active power (p<0.001) and reactive power (p < 0.001). This justifies node-specific modelling throughout.

Pairwise tests were then conducted to confirm which node pairs differ. For active power, all 10 pairwise comparisons are significant ($p < 0.001$). This means every pair of active power nodes has a statistically distinct covariance structure. For reactive power, 8 of the 10 node pairs show significant differences in their covariance structure. The only non-significant comparisons are Q_node1 (agricultural) vs Q_node2 (agricultural) ($p = 0.216$) and Q_node11 (residential) vs Q_node4 (residential) ($p = 0.056$). The lack of significance for Q_node1 (agricultural) and Q_node2 (agricultural) suggests their intra-day reactive power variability is similar, and that both nodes represent agricultural clusters.

In summary, three main conclusions arise from the covariance analysis. Firstly, most nodes exhibit significant seasonal effects, meaning the variability structure of daily load profiles changes across seasons. The only exceptions are P_node1 (agricultural) and Q_node4 (residential), which show no significant seasonal differences. Secondly, weekday effects are limited in the categorical model. Among active power nodes, only P_node2 (agricultural) shows significant weekday-related differences, mainly driven by contrasts between weekdays and weekends. No reactive power node exhibits a significant weekday effect. Thirdly, covariance structures differ significantly across nodes. This confirms strong node-specific behaviour and justifies separate modelling for each node. Based on these findings, seasonal covariance models should be used for all nodes except P_node1 (agricultural) and Q_node4 (residential). For P_node2 (agricultural), both seasonal and weekday effects should be incorporated. For P_node1 (agricultural) and Q_node4 (residential), a single pooled covariance structure is sufficient. Node-specific modelling remains necessary throughout.

5.2 Covariance Testing with Sine-Corrected Profiles

Since seasonal effects are present for most nodes (Section 5.1), this section examines whether a continuous sine-based seasonal correction can eliminate these effects. After correction, any remaining variation should be due only to weekday patterns or node-specific behaviour. In this approach, the estimated seasonal component is subtracted from the original data. The covariance tests are then repeated. If the correction is successful, the covariance matrices should no longer vary across seasons. This would indicate that the seasonal effect has been fully captured by the sine model.

The sine model from Section 4.4.3 fits separate coefficients per weekday and hour. The season-corrected profile is then obtained by subtracting the full fitted model. This includes both intercept and slope:

$$P^*_{n,d,H} := \tilde{P}_{n,d,H} - \left[\hat{\beta}^1_{n,w(d),H} + \hat{\beta}^2_{n,w(d),H} \cdot \sin\left(\frac{2\pi(d-80)}{365}\right)\right].$$

In R:

```
season_corrected <- predicted_profiles_sine_df %>%
  mutate(P_star = Smoothed - Predicted)
```

where `Predicted` $= \hat{\beta}^{P,1} + \hat{\beta}^{P,2} \times$ SineTerm as fitted in Section 4.4.3.

After correction, the model reduces to

$$P^*_{n,d,H} = \hat{\mu}_{n,d,H} + E_{n,d,H}, \quad E_{n,d,H} \sim \mathcal{N}_{24}\left(0_{24}, \Sigma_{n,w(d),H}\right) \quad \text{for } d \text{ with } w(d) = w.$$

Here, $\hat{\mu}_{n,d,H} \in \mathbb{R}^{24}$ is the 24-dimensional daily hourly vector of season-corrected profiles, aggregated from the 355 clean days.

Since the full regression model is subtracted, the residuals $P^*_{n,d,h}$ are centered near zero by construction. Covariance matrices are estimated directly without additional centering:

$$\hat{\Sigma}_{n,w} := \frac{1}{|D_w|-1} \sum_{d \in D_w} \left(P^*_{n,d} - \hat{\mu}_{n,w}\right)\left(P^*_{n,d} - \hat{\mu}_{n,w}\right)^\top,$$

where $P^*_{n,w} = \frac{1}{|D_w|}\sum_{d \in D_w} P^*_{n,d}$ is the set of all days with weekday w, regardless of season. The notation $|D_w|$ denotes the count of days.

Testing then proceeds as in Section 5.1. Permutation tests with 1000 permutations assess the equality of covariance matrices. The same construction applies to reactive power, replacing P with Q throughout.

Global Weekday Effect: The first test checks if the covariance matrix of season-corrected daily profiles differs by weekday for each node. The null hypothesis is

$$H_0^w : \Sigma_{n,1} = \Sigma_{n,2} = \cdots = \Sigma_{n,7}.$$

Results are shown in Table 18. Only P_node2 (agricultural) for active power ($p = 0.004$) and Q_node10 (residential) for reactive power ($p = 0.029$) show significant weekday effects. The other nodes show no significant variation.

Node	Active Power p-value	Reactive Power p-value
Node 1	0.830	0.349
Node 2	0.004*	0.436
Node 4	0.610	0.553
Node 10	0.191	0.029*
Node 11	0.943	0.148

Table 18: Global weekday effect p-values for season-corrected active and reactive power profiles. Asterisks denote $p < 0.05$.

Since P_node2 (agricultural) showed a significant global weekday effect, pairwise permutation tests identified which weekday contrasts drive this result. Sunday consistently differs from Monday through Thursday. Saturday differs from Tuesday and Wednesday, though to a lesser extent. Friday does not significantly differ from any other day and sits between weekdays and the weekend. This aligns with Section 5.1 and confirms that the weekday covariance effect for P_node2 (agricultural) remains after seasonal correction.

