TU Delft

Delft University of Technology

Document Version
Final published version

Citation (APA)
Duque, E. M. S., Holst van der, B., Vergara, P. P., Giraldo, J. S., Nguyen, P. H., Van der Molen, A., & Slootweg, H. J. G. (2025). Lower Dimensional Spherical Representation of Medium Voltage Load Profiles for Visualization, Outlier Detection, and Generative Modelling. *IEEE Transactions on Smart Grid*, *16*(6), 5170-5184. https://doi.org/10.1109/TSG.2025.3597451

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
In case the licence states “Dutch Copyright Act (Article 25fa)”, this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.

Green Open Access added to TU Delft Institutional Repository
as part of the Taverne amendment.

More information about this copyright law amendment
can be found at https://www.openaccess.nl.

Otherwise as indicated in the copyright section:
the publisher is the copyright holder of this work and the
author uses the Dutch legislation to make this work public.

Lower Dimensional Spherical Representation of Medium Voltage Load Profiles for Visualization, Outlier Detection, and Generative Modelling

Edgar Mauricio Salazar Duque, *Graduate Student Member, IEEE*, Bart van der Holst, *Graduate Student Member, IEEE*, Pedro P. Vergara, *Senior Member, IEEE*, Juan S. Giraldo, *Senior Member, IEEE*, Phuong H. Nguyen, *Member, IEEE*, Anne Van der Molen, *Member, IEEE*, and Han J. G. Slootweg, *Senior Member, IEEE*

***Abstract*—This article presents the theoretical and practical foundation of a spherical lower dimensional representation for daily medium voltage load profiles, based on principal component analysis. The objective is to unify and simplify the tasks for (i) clustering visualisation, (ii) outlier detection and (iii) generative profile modelling under one concept. The lower dimensional projection of standardised MV load profiles unveils a latent distribution in a three-dimensional sphere. This spherical structure allows us to detect outliers by fitting probability distribution models in the spherical coordinate system, identifying measurements that deviate from the spherical shape. The same latent distribution exhibits an arc shape, suggesting an underlying order among load profiles. We develop a principal curve technique to uncover this order based on similarity, offering new advantages over conventional clustering techniques. This finding reveals that energy consumption in a wide region can be seen as a continuously changing process. Furthermore, we combined the principal curve with a von Mises-Fisher distribution to create a model capable of generating profiles with continuous mixtures between clusters. The presence of the spherical distribution is validated with data from four municipalities in the Netherlands. The uncovered spherical structure implies the possibility of employing new mathematical tools from directional statistics and differential geometry for load profile modelling.**

***Index Terms*—Dimensionality reduction, principal component analysis, principal coordinate analysis, principal curve analysis, electricity profiles, outlier detection.**

Received 22 November 2024; revised 1 April 2025 and 18 July 2025; accepted 4 August 2025. Date of publication 11 August 2025; date of current version 23 October 2025. Paper no. TSG-02019-2024. *(Corresponding author: Edgar Mauricio Salazar Duque.)*

Edgar Mauricio Salazar Duque and Han J. G. Slootweg are with the Electrical Energy Systems Group, Eindhoven University of Technology, 5612 AE Eindhoven, The Netherlands, and also with the Department of Asset Information Management, Enexis, 5223 MB 's-Hertogenbosch, The Netherlands (e-mail: e.m.salazar.duque@tue.nl; j.g.slootweg@tue.nl).

Bart van der Holst and Phuong H. Nguyen are with the Electrical Energy Systems Group, Eindhoven University of Technology, 5612 AE Eindhoven, The Netherlands (e-mail: b.v.d.holst@tue.nl; p.nguyen.hong@tue.nl).

Pedro P. Vergara is with the Intelligent Electrical Power Grids Group, Delft University of Technology, 2628 CD Delft, The Netherlands (e-mail: p.p.vergarabarrios@tudelft.nl).

Juan S. Giraldo is with the Techno-Economic Energy Transition Studies Group, Netherlands Organization for Applied Scientific Research, 1043 NT Amsterdam, The Netherlands (e-mail: juan.giraldo@tno.nl).

Anne van der Molen is with the Department of Electrical Engineering, Eindhoven University of Technology, and also with the Department of System and Network Strategy, Stedin, 3011 TA Rotterdam, The Netherlands (e-mail: a.e.v.d.molen@tue.nl).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TSG.2025.3597451.

Digital Object Identifier 10.1109/TSG.2025.3597451

I. Introduction

A. Background on Load Profile Analysis

THE WIDESPREAD installation of advanced metering infrastructure (AMI) in the medium voltage (MV) distribution grid has enabled distribution network operators (DNOs) to gather enormous volumes of data; within the variety of data collected by the AMI, there are daily load profiles that opens new possibilities for load analysis, forecasting, and load management [1]. These load profiles can be collected at different time resolutions. The resolution refers to the discretisation steps of the time axis at equal intervals, e.g., minutes, quarterly, and hourly. When each discrete step is considered an independent variable, any load profile modelling becomes a multivariate analysis problem. The dimensionality of the model depends on the discretisation resolution. e.g., a daily profile with a quarterly resolution is a 96-dimensional model.

The large number of load profiles collected by the DNOs creates challenges for the multivariate analysis, storage, and efficient use of computing processing power. One technique to aid these challenges is to reduce the dimensionality of the data. In this article, we refer to dimensionality reduction (DR) as the broad category of linear and nonlinear space embedding methods [2], which reduces the dataset to a lower dimensionality space while maintaining most of its information.

B. Dimensionality Reduction Applications in Power Systems

In general, four major tasks can benefit from the application of dimensionality reduction (DR) techniques: (i) Clustering can be potentially improved when most of the information is contained in few dimensions [3], (ii) Visualisation and interpretation of the clustering can be made if the data is represented in less than three dimensions, (iii) Anomalous readings can be spotted when the data does not follow the structure in the lower-dimensional representation (outlier detection), and (iv) simple generative models can be designed to create load profiles with the same statistical properties as the original data set. For each of the above-mentioned tasks, the data analyst should employ various methods.

1949-3053 © 2025 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: TU Delft Library. Downloaded on October 30,2025 at 11:36:28 UTC from IEEE Xplore. Restrictions apply.

Each task has its own family of methods, which constitutes a considerable area of research in its own right. Here, we mention the relevant literature on the applications of DR for each task, highlighting the relevance of load profile modelling for power systems.

C. Dimensionality Reduction for Clustering and Visualisation

Clustering techniques are used to understand and allocate consumer behaviour in groups [4], which can be used to customise demand response energy efficiency programmes [5]. Extensive research has been dedicated to studying the clustering of residential load demand profiles [6]. Examples of clustering algorithms used are k-means, self-organising maps, hierarchical and spectral clustering [7], load decomposition [8], c-vine copulas [9], and probability distribution mixture models [10]. It is common to apply these methodologies considering all dimensions and achieving successful results. However, it creates a scalability problem when the number of profiles increases. Applying the DR technique before clustering is a different approach to improve results and reduce computational time. Principal component analysis (PCA) is generally used as a data preprocessing step as a linear DR technique that preserves most of the variance of the data in a lower-dimensional projection. An important observation from the works that apply DR before clustering is that approximately 85-90% of the variance for the load profiles can be retained in less than four dimensions using PCA, i.e., [7], [11], [12], [13], [14]. This reduction implies that PCA shows great potential for visualising and analysing clustering results for load profile data with minimal computational effort. However, the study of data structure and lower-dimensional embeddings created by PCA is often neglected. This includes examining common shapes and patterns in the projections of various load profile datasets, which can be valuable for anomaly detection and generative modelling tasks. Combining DR with clustering techniques offers the additional benefit of analysing and visualising the resulting groupings if the reduced dimensions are three or fewer.

Recent research shows applications of nonlinear DR techniques on energy data for visualisation and clustering using neighbour-based techniques, such as lower linear embedding (LLE) [15], Isomap [16], and t-SNE [17]. The main drawback of these techniques is the number of parameters needed to tune, e.g., the number of neighbours, components, learning rate, and regularisation, which can significantly impact the projection results. Additionally, nonlinear DR techniques based on neural networks, e.g., UMAP [18], and convolutional autoencoders [19], show good performance in compression and clustering measure statistics. Nevertheless, creating the non-linear model requires significant time and computational power, often needing data augmentation techniques. Also, they are very flexible models susceptible to random initialisation in the neural network. Consequently, when an autoencoder is trained with a different random initialisation, we can expect different projections or embeddings for the same input data. This randomness creates a potential problem in defining anomaly scores in the reduced space and requires reinterpreting the clustering results between tests, restricting its use only to visualisation purposes. In general, nonlinear techniques have great compressibility power. However, they sacrifice stable and consistent projections compared to PCA. In this article, we present a comparison of the aforementioned non-linear techniques and our proposed method in Section V-E in order to assess the deterministic and reliable projection of our PCA-based method versus state-of-the-art techniques.

D. Dimensionality Reduction for Outlier Detection

Additionally, DR techniques can also be applied for anomaly detection. Load profile data can have outliers, which are corrupted, abnormal, noisy, and missing data due to various causes. The typical approach to find outliers of the load profile is to use all dimensions and apply kernel smoothing techniques [20], [21], a nearest-neighbours approach using local outlier factor (LOF) [22], or tree-based methods such as isolation forest [23]. These techniques are commonly applied in power system contexts such as electricity theft detection [24] and synchrophasor anomaly detection for grid monitoring [25], [26]. Recent frameworks increasingly adopt ensemble methods, combining multiple unsupervised models to improve detection performance. For example, [25] integrates Isolation Forest, KMeans, and Local Outlier Probabilities (LoOP) [27], while [26] employs DBSCAN [28], KMeans, LOF, Feature Bagging, and Robust Random Cut Forest (RRCF) [29]. Although effective, these ensembles introduce a large number of hyperparameters, limiting scalability across large-scale deployments involving thousands of sensors. Furthermore, those methods are successfully tested and applied for single-metre readings. Usually, outlier detection on load profiles is a laborious process for any data analyst, since an outlier model for each meter should be created. However, in a lower-dimensional space, it is possible to identify outlier profiles from a group of sensors analysing the latent distribution of the data, i.e., probability distribution in the lower-dimensional projection. Simpler models with consistent projections are desired, allowing data analysts to spot anomalous readings quickly.

E. Generative Modelling of Load Profiles

Generative models for electricity load profiles have been proven to be effective tools to support different operational processes within DNOs. These models enable the creation of large synthetic databases for diverse consumption patterns, which can be used to improve rare event risk assessments [30], train data-intensive control algorithms such as reinforcement learning [31], and evaluate the efficiency of photovoltaic control mechanisms [32].

In the literature. Two principal modelling paradigms are commonly employed for *daily load profile* generation. The first approach directly models the data in its original high-dimensional space. e.g., the daily load profile at 15-minute resolution is treated as a 96-dimensional vector. Techniques in this category include copula-based models [9], [33], normalizing flows [34], and diffusion models, all of which aim to learn

Authorized licensed use limited to: TU Delft Library. Downloaded on October 30,2025 at 11:36:28 UTC from IEEE Xplore. Restrictions apply.

the full data distribution without performing dimensionality reduction.

The second approach involves latent-space modelling, where the objective is to learn a compact, lower-dimensional embdedding of the data. Synthetic profiles are then generated by sampling from this latent space and decoding the results back into the original high-dimensional space. Representative approaches include Variational Autoencoders (VAE) [35] and generative adversarial networks (GANs) [34], which are optimized to replicate the statistical distribution of the original data with high fidelity.

Although these latent spaces are useful for generation, they are not necessarily interpretable or suitable for visualisation. This contrasts with dimensionality reduction techniques such as Isomap, t-SNE, and UMAP, which are explicitly designed to preserve key local, and in some cases global—data relationships, in lower-dimensional projections, making them more appropriate for exploratory data analysis and visualization tasks.

