Electric Power Systems Research 187 (2020) 106431

Contents lists available at ScienceDirect

Electric Power Systems Research

journal homepage: www.elsevier.com/locate/epsr

Three-way unsupervised data mining for power system applications based on tensor decomposition

Betsy Sandoval[a], Emilio Barocio[a,*], Petr Korba[b], Felix Rafael Segundo Sevilla[b]

[a] Electrical Engineering Department, Universidad of Guadalajara, Guadalajara, MX 44430, Mexico
[b] Electrical Power System and Smart Grid Lab, Zurich University of Applied Science, Winterthur, CH-8401, Switzerland

ARTICLE INFO

Keywords:
Three-way tensor decomposition
PARAFAC
Clustering
Compression
Missing data
Electrical load data

ABSTRACT

Sophisticated geospatial metering devices used in today's networks such as the advanced metering infrastructure (AMI), wide area measurement system (WAMS) and supervisory control and data acquisition (SCADA) open new opportunities to monitor the security of the system in real time. Consequently, these metering infrastructures have received significant attention in recent years from data mining communities because of the new challenges involved on managing this information. One of the main challenges is the analysis of multivariable data, which represents datasets containing variables of different nature, which are linked. In this document a data mining technique that allows the analysis of multivariate data is presented. Moreover, an innovative application of an unsupervised data mining algorithm for smart meters data, particularly to Electrical Load Profile using tensor decomposition is presented. Since the proposed tensor representation allows to assign a given dimension to a particular variable involved; data reduction, data compression, data visualization and data clustering is archived separately for every variable. To validate the effectiveness of the proposed methodology, a three-way tensor built with data from the Electrical Reliability Council of Texas (ERCOT) is presented. The results demonstrate that is possible to extract more information than using conventional approaches based on 2-way arrangements (matrices). Additionally, the proposed algorithm is solved using an iterative approach, which indirectly enable to estimate missing data.

1. Introduction

The main challenges existing moving towards the paradigm of industry 4.0 is to integrate, manage and exploit process data to benefit business and society [1]. Based on this trend, digitalization of power systems is becoming an strategic pillar for the development of the new market energy models [1,2]. As result, the sophisticated geospatial metering devices that are currently used in the network such as the advanced metering infrastructure (AMI), wide area measurement system (WAMS) and supervisory control and data acquisition (SCADA) are not just providing the basis of the new digital era on power systems but are in fact providing more visibility of the network variables and open new opportunities to monitor the security of the system in real time. Consequently, metering infrastructures has lately received significant attention from data analytic communities such as data mining because of the new challenges involved on managing this information.

This document presents an innovative application of an unsupervised data mining algorithm for smart meter (SM) data, particularly to Electric Load Data (ELD) using tensor decomposition. The motivation of choosing the aforementioned data analytic tool on this specific power system application is to face the challenges arising from these type of devices such as: *1) Multivariable data:* SM data are intrinsically multivariable, these means that one stream can contain variables of different nature, which are related to each other. A sensitivity analysis of the electricity and gas consumption to climate presented in [3] shows how SM can provide a mix of variables such as temperature, relative humidity and wind speed simultaneously and this document emphasis the need to deal with multivariable data. Similarly [4], highlight the relevance of combining SM information with economic growth and financial development in order to derive dynamic links of energy consumption. *2) Big Data:* Sophisticated monitoring infrastructure, such as SM, generate significant amounts of data that needs to be stored on servers. Until 2018, more than 86.5 millions of SM have been installed only in the USA [52], generating monthly more than 23.7 Tb of information per million of SM installed [5]. *3) Visualization:* Plotting data to fit models, make predictions and derive conclusions is a crucial component of data analytics [6]. The more data available the more complex becomes to visualize this information. For

* Corresponding author.
E-mail address: emilio.barocio@cucei.udg.mx (E. Barocio).

https://doi.org/10.1016/j.epsr.2020.106431
Received 13 November 2019; Received in revised form 18 March 2020; Accepted 24 May 2020
Available online 09 June 2020

this reason, ensuring accurate visual representations irrespective of the complexity of the problem is an important task.

These constraints open the opportunity to explore data analytic tools on this field. In this case, a data driven approach based on tensor decomposition is used for analysis and process of datasets collected on distribution systems.

Given that, tensor decomposition is a dimensionality reduction technique, additionally to this property, the method allows to achieve at the same time data compression, data visualization, and data clustering. Moreover, in contrast with traditional techniques that work with 2D arrays such as PCA and SVD, tensor decomposition allows to work with multivariate data. As result, the proposed technique allows to extract more information after application of the data mining process and thus, more flexibility than working with traditional techniques is obtained.