Non-Weekend Effect: To assess the persistence of the weekday effect within working days, the analysis was limited to the period from Monday to Friday. The null hypothesis is given by

$$\tilde{H}_0^w : \Sigma_{n,1} = \cdots = \Sigma_{n,5}.$$

Node	Active Power p-value	Reactive Power p-value
Node 1	0.766	0.34
Node 2	0.304	0.749
Node 4	0.669	0.79
Node 10	0.087	0.041*
Node 11	0.832	0.115

Table 19: Weekday effect (Monday-Friday only) p-values for season-corrected active and reactive power profiles. Asterisks denote $p < 0.05$.

Results are presented in Table 19. For P_node2 (agricultural), $p = 0.304$, indicating no significant effect. The weekday effect arises only from the distinction between working days and Sunday, not from variation within the working week. No other active power node shows significance in this restricted test. For reactive power, Q_node10 (residential) yields $p = 0.041$, remaining significant across Monday to Friday. The only significant pairwise contrast within working days for Q_node10 (residential) is between Tuesday and Thursday ($p = 0.018$). This indicates a specific difference in the reactive power covariance structure for these days at this node after seasonal correction. This is a novel finding relative to Section 5.1 and merits attention, though the effect is limited to a single node and pair.

Node Effect per weekday: The node effect test checks whether the covariance matrices differ across nodes for each weekday, using the season-corrected profiles. The null hypothesis is

$$H_0^n : \Sigma_{1,w} = \cdots = \Sigma_{N,w}.$$

After the node effect test, the global test shows statistical significance ($p \approx 0$) across all seven weekdays for both active and reactive power. This confirms the persistence of node-specific covariance structures after seasonal correction.

For pairwise comparisons, P_node1 (agricultural) differs significantly from all other nodes on every weekday for active power. The only consistent non-significant pair is P_node10 (residential) vs P_node11 (residential). For reactive power, Q_node1 (agricultural) vs Q_node2 (agricultural) remains non-significant across all weekdays. Q_node11 (residential) vs Q_node4 (residential) is significant on only two weekdays, Tuesday and Thursday. Both results are consistent with the findings of Section 5.1.

In summary, the covariance analysis shows that node-specific effects predominate in shaping the variability structure of daily load profiles. For both active and reactive power, covariance matrices differ significantly across nodes for every weekday, indicating substantial structural differences. The influence of weekdays is limited. Only P_node2 (agricultural, active power) and Q_node10 (residential, reactive power) show significant variation across weekdays. These distinctions mainly arise from contrasts between weekdays and weekends rather than among individual weekdays.

Overall, the results under sine-based seasonal correction align with those from Section 5.1. This corroborates the conclusion that the identified covariance structures are robust to the method of seasonal adjustment. These findings show that covariance modelling is primarily node-driven, with minor weekday effects limited to P_node2 (agricultural) and Q_node10 (residential). For all other nodes, a single covariance structure suffices.

5.2.1 Validation: Residual Season Effects After Sine Correction

An important question is whether the sine-based model successfully removes seasonal effects. To assess this, residuals are examined after sine correction. The goal is to determine whether differences persist across season categories in their means or covariance structures.

Mean-level test: The MANOVA procedure described in Section 4.4.3.1 is applied to the residual vectors $\mathbf{P}^*_{n,d,H}$, treating season as the categorical predictor. The null hypothesis is

$$H_0^1 : \text{Season means of residuals are equal for each node.}$$

Rejection of this hypothesis indicates that the sine model does not fully remove seasonal structure in the mean.

Covariance-level test: Permutation-based covariance tests are also applied to the residuals to assess whether covariance structures differ across seasons. The null hypothesis is

$$H_0^{s|\text{resid}} : \mathbf{\Sigma}^{P*}_{n,Sum} = \mathbf{\Sigma}^{P*}_{n,Win} = \mathbf{\Sigma}^{P*}_{n,Rest}.$$

Rejection indicates that residual seasonal differences persist in the covariance structure after sine correction.

Table 12 reports permutation test p-values for seasonal effects on the covariance structure of sine-corrected residuals. For active power, seasonal effects remain significant for all five nodes after correction. p-values range from 0.047 (P_node1) to below 0.001 (P_node4, P_node10, P_node11). For reactive power, four of five nodes continue to show significant effects. Q_node4 does not exhibit a remaining seasonal effect ($p = 0.146$).

The sine model captures the main seasonal trend, yielding near-zero mean residuals, but does not fully remove season-specific covariance differences. Section 5.4 confirms that cross-season correlations are generally close to zero, with residual structure persisting for Q_node2 (Summer–Rest correlation: -0.297). Thus, while the sine model offers a parsimonious detrending option, season-specific covariance modelling remains necessary for high precision near seasonal transitions.