F. Towards a Unified Framework

The four tasks of clustering, visualisation, outlier detection, and generative modelling are highly interlinked in the DR domain, and it is always on great interest to look for models that bring the different tasks closer together. Generally, data analysts address each task separately, requiring an extensive collection of methods as building blocks to create a comprehensive model for load profile analysis. This work proposes projecting load profiles into a simplified lower-dimensional structure—a sphere. The spherical representation of the data is based on the computationally efficient PCA method, which has the property of consistent projections between different load profile datasets. This property is studied and confirmed with data from four municipalities in the Netherlands. With the spherical representation, we can define simple outlier detection models in the spherical coordinate system. Moreover, the results obtained from unsupervised clustering algorithms, designed to characterise different consumption patterns such as residential, commercial, and mixed zones, can be readily visualised using this spherical representation. We can also create simple generative models based on probability distributions from directional statistics. This spherical model allows us to unify the four previously mentioned topics into one model, offering new research directions under one umbrella.

G. Contributions

The contributions of this article are as follows:

- We show that standardised daily MV load profiles lie in a hypersphere. Moreover, we showed that applying a PCA-based dimensionality reduction to such standardised profiles results in a sphere projection.
- We exploit this spherical model to unify the tasks of clustering, visualisation, anomaly detection, and generative modelling into one concept. This representation allows the development of outlier detection models in the spherical coordinate system, e.g., anomalous patterns and sensors with faulty readings.
- We introduce the concept of *order of load profiles* using principal curves, which can be used to perform clustering using a one-dimensional parameter. Based on this ordering concept, a generative model is developed for MV load profiles exploiting the spherical properties of the model, allowing us to control the sampling between clusters in a continuous form.

Additionally, it should be noted that the goal of this article is not to present a model that outperforms all the techniques for each task at the same time, i.e., clustering, anomaly detection, visualisation, and generative modelling, but rather to offer an alternative approach to load profile modelling that can be simpler and practical when dealing with all the tasks at the same time. Hence, this article focuses on establishing the theoretical and practical foundation of the spherical representation and demonstrating its direct application to MV load profiles.

II. Lower Dimensional Space Modelling

This section aims to create a lower-dimensional spherical representation of the dataset to identify outliers and discuss the mathematical properties of the DR technique.

The daily load profiles in a large area can be collected in the form of a matrix $\boldsymbol{P} \in \mathrm{R}^{M\times D}$ described as

$$\boldsymbol{P} = \left[\boldsymbol{p}_1, \ldots, \boldsymbol{p}_M\right], \tag{1}$$

where each row-vector $\boldsymbol{p}$ are the readings for each transformer (M), and columns represent the power value of discretised time step (D), e.g., for 15min resolution daily profile $D = 96$. The idea of DR is to represent such a data set in a decreased N-number of variables, i.e., $N \ll D$,

A. Profile Standardisation and Dimensionality Reduction

Standardisation of electricity profiles is a common practise for applications concerned with the load profile's shape rather than the absolute values in consumption. The profile shape depicts the type of consumption in the serviced areas by the MV distribution transformer.

Standardisation of daily MV load profiles is common for applications concerned with the shape of the load profile rather than the absolute values in consumption. The shape depicts the type of consumption in the MV transformer's serviced areas. Normalisation scales the profiles with different active power values into a common range to group profiles with similar shapes.

An essential observation is that standardisation of load profiles brings spherical properties to the dataset in $\boldsymbol{P}$. When we standardise each profile of $\boldsymbol{P}$ using

$$\hat{p}_{i,j} = \frac{p_i - \bar{\mu}_i}{\sigma_i} \quad \forall \quad i = \{1, \ldots, M\},\ j = \{1, \ldots, D\} \tag{2}$$

where

$$\bar{\mu}_i = \frac{\sum_{j=1}^{D} p_{i,j}}{D} \quad \text{and} \quad \sigma_i = \sqrt{\frac{\sum_{j=1}^{D} \left(p_{i,j} - \bar{\mu}_i\right)^2}{D}}, \tag{3}$$

Authorized licensed use limited to: TU Delft Library. Downloaded on October 30,2025 at 11:36:28 UTC from IEEE Xplore. Restrictions apply.

a standardised matrix $\hat{\boldsymbol{P}}$ is created. It is critical to notice that the distance of the points of $\hat{\boldsymbol{P}}$ with respect to the origin is also constant because

$$||\hat{\boldsymbol{p}}_i||^2 = \hat{\boldsymbol{p}}_i \hat{\boldsymbol{p}}_i^\intercal \tag{4}$$

$$= \hat{p}_{i,1}^2 + \hat{p}_{i,2}^2 + \ldots + \hat{p}_{i,D}^2 \tag{5}$$

using (2) and (3),

$$||\hat{\boldsymbol{p}}_i||^2 = \left[\frac{\sum_{j=1}^{D} (p_{i,j} - \bar{\mu}_i)}{\sum_{j=1}^{D} (p_{i,j} - \bar{\mu}_i)/D} \right] = D, \tag{6}$$

$$||\hat{\boldsymbol{p}}_i|| = \sqrt{D}. \tag{7}$$

This means that when the dataset profiles $\boldsymbol{P}$ are standardised by rows, the rows have a constant sum and a constant sum of squares. The constant sum of squares means that the sample points in $\hat{\boldsymbol{P}}$ lie on a hypersphere. For simplicity and to keep $||\hat{\boldsymbol{p}}_i|| = 1$, we normalize the standardised data set as

$$\boldsymbol{X} = \frac{1}{\sqrt{D}} \hat{\boldsymbol{P}}. \tag{8}$$

From now on, we assume that the data set is centered *column-wise*, meaning that the mean of each time step across all load profiles is subtracted from the corresponding column values, i.e., $\boldsymbol{X} \leftarrow \mathrm{C}_M \boldsymbol{X}$, where C_M is the centering matrix for the M samples. The centering process effectively translates the hypersphere so that its new center aligns with the mean of the columns, without changing the overall geometric structure.

Principal component analysis is a linear transformation technique that transforms the data into a new coordinate system so that the variance of the projected data is maximised. This is achieved when the linear transformation comprises the eigenvectors from the covariance matrix of $\boldsymbol{X}$ [36]. Such eigenvectors can be computed using the spectral decomposition of the covariance matrix $\boldsymbol{C} \in \mathbb{R}^{D \times D}$ as

$$\boldsymbol{C} = \frac{1}{M-1} \boldsymbol{X}^\intercal \boldsymbol{X} = \frac{1}{M-1} \boldsymbol{V} \boldsymbol{\Lambda} \boldsymbol{V}^\intercal, \tag{9}$$

where $\boldsymbol{V}$ is the matrix with the orthonormal eigenvectors (column vectors $\boldsymbol{v}_i \in \mathrm{R}^D$) and $\boldsymbol{\Lambda}$ is a diagonal matrix with the eigenvalues of S, i.e., $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_D)$. The eigenvalues and eigenvectors pairs are sorted so that $\lambda_1 > \lambda_2 > \ldots > \lambda_D$. This sorting specifies each eigenvector's importance order, creating the subspace aligned with the largest data variance. The projection to the new coordinates created by the eigenvectors is given by

$$\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{V}. \tag{10}$$

The new projected values, $\boldsymbol{Z}$, can be seen as the weights for each eigenvector necessary to recreate $\boldsymbol{X}$. A convenient way to make this clearer for load profile analysis is writing $\boldsymbol{X}$ using the inverse of the transformation of (10) in terms of the weighted sum of eigenvectors as

$$\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{V}^\intercal = \begin{bmatrix} | & | & & | \\ \boldsymbol{z}_1 & \boldsymbol{z}_2 & \cdots & \boldsymbol{z}_D \\ | & | & & | \end{bmatrix} \begin{bmatrix} - \boldsymbol{v}_1 - \\ - \boldsymbol{v}_2 - \\ \vdots \\ - \boldsymbol{v}_D - \end{bmatrix} \tag{11}$$

$$= \boldsymbol{z}_1 \boldsymbol{v}_1^\intercal + \boldsymbol{z}_2 \boldsymbol{v}_2^\intercal + \cdots + \boldsymbol{z}_D \boldsymbol{v}_D^\intercal \tag{12}$$

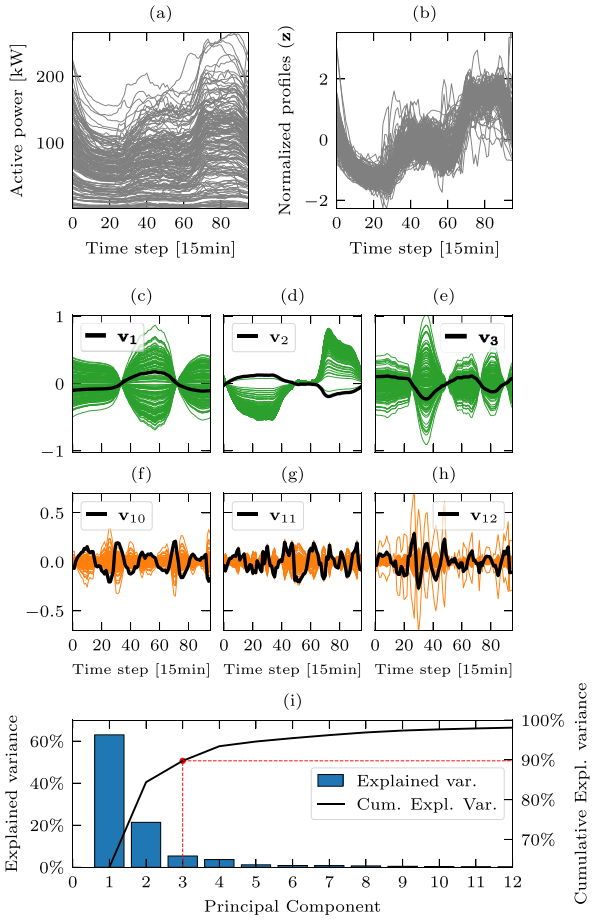

Fig. 1. Decomposition of subset of $\boldsymbol{P}$ in its elementary matrices (13). (a) Original subset $\boldsymbol{P}$. (b) Standardised profiles $\hat{\boldsymbol{P}}$ using (2). (c)-(e) The first three most significant elementary matrix profiles are green. i.e., $\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3$, with their respective eigenvector components in a solid black line. Less significant elementary matrices, i.e., $\boldsymbol{X}_{10}, \boldsymbol{X}_{11}, \boldsymbol{X}_{12}$, are shown in orange, for the eigenvectors (f) $\boldsymbol{v}_{10}$, (g) $\boldsymbol{v}_{11}$, and (h) $\boldsymbol{v}_{12}$. (i) Blue bars show the explained variance by the most important eigenvalues; the solid line is the CEV. The plot is truncated to 12 eigenvalues out of 96.

$$= \boldsymbol{X}_1 + \boldsymbol{X}_2 + \ldots + \boldsymbol{X}_D \tag{13}$$

where matrices $\boldsymbol{X}_i \in \mathbb{R}^{M \times D}$ $\forall\, i = \{1, \ldots, D\}$ are elementary matrices. The elementary matrices can be interpreted as the deconstruction of the profiles $\boldsymbol{X}$ in basic profiles $\boldsymbol{x}_{i,(j,*)} \in \mathbb{R}^D$ $\forall\, j = \{1, \ldots, M\}$, which are the row-vector profiles of each elementary matrices. For instance, the subset of residential profiles of dataset $\boldsymbol{P}$ shown in Fig. 1(a), has the standardised profiles $\hat{\boldsymbol{P}}$ which shows a clearer residential pattern in Fig. 1(b). After applying PCA on $\hat{\boldsymbol{P}}$, the three most important eigenvectors (higher eigenvalues) have the elementary matrix profiles plotted in green, and their respective eigenvectors in a solid black line shown in subplots Fig 1(c)-(e). Three less significant eigenvectors are shown in Fig. 1(f)-(h). We observe that the most important eigenvectors have a lower frequency, carrying the largest variance and capturing the main shape structure of the profiles. On the other hand, the least important eigenvectors capture the higher frequency component of the load profile, resembling noise-like behaviour. We also see that in the case of load profiles, such

Authorized licensed use limited to: TU Delft Library. Downloaded on October 30,2025 at 11:36:28 UTC from IEEE Xplore. Restrictions apply.