1.1. Background

The most popular methods to process and analyze ELD can be divided in three categories, namely: *direct data analytic approaches, model-based data analytic techniques* and *indirect data analytic techniques*. Some of the direct data analytic technics include the minimization of various types of Euclidean distances [7,8], unsupervised neural network [9], swarm optimization techniques [10,11] and k-means [12,13]. Each of these methods allows the algorithm to extract the implicit knowledge of the energy data. Hierarchical clusters depicted using dendrograms, present visualization problems when large volumes of high-dimensional data are processed. Model-based data analytic methods have been used to overcome ELD customer clustering, such is the case of Gaussian mixture model (GMM), which is one of the most widely used model-based clustering [14–16]. However, the performance of this type of algorithms is directly affected from the inherent limitations existing on standard multivariate functions; all mixture model components follow a predefined marginal distribution function and dependency structure. To circumvent these limitations new approaches such as finite mixture modeling frameworks have been proposed in [17], which is a method based on C-vine copulas (CVMM) for carrying out consumer categorization. The advantage of [17] over other approaches, resides in the eminent flexibility of pairing copulas toward identifying multi-dimensional dependency structures present in load profiling data.

Conversely, indirect data analytic techniques offer two sub-categories: time series grouped and feature extraction. The first sub-category assumes that electric profile data are essentially time series [18,19]. Both sub-categories can reduce the dimensionality of the time series while maintaining some of the original characteristic of the consumption profiles. Additionally, the tuning parameters required and the high computational cost, impose limitations to process large multivariable data sets. Therefore, a well-known family of statistical and embedding methods such as: principal component analysis (PCA), canonical component analysis (CCA), and more recently Sammon, stochastic neighbors embedding (SNE) and local linear embedding (LLE) can be applied directly to process a multivariate time-series due to its ability to handle and analyze large volume of data sets [7,8]. These methods can be used to obtain data dimensional reduction, data visualization and indirect data clustering analysis. Similarly, these methods transform high-dimensional space data into a low dimensional space while retaining the most significant information, allowing to obtain significant data compression.

Irrespective of the popularity and advantages of these approaches for analyzing ELD data, a negative aspect related to these methods is the processing of the input data as two-dimensional arrays (matrices). The 2D representation neglects the fact this type of data, are intrinsically multivariable. If a process with more than one variable, like is the case for SM data, is represented on a 2D array, after application of the mining process the different variables will mix and important information will be lost, which will not be possible to retrieve.

To cope with these challenges, a novel application of data mining based on tensor decomposition for power systems applications is proposed. The benefit of using multidimensional arrays on SM data is the designation of a particular dimension to a given variable. As result, after the data mining process, the physical nature of the involved variables remains the same and as result, dimensionality reduction, data compression, data visualization and data clustering is possible to obtain.

The mathematical background of tensor decomposition can be found in [20–22], and the resume of the particular algorithm used on this work, which is referred as Parallel Factor model (PARAFAC) can be found in [23,24].

Tensors and their decompositions is a truly established and documented methodology that has been successfully applied on different disciplines of science such as: Phonetics [25], Psychometrics [26], Chemometrics [27,23] and Neuroscience [28,29] just to mention some. However, on these applications it has not been proposed the use of tensor as a tool for data mining. This alternative use of tensors and how the idea was first introduced is summarized and documented in works such as [30,20] and [21], respectively. Although these documents report and categorize together different procedures on how to profit using tensors, the authors do not provide a clear description about the formulation and information extracted from the decomposition. Moreover, these documents lack to explain what the implications are when the input data to build the tensors are time series, like is the case in this work. In power systems applications, the use of tensors has not been fully exploited and very limited reported cases that can be found in the literature. Such is the case of [31,32], where tensors are used to forecast power grids sequences and to find energy disaggregation, respectively. Therefore, this work represents the first actual application of tensors for data mining on electrical power systems. Here, the problem is reformulated for the first time as a 3-Dimensional array or tensor ($I \times J \times K$) and introduces a new multivariate data mining approach with the main objective to achieve data dimensionality reduction, data compression, data visualization and indirectly, also data clustering.

It is worth noticing that as mentioned before, tensors and their decompositions are well stablished methodologies; however, combining them for power systems applications is not trivial and there are not similar applications available in the literature. Thus, the contribution of this work is in the form of establishing the basis for the combination of these tools and the application itself.