5.3 Covariance testing by Week vectors

Having established that discrete season categories and sine-based corrections produce consistent node-specific covariance results (Sections 5.1–5.2), this section switches to a weekly representation. The aim is to study the within-week dependence structure directly. This switch is motivated by computational and methodological concerns. Using

7-dimensional week vectors (one per weekday) removes the hour dimension. This avoids numerical issues linked to 24×24 covariance matrices with limited group samples. Tests for hour effects (Section 5.3, first test) show the covariance structure does not vary across hours for any node ($p = 1.0$ for all nodes). Thus, pooling all 24 hours into a single 7×7 covariance matrix per node is justified. The resulting matrices capture the correlation structure between weekdays, providing patterns directly interpretable for forecasting and grid analysis.

The season-corrected profiles $P^*_{n,d,H}$ from Section 5.2 are used as input. For each clean week ω with $w(\omega)$ clean days, where $\omega \in \{1, ..., W_{clean}\}$, let d_i be the day in week ω with $w(d_i) = i$. This ensures the components correspond to Monday through Sunday. The week vector at hour H is:

$$\mathbf{P}^*{}_{n,\omega,H} = \left(P^*_{n,d_1,H}, \ldots, P^*_{n,d_7,H}\right)^\top \in \mathbb{R}^7,$$

Incomplete weeks and weeks with changepoints are excluded. Out of 54 total weeks, 2 are incomplete, and 9 contain anomalous days. This leaves $W_{clean} = 43$ clean weeks. Each node contributes $43 \times 24 = 1,032$ week vectors. The model assumed is

$$P^*_{n,\omega,H} = \hat{\mu}_{n,H} + E_{n,\omega,H}, \quad E_{n,\omega,H} \sim \mathcal{N}_7\left(0_7, \Sigma_{n,H}\right), \quad \omega = 1, \ldots, W.$$

Here, $\hat{\mu}_{n,H} = \left(\hat{\mu}_{n,1,H}, \ldots, \hat{\mu}_{n,7,H}\right)^\top$ is the mean week profile for node n at hour H. $\Sigma_{n,H}$ is the 7×7 covariance matrix capturing week-to-week variability in the within-week pattern at hour H. The estimators are

$$\hat{\mu}_{n,H} := \frac{1}{W_{clean}} \sum_{\omega=1}^{W} P^*_{n,\omega,H}, \quad \hat{\Sigma}_{n,H} := \frac{1}{W_{clean} - 1} \sum_{\omega=1}^{W} \left(P^*_{n,\omega,H} - \hat{\mu}_{n,H}\right) \left(P^*_{n,\omega,H} - \hat{\mu}_{n,H}\right)^\top.$$

The same construction applies to reactive power, replacing P with Q throughout.

Hour effects: This test examines whether the covariance matrix $\Sigma_{n,H}$ differs across the 24 hours of the day for each node. The null hypothesis is

$$H_0^H : \Sigma_{n,1} = \cdots = \Sigma_{n,24}.$$

This is testing using the model

$$P^*_{n,\omega,H} = \hat{\mu}_n + E_{n,\omega,H}, \quad E_{n,\omega,H} \sim \mathcal{N}_7\left(0_7, \Sigma_n\right),$$

where Σ_n is a single 7×7 matrix for node n, estimated by

$$\hat{\mu}_n := \frac{1}{H * W_{clean}} \sum_{h=1}^{H} \sum_{\omega=1}^{W} P^*_{n,\omega,H},$$

$$\hat{\Sigma}_n := \frac{1}{H * W_{clean} - 1} \sum_{h=1}^{H} \sum_{\omega=1}^{W} \left(P^*_{n,\omega,H} - \hat{\mu}_n \right) \left(P^*_{n,\omega,H} - \hat{\mu}_n \right)^\top.$$

A significant result would suggest that week-to-week variability in within-week patterns changes with the time of day, requiring hour-specific covariance models. However, for all five nodes (P and Q), the permutation test yields $p \approx 1.0$, indicating hour-to-hour variation in covariance is negligible compared to overall week-to-week variability in $P^*_{n,\omega,H}$. Consequently, this justifies pooling all 24 hours into a single covariance matrix $\hat{\Sigma}_n$ per node.

The result of $p = 1$ for P_node1 (agricultural) may seem surprising since this node has near-zero power values at night and active consumption during the day, which might suggest different covariance structures at different hours. Nevertheless, P_node1 also has the highest overall variability among all active power nodes, with an observed total Frobenius distance of 0.030 across hours compared to at most 0.003 for the other nodes. This high variability means randomly reassigning hour labels produces permuted distances as large as the observed one on every permutation, so the test never finds significance. Thus, hour-to-hour differences in P_node1's covariance structure exist but are small relative to its overall variability and cannot be statistically distinguished from random noise.

The covariance structure of the 7-dimensional week vectors is the same across all 24 hours of the day, for every node.

Node Effects per Hour: If hour effects are absent, the node effect tests whether pooled covariance matrices differ across nodes. At each hour H, the null hypothesis is

$$H_0^n : \Sigma_{1,H} = \cdots = \Sigma_{N,H}.$$

This is tested with the model

$$P^*_{n,\omega,H} = \hat{\mu}_H + E_{n,\omega,H}, \quad E_{n,\omega,H} \sim \mathcal{N}_7 \left(0_7, \Sigma_H \right),$$

and is estimated as

$$\hat{\mu}_H := \frac{1}{N * W_{clean}} \sum_{n=1}^{N} \sum_{\omega=1}^{W} P^*_{n,\omega,H},$$

$$\hat{\Sigma}_H := \frac{1}{N * W_{clean} - 1} \sum_{n=1}^{N} \sum_{\omega=1}^{W} \left(P^*_{n,\omega,H} - \hat{\mu}_H\right) \left(P^*_{n,\omega,H} - \hat{\mu}_H\right)^\top.$$

The test is significant at all 24 hours for both active and reactive power ($p \approx 0$), confirming each node consistently follows its own daily load pattern. Thus, node-specific covariance modelling is justified.