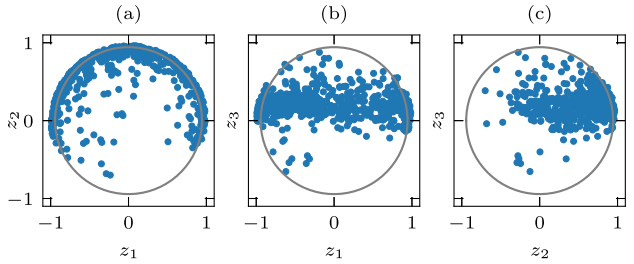

Fig. 2. The values of the projection $\boldsymbol{Z}$ in a 3-dimensional space. (a-c) Orthographic projection of the sphere. Each blue point represents a single transformer's daily profile. The sphere overlayed in the data is found via (15).

eigenvectors can be seen as *eigenprofiles*, which are the basic signals needed to rebuild the dataset again.

To determine the number of reduced variables, i.e., the number of *eigenprofiles* to reduce the dataset's dimensionality, we look at the cumulative explained variance ratio (CEV), which is expressed as

$$\text{CEV}(\lambda_n) = \frac{\sum_{j=1}^{n} \lambda_j}{\sum_{k=1}^{D} \lambda_k}. \tag{14}$$

Selecting the N number of major eigenvalues (λ) means we re-construct the data with the N principal elementary matrices. The bar plot in Fig. 1(i) shows the proportion of total variance explained by each eigenvalue, and the solid line is the CEV. It is then shown that 90% of the data can be explained by the first three *eigenprofiles*. From here, the reduced form for the eigenvector or eigenvalue matrices will have a *hat* notation, e.g., matrix $\boldsymbol{V}$ that uses only the three principal *eigenprofiles* is referred to as $\bar{\boldsymbol{V}} \in \mathbb{R}^{D\times 3}$.

From the analysis of (4)-(7), we know that the data set $\boldsymbol{X}$ lies in a hypersphere. Now, if we assume that the data points are distributed along different orthants in the hypershpere, the linear projection of the PCA in (10) using three principal eigenvectors ($\bar{\boldsymbol{V}}$) results at least in an ellipsoid-shaped projection. For simplicity, we will now assume that the projected data in the three-dimensional space have a spherical-shaped structure instead of an ellipsoid. This assumption allows us to apply mathematical tools from circular statistics to perform outlier analysis and generative modelling later in Section IV.

To visually confirm the spherical-shaped structure of the projected data, Fig. 2 shows the results of the PCA using three principal components for daily profiles from 560 MV distribution transformers from a municipality in the Netherlands. Each blue point in three dimensions corresponds to a *lower-dimensional* representation of the daily load profile of 96 dimensions (15-minute resolution daily profile).

Interestingly, most projected data points are agglomerated on one side of the sphere revealing a lower dimensional latent distribution, which forms an arc-shaped structure, clearly seen in Fig. 2(a). Some points lie outside of the latent distribution, meaning they represent a weighted sum of *eigenprofiles* that are not usual from the dataset, creating an outlier-type of profiles. The interest is in labelling those outliers to study the source for such anomalous behaviour.

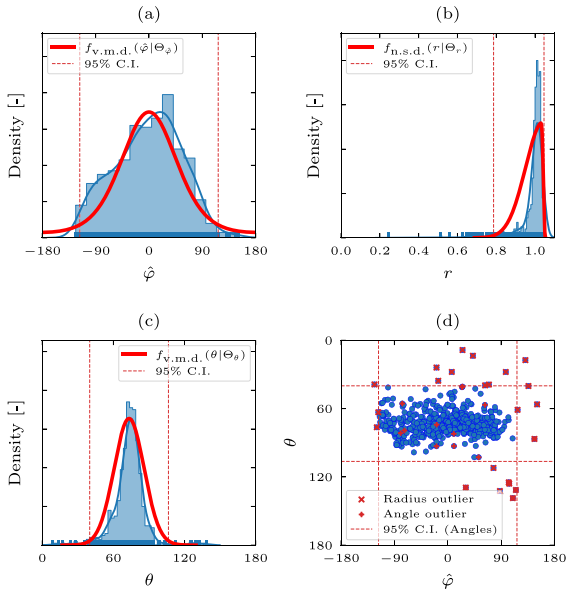

Fig. 3. Probability distributions of the spherical projection variables for the dataset $\boldsymbol{X}$. (a) Azimuthal angle distribution centred around the mean. (b) Radius. (c) Polar angle. (d) 2D projection using angle values. Flagged points in red are outliers, which are the points that fall outside the rejection region created by the 95% confidence interval (CI) from each fitted distribution (delineated by the vertical and horizontal red dotted lines).

III. Sphere Modelling and Outlier Detection

After the previous findings, the natural step is to create a sphere as a mathematical model that helps us visualise and discern whether a profile is an outlier. The sphere parameters are found by solving the following optimisation problem

$$\underset{\boldsymbol{c},\rho}{c\min} \quad \sum_{j=1}^{3}\sum_{i=1}^{M}\left(z_{i,j} - c_j\right)^2 - \rho, \tag{15}$$

using the change of variable $\rho = r^2$ for the radius, and vector $\boldsymbol{c} = [c_1, \ldots, c_3]$ as the sphere's centroid. The solution of the sphere parameters allows us to centre the dataset around the origin. i.e., $\boldsymbol{z} \leftarrow \boldsymbol{z} - \boldsymbol{c}$, to apply a change of coordinate system to build models for outlier detection. We convert centred points to the spherical coordinate system using

$$\varphi_i = \text{sgn}(\hat{z}_{2,i}) \arccos\left(\hat{z}_{1,i}/\sqrt{\hat{z}_{1,i}^2 + \hat{z}_{2,i}^2}\right), \tag{16}$$

$$\theta_i = \arccos\left(\hat{z}_{3,i}/\sqrt{\hat{z}_{1,i}^2 + \hat{z}_{2,i}^2 + \hat{z}_{3,i}^2}\right), \tag{17}$$

$$r_i = \sqrt{\hat{z}_{1,i}^2 + \hat{z}_{2,i}^2 + \hat{z}_{3,i}^2}, \qquad \forall\, i = \{1, \ldots, N\}. \tag{18}$$

The azimuthal angle φ_i has the positive x-axis as reference (z_1), and the polar angle θ_i is measured from the positive z-axis (z_3). The coordinate system change makes it easier to fit parametric distributions for each variable as shown in Fig. 3, where the angle φ is centered around its mean, $\hat{\varphi} = \varphi - \bar{\varphi}$, to aid the visualisation. For the angles, we fit a von Mises distribution (VMD) as a continuous probability distribution used to describe angles in circular statistics as shown in Fig. 3 (a), and (c). The radius variable is modelled with a skew-normal distribution (SND) as the data is concentrated on a thin shell of the sphere, creating a skewed distribution

Authorized licensed use limited to: TU Delft Library. Downloaded on October 30,2025 at 11:36:28 UTC from IEEE Xplore. Restrictions apply.

(Fig. 3 (b)). The parameters for each distribution, i.e., $\Theta_{\hat{\varphi}}$, Θ_{θ}, Θ_{r}, are found via maximum likelihood estimation.

The outlier profiles are deduced using the rejection region created from the 95% confidence interval for each fitted distribution shown as vertical dotted red lines in Fig. 3. The flagged data points are discriminated between outliers from the angle models (φ_i and θ_i) and radius (r) as shown in Fig. 3(d). Details about the characteristics of those outliers are discussed in the case study in Section V.

IV. Ordering and Generative Modelling

This section dives into the load profiles' lower dimensional structure. It discusses the arc-shape pattern in the sphere shown in Fig. 2(a), suggesting that the load profiles of a large area have a latent ordering, which can be used for clustering and generative modelling purposes.

A. Ordering by Principal Coordinate Analysis

The proximity of the points in the sphere on Fig. 2 implies a great degree of similarity between the load profiles. Meaning that the profiles are constructed with similar weights (z-values) combinations of *eigenprofiles*. Such similarity between load profiles can be quantified using the dot product between them, creating a similarity matrix of the form

$$\boldsymbol{S} = \boldsymbol{X}\boldsymbol{X}^{\intercal} = \begin{bmatrix} \cos\left(\gamma_{1,1}\right) & \dots & \cos\left(\gamma_{1,M}\right) \\ \vdots & \ddots & \vdots \\ \cos\left(\gamma_{M,1}\right) & \dots & \cos\left(\gamma_{M,M}\right) \end{bmatrix}, \tag{19}$$

this matrix is also called the Gram matrix. The dot product between vectors (profiles) of $\boldsymbol{X}$ is determined by $\boldsymbol{x}_i\boldsymbol{x}_j^{\intercal} = ||\boldsymbol{x}_i|| \cdot ||\boldsymbol{x}_j|| \cos(\gamma_{i,j})$, and knowing from (8) that the profiles are normalised vectors, i.e., $||\boldsymbol{x}|| = 1$, the similarity matrix is reduced to the cosines of the angles between the vectors, ranging between $[-1, 1]$. The dimensionality of the similarity matrix $S \in \mathbb{R}^{M\times M}$ is high for large samples ($M \gg D$). Therefore, it is required to apply a DR technique to represent $\boldsymbol{S}$ in a low-dimensional space. Reducing similarity matrices into a lower dimensional projection is common in biology, archaeology, and psychology using multidimensional scaling techniques (MDS).

Principal Coordinate Analysis (PCoA), or classical MDS, is when the PCA technique is applied to the double-centred Gram matrix, i.e., $\boldsymbol{G} = \mathrm{C}_M\boldsymbol{S}\mathrm{C}_M$, which has a spectral decomposition $\boldsymbol{G} = \boldsymbol{\xi}\Lambda\boldsymbol{\xi}^{\intercal}$. The objective of PCoA is to preserve the similarity structure (distances between points) in a lower dimensional space.

When the matrix $\boldsymbol{S}$ has a *band matrix* structure, the PCoA projection has an interesting arc-shaped effect pattern [37], like the one shown in Fig. 2(a) for the load profiles. A *band matrix* in our context, means the highest similarity values are closer to the diagonal. The arc effect has been studied and discussed in other study fields in the past for different types of similarity matrices [38], and is often seen when the sample set $\boldsymbol{X}$ has an underlying latent ordering [39]; or the data comprises a gradual rate of change of a process [40]. The parameterisation of these lower dimensional arcs or paths can be used to recover the underlying ordering from the dataset. In our work, knowing the load profile order helps us to determine threshold values to determine clusters for different consumption types and create generative models in the lower dimensional space.

Illustrative Example for PCoA: As an illustrative example of how PCoA is used, consider the following non-linear function that gradually changes for each sample:

$$h_i(t) = \tau(i)g(t, 1) \ + (1 - \tau(i))g(t, \nu(i)) + 0.03\ \varepsilon \tag{20}$$

$$g(t, \mu) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(t - \mu)\right) \tag{20a}$$

$$\nu(i) = 0.08i - 4 \tag{20b}$$

$$\tau(i) = 0.02i - 1 \qquad \forall\, i = \{1, \dots, 100\}, \tag{20c}$$

$$\varepsilon \sim \mathcal{N}(0, 1) \tag{20d}$$

variable i is the sample index, and each sample profile $h_i(t)$ is discretized by 20 steps, creating a row-vector $\boldsymbol{h}_i \in \mathbb{R}^{20}$. The row-vectors builds the data matrix $\boldsymbol{H} = [\boldsymbol{h}_1, \dots, \boldsymbol{h}_{100}] \in \mathbb{R}^{100\times 20}$. The samples $\boldsymbol{h}_i$ are labelled and coloured in the increasing sample index order shown in Fig. 4(a1). The gradual change can be seen as a heatmap in Fig. 4(a2) for the standardised row-vectors applying (2), i.e., $\boldsymbol{H} \xrightarrow{\text{std}} \hat{\boldsymbol{Y}}$.