1.2. Summary of contributions

The innovation of using tensor decomposition on smart meters (SM) data results on the following contributions:

- *Formulation of the problem from a different perspective.* Until now, measurements from SM were expressed only as vectors (1D) or matrices (2D). In this work, a higher order representation, 3D in this case, is proposed. The new format allows to store multivariate data in a more natural form, and at the same time is possible to retrieve more information about a certain event. It should be stressed that, although in other disciplines such as data science the introduction of more dimensions is not required and it actually complicates the formulation of the problem, in power systems applications where data have multivariate nature, working with two dimensions leads to insufficient visibility of the problem and consequently to misinterpretation of the phenomena under investigation. Thus, although the proposed approach would certainly add complexity to

find the solution when working with conventional algorithms such as PCA or LLE, the higher dimension does not represent an obstacle for the proposed approach and guarantees a more accurate interpretation of the solution.
- *Significant data compression*. After the tensor decomposition has been performed, the original data sets are kept on new formats of the decomposition, which require significant less memory space.
- *Reduction and Visualization*: The reduction is carried out for each dimension and thus, visualization and clustering of each variable is achieved.
- *Reconstruction of missing data.* The iterative nature of the algorithm for solving the objective function during the tensor decomposition process, indirectly allows the estimation and reconstruction of missing data.

To validate the effectiveness of the proposed methodology, the results are compared against classical dimensional data mining tools such as PCA and SVD. However, since these basic approaches work only for matrices and vectors, the 3D arrays resulting from the ELD are unfolded into equivalent 2D arrays to perform the comparing [30,33].

2. Tensor decomposition

2.1. Multidimensional representation (tensors)

A tensor is a multidimensional array. More formally, a *N-way* or *N*-order tensor is an element of the tensor product of *N* vector spaces, each of which has its own coordinate system [34]. In this work, we use three-way or three-dimensional (cubes) tensors to represent smart meter information. Up to now, vectors and matrices have been the most traditional way to handle these data types, which are in principle tensors representations of first and second order, respectively. However, although these low order tensor representations allow to work with almost any type of data and must of the classical techniques could be applied, some properties of the original set of data could be lost during the transformation process. In this work, the benefits of adding one extra dimension will be demonstrated in the context of power systems. Before providing the general background of tensor decomposition, the mathematical notation used through the document is first introduced.

The standardized notation and terminology for multivariate analysis used in this document is adopted from reference [35]. Therefore, the tensors are represented with bold letters underlined $\underline{\mathbf{X}}$. Matrices, vectors and scalars are represented by uppercase letters in bold, lowercase italics in bold and lowercase italics, respectively ($\mathbf{X}$, $\boldsymbol{x}$, x). The element (i, j, k) from tensor $\underline{\mathbf{X}} \in \mathbb{R}^{I \times J \times K}$ is denoted as x_{ijk}, meanwhile the element (i, j) from matrix $\mathbf{X} \in \mathbb{R}^{I \times J}$ is denoted as x_{ij}, in a similar way the i-th input of vector $\boldsymbol{x} \in \mathbb{R}^{I}$ is x_i. Note that a vector $\boldsymbol{x}$ and matrix an $\mathbf{X}$ can be denoted as tensor of order one ($\underline{\mathbf{X}} \in \mathbb{R}^{I}$) and tensor of second order ($\underline{\mathbf{X}} \in \mathbb{R}^{I \times J}$).

2.2. PARAFAC for tensor decomposition

Tensor decomposition is an established and documented area of multilinear algebra. This subsection provides the basis of the definitions that are going to be used on subsequent sections and it is worth noticing that no scientific contribution has been done in relation to this methodology itself.

Following the standardized notation, assume that $\underline{\mathbf{X}} \in \mathbb{R}^{I \times J \times K}$ is defined as 3D tensor. Assume that $I = 1, \ldots, m$ denotes an element of observation, at J measurement points at t_i, $i = 1, 2 \ldots l, \ldots, w$ is the time at which the observations are made and K is the different operation condition at which each $\underline{\mathbf{X}} \in \mathbb{R}^{I \times J \times K}$ data are collected or stored. Applying the PARAFAC decomposition, a triadic vector decomposition is represented as follows:

$$\underline{\mathbf{X}} = \sum_{r=1}^{R} \boldsymbol{a}_r \circ \boldsymbol{b}_r \circ \boldsymbol{c}_r + \underline{\mathbf{E}} \tag{1}$$

where $r = 1, 2, \ldots, R$ is the tensor's rank that allows rebuilt the tensor $\underline{\mathbf{X}}$. Fig. 1a illustrate, how the outer vector product of the three vectors $\boldsymbol{a}_r$, $\boldsymbol{b}_r$ and $\boldsymbol{c}_r$, generate an approximation to generate the tensor $\underline{\mathbf{X}}$, and the symbol "$\circ$" represent the vector outer product.

A different way to represent the PARAFAC model decomposition is denoted as the sum of R tensors as follow

$$\underline{\mathbf{X}} = \sum_{r=1}^{R} \underline{\mathbf{X}}_r + \underline{\mathbf{E}} \approx \underline{\mathbf{X}} \tag{2}$$

where (2) is illustrated at Fig. 1b.