Overall Node Effect: Since the hour effect is not significant, the pooled model $E_{n,\omega,H} \sim \mathcal{N}_7(0_7, \Sigma_n)$ is used. The overall node effect is then tested by

$$H_0^n : \Sigma_1 = \cdots = \Sigma_5,$$

This is tested using the model

$$P^*_{n,\omega,H} = \hat{\mu} + E_{n,\omega,H}, \quad E_{n,\omega,H} \sim \mathcal{N}_7\left(0_7, \Sigma\right),$$

and is estimated by

$$\hat{\Sigma} := \frac{1}{N * H * W_{clean} - 1} \sum_{n} \sum_{H} \sum_{\omega} \left(P^*_{n,\omega,H} - \hat{\mu}\right) \left(P^*_{n,\omega,H} - \hat{\mu}\right)^\top,$$

using all weeks and hours.

The test gives $p < 0.001$ for active and reactive power, showing that within-week covariance structures differ significantly across nodes after pooling over hours. Node-specific covariance matrices $\hat{\Sigma}_n$ are used for all further analysis.

Pooled Covariance Matrices and Inter-weekday Correlations

The pooled covariance is estimated by averaging the covariance matrices for each of the 24 hours

$$\hat{\Sigma}_{n,H} := \frac{1}{W_{clean} - 1} \sum_{\omega=1}^{W_{clean}} \left(P^*_{n,\omega,H} - \hat{\mu}_{n,H}\right) \left(P^*_{n,\omega,H} - \hat{\mu}_{n,H}\right)^\top,$$

Since each hour H has the same number of clean weeks ($W_{clean} = 43$), the pooled covariance is computed as an unweighted average. This transitions from hour-specific estimation to aggregation:

$$\hat{\Sigma}_n := \frac{1}{24} \sum_{H=0}^{23} \hat{\Sigma}_{n,H}.$$

This assigns equal weight to each hour and pools weekly patterns across all 24 hours.

To summarise the inter-weekday relationships across nodes, Table 20 presents the mean, minimum, and maximum inter-weekday correlation for each node.

Node	**Mean**	**Min**	**Max**	**Range**
P_node1	0.352	0.082	0.652	0.570
P_node2	0.708	0.440	0.885	0.445
P_node4	0.662	0.497	0.812	0.314
P_node10	0.546	0.307	0.777	0.470
P_node11	0.753	0.577	0.880	0.302
Q_node1	0.654	0.468	0.792	0.324
Q_node2	0.881	0.781	0.941	0.160
Q_node4	0.558	0.410	0.657	0.246
Q_node10	0.293	0.152	0.508	0.356
Q_node11	0.242	0.105	0.458	0.354

Table 20: Summary of pooled inter-weekday correlations averaged across all 24 hours. Mean correlation indicates stability: high values suggest consistent weekly patterns; low values indicate day-to-day variability.

The results reveal major node-specific differences in within-week dependence. For active power, P_node11 (residential) is the most stable node, with a high mean inter-weekday correlation of 0.753 and a narrow range of 0.302, indicating consistent daily patterns across all weekday pairs. In contrast, P_node1 (agricultural) is the most variable, with a low mean correlation of 0.352 and the widest range of 0.570, showing inconsistent weekday relationships between weeks. P_node2 (agricultural) and P_node4 (residential) are intermediate, with moderately high mean correlations between them.

For reactive power, Q_node2 (agricultural) is the most stable node, with a very high mean correlation of 0.881 and a narrow range of 0.160, the tightest among all ten nodes. In comparison, Q_node10 (residential) and Q_node11 (residential) are the least stable: Q_node10 has a mean correlation of 0.293, and Q_node11 has 0.242, both with broader ranges. This shows that while Q_node2's pattern is highly consistent week to week, Q_node10 and Q_node11 exhibit weaker, more variable patterns over time.

The pooled correlation matrices for selected nodes are shown in Tables 21, 22, 23, and 24, illustrating the range of within-week dependence structures across agricultural and residential nodes for both P (active power) and Q (reactive power). For P_node2 (agricultural), the highest correlations occur between Saturday and Sunday ($r = 0.865$) and between these and Monday ($r = 0.831$), while the weakest is between Thursday and Saturday ($r = 0.480$), reflecting a clear weekday-weekend contrast. In comparison, P_node11 (residential) shows correlations ranging from 0.577 to 0.880, with only Friday-Sunday below 0.6 ($r = 0.577$), suggesting a more consistent weekly pattern across all days. For Q_node2 (agricultural), all weekday pairs show very high correlations with a minimum of 0.781 (Thursday vs Saturday), indicating the most stable within-week structure and reactive power among all nodes. On the other hand, Q_node11 (residential) stands out with many pairs below 0.2, such as Tuesday-Thursday ($r = 0.105$), Wednesday-Monday ($r = 0.110$), and Tuesday-Sunday ($r = 0.111$), confirming a highly variable and irregular inter-weekday dependence as implied by its low mean correlation of 0.242.