Using (8), i.e., $\hat{\boldsymbol{Y}} \xrightarrow{\text{norm}} \boldsymbol{Y}$ and computing the similarity matrix, $\boldsymbol{S_Y} = \boldsymbol{Y}\boldsymbol{Y}^{\intercal}$, a clear band matrix structure is appreciated in Fig. 4(a3). The PCoA over the double centered Gram matrix $\boldsymbol{G_Y}$, and using the three principal eigenvectors, i.e., $\boldsymbol{W} = \bar{\boldsymbol{\Lambda}}^{\frac{1}{2}}\bar{\boldsymbol{\xi}}$, is shown in Fig. 4(a4),(a5) with a spherical orthographic projection. The projections show a clear arc-shaped latent ordering, i.e., the colours of the sample index follow the arc in a clockwise motion in Fig. 4(a4).

The second column of Fig. 4 shows the same procedure, only that this time the sample index is shuffled, i.e., the row-vector of $\boldsymbol{H}$ are shuffled (Fig. 4)(b1). This is more likely the real case scenario for electricity load profiles, where we do not know the correct order beforehand of similarity between areas serviced by MV transformers. Notice that without order, there is no discernible pattern change in Fig. 4(b2), nor a band-matrix structure as shown in Fig. 4(b3). However, the latent arc shape remains unchanged as shown in Fig. 4(b4). To recover the ordering, we create a parametrised curve that passes through the middle of the lower dimensional points. Then, we re-label the samples following the projection of the points to the curve.

The reason for choosing the dot product as the similarity matrix over other types of distance matrix options is because reducing (19) and (8) produce the same lower dimensional projection [41]. Nevertheless, it is helpful to compute the matrix $\boldsymbol{S}$ after the sample ordering procedure to visualise and verify the existence of a band matrix structure.

In the case of load profiles data, the arc-shape pattern can be seen in Fig. 2(b), indicating a latent ordering of the dataset. The following subsection discusses the procedure for finding the parametrised curve.

B. Principal Curve Model

The principal curve is a technique to summarise complex N-dimensional distributions into one dimension [42]. The

Authorized licensed use limited to: TU Delft Library. Downloaded on October 30,2025 at 11:36:28 UTC from IEEE Xplore. Restrictions apply.

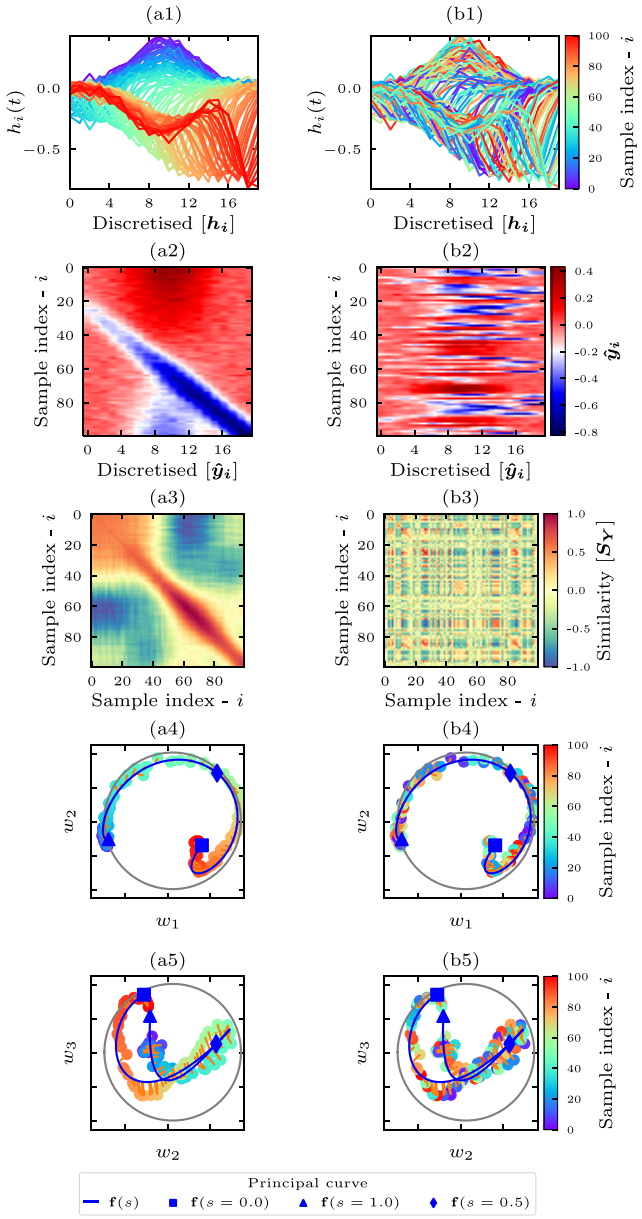

Fig. 4. Example of latent space ordering for the process in (20). (a1) Data matrix $\boldsymbol{H}$ created by discretisation of (20) using 20 time steps. (a2) Heatmap for standardised matrix $\hat{\boldsymbol{Y}}$. (a3) Heatmap similarity matrix $\boldsymbol{S_Y}$. (a4) and (a5) spherical orthographic projection for the PCoA applied to $\boldsymbol{S_Y}$ using three principal components which shows a clear arc-shape structure. (b1) $\boldsymbol{H}$ with shuffled rows. (b2) The gradual change pattern is lost due to shuffling. (b3) The banded structure of $\boldsymbol{S_Y}$ is absent. (b4) and (b5) The latent ordering of the samples still exists, and the original sample labels (i) can be recovered following the parametrised curve that passes through the middle of the points.

technique consists in creating a parametric curve $\boldsymbol{f}(s)$ that passes through the middle of the data points in a smooth way. Here, we use this concept to find the latent ordering for the load profiles. The curve $\boldsymbol{f}(s)$ comprises N functions with a single variable s. Precisely is defined as

$$\boldsymbol{f}(s) = \left[f_1(s), \ldots, f_N(s)\right]^{\mathsf{T}}, \quad s \in [0, 1]. \tag{21}$$

Each function $f_k(s) \mapsto \mathbb{R} \;\; \forall k = \{1, \ldots, N\}$ can be parametrised by nearest-neighbours, kernel or spline smoothers. In this study, we used splines due to the smooth spherical surface, and the number of functions is three (N=3). i.e., one function per each cartesian coordinate. A general procedure to find $\boldsymbol{f}(s)$ is described in [42]. However, a more robust version is developed in [43], where the initial iteration exploits the fact that most data reside in a shell in a sphere, increasing the correct convergence likelihood. The iterative algorithm's objective is to find the spline parameters that minimise the orthogonal projections of the data into the curve. The projection index is defined as

$$\lambda_{\boldsymbol{f}(\cdot)}(\boldsymbol{z}) = \sup_s \left\{ s : ||\boldsymbol{z} - \boldsymbol{f}(s)|| = \inf_{\mu} ||\boldsymbol{z} - \boldsymbol{f}(\mu)|| \right\} \tag{22}$$

The projection index $\lambda(\cdot)$ of the PCA point $\boldsymbol{z_i}$ is the s_i value for which $\boldsymbol{f}(s_i)$ is closest to $\boldsymbol{z_i}$, i.e., $\lambda(\boldsymbol{z_i}) \mapsto s_i$. This means that each point in the sphere, representing a profile, has a corresponding s-value between 0 and 1 for the parametrised curve. For instance, the curve fitted for the points in Fig. 4(a4,b4,a5,b5) shown as a solid blue line, is used to sort the projected values of $\boldsymbol{w}$ on the curve. Once the $\boldsymbol{w}$ values are sorted, we recovered the original ordering of the process in (20).

Using a single-valued variable s to represent profiles can aid in creating data clusters by dividing it into groups or bins, with each bin representing a cluster.

C. Generative Modelling

The benefit of having the principal curve is combining it with a normal probability distribution to create a generative model. In the field of directional statistics, the generalisation of the normal probability distribution for a $(N-1)$-sphere is the von-Mises-Fisher (VMF) distribution. For the specific case of a three-dimensional sphere, the VMF distribution is described as

$$\boldsymbol{f}_{\text{vmf}}(\boldsymbol{z}; \boldsymbol{\mu}_{\text{vmf}}, \kappa) = \frac{\kappa}{2\pi\left(e^{\kappa} - e^{-\kappa}\right)} \exp\left(\kappa \boldsymbol{\mu}_{\text{vmf}}^{\mathsf{T}} \boldsymbol{z}\right), \tag{23}$$

where κ is the concentration parameter and controls the spread of the samples over the sphere. Parameter $\boldsymbol{\mu}_{\text{vmf}} \in \mathbb{R}^3$ is the directional vector where the normal distribution is centred. Here, we use $\boldsymbol{f}(s) \equiv \boldsymbol{\mu}_{\text{vmf}}$, to dynamically and continuously move the VMF distribution over the sphere creating data from the desired section. The synthetic standardised shape profiles $\bar{\boldsymbol{P}}$ can be obtained by sampling from (23) and using the inverse process described in the previous sections as

$$\bar{\boldsymbol{Z}} \sim \boldsymbol{f}_{\text{vmf}}(\cdot; \kappa, \boldsymbol{f}(s)) \mapsto \bar{\boldsymbol{X}} = \bar{\boldsymbol{Z}}\bar{\boldsymbol{V}}^{\mathsf{T}} \mapsto \bar{\boldsymbol{P}} = \sqrt{D}\bar{\boldsymbol{X}} \tag{24}$$

V. Case Study

This section discusses (i) The meter readings labelled as outliers in Section III and the visualisation of load profile clustering results on the sphere. (ii) The ordering of load profiles created by the principal curve techniques. (iii) Generative modelling benefits and limitations, and (iv) generalisation of the spherical modelling for other MV load profile datasets. Results presented in this Section use the profile dataset introduced earlier (560 MV load profiles) to keep consistency between the results presented in the previous sections.

Authorized licensed use limited to: TU Delft Library. Downloaded on October 30,2025 at 11:36:28 UTC from IEEE Xplore. Restrictions apply.

The implementation of the spherical representation of load profiles is available as a Python package called `spheral` [44]. Similarly, the principal curve method described in Section IV-B is implemented in the `procurve` Python package [43], which also includes a pseudo-code description. Additionally, an interactive version for the spherical visualisation of the datasets discussed in this Section V-E can be viewed at [45].

A. *Profile Clustering and Anomalous Meter Readings*

The outlier meter readings labelled by the rejection region from the probability distributions presented in Section III are shown with x markers in Fig. 5(a-c). It has been found that from the three spherical variables, the radius is the most effective descriptor of the defective meters, and its labelled profiles are shown in Figure 5(h,i,k), where subplot (i) shows 5 profiles for meters that are recording noise.

A more challenging failure to detect is shown in Figure 5(h), where 6 meters are only storing absolute values of active power, meaning that the power injection from PV panels was recorded as consumption and not generation, explaining the increased consumption bump seen in daylight hours. This failure is challenging because outlier detection techniques that rely on past data, like smoothing kernels [20], [21], can not detect similar and repetitive failure readings. Nevertheless, with the spherical modelling, those profiles stand out from the latent distribution of the complete load profile set, making its identification easier. The profiles shown in Fig. 5(k) are two large areas with midnight operations. These profiles are in fact, outliers, as their unique shape are only 2 out of 560, but do not correspond to meter failure.

Profiles labelled as outliers by the azimuth and polar angle distribution models are grouped in Fig. 5(j), which are the points scattered at the top of the sphere (purple circle markers). These profiles have a pronounced high load consumption early in the morning, and PV generation creates a valley in the middle of the profile. This particular shape structure requires higher weight (z_3) for the third *eigenprofile* shown in Fig. 8(c) (Municipality 1), to recreate the morning load spike consumption. This group of 14 profiles can be considered a cluster independently and are not defective meter readings.

We excluded the data points associated with defective meters on the sphere and employed various clustering techniques, namely k-means, Gaussian mixtures, Hierarchical, and Spectral clustering. It should be noted that none of the clustering techniques could detect the profiles in Fig. 5(k) as a unique cluster, as they incorporated those points into different groups. Therefore, we isolated those readings from the clean profiles set to repeat the clustering for a total of 533.