Equations (1) and (2), are equivalent. The triadic outer product among $\boldsymbol{a}_1 \circ \boldsymbol{b}_1 \circ \boldsymbol{c}_1$ allows to build the tensor $\underline{\mathbf{X}}_1$, and the outer product between $\boldsymbol{a}_R$, $\boldsymbol{b}_R$ and $\boldsymbol{c}_R$ generates the tensor $\underline{\mathbf{X}}_R$. Additionally, the error denoted by $\underline{\mathbf{E}}$ in equations (1) and (2), are the same.

The foremost tensor decomposition will be that which minimize the cost function and expresses the least square difference between the original tensor $\underline{\mathbf{X}}$ and the calculated approximation $\hat{\underline{\mathbf{X}}}$:

$$f(\boldsymbol{a}_r, \boldsymbol{b}_r, \boldsymbol{c}_r) = \|\underline{\mathbf{X}} - \hat{\underline{\mathbf{X}}}\|^2 \tag{3}$$

To carry out the solution of (3) the alternating least squares (ALS) [36] algorithm is used. Considering, that the accuracy of the PARAFAC model is determined by the rank R, this parameter has been developed in such a way that is possible to solve this challenging problem with the following criterions: Assessment error [36], the Euclidean norms of successive estimates of **A, B** and **C** [36], and CORCONDIA algorithm [37]. In this study is used the CORCONDIA algorithm, unlike the other two, determines the number of components based in analyze if the tensor reflects a trilinear variation in the data; a necessary condition to be decomposed by PARAFAC.

2.3. Interpretation of the tensor decomposition

Since there are no literature available regarding time series as raw data used to build 3-dimensional tensors, it was not possible to correlate the results produced from the tensor decomposition with the physical significance of the smart meter data. For this reason, the relations are derived first and represent some of the main contributions on this work. To gain insight on the meaning of the tensor decomposition let us analyze a third order tensor denoted as $\underline{\mathbf{X}} \in \mathbb{R}^{I \times J \times K}$. Note that for this application, the tensor was built using measurements of I days, measured J samples of time, in K locations.

The tensor decomposition of $\underline{\mathbf{X}}$ is carried out using the PARAFAC model with rank R, which is described by three load matrices: **A, B** and **C**, which are composed by the vectors $\boldsymbol{a}_r$, $\boldsymbol{b}_r$ and $\boldsymbol{c}_r$ as depicted on Eq. (4):

$$\begin{aligned} \mathbf{A} &= [\boldsymbol{a}_1 \ldots \boldsymbol{a}_R] \in \mathbb{R}^{I \times R}, & \boldsymbol{a}_r &= [a_1 \ldots a_i]^T \in \mathbb{R}^I \\ \mathbf{B} &= [\boldsymbol{b}_1 \ldots \boldsymbol{b}_R] \in \mathbb{R}^{J \times R}, & \boldsymbol{b}_r &= [b_1 \ldots b_j]^T \in \mathbb{R}^J \\ \mathbf{C} &= [\boldsymbol{c}_1 \ldots \boldsymbol{c}_R] \in \mathbb{R}^{K \times R}, & \boldsymbol{c}_r &= [c_1 \ldots c_k]^T \in \mathbb{R}^K \end{aligned} \tag{4}$$

The information contained in vectors $\boldsymbol{a}_r$, $\boldsymbol{b}_r$ and $\boldsymbol{c}_r$, has relation with the physical significance of the input data and at the same time could be used to reconstruct part or all the original tensor. After Eq. (3) has been solved using the ALS algorithm, the reconstruction of the tensor denoted by $\hat{\underline{\mathbf{X}}}$ can be performed using the load matrices introduced in Eq. (4). Frontal slices are denoted by $\hat{\underline{\mathbf{X}}}(:,:,k)$, where each reconstructed frontal slice is described as follows:

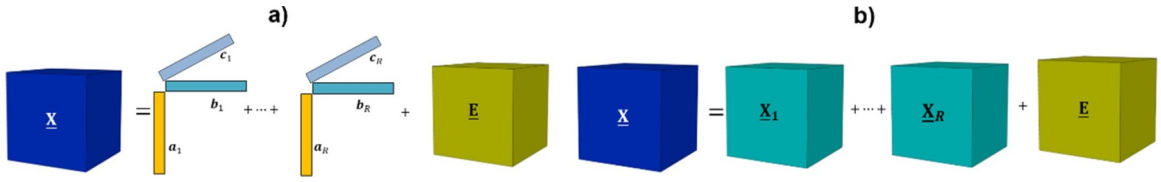

Fig. 1. (a)Tensor decomposition: Sum of triad outer vector products. (b) Tensor decomposition: Sum of rank one tensors

$$\hat{\underline{\mathbf{X}}}(:,:,1) = \begin{bmatrix} \alpha^r_{111} & \alpha^r_{121} & \cdots & \alpha^r_{1j1} \\ \alpha^r_{211} & \alpha^r_{221} & \cdots & \alpha^r_{2j1} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha^r_{i11} & \alpha^r_{i21} & \cdots & \alpha^r_{ij1} \end{bmatrix}$$