P_node2 (Agricultural)							
	Mon	**Tue**	**Wed**	**Thu**	**Fri**	**Sat**	**Sun**
Mon	1.000	0.831	0.719	0.583	0.603	0.773	0.831
Tue	0.831	1.000	0.885	0.735	0.734	0.674	0.773
Wed	0.719	0.885	1.000	0.880	0.777	0.615	0.721
Thu	0.583	0.735	0.880	1.000	0.828	0.480	0.585
Fri	0.603	0.734	0.777	0.828	1.000	0.440	0.539
Sat	0.773	0.674	0.615	0.480	0.440	1.000	0.865
Sun	0.831	0.773	0.721	0.585	0.539	0.865	1.000

Table 21: Pooled inter-weekday correlation matrix for P_node2, averaged across all 24 hours.

P_node11 (Residential)							
	Mon	**Tue**	**Wed**	**Thu**	**Fri**	**Sat**	**Sun**
Mon	1.000	0.870	0.744	0.755	0.723	0.750	0.804
Tue	0.870	1.000	0.829	0.771	0.731	0.758	0.769
Wed	0.744	0.829	1.000	0.880	0.775	0.750	0.679
Thu	0.755	0.771	0.880	1.000	0.842	0.730	0.695
Fri	0.723	0.731	0.775	0.842	1.000	0.645	0.577
Sat	0.750	0.758	0.750	0.730	0.645	1.000	0.744
Sun	0.804	0.769	0.679	0.695	0.577	0.744	1.000

Table 22: Pooled inter-weekday correlation matrix for P_node11 (residential), averaged across all 24 hours.

In summary, no statistically significant hour effect is detected for any node at $\alpha = 0.05$, so a single pooled covariance matrix $\hat{\Sigma}_n$ is used per node. Node effects are significant throughout, confirming that within-week covariance structures are node-specific. The pooled correlation matrices show a wide range of inter-weekday dependence across

Q_node2 (Agricultural)							
	Mon	**Tue**	**Wed**	**Thu**	**Fri**	**Sat**	**Sun**
Mon	1.000	0.937	0.887	0.862	0.865	0.903	0.931
Tue	0.937	1.000	0.934	0.893	0.894	0.871	0.909
Wed	0.887	0.934	1.000	0.938	0.905	0.833	0.868
Thu	0.862	0.893	0.938	1.000	0.922	0.781	0.817
Fri	0.865	0.894	0.905	0.922	1.000	0.791	0.816
Sat	0.903	0.871	0.833	0.781	0.791	1.000	0.941
Sun	0.931	0.909	0.868	0.817	0.816	0.941	1.000

Table 23: Pooled inter-weekday correlation matrix for Q_node2, averaged across all 24 hours.

Q_node11 (Residential)							
	Mon	**Tue**	**Wed**	**Thu**	**Fri**	**Sat**	**Sun**
Mon	1.000	0.174	0.110	0.324	0.356	0.256	0.458
Tue	0.174	1.000	0.256	0.105	0.220	0.223	0.111
Wed	0.110	0.256	1.000	0.297	0.287	0.275	0.118
Thu	0.324	0.105	0.297	1.000	0.232	0.171	0.313
Fri	0.356	0.220	0.287	0.232	1.000	0.244	0.308
Sat	0.256	0.223	0.275	0.171	0.244	1.000	0.238
Sun	0.458	0.111	0.118	0.313	0.308	0.238	1.000

Table 24: Pooled inter-weekday correlation matrix for Q_node11, averaged across all 24 hours.

nodes, from the highly stable Q_node2 (agricultural) to the highly irregular Q_node11 (residential) and P_node1 (agricultural).

5.4 Cross-Season Covariances and Correlations

This section examines correlations between residuals across seasons in the sine-corrected profiles. The aim is to validate the seasonal correction approach used in Sections 5.2–5.3. If the sine model from Section 4.4.3 has removed the seasonal trend, residuals across seasons should be uncorrelated. Cross-season correlations near zero provide evidence of effective sine-based correction. This justifies pooling covariance matrices across seasons for the weekly analysis. Identifying residual seasonal structure in specific nodes guides interpretation of those nodes' load patterns and indicates where more sophisticated modelling may be needed.

For each node n, weekday w, and hour H, the Pearson correlation coefficient between sine-corrected residuals from seasons s_1 and s_2 is computed. Let $X_{s,n,w,H}$ denote the collection of sine-corrected residuals from season s:

$$X_{s,n,w,H} := \{P^*_{n,d,H} : d \in S_s,\ w(d) = w\},$$

where $P^*_{n,d,H}$ are the sine-corrected residuals from Section 5.2, and $S_s = \{d : s(d) = s\}$ is the set of all days in season s. The seasons are Summer: 15 May to 14 September; Winter: 1 November to 20 March; Rest: 21 March to 14 May and 15 September to 31 October.

The correlation coefficient between these two sets is

$$\hat{\rho}^{n,w,H}_{s_1,s_2} := \mathrm{corr}(X_{s_1,n,w,H}, X_{s_2,n,w,H}).$$

This individual correlation measures the degree to which residuals from season s_1 covary with residuals from season s_2 when both are observed at the same temporal location (weekday and hour) and physical location (node). Near-zero values indicate independence between seasonal residuals; large absolute values would signal that seasonal effects persist despite the sine-based correction.