The coloured visualisation for the Agglomerative Hierarchical clustering (AggHC) results with Ward distance is shown in Fig. 5(a-c). Four representative groups are found and correspond to different consumption activity zones. From Fig. 5 subplot (d) is C1: Commercial/Offices, (e) C2: Mix of residential and commercial, (f) C3: Residential, and (g) C4: Residential with high PV penetration. The spherical visualisation of the clusters offers a different view of the proximity between clusters in the lower dimensional domain. The clear case is for the cluster with a mixture of residential and commercial zones (C2). This cluster lies between commercial and residential areas clusters, indicating that C2 is a transition section on the sphere between the adjacent clusters. Likewise, C3 is viewed as the transition region between C4 and C2.

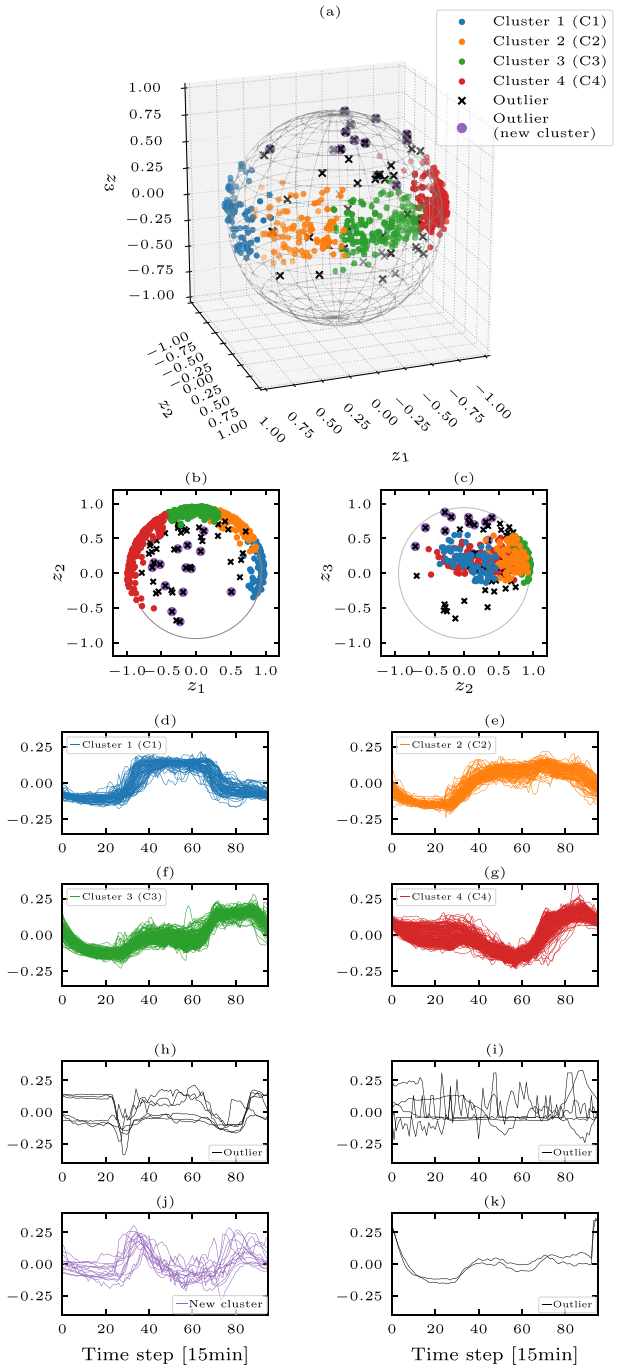

Fig. 5. Clustered profiles from Municipality 1 and outlier identification. (a-c) Visualisation for clustering results and outliers using spherical modelling. Daily Profiles that are not considered outliers are grouped into four: (d) Commercial/Offices (e) Mix of residential and commercial (f) Residential (g) Residential with high PV penetration. (h-k) Anomalous reading labelled by outlier models from Section III. (j) The scatter points on the top of the sphere marked as outliers are a cluster of their own.

Authorized licensed use limited to: TU Delft Library. Downloaded on October 30,2025 at 11:36:28 UTC from IEEE Xplore. Restrictions apply.

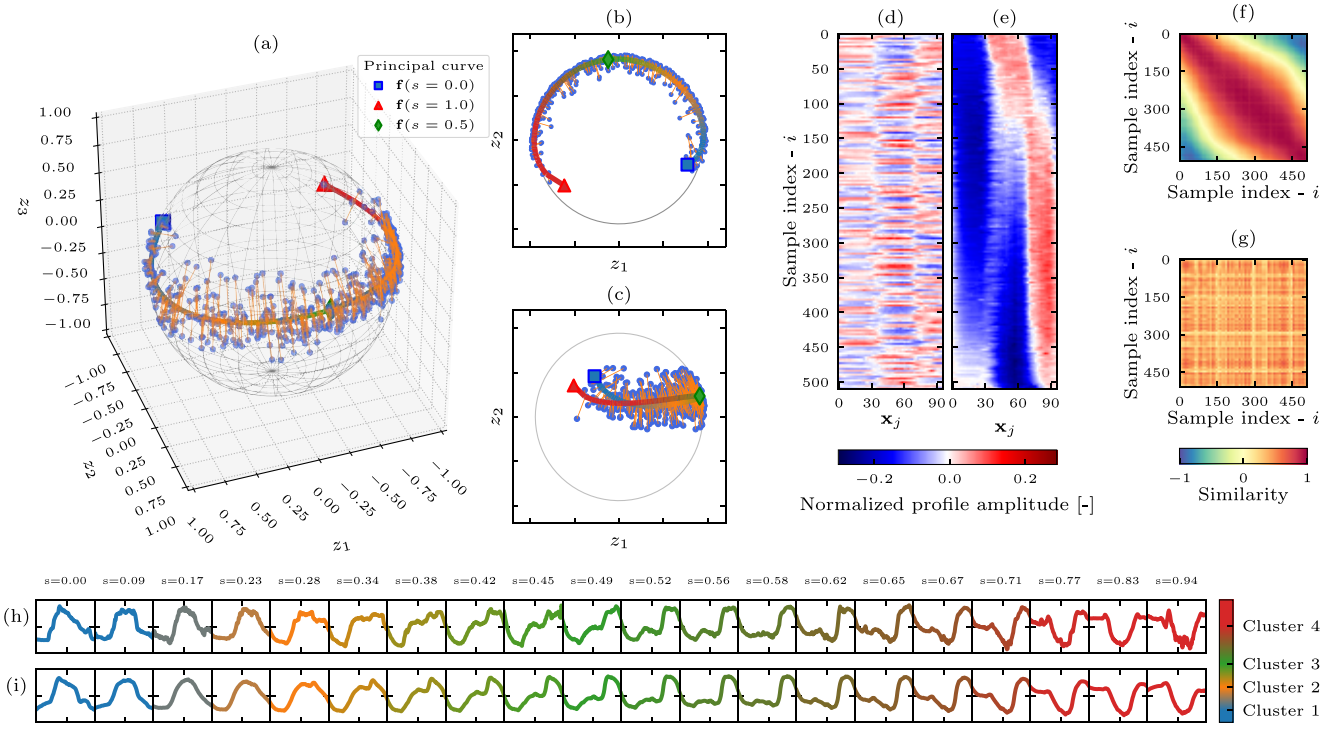

Fig. 6. Ordered dataset for Municipality 1. (a-c) The principal curve model passes through the middle of the data points representing the load profiles. It shows a latent ordering due to the compactness of the cloud of points. Orange lines are the orthogonal projection of the points to the principal curve. (d) Heat map of the original load profile data set matrix $\boldsymbol{P}$. (e) The ordered matrix $\boldsymbol{P}$ using the principal curve uncovers the latent ordering of the dataset, confirmed in (f) by the banded similarity matrix. (g) The disorganised original dataset $\boldsymbol{P}$ produces a non-structured similarity matrix. (h) Gradual change between clusters is exposed by plotting twenty original profiles ordered by the principal curve. The change goes from commercial areas (blue) to residential areas with high PV penetration (red). (i) The same twenty profiles are recovered using its orthogonal projection to the principal curve.

B. Latent Ordering by Principal Curve

The principal curve fitted in the cleaned dataset after clustering is shown in Fig. 6(a-c). The orange lines represent the orthogonal projections of each data point into the principal curve, i.e., $\lambda(\boldsymbol{z}_i) \mapsto s_i \mapsto \boldsymbol{f}(s_i)$. Figure 6(h) shows 20 profiles from the matrix (8) ordered by s_i values, which gives a clear view of how the profiles gradually change between clusters, confirming a latent ordering for the profiles set. The projected points in the curve can also be transformed back into standardised values, i.e., $\bar{\boldsymbol{P}} = \sqrt{D}\boldsymbol{f}(s_i)\bar{\boldsymbol{V}}^{\mathsf{T}}$, and is shown in Figure 6(i). These transformed profiles have a smoother shape than the original profiles because the reconstruction is done using the three principal *eigenprofiles*, and the high-frequency components, characterised by the least important *eigenprofiles*, are not used as discussed in Section II-A.

The heat map in Fig. 6(d) shows the matrix (1) as it was initially collected from the AMI database, without any particular order, and its similarity matrix (19) is plotted in Fig. 6(g). Both heat maps do not show any structure. After re-ordering (1) with the principal curve, Fig. 6(e) depicts a gradual change in the profiles. Also, the similarity matrix of the ordered dataset, in Fig. 6(f), shows a banded structure, explaining the arc-shape pattern discussed in Section IV-A.

The s_i is a single-valued descriptor for each ordered profile. In our case, $s = 0.0$ corresponds to the first sample, $\boldsymbol{f}(s = 0.0)$ in a blue square marker in Fig. 6(a); and $s = 1.0$ for the last sample, $\boldsymbol{f}(s = 1.0)$ in a red triangle marker in the same subplot. The principal curve can be split into segments to create hard clusters over the dataset. Each segment can be seen as bins to group the profiles. In this example, the bins range that concur with AggHC results are C1 $\in [0.0, 0.2)$, C2 $\in [0.2, 0.4)$, C3 $\in [0.4, 0.6)$ and C4 $\in [0.6, 1.0]$.

The ordering with the principal curve brings two significant advantages compared to other clustering techniques: (i) The profiles within each cluster are ordered, which gives the DNO additional information about ranked areas within the city based on their similarities. Consequently, this data offers detailed insights into the synchronisation of peak and valley consumption times in these areas. (ii) Principal curve conveys a sense of continuity between clusters, giving extra information to quantify mixtures of areas. For instance, the centroid of C1, i.e., $s_{c1} = 0.1$, is the representative value for commercial areas, and C3 centroid, i.e., $s_{c3} = 0.5$ for the residential areas. The profiles that lie in the curve segment created by $[s_{c1}, s_{c3}]$ can be treated as a mixture between both activity areas; for example, $s = 0.22$ would be a profile for an area with 30% residential and 70% commercial consumption.

C. Generative Modelling

The principal curve acts as the probability distribution centre ($\boldsymbol{\mu}_{\text{vmf}}^{\mathsf{T}}$) in the model (23), allowing the generation of points in different sections on the sphere depending on the variable s. The spherical generative model, labelled as (VMF), is tested by creating (sampling) 10 profiles using (24) for 25 points over the principal curve. i.e., $\bar{\boldsymbol{Z}}_k \sim \boldsymbol{f}_{\text{vmf}}(\cdot; \kappa, \boldsymbol{f}(s_k))$, where $s_k = s_{k-1} + \Delta s$, $\Delta s = 1/25$, $s_0 = 0.0$, $\forall\ k = \{1, \ldots, 25\}$.

Authorized licensed use limited to: TU Delft Library. Downloaded on October 30,2025 at 11:36:28 UTC from IEEE Xplore. Restrictions apply.