$$\hat{\underline{\mathbf{X}}}(:,:,2) = \begin{bmatrix} \alpha^r_{112} & \alpha^r_{122} & \cdots & \alpha^r_{1j2} \\ \alpha^r_{212} & \alpha^r_{222} & \cdots & \alpha^r_{2j2} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha^r_{i12} & \alpha^r_{i22} & \cdots & \alpha^r_{ij2} \end{bmatrix}$$

$$\hat{\underline{\mathbf{X}}}(:,:,k) = \begin{bmatrix} \alpha^r_{11k} & \alpha^r_{12k} & \cdots & \alpha^r_{1jk} \\ \alpha^r_{21k} & \alpha^r_{22k} & \cdots & \alpha^r_{2jk} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha^r_{i1k} & \alpha^r_{i2k} & \cdots & \alpha^r_{ijk} \end{bmatrix} \tag{5}$$

where α^r_{ijk} is an element of the reconstructed tensor $\hat{\underline{\mathbf{X}}}$, conformed by the scalar products between a_i, b_j and c_k, which correspond to vectors $\boldsymbol{a}_r$, $\boldsymbol{b}_r$ and $\boldsymbol{c}_r$, respectively; where each element α^r_{ijk} is described as:

$$\alpha^r_{ijk} = \sum_{r=1}^{R} a_{ir} b_{jr} c_{kr} \tag{6}$$

Consequently, each element of the matrices depicted on Eq. (5) are the result of multiple product combinations on Eq. (6); and thus, the following remarks can be derived:

- Assume that each row of the frontal slice $\hat{\underline{\mathbf{X}}}(:,:,k)$, represents a time series and assume that the subscript "i" of α^r_{ijk} is fixed; it can be concluded that the *i-th* scalar a_i of the vector $\boldsymbol{a}_r$ influence only the time series of the *i-th* row of each frontal slice. Thus, each vector $\boldsymbol{a}_r$ represents a particular day of the week.
- Analogously, assume that each column of the frontal slice $\hat{\underline{\mathbf{X}}}(:,:,k)$ represents the measurements at each instant of time and assume that the subscript "j" of α^r_{ijk} is fixed; it can be concluded that the *j-th* scalar b_j of the vector $\boldsymbol{b}_r$ influence the specific *j-th* sampling time.
- Finally, each frontal slice represents data collected under different operation conditions. Now, assume that the subscript "k" associated to each frontal slice is fixed; then each element c_k of the vector $\boldsymbol{c}_r$ influence the time series of the k point of each frontal slice. As result, each vector $\boldsymbol{c}_k$ represents a particular condition.

2.4. Analysis of slices as result of PARAFAC model decomposition

To gain more in depth in the analysis of tensor decomposition given in Eqs. (1) and (2), this subsection introduces two subarrays within this theory: Slices and Fibers.

Slices: Different subarrays referred as *slices* can be used to build a tensor $\underline{\mathbf{X}}$. Three different projections are illustrated on Fig. 2: *horizontals, laterals* and *frontals*.

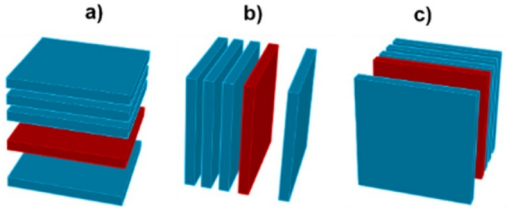

Fig. 2. Tensor partition for slices: (a) Horizontal slices. (b) Lateral slices. (c) Frontal slices

- The *i-th horizontal slice* (depicted on Fig. 2a in red) represent *i-th* sensor measurements for the J instant of time for different K regions. The *i-th* horizontal slice is denoted as follow:

$$\hat{\underline{\mathbf{X}}}_{\boldsymbol{h}}(i,:,:) = \mathbf{C}\ diag(\boldsymbol{a}_i)\ \mathbf{B}^{\mathrm{T}} \tag{7}$$

where $\mathbf{C}$ and $\mathbf{B}$ are the load matrices obtained from tensor decomposition (2). The term $diag(\cdot)$ is an algebraic operation that transform a vector on diagonal matrix. The row vector $\boldsymbol{a}_i$ obtained from $\mathbf{A}$, contains the $\boldsymbol{a}_r$ values for the *i-th* sensor.

- The *j-th lateral slice*, (Fig. 2b in red) represent the *j-th* measurement at each instant of time j for all I located through K conditions:

$$\hat{\underline{\mathbf{X}}}_{\boldsymbol{l}}(:,j,:) = \mathbf{A}\ diag(\boldsymbol{b}_j)\ \mathbf{C}^{\mathrm{T}} \tag{8}$$

where $\boldsymbol{b}_j$ is a row vector from $\mathbf{B}$, which contains the $\boldsymbol{b}_r$ values for the *j-th* interval of time.