Individual correlations are aggregated in two stages to produce summary statistics suitable for interpretation. This two-stage approach enables the identification of both global seasonal patterns (via overall pooled correlations) and node-specific deviations (via node-specific pooled correlations).

First, each node's individual correlations are averaged across all 7 weekdays and 24 hours to obtain a node-specific summary

$$\bar{\rho}^{(n)}_{s_1,s_2} := \frac{1}{W \cdot H} \sum_{w=1}^{W} \sum_{H=0}^{23} \hat{\rho}^{n,w,H}_{s_1,s_2},$$

where $W = 7$ is the number of weekdays and $H = 24$ is the number of hours. This aggregation smooths hour-to-hour and weekday-to-weekday variation, yielding a single correlation value per node per season pair. These node-specific pooled correlations appear in Tables 26 and 27 and identify which nodes exhibit residual seasonal structure after sine correction.

Then further aggregation across all $N = 5$ nodes produces a global summary

$$\bar{\rho}_{s_1,s_2} := \frac{1}{N}\sum_{n=1}^{N}\bar{\rho}^{(n)}_{s_1,s_2}.$$

Equivalently, this overall correlation can be computed directly as

$$\bar{\rho}_{s_1,s_2} := \frac{1}{N \cdot W \cdot H}\sum_{n=1}^{N}\sum_{w=1}^{W}\sum_{H=0}^{23}\hat{\rho}^{n,w,H}_{s_1,s_2},$$

which shows that all 840 individual correlations (5 nodes × 7 weekdays × 24 hours) are equally weighted in the global summary. The three overall correlations, Summer–Winter, Summer–Rest, and Winter–Rest, are assembled into the 3×3 symmetric correlation matrix in Table 25, with diagonal entries set to 1.0 by convention.

Active Power (P)

	Summer	Winter	Rest
Summer	1.0000	-0.0111	-0.0343
Winter	-0.0111	1.0000	-0.0905
Rest	-0.0343	-0.0905	1.0000

Reactive Power (Q)

	Summer	Winter	Rest
Summer	1.0000	-0.0303	-0.0539
Winter	-0.0303	1.0000	-0.0662
Rest	-0.0539	-0.0662	1.0000

Table 25: Overall pooled correlation matrices between seasonal residuals for active (P) and reactive (Q) power. All cross-season correlations are near zero, indicating that residuals from different seasons are nearly independent.

Table 25 shows the overall pooled cross-season correlations for active and reactive power. These 3×3 matrices summarise the dependence structure across the three seasons, enabling assessment of the global effectiveness of the sine-based seasonal correction.

Node	**Summer–Winter**	**Summer–Rest**	**Winter–Rest**
P_node1	0.017	−0.063	−0.032
P_node2	0.039	−0.098	−0.127
P_node4	−0.032	0.021	−0.097
P_node10	−0.079	−0.026	0.007
P_node11	−0.001	−0.006	−0.204

Table 26: Mean cross-season correlations for active power, averaged across all hours and weekdays.

For active power, all cross-season correlations are small, ranging from -0.0905 (Winter vs Rest) to -0.0111 (Summer vs Winter), with a mean absolute correlation of 0.053. For reactive power, the range is -0.0662 (Winter vs Rest) to -0.0303 (Summer vs Winter), with a mean absolute correlation of 0.047. These near-zero values across all season pairs indicate that the sine-based correction has effectively removed seasonal effects. The small

Node	Summer–Winter	Summer–Rest	Winter–Rest
Q_node1	−0.100	0.042	−0.061
Q_node2	0.047	−0.297	−0.084
Q_node4	−0.048	0.016	−0.107
Q_node10	−0.054	−0.009	−0.014
Q_node11	0.004	−0.021	−0.065

Table 27: Mean cross-season correlations for reactive power, averaged across all hours and weekdays.

negative correlations suggest weak inverse relationships between seasons, consistent with the residual noise structure rather than systematic seasonal dependence.

The near-zero overall pooled correlations in Table 25 support pooling weeks across all seasons in weekly covariance analysis (Section 5.3). For most node–season pairs, the sine correction eliminates seasonal dependence. This allows a single pooled covariance matrix $\hat{\Sigma}_n$ to describe the within-week correlation structure.

While most nodes show near-zero cross-season correlations, two nodes exhibit notable residual seasonal structure that warrants attention. Tables 26 and 27 break down results by node.

For active power (Table 26), only P_node11 (residential) stands out, displaying a Winter–Rest correlation of -0.204. This exceeds typical noise levels and shows the sine model does not fully explain the seasonal transition for this node. The negative correlation means higher winter consumption at P_node11 tends to coincide with lower consumption during the rest of the year, revealing a residual seasonal pattern unaccounted for by the sine model. In contrast, P_node1, P_node2, P_node4, and P_node10 have all cross-season correlations below $|r| = 0.10$, confirming successful seasonal correction at those nodes.

For reactive power (Table 27), Q_node2 (agricultural) shows a pronounced Summer–Rest correlation of -0.297, the highest among all node–season pairs. This strong residual pattern at Q_node2 after sine correction mainly affects the summer-to-intermediate transition. It shows the sinusoidal model does not fully capture seasonal effects specific to this agricultural node. All remaining nodes have correlations below $|r| = 0.11$, indicating effective removal of seasonal structure.