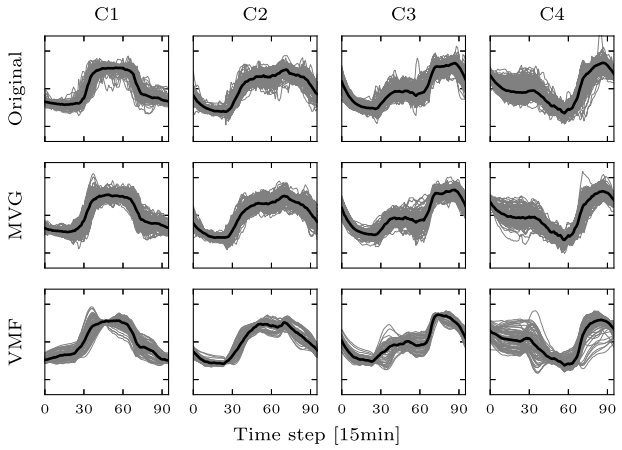

Fig. 7. Comparison between original and synthetic profiles generated by spherical modelling (23), labelled as VMF and MVG. The solid black line represents the median for the profiles in each cluster.

TABLE I
GENERATED LOAD PROFILES COMPARISON METRICS FOR EACH CLUSTER

Metric	Model	C1	C2	C3	C4
KS	VMF	.054	.032	.031	.034
	MVG	.033	.022	.020	.023
RMSE	VMF	.022	.019	.018	.020
	MVG	.008	.005	.007	.010

The concentration parameter $\kappa = 7.1$ is taken from the polar angle model parameters (Θ_θ). The total generated profiles are 250, and using their projected s-values into the principal curve, they are clustered using the bin values from Section V-B.

For comparison purposes, we generated the same number of profiles per cluster using a multivariate Gaussian distribution (MVG), shown in [33] as a suitable model to generate MV load profiles. The synthetic profiles from the models are plotted in Fig. 7 and compared against the original dataset. It is observed that the profiles generated by VMF are smoother than the original profiles due to the limited reconstruction of the profile with three components. This could limit the VMF application for analyses that require high accuracy for active power values every fifteen minutes. This is expected to be less critical for profiles recorded at lower resolutions, e.g., hourly. However, the main VMF model advantage is that it can control the generation of profiles between clusters, as shown in Fig. 6(i). This is not feasible with any other mixture model technique that typically treats each cluster as a categorical variable (discrete). The profile quality is quantified in Table I by using the Kolgomorov-Smirnov (KS) probability distance metric, and the root mean squared error (RMSE) between the original and the synthetic data, which shows that VMF and MVG can closely model the original profiles.

D. Spherical Modelling for Different Datasets

Most of the results in this article are based on data from one municipality in the Netherlands (Municipality 1). Even though the mathematical formulation of the spherical modelling is robust, the natural question arises about generalising the model for different MV load profile datasets and, specifically, inquiring if other datasets show the same latent distribution (concentrated cloud of points in a sphere). Here, we use the profiles from four different municipalities in the Netherlands. Instead of repeating all figures and analyses on each one, we argue that all datasets have similar properties in the latent space, and spherical modelling can be applied to any of them.

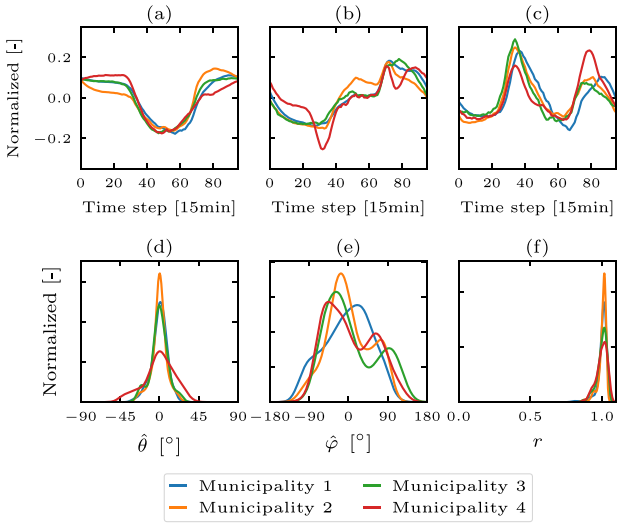

Fig. 8. Data properties for four different municipalities in the Netherlands. (a,b,c) Three most important *eigenprofiles*. Probability distributions for the spherical coordinates variables: (d) Polar and (e) Azimuth angles (centred around the mean), and (f) radius. The datasets show similar *eigenprofiles* and the lower dimensional spherical coordinates characteristics.

TABLE II
CUMULATIVE EXPLAINED VARIANCE (%) OF PRINCIPAL COMPONENTS FOR DIFFERENT MUNICIPALITIES IN THE NETHERLANDS

Municipality	PC1	PC2	PC3	PC4	PC5	Transf. count	Outilers (%)
Municipality 1	68.7	88.4	92.5	95.1	96.0	560	27 (5.0%)
Municipality 2	69.9	88.3	92.1	94.3	95.9	375	9 (2.4%)
Municipality 3	71.8	87.1	91.0	93.4	94.5	354	29 (8.2%)
Municipality 4	61.1	73.0	81.8	85.8	88.2	392	61 (16.9%)

Table II describes the CEV for each municipality, and for all datasets, three components represent most of the variance, validating the use of a three-dimensional projection. Additionally, the three most important *eigenprofiles* are plotted in Fig. 8(a-c), showing a clear similarity pattern between them, meaning that elementary matrices (13) are similar.

The most important observations are in the probability distributions for the spherical coordinates, shown in Fig. 8(d-f) with their respective moments quantified in Table III. The distributions show that most data are concentrated in the sphere's shell. The radius distribution has negative skewness and mean closer to 1.0. Also, the data is concentrated in the polar angle ($\hat{\theta}$), which shows a single-mode distribution. The biggest difference between the distributions lies in the azimuthal angle with a higher standard deviation. This is expected because this angle covers the axes that keep most of the dataset variance, e.g., first and second *eigenprofiles*, where most of the profile clusters are identified.

Authorized licensed use limited to: TU Delft Library. Downloaded on October 30,2025 at 11:36:28 UTC from IEEE Xplore. Restrictions apply.

TABLE III
SPHERICAL MODELLING SUMMARY STATISTICS

Variable	Municipality	Mean	Std	Skewness	Kurtosis
$\hat{\theta}$	Municipality 1	0.000	9.149	-0.512	0.960
	Municipality 2	0.000	8.660	-0.321	1.396
	Municipality 3	0.000	10.039	-0.181	0.615
	Municipality 4	0.000	17.236	-0.449	-0.013
$\hat{\varphi}$	Municipality 1	0.000	52.811	-0.229	-0.787
	Municipality 2	0.000	49.195	0.146	-0.674
	Municipality 3	0.000	58.365	0.543	-0.989
	Municipality 4	0.000	55.126	0.347	-1.092
r	Municipality 1	1.002	0.036	-2.640	9.205
	Municipality 2	1.003	0.029	-2.221	5.773
	Municipality 3	0.999	0.036	-1.203	1.384
	Municipality 4	0.998	0.039	-0.815	0.310

E. Dimensionality Reduction Techniques Comparison

This subsection discusses the advantages and limitations of spherical modelling in comparison with other state-of-the-art dimensionality reduction techniques for daily load profile representation. To provide a consistent and fair comparison, five methods: Spherical modelling (ours), Isomap, UMAP, t-SNE, and Variational Autoencoder (VAE) were applied to datasets from July and December for the same municipality and year. Interactive 3D visualizations of the projections are available at [45]. Each method projects the MV daily load profiles into a three-dimensional space. The results are shown in Fig. 9, where each embedding includes an orthographic view to better visualize the spatial structure of the projection.

To aid the visual comparison of cluster organization across techniques, each daily profile was assigned a cluster label using k-means clustering applied in the original high-dimensional space. These clusters are color-coded consistently across all subplots, following the same color scheme as used in Fig. 5 (d–g). It useful to notice that the red cluster (cluster 4), characterized by high photovoltaic (PV) penetration and a distinct duck-curve shape, is prominent in July but significantly reduced in December due to seasonal declines in solar irradiance.

Figure 9 also highlights two reference profiles to illustrate the interpretability of the projections. A purple circle denotes an anomalous profile attributed to a sensor installation error, specifically, a wiring misconfiguration where one phase was inverted. This leads to an inverted residential load shape. The anomaly persists throughout the year and is consistently visible in both July and December embeddings. In contrast, a black square indicates a typical profile with high PV penetration, representative of cluster 4. It retains its relative position in the red cluster during both months, although its duck-curve shape is more pronounced in July due to higher irradiance.

Among all methods, the spherical projection demonstrates a consistent spatial organization of the clusters between July and December. Points are distributed on the crust of the sphere, reflecting the natural ordering of profiles, explained in Section V-B, and preserving the global structure of the data. This contrasts with UMAP, which produces compact embeddings that distort the proximity of anomalous points. For instance, in Fig. 9 (c,j), the anomalous purple point appears close to normal profiles, potentially reducing its utility for heuristic anomaly detection. UMAP also lacks stability across runs due to random initialization and its focus on preserving local neighborhoods, its embeddings vary in global geometry, complicating their use for modelling tasks for a large number of sensors.

The t-SNE model exhibits similar limitations: although it offers a visually consistent layout between July and December, it does not preserve meaningful distances—outliers often appear embedded within clusters. Additionally, t-SNE incurs the highest computational cost among the tested methods (Table IV). Isomap, on the other hand, delivers a stable and interpretable projection. Its embedding preserves the curved manifold structure of the original dataset and effectively separates clusters and outliers. This behavior can be attributed to its use of geodesic distances in constructing the similarity matrix, in contrast to the Euclidean distances used in PCA or the spherical model.

When considering reconstruction capability, it is important to note that Isomap, UMAP, and t-SNE are primarily designed for visualization and do not support decoding from low to high-dimensional space. In contrast, the proposed spherical model, being linear, supports inverse projection in a straightforward and computationally efficient manner. Among the tested models, only the VAE natively supports decoding. As expected, the VAE achieves high reconstruction accuracy, with average errors below 5%. However, this comes at the cost of reduced interpretability. For instance, in Figure 9(e,l), clusters are tightly packed and largely indistinguishable. This is a direct consequence of the VAE's latent space regularization, which constrains the learned distribution to be approximately Gaussian, causing overlap between distinct clusters. Generative adversarial networks (GANs) proved to be less scalable, as manual tuning was required for each model. Model collapse frequently occurred during training, making GANs impractical for this study, especially considering the large number of generative models required (one per municipality and month).

An important feature for practical deployment is out-of-sample transformation, which is the ability to project new profiles into the existing low-dimensional space without retraining. This can be done by the spherical model, UMAP, and VAE. In the spherical model, the PCA transformation in (10) is directly applied to new data. UMAP and VAE enable this through approximate neighbors and an encoder network, respectively. In contrast, Isomap and t-SNE lack native support for this feature, requiring full retraining to incorporate new points, which limits their scalability.

It is theoretically possible to extend methods like Isomap, UMAP, and t-SNE with neural-network-based decoders to enable data generation, this would require stacking two separate models, increasing implementation complexity and training overhead. By contrast, the spherical model offers a balanced solution: it preserves geometric structure, enables anomaly detection through spatial interpretation, and supports straightforward decoding, all within a simple and computationally lightweight framework.

Authorized licensed use limited to: TU Delft Library. Downloaded on October 30,2025 at 11:36:28 UTC from IEEE Xplore. Restrictions apply.