- The *k-th frontal slice* (Fig. 2c in red) represent the measurement of I sensors, during J instances of time, for the k-th condition:

$$\hat{\underline{\mathbf{X}}}_{\boldsymbol{f}}(:,:,k) = \mathbf{A}\ diag(\boldsymbol{c}_k)\ \mathbf{B}^{\mathrm{T}} \tag{9}$$

where $\boldsymbol{c}_k$ is a row vector obtained from the $\mathbf{C}$, which contain the $\boldsymbol{c}_r$ values for the *k-th* condition. Note Eqs. (7)–(9) where deduced from the definition presented on Eq. (4), which is not trivial.

2.5. Analysis of fibers as result of PARAFAC model decomposition

Different subarrays based on fibers can be used to obtain a specific information about tensor $\hat{\underline{\mathbf{X}}}(i,j,k)$. To compute the fibers, two indexes of tensor $\hat{\underline{\mathbf{X}}}(i,j,k)$ must remain fixed and one of them must be shifted. Based on this, three types of fibers can be computed: mode-1 where only the subscript "i" is varying, mode-2 where only the subscript "j" is varying and mode-3 where only the subscript "k" is varying. These modes are illustrated on Fig. 3a–c, respectively.

Once the tensor is decomposed by PARAFAC, the approximation of any mode fiber can be calculated.

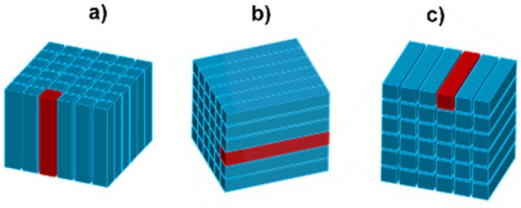

Fig. 3. Fibers-of mode: (a) Mode-1. (b) Mode-2. (c) Mode-3.

Fiber of mode-1, represent the measurement of I sensors at the time instant j, for the specific condition k. The fiber of mode-1 is computed as follow:

$$\hat{\underline{\mathbf{X}}}(:j,\ k) = \sum_{r=1}^{R} b_{jr} c_{kr}\ \boldsymbol{a}_r \tag{10}$$

Fiber of mode-2, represent the measurement of the sensor i, during the time J, for the specific condition k. The fiber of mode-2 is computed as follow:

$$\hat{\underline{\mathbf{X}}}(i,\ :,k) = \sum_{r=1}^{R} a_{ir} c_{kr}\ \boldsymbol{b}_r \tag{11}$$

Fiber of mode-3 represent the measurement of K conditions for the specific sensor i, and specific time instant j.The fiber of mode-3 is computed as follow:

$$\hat{\underline{\mathbf{X}}}(i,\ j,\ :) = \sum_{r=1}^{R} a_{ir} b_{jr}\ \boldsymbol{c}_r \tag{12}$$

Similar to slices, the deduction of Eqs. (10)–(12) is derived from the definition of Eq. (4) and are not obvious.

3. Added value of PARAFAC in power systems

The following stages describe the advantages of using PARAFAC as an algorithm for data mining in power systems applications over the existing techniques.

Stage 1: Tensor design:

- The multivariate data is stored in a 3D-tensor, where days, time and the geographic region represent each of the dimensions. As result, the final tensor hold more information than that stored on conventional matrices.

Stage 2: Tensor decomposition:

- Data compression: As result of the tensor decomposition using PARAFAC, the load matrices **A, B y C** are calculated and are used to build the tensor approximation $\hat{\underline{\mathbf{X}}}$. Hence, the new order of the compressed data is the sum of the orders of the individual load matrices. **A, B** and **C**.
- Data visualization: Every load matrix contains information related to a particular variable in a reduced dimension, because of that, the following is possible: 1) Matrix **A** allows you to create a low-dimension score plot $\boldsymbol{R} > \boldsymbol{I}$ to visualize a daily dynamic behaviour. 2) Similarly, **C** matrix outputs are used to build a low dimensional score plot, used to visualize information related to the geographical location for a reduce-dimensional space $\boldsymbol{R} > \boldsymbol{K}$. Note that using a formulation of the problem based on conventional matrix representations, it is not possible to obtain such data groups.
- *Clustering data*: The optimal number of clusters in the Euclidean score plots, built with matrices **A** and **C**, can be determined using *K*-means. Additionally, a validity index to confirm the geographical clusters is proposed.