The two exceptions (P_node11 Winter–Rest for active power and Q_node2 Summer–Rest for reactive power) pinpoint nodes where residual seasonal structure persists after sine correction. For P_node11, forecasts near the winter-to-spring transition should use season-specific adjustments. For Q_node2, attention is needed during the summer-to-intermediate season shift. These findings highlight where latent season-dependent effects remain and reinforce the need for tailored interpretations. These patterns align with residual effects in Table 12, confirming that while the sine model is broadly effective, fine-scale seasonal structure may still influence certain nodes.

6 Conclusion

This thesis analysed the covariance and correlation structure of active and reactive power load profiles at five representative distribution grid nodes using one year of 15-minute SimBench measurements. The central finding is that node identity dominates covariance variation, while seasonal and weekday effects are secondary and inconsistently present across nodes.

Dimensionality reduction via pairwise regression identified two clusters. The agricultural cluster includes P_node1, P_node2, Q_node1, and Q_node2. The residential cluster includes P_node4, P_node10, P_node11, Q_node4, Q_node10, and Q_node12. These clusters differ mainly by scaling. PELT changepoint detection flagged 11 atypical days, leaving 355 for analysis. Two complementary seasonal modelling approaches yielded consistent results. The discrete approach estimated mean load profiles per season-weekday combination using MANOVA with Pillai's trace. For P-nodes, model structures were mixed: P_node2 and P_node11 required a Season × Weekday interaction, P_node1 and P_node10 required additive effects, and P_node4 required weekday-only effects. For Q-nodes, seasonality was minimal except for Q_node1, Q_node2, and Q_node10. The continuous sine regression approach, with weekday intercepts and node-specific slopes, achieved R^2 values from 0.57 to 0.66. It also confirmed that sine model parameters differ significantly across weekdays for P_node1, P_node2, and P_node10, but not for P_node4 or P_node11. Both approaches support robust model structures and provide confidence in the identified seasonal patterns.

Permutation-based Frobenius distance tests on 24-dimensional daily vectors revealed a clear hierarchy of effects. Node effects dominated, with significant covariance matrix differences between all pairs of active power nodes ($p < 0.001$). For reactive power, 8 of 10 pairs were significant. Only Q_node1 vs Q_node2 ($p = 0.216$) and Q_node11 vs Q_node4 ($p = 0.056$) were not significant, suggesting some structural homogeneity within the clusters. This node-driven pattern persisted through all seasons and aggregation. Seasonal effects were present but less consistent. Most nodes showed significant seasonal effects on covariance. P_node11 (residential) had the most, with 6 of 7 weekdays significant ($p < 0.05$). Global season tests confirmed significance for P_node2, P_node4, P_node10, P_node11 and Q_node1, Q_node2, Q_node10 (all $p < 0.05$). However, P_node1 ($p = 0.062$) and Q_node4 ($p = 0.242$) remained stable across seasons. After a sine-based seasonal correction, results matched those of the discrete approach, indicating the adjustment was robust. Weekday effects were minimal. Only P_node2 (active power, $p < 0.001$) showed global weekday significance, due to the difference between weekday and weekend, not within working days (Monday–Friday $p = 0.062$). No reactive power node showed significant weekday covariance effects. After correction, only P_node2 ($p = 0.004$) and Q_node10 ($p = 0.029$) remained significant, with Q_node10 showing an effect between Tuesday and Thursday ($p = 0.018$). Thus, when present, weekday effects are neither widespread nor primary drivers.

Analysis shifted to a week-based model using 7-dimensional vectors for Monday to Sunday. Hour effects, tests for covariance change across 24 hours, were not significant for any node ($p = 1.0$). This justified averaging to a single pooled covariance matrix, $\hat{\Sigma}_n$, per node across all hours. Pooled inter-weekday correlations showed node-specific within-week stability. Q_node2 (agricultural, reactive power) had the highest stability: mean correlation 0.881, with all weekday pairs $r > 0.78$. Q_node11 (residential) and P_node1 (agricultural) were the least stable, with mean correlations of 0.242 and 0.352, respectively, indicating high variability. P_node2, P_node4, and P_node11 were intermediate, with correlations from 0.66 to 0.75. These pooled correlations summarise within-week dependence. They help with forecast model design and reserve scheduling, and show that some nodes exhibit predictable weekly patterns while others are unpredictable.

Cross-season validation examined pairwise correlations between sine-corrected residuals from different seasons. Pooled cross-season correlations were close to zero: $|r| = 0.01$–0.09 for active power and $|r| = 0.03$–0.07 for reactive power. This shows the sine model removes most seasonal dependence and supports pooling weekly covariance across seasons. However, P_node11 (residential, Winter–Rest $r = -0.204$) and Q_node2 (agricultural, Summer–Rest $r = -0.297$) showed exceptions, suggesting seasonal correction was incomplete for some nodes due to non-sinusoidal or asymmetric effects.

Results from discrete MANOVA, sine-parametric, and permutation-based approaches are consistent. This confirms that identified patterns reflect real data properties. MANOVA assumptions (multivariate normality and equal covariance) are not tested, and small sample sizes (6–8 days per cell) may increase Type I errors. Still, distribution-free permutation tests based on the Frobenius distance independently confirm the findings and mitigate concerns about temporal dependence, demonstrating methodological robustness.