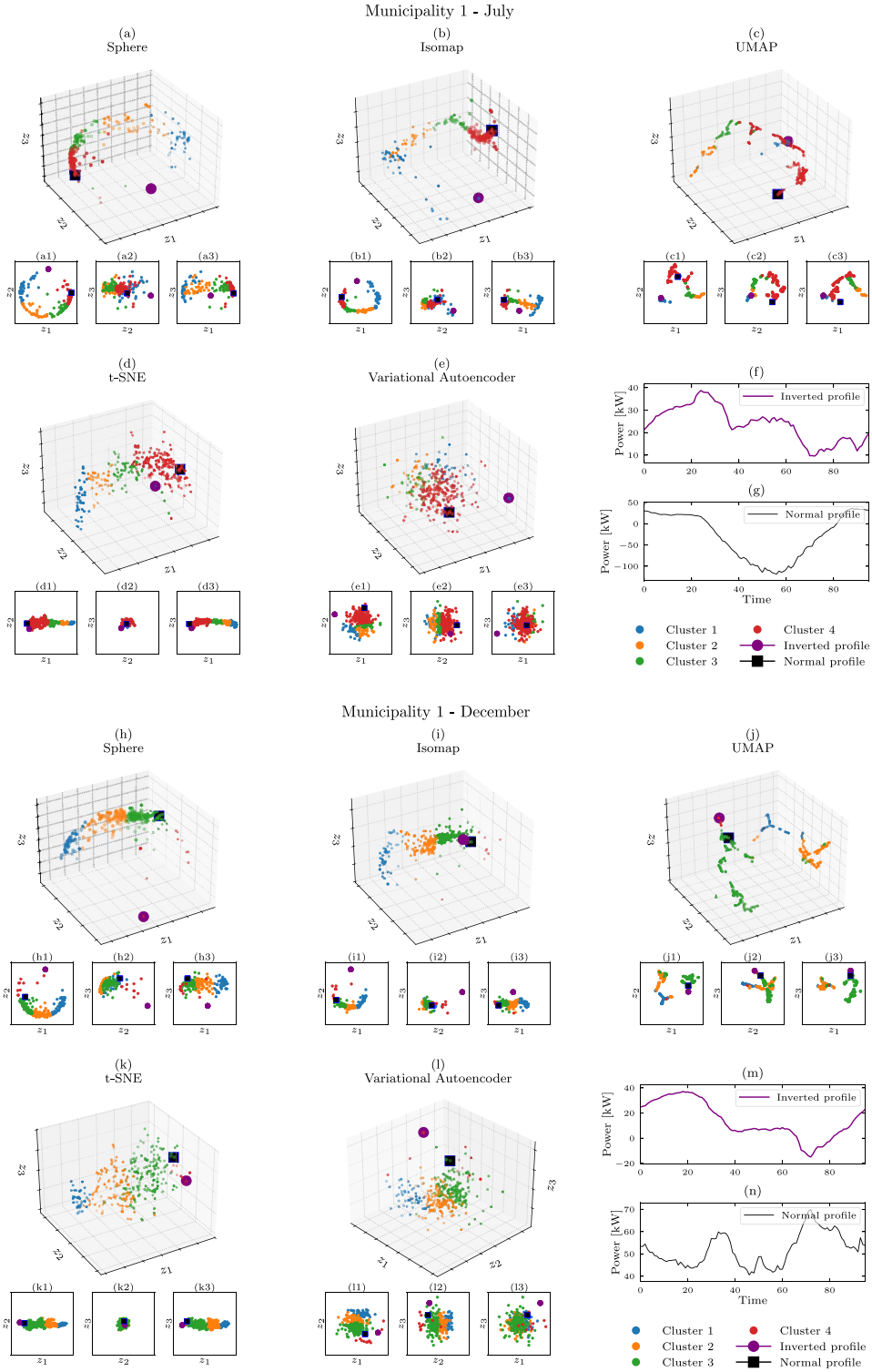

Fig. 9. Lower-dimensional embeddings of MV daily load profiles for one municipality, shown for July (upper pane) and December (lower pane), using five methods: (a,f) Spherical modelling (proposed), (b,g) Isomap, (c,h) UMAP, (d,i) t-SNE, and (e,j) Variational Autoencoder. All methods project the same high-dimensional dataset into three dimensions. Colors correspond to k-means clustering results computed once in the high-dimensional space and consistently applied across all embeddings. The purple dot highlights an anomalous sensor affected by an inverted wiring configuration, while the black square denotes a normal profile from an area with high PV penetration. Interactive version of the datasets is available at [45].

Table IV summarizes the qualitative characteristics and computational efficiency of each method. The results are averaged over 120 models (10 municipalities across 12 months). The spherical model shows the fastest execution time, due to its PCA-based formulation, and is particularly well suited for scalable deployment across large sensor networks. On the other hand, the VAE, while effective for generative tasks, is computationally demanding and less suitable for visualization

Authorized licensed use limited to: TU Delft Library. Downloaded on October 30,2025 at 11:36:28 UTC from IEEE Xplore. Restrictions apply.

TABLE IV
COMPARISON BETWEEN LOW-DIMENSIONAL PROJECTION METHODS

Algorithm	Type (NL: Non-linear)	Encoder	Geometrical embedding (stable structure)	Decoder	Susceptible to random initialization	Out-of-sample transform	Computational time (seconds)
Spherical (PCA)	Linear	Yes	Yes	Yes	No	Yes	0.0037 ± 0.0030
Isomap	NL (Geodesic)	Yes	Yes	No	No	No	0.0702 ± 0.0531
UMAP	NL	Yes	No	No	Yes	Yes	0.7029 ± 1.0392
t-SNE	NL	Yes	No	No	Yes	No	4.3986 ± 1.0499
VAE	NL (Neural Network)	Yes	No	Yes	Yes	Yes	15.0274 ± 6.5563

or exploratory analysis. Overall, the spherical projection has a practical balance across reconstruction, interpretability, and computational cost, offering a robust tool for data analysts working with high-resolution load profile datasets.

VI. DISCUSSION AND FUTURE RESEARCH

The spherical modelling approach offers a well-balanced trade-off across interpretability, reconstruction capability, scalability, and computational cost, as summarized in Table IV. However, several limitations and avenues for future work remain:

First, the interpretability of the lower-dimensional embedding relies on the CEV of the dataset being sufficiently high, e.g., ideally above 80% to retain the spherical structure in the lower dimension. The spherical model has a straightforward inverse projection for generative modelling. However, the reconstruction quality of profiles generated from the Von Mises distribution in three dimensions is limited when compared to specialized generative models such as variational autoencoders (VAEs).

A promising direction for future research is to extend the current use of the Von Mises distribution to the Von Mises–Fisher distribution, which generalizes the model from the unit circle to the hypersphere. This would allow to use the spherical structure for more than three dimensions and potentially improve reconstruction quality while continuing to exploit the geometric structure of the data.

One of the most notable observation of the spherical model is the smooth latent ordering among the daily MV profiles. The clusters transition gradually from commercial to residential consumption, and finally to profiles with high photovoltaic (PV) penetration. This continuum supports the idea that consumption behaviors construct a latent spectrum, rather than strictly separable categories or clusters.

Future work may also take advantage of simple structure of the sphere and correlate its sections with socio-demographic data. By mapping known demographic characteristics to specific regions of the sphere, it may be possible to generate synthetic daily profiles in areas lacking advanced metering infrastructure. This could support data-driven planning and forecasting by enabling DNOs to infer expected consumption patterns based on population or urban characteristics.

Currently, the spherical representation identifies outliers but cannot distinguish between valid rare consumption behaviour and meter faults. This limitation is crucial in practice. Ongoing work explores rule-based heuristics and semi-supervised learning to enable fault classification and support automatic fault recognition in operational settings.

Finally, the application of the proposed spherical model to low-voltage smart meter datasets remains an open question due to the high variability and uncertainty of individual consumption patterns, which prevents a low-dimensional spherical representation from capturing sufficient variance. An important direction for future research is determining the level of aggregation required for LV smart meter data so that spherical modelling explains 80% or more of the variance. In this article, we have focused on establishing the theoretical and practical foundation of the spherical representation and demonstrating its direct application to MV load profiles, where these conditions are met.

VII. CONCLUSION

This article presented a spherical model to represent MV load profiles using a PCA dimensionality reduction technique. The proposed modelling aids in the detection of outlier load profiles using a spherical coordinate system. It was found that the radius of the PCA projection is the most effective variable for detecting defective meters. The spherical profile visualisation uncovers a continuous underlying latent ordering within the load profiles. The ordering is carried out by applying a principal curve technique. A generative model was created by combining the principal curve and the von Fisher-Misses distribution from the directional statistics. Analyses of four different load profile datasets from different municipalities in the Netherlands show that the spherical structure with latent ordering exists, validating the spherical modelling for multiple datasets. An extended analysis was conducted using a dataset for one year, covering ten municipalities, where four state-of-the-art dimensionality reduction techniques were applied. This comparison highlights the consistency, interpretability, and low computational cost of the proposed spherical modelling approach. The resulting projections from the various lower-dimensional modelling techniques applied to the MV daily load profiles in this article can be interactively explored at [45].

REFERENCES

[1] Y. Wang, Q. Chen, T. Hong, and C. Kang, "Review of smart meter data analytics: Applications, methodologies, and challenges," *IEEE Trans. Smart Grid*, vol. 10, no. 3, pp. 3125–3148, May 2019.

[2] I. Guyon, Ed., *Feature Extraction* (Studies in Fuzziness and Soft Computing). New York, NY, USA: Springer-Verlag, 2006.

Authorized licensed use limited to: TU Delft Library. Downloaded on October 30,2025 at 11:36:28 UTC from IEEE Xplore. Restrictions apply.