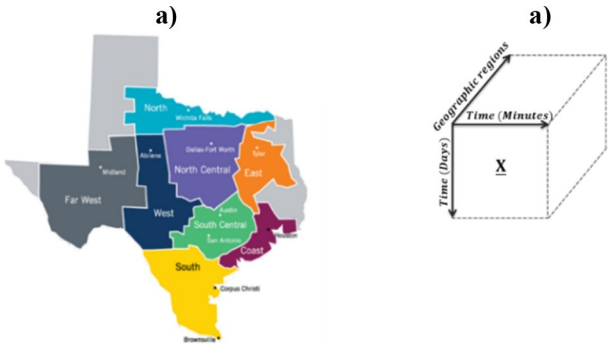

Fig 5. (a) Geographic areas of ERCOT system. (b) Designed tensor $\underline{\mathbf{X}} \in {}^{828\times96\times8}$

Stage 3: Data reconstruction:

- *Handling Missing Data*: In the face of lack of data in the original time series, PARAFAC decomposition allows to estimate these missing information through the ALS algorithm. It is worth noticing that neither the 2-way SVD nor the 2-way PCA, can deal with this constraint.

The general diagram depicted on Fig. 4 provides a comprehensive graphical description of the information related to the different stages. This procedure was implemented using the toolbox *N-WAY* for Matlab [38] in the version *R2015*. The main results of this work are presented in the following section.

4. Main results

To test the effectiveness of processing multivariable data using the PARAFAC decomposition for power systems applications, the open database from the ERCOT system in the USA was used as input and the results are presented in 3 different subsections divided as follows: Tensor design for the ERCOT system, tensor decomposition of the same system, and data reconstruction.

4.1. Tensor design for the ERCOT system

To validate the methodology presented on Section 3, the repository open source data provided on the website of the ERCOT system is used as input [39]. For sake of simplicity, in the following subsections the proposed methodology is evaluated using historical data collected from eight different areas: 1) coast area, 2) east zone, 3) far west zone, 4) north central zone, 5) north zone, 6) south central zone, 7) south zone, and 8) west zone. These zones are graphically displayed on Fig. 5a and the repository database contains historical information from 1998 up today, for 8 types of electric consumption profiles.

On this paper, the summers from 1998 to 2006 of the business low load profile (BLLP) type, disperse on 8 zones of the ERCOT system are analysed. The BLLP with time of use electricity rate scheme may be

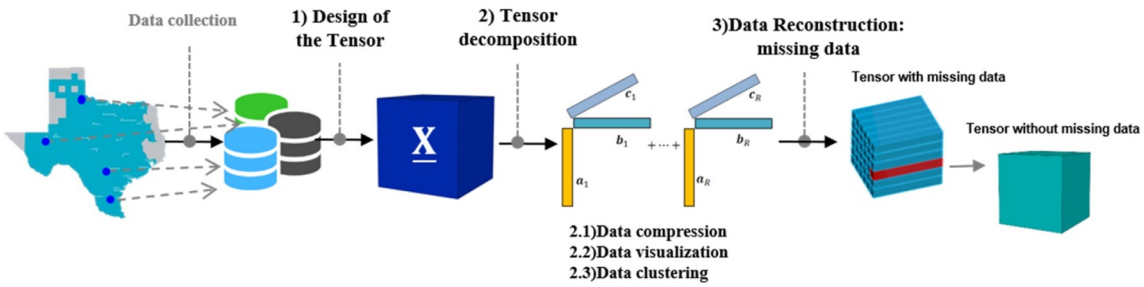

Fig. 4. Flow-chart for the proposed tensor data mining technique based on PARAFAC decomposition

used to shift or reduce load in times of high prices or increment in times of load prices. Therefore, a more detail segmentation of the BLLP consumption provides a wider range of related analysis such as the need of grid expansion and load prediction of electricity markets that allow to design better electricity price schemes.

Summer is one of the most challenges periods for the utility ERCOT, because it must keep the balance among high temperature (33 Celsius degrees) and the increment of the energy demand with low energy prices. The summer is composed of 92 days. The difference between working days (Monday-Friday) and weekends (Saturday-Sunday) is also analyzed in this research. The ELD were collected using 15 minutes sampling time until complete a period of 24 hrs. These information was organized on a 3D tensor of dimensions $\underline{\mathbf{X}} \in {}^{828\times96\times8}$. The subscript "828" is the total number of days over nine years, "96" is the number of samples per day and the subscript "8" is the number of geographic zones. The resulting tensor is illustrated on Fig. 5b.

4.2. Tensor decomposition of the ERCOT system

After the decomposition has been calculated, the stored information in tensor $\hat{\underline{\mathbf{X}}}$ is now represented by the three load matrices **A, B** and **C**. Now, the subsequent section introduce the data mining process based on PARAFAC decomposition for: 1) data reduction and compression, 2) data visualization and 3) data clustering.