Operationally, node-specific covariance models suffice. Separate models per node are justified, with season-specific adjustments for nodes with strong seasonal effects (P_node11, Q_node2) and simpler models for stable nodes. The range of inter-weekday correlations (0.24–0.88) highlights heterogeneity: low-correlation nodes (Q_node11, P_node1) require larger reserves; high-correlation nodes (Q_node2, P_node11) allow more aggressive demand-response, supporting tailored scenario analysis and adaptive control.

Future research should extend the analysis over multiple years to assess whether the identified node-specific patterns hold consistently and to understand the reasons for the strong seasonal effects observed in certain nodes. Methods such as customer surveys or sub-metering can provide deeper insight. Implementing time-varying covariance models (e.g., rolling-window or dynamic factor approaches) would help model seasonal transitions more smoothly than fixed categories. Including network topology metrics can clarify whether spatial proximity influences covariance. Additionally, developing and validating node-specific forecast models is essential to quantify practical improvements. The observed large Summer–Rest correlation for Q_node2 signals a need to investigate its equipment-level reactive compensation mechanisms as a possible explanatory factor.

In summary, node identity dominates covariance variation in this distribution grid, a finding robust across all approaches. This implies an operational principle: use node-specific covariance models for forecasting, reserves, and scenario analysis. Seasonal effects are present but inconsistent; weekday effects are minimal. Each node has a characteristic, stable variability pattern. The pooled covariance matrices enable uncertainty quantification and reserve calculations, allowing operators to move from generic to node-centric models.

Bibliography

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society series b-methodological*, 57:289–300, 1995. URL `https://api.semanticscholar.org/CorpusID:45174121`. (visited on 23/01/26).

Jedelyn Cabrieto, Francis Tuerlinckx, Peter Kuppens, Borbála Hunyadi, and Eva Ceulemans. Testing for the presence of correlation changes in a multivariate time series: A permutation based approach. *Scientific Reports*, 8(1):769, 2018. doi: 10.1038/s41598-017-19067-2. URL `https://doi.org/10.1038/s41598-017-19067-2`. (visited on 23/01/26).

Gianfranco Chicco. Overview and performance assessment of the clustering methods for electrical load pattern grouping. *Energy*, 42:68–80, 2012. URL `https://api.semanticscholar.org/CorpusID:109972335`. (visited on 01/03/26).

Ludwig Fahrmeir, Thomas Kneib, Stefan Lang, and Brian Marx. *Regression: Models Methods and Applications*. Springer, 2013. ISBN ISBN 978-3-642-34332-2, ISBN 978-3-642-34333-9 (eBook).

Garrett Grolemund and Hadley Wickham. *lubridate: Make Dealing with Dates a Little Easier*. CRAN, 2024. URL `https://CRAN.R-project.org/package=lubridate`. (visited on 15/05/25).

Wolfgang Karl Härdle, Matthias R. Fengler, and Léopold Simar. *Applied Multivariate Statistical Analysis*. Springer Nature Switzerland AG, Cham, 6th edition, 2024. ISBN 978-3-031-63832-9. doi: 10.1007/978-3-031-63833-6.

Richard A. Johnson and Dean W. Wichern. *Applied Multivariate Statistical Analysis*. Pearson Prentice Hall, Upper Saddle River, NJ, 6 edition, 2007.

Rebecca Killick and Idris A. Eckley. *changepoint: An R package for changepoint analysis*. CRAN, 2023. URL `https://CRAN.R-project.org/package=changepoint`. (visited on 27/06/25).

Rebecca Killick, Paul Fearnhead, and Idris Arthur Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107:1590 – 1598, 2011. URL `https://api.semanticscholar.org/CorpusID:5627005`. (visited on 28/06/25).

Steffen Meinecke, Dzanan Sarajlic, Simon Drauz, Annika Klettke, Lars-Peter Lauven, Christian Rehtanz, Albert Moser, and Martin Braun. Simbench—a benchmark dataset of electric power systems to compare innovative solutions based on power flow analysis. *Energies*, 2020. URL `https://api.semanticscholar.org/CorpusID:225647452`. (visited on 01/04/25).

Erich Neuwirth. *RColorBrewer: ColorBrewer Palettes*. CRAN, 2022. URL `https://CRAN.R-project.org/package=RColorBrewer`. (visited on 15/05/25).

Jakub Nowotarski and Rafał Weron. Recent advances in electricity price forecasting: A review of probabilistic forecasting. *HSC Research Reports*, 2016. URL `https://api.semanticscholar.org/CorpusID:157301643`. (visited on 01/03/26).

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL `https://www.R-project.org/`. (visited on 15/05/25).

SimBench Project. Simbench datasets, 2024. URL `https://simbench.de/en/download/datasets/`. (visited on 01/04/25).

J. Tukey. *Exploratory data analysis*. Addison-Wesley Publishing Company, Don Mills, 1977.

Taiyun Wei and Viliam Simko. *corrplot: Visualization of a Correlation Matrix*. CRAN, 2024. URL `https://CRAN.R-project.org/package=corrplot`. (visited on 15/05/25).

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL `https://ggplot2.tidyverse.org`. (visited on 15/05/25).

Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019. doi: 10.21105/joss.01686.