[3] C. Bouveyron and C. Brunet-Saumard, "Model-based clustering of high-dimensional data: A review," *Comput. Stat. Data Anal.*, vol. 71, pp. 52–78, Mar. 2014.
[4] G. Chicco and A. Mazza, "Load profiling revisited: Prosumer profiling for local energy markets," in *Local Electricity Markets*. Amsterdam, The Netherlands: Elsevier, 2021, pp. 215–242.
[5] S. Lin, F. Li, E. Tian, Y. Fu, and D. Li, "Clustering load profiles for demand response applications," *IEEE Trans. Smart Grid*, vol. 10, no. 2, pp. 1599–1607, Mar. 2019.
[6] K.-L. Zhou, S.-L. Yang, and C. Shen, "A review of electric load classification in smart grid environment," *Renew. Sustain. Energy Rev.*, vol. 24, pp. 103–110, Aug. 2013.
[7] M. Jain, T. AlSkaif, and S. Dev, "Validating clustering frameworks for electric load demand profiles," *IEEE Trans. Ind. Informat.*, vol. 17, no. 12, pp. 8057–8065, Dec. 2021.
[8] J. Kwac, J. Flora, and R. Rajagopal, "Household energy consumption segmentation using hourly data," *IEEE Trans. Smart Grid*, vol. 5, no. 1, pp. 420–430, Jan. 2014.
[9] M. Sun, I. Konstantelos, and G. Strbac, "C-vine copula mixture model for clustering of residential electrical load pattern data," *IEEE Trans. Power Syst.*, vol. 32, no. 3, pp. 2382–2393, May 2017.
[10] R. Granell, C. J. Axon, and D. C. Wallom, "Clustering disaggregated load profiles using a Dirichlet process mixture model," *Energy Convers. Manag.*, vol. 92, pp. 507–516, Mar. 2015.
[11] A. Aleshinloye, M. A. Manzoor, and A. Bais, "Evaluation of dimensionality reduction techniques for load profiling application in smart grid environment," *IEEE Can. J. Elect. Comput. Eng.*, vol. 44, no. 1, pp. 41–49, May 2021.
[12] A. Bosisio et al., "A method to analyzing and clustering aggregate customer load profiles based on PCA," in *Proc. 5th Int. Conf. Green Technol. Sustain. Dev. (GTSD)*, Nov. 2020, pp. 41–47.
[13] A. Aleshinloye, A. Bais, and I. Al-Anbagi, "Performance analysis of dimensionality reduction techniques for demand side management," in *Proc. IEEE Elect. Power Energy Conf. (EPEC)*, Oct. 2017, pp. 1–6.
[14] G. Chicco, R. Napoli, and F. Piglione, "Comparisons among clustering techniques for electricity customer classification," *IEEE Trans. Power Syst.*, vol. 21, no. 2, pp. 933–940, May 2006.
[15] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, Dec. 2000. [Online]. Available: https://www.science.org/doi/10.1126/science.290.5500.2323
[16] J. B. Tenenbaum, V. D. Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, Dec. 2000.
[17] A. Arechiga, E. Barocio, J. J. Ayon, and H. A. Garcia-Baleon, "Comparison of dimensionality reduction techniques for clustering and Visualization of load profiles," in *Proc. IEEE PES Transm. Distrib. Conf. Expo. Latin America (PES T D-LA)*, Sep. 2016, pp. 1–6.
[18] W. Khan, S. Walker, and W. Zeiler, "A bottom-up framework for analysing city-scale energy data using high dimension reduction techniques," *Sustain. Cities Soc.*, vol. 89, Feb. 2023, Art. no. 104323.
[19] S. Ryu, H. Choi, H. Lee, and H. Kim, "Convolutional autoencoder based feature extraction and clustering for customer load analysis," *IEEE Trans. Power Syst.*, vol. 35, no. 2, pp. 1048–1060, Mar. 2020.
[20] J. Chen, W. Li, A. Lau, J. Cao, and K. Wang, "Automated load curve data cleansing in power systems," *IEEE Trans. Smart Grid*, vol. 1, no. 2, pp. 213–221, Sep. 2010.
[21] Z. Guo, W. Li, A. Lau, T. Inga-Rojas, and K. Wang, "Detecting X-outliers in load curve data in power systems," *IEEE Trans. Power Syst.*, vol. 27, no. 2, pp. 875–884, May 2012.
[22] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," *ACM SIGMOD Rec.*, vol. 29, no. 2, pp. 93–104, Jun. 2000.
[23] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proc. 8th IEEE Int. Conf. Data Min.*, Dec. 2008, pp. 413–422. [Online]. Available: http://ieeexplore.ieee.org/document/4781136/
[24] Y. Peng et al., "Electricity theft detection in AMI based on clustering and local outlier factor," *IEEE Access*, vol. 9, pp. 107250–107259, 2021.
[25] E. Khaledian, S. Pandey, P. Kundu, and A. K. Srivastava, "Real-time synchrophasor data anomaly detection and classification using *isolation forest, KMeans, and LoOP*," *IEEE Trans. Smart Grid*, vol. 12, no. 3, pp. 2378–2388, May 2021.
[26] A. Gholami, A. Tiwari, C. Qin, S. Pannala, A. K. Srivastava, and R. Sharma, "Detection and classification of anomalies in power distribution system using outlier filtered weighted least square," *IEEE Trans. Ind. Informat.*, vol. 20, no. 5, pp. 7513–7523, May 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10438880/
[27] H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek, "LoOP: Local outlier probabilities," in *Proc. 18th ACM Conf. Inf. Knowl. Manag.*, Nov. 2009, pp. 1649–1652.
[28] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. 2nd Int. Conf. Knowl. Disc. Data Min. (KDD)*, 1996, pp. 226–231.
[29] S. Guha, N. Mishra, G. Roy, and O. Schrijvers, "Robust random cut forest based anomaly detection on streams," in *Proc. 33rd Int. Conf. Mach. Learn.*, vol. 48, Jun. 2016, pp. 2712–2721.
[30] I. Konstantelos, M. Sun, S. H. Tindemans, S. Issad, P. Panciatici, and G. Strbac, "Using vine copulas to generate representative system states for machine learning," *IEEE Trans. Power Syst.*, vol. 34, no. 1, pp. 225–235, Jan. 2019.
[31] E. M. S. Duque, J. S. Giraldo, P. P. Vergara, P. Nguyen, A. Van Der Molen, and H. Slootweg, "Community energy storage operation via reinforcement learning with eligibility traces," *Elect. Power Syst. Res.*, vol. 212, Nov. 2022, Art. no. 108515.
[32] P. P. Vergara, E. M. S. Duque, T. T. Mai, P. H. Nguyen, and H. Slootweg, "A comprehensive assessment of PV inverters operating with droop control for overvoltage mitigation in LV distribution networks," *Renew. Energy*, vol. 159, pp. 172–183, Oct. 2020.
[33] E. M. S. Duque, P. P. Vergara, P. H. Nguyen, A. Van Der Molen, and J. G. Slootweg, "Conditional multivariate elliptical copulas to model residential load profiles from smart Meter data," *IEEE Trans. Smart Grid*, vol. 12, no. 5, pp. 4280–4294, Sep. 2021.
[34] W. Xia, C. Wang, P. Palensky, and P. P. Vergara, "A flow-based model for conditional and probabilistic electricity consumption profile generation and prediction," *Energy AI*, vol. 21, Aug. 2025, Art. no. 100586, doi: 10.1016/j.egyai.2025.100586.
[35] C. Wang, E. Sharifnia, Z. Gao, S. H. Tindemans, and P. Palensky, "Generating multivariate load states using a conditional variational autoencoder," *Elect. Power Syst. Res.*, vol. 213, Dec. 2022, Art. no. 108603.
[36] C. M. Bishop, *Pattern Recognition and Machine Learning* (Information Science and Statistics). New York, NY, USA: Springer, 2006.
[37] J. T. Morton, L. Toran, A. Edlund, J. L. Metcalf, C. Lauber, and R. Knight, "Uncovering the horseshoe effect in microbial analyses," *mSystems*, vol. 2, no. 1, Feb. 2017, Art. no. e00166.
[38] J. Podani and I. Miklós, "Resemblance coefficients and the horseshoe effect in principal coordinates analysis," *Ecology*, vol. 83, no. 12, pp. 3331–3343, 2002.
[39] D. G. Kendall, "A mathematical approach to seriation," *Philos. Trans. R. Soc. A*, vol. 269, no. 1193, pp. 125–134, 1970.
[40] K. R. Campbell and C. Yau, "Uncovering pseudotemporal trajectories with covariates from single cell and bulk expression data," *Nat. Commun.*, vol. 9, no. 1, p. 2442, Jun. 2018.
[41] M. A. A. Cox and T. F. Cox, "Multidimensional scaling," in *Handbook of Data Visualization*. Berlin, Heidelberg: Springer, 2008, pp. 315–347.
[42] T. Hastie and W. Stuetzle, "Principal curves," *J. Amer. Stat. Assoc.*, vol. 84, no. 406, pp. 502–516, 1989.
[43] M. Salazar. "Procurve: Principal curve method (python package)." Accessed: Jul. 2025. [Online]. Available: https://github.com/MauricioSalazare/procurve
[44] M. Salazar. "Spheral: Spherical representation of load profiles (python package)." Accessed: Jul. 2025. [Online]. Available: https://github.com/MauricioSalazare/spheral
[45] M. Salazar. "Interactive visualisation for spherical embeddings." Accessed: Jul. 2025. [Online]. Available: https://embedding.datamachine.app/

Edgar Mauricio Salazar Duque (Member, IEEE) received the B.E. degree in electrical and electronic engineering from the Universidad de Los Andes, Bogotá, Colombia, in 2008, the M.Sc. degree (cum laude) in smart electrical grids and systems from the KTH Royal Institute of Technology, Stockholm, Sweden, and the Eindhoven University of Technology, The Netherlands, in 2018, and the Ph.D. degree from the Eindhoven University of Technology in 2024. He is currently appointed as a Grid Analytics Scientist, Enexis. His research focuses on data analysis and the application of machine learning techniques to power distribution grids for planning and operation.

Authorized licensed use limited to: TU Delft Library. Downloaded on October 30,2025 at 11:36:28 UTC from IEEE Xplore. Restrictions apply.

Bart van der Holst (Member, IEEE) received the B.E. degree (cum laude) in applied physics and the master's degree (both cum laude) in applied physics and applied mathematics from the Technical University of Eindhoven in 2017 and 2020, respectively, where he is currently pursuing the Ph.D. degree with the Electrical Energy Systems Group, after working as a Data Scientist in the Energy Domain. His research focuses on the application of data analysis and optimization techniques for congestion management in power distribution networks.

Pedro P. Vergara (Senior Member, IEEE) was born in Barranquilla, Colombia, in 1990. He received the B.Sc. degree (with hons.) in electronic engineering from the Universidad Industrial de Santander, Bucaramanga, Colombia, in 2012, the M.Sc. degree in electrical engineering from the University of Campinas, UNICAMP, Campinas, Brazil, in 2015, and the Ph.D. degree from the University of Campinas, UNICAMP, Brazil, and the University of Southern Denmark, SDU, Denmark, funded by the Sao Paulo Research Foundation (FAPESP), in 2019. In 2019, he joined the Eindhoven University of Technology (TU/e), The Netherlands as a Postdoctoral Researcher. In 2020, he was appointed as an Assistant Professor with the Intelligent Electrical Power Grids (IEPG) Group, Delft University of Technology, The Netherlands. He is currently an Associate Professor with the Delft University of Technology. His main research interests include the development of algorithms for the control, planning, and operation of electrical distribution systems with high penetration of low-carbon energy resources (e.g., electrical vehicles, PV systems, electric heat pumps) using optimization and machine learning approaches. He received the Best Presentation Award at the Summer Optimization School in 2018 organized by the Technical University of Denmark (DTU) and the Best Paper Award at the 3rd IEEE International Conference on Smart Energy Systems and Technologies (SEST), in Turkey, in 2020.

Juan S. Giraldo (Senior Member, IEEE) received the B.Sc. degree in electrical engineering from the Universidad Tecnológica de Pereira, Pereira, Colombia, in 2012, and the M.Sc. and Ph.D. degrees in electrical engineering from the University of Campinas, Campinas, Brazil, in 2015 and 2019, respectively. From October 2019 to May 2021, he was a Postdoctoral Fellow with the Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands. From June 2021 to August 2022, he was a Researcher with the Mathematics of Operations Research Group, University of Twente, Enschede, The Netherlands. He is currently a Researcher with the Techno-Economic Energy Transition Studies Group, Netherlands Organization for Applied Scientific Research (TNO). His current research interests include optimization, planning, and control of energy systems, energy transition pathways, and machine learning applications to energy systems.

Phuong H. Nguyen (Member, IEEE) received the Ph.D. degree from the Eindhoven University of Technology (TU/e), The Netherlands, in 2010. During his one-year sabbatical leave in 2019, he took up a Group Leader position of the Sustainable Energy Systems (SES) Group, Luxembourg Institute of Science and Technology (LIST). Since January 2020, he has been back to TU/e as an Associate Professor with the Electrical Energy System (EES) Group. He as committed his research effort to realize synergies of advanced monitoring and control functions for the distribution networks along with emerging digital technologies. This distinctive combination of competencies allows him to develop a research pathway crossing over various domains of mathematical programming, stochastics, data mining, and communication networks. His research of interests includes data analytics with deep learning, real-time system awareness using (IoT) data integrity, as well as predictive and corrective grid control functions.

Anne van der Molen (Member, IEEE) received the M.S. degree in electrical engineering from Twente University, in 1997. He is presently working for Dutch Distribution System Operator Stedin where he is engaged with smart grids strategy and technology planning. His areas of interest include system/market operations, operational technology and flexibility. Next to that, he is a part-time Research Associate with the Eindhoven University of Technology in the area of intelligent energy systems. He is also a Chair of the Dutch Network Operators Association's working group on flexibility and storage, which connects the activities of the Dutch network operators and which works closely together with government, research institutes and industry on capability- and technology development. He is also member of the Technology Committee of the Association of European Distribution System Operators (E.DSO).

Han J. G. Slootweg (Senior Member, IEEE) received the M.Sc. degree (cum laude) in electrical power engineering and the Ph.D. degree from Delft University of Technology, Delft, The Netherlands, in 1998 and 2003, respectively, and the M.Sc. degree in business administration. He is currently a member of the board and COO of Enexis Netbeheer B.V., Hertogenbosch, The Netherlands, one of the largest Distribution Network Operators of the Netherlands. His main focus is accelerating the extension of Enexis' LV and MV grids by boosting operational efficiency and developing the cooperation with external contractors. From 2009 to 2025, he held a part-time Professorship in Smart Grids, Electrical Energy Systems Group, Eindhoven University of Technology. He has supervised more than 20 Ph.D. projects and (co-)authored more than 200 papers, covering a broad range of various aspects of electrical power systems.

Authorized licensed use limited to: TU Delft Library. Downloaded on October 30,2025 at 11:36:28 UTC from IEEE Xplore. Restrictions apply.