4.2.1. Data compression

To provide a more efficient storage and data processing and to preserve high resolution as much as possible, an efficient compression data algorithm is required. In [40], different data compression methods have been classified in two categories: lossless compression and lossy compression. The difference between these categories is the quality of the original data recovered after compression. In lossless compression, the reconstructed data are identical to the original; meanwhile lossy compression is essentially used to preserve the most relevant details and extract punctual characteristics. Data compression techniques are chosen base on the characteristics of the datasets such as coarse granularity and communication data for storage. In this regard the lossless compression have demonstrated to be more effective for handling information with high coarse granularity [41] and less popular to store and transmit compressed data. On the other hand, lossy data compression is the best option to transmit, store and analyze consumption profiles, because of the characteristics (sparse and diverse) of the smart meter data. The ability to compress data of the proposed PARAFAC methodology is compared against two additional lossy compression techniques: the 2-way singular value decomposition (2W-SVD) and the 2-way principal component analysis (2W-PCA). Since the proposed approach works with a 3D tensor arrangement $\underline{\mathbf{X}} \in {}^{828\times96\times8}$, the data is unfold on its mode-1 [20] into a 2D arrangement $\underline{\mathbf{X}}_{(1)} \in {}^{828\times748}$ so it can be used for comparison. It is worth noticing that the PARAFAC tensor requires 4.8516 Mb of storage, while the 2W-SVD and 2W-PCA matrices represent a storage demand of 4.8516 Mb each, respectively. To represent 99% of the original data 2W-SVD uses 191 singular values and 2W-PCA uses 6 principal components. Fig. 6 shows the percentage variability of the data resulting from the 2D techniques as follows: $\upsilon_{USVD} = \sum_{i=1}^{N-1} \boldsymbol{\sigma_i}/\boldsymbol{max}(\boldsymbol{\sigma})$*100 y $\upsilon_{UPCA} = \sum_{i=1}^{N-1} \boldsymbol{\lambda_i}/\boldsymbol{max}(\boldsymbol{\lambda})$*100

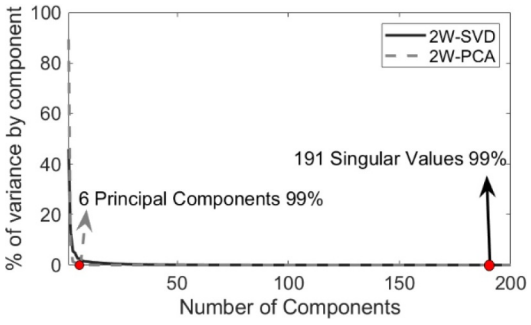

Fig. 6. Percentages of variability associated at 2W-SWD and 2W-PCA.

Table 1 summaries the results of the comparison. First column displays the methodology used, second column depicts the size in bytes of the matrices involved on the decomposition process, third column shows the total size of the matrix decomposition and the last column depicts the compression ratio (CR) [42], which is a measure to standardize and quantify the individual level of compression achieved. The CR index is defined as the ratio between the original data and the compressed data as depicted on Eq. (13)

$$\mathrm{CR} = S_0/S \tag{13}$$

where S_0 is the size of the uncompressed data and S is the size of the compressed data. The higher the value of CR, the smaller the magnitude of the compressed file.

The results show that PARAFAC achieve the maximum compression with up to 99.69% packing of the original data, followed by 2W-PCA with 98.39% and 2W-SVD in last with only 46.18%, respectively. Although the difference between the 2W-PCA and PARAFAC is marginal, the magnitude of the CR index shows a significant difference between the two approaches, demonstrating the relevance and the potential of using PARAFAC for compression of large volumes of data.

4.2.1. Data visualization and data clustering

In this section, the number of groups existing in tensor $\underline{\mathbf{X}}$ are determined. The groups are identified based on the following variables: 1) clustering of days for the ELD and 2) clustering of geographical areas.

1) ***Grouping ELD:*** In Section 2.3 was shown that the *i-th* scalar a_i of the vector $\boldsymbol{a_r}$ is used to weight exclusively time series of the *i-th* row of each frontal slice. Therefore, vectors a_i represent a given day of the week. The next step is to consider that the load matrix **A** captures the grouping of the data associated with the load profiles stored in tensor $\underline{\mathbf{X}}$. Then, the *r-v*ectors $\boldsymbol{a_r}$ can be used to visualize the original profile in a reduced-dimensional space; since the dimension of the decomposition is smaller than the original: $R < I$. Fig. 7 shows the score plots of the different approaches. Fig. 7a depicts vectors $\boldsymbol{a}_1$ vs $\boldsymbol{a}_2$ resulting from PARAFAC where two groups are clearly observed. Following an iterative procedure, three right eigenvectors ($\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{u}_3$ and $\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{w}_3$) were considered, which are associated to the corresponding dominant singular values ($\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2, \boldsymbol{\sigma}_3$) and eigenvalues ($\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \boldsymbol{\lambda}_3$) resulting from the 2W-SVD and 2W-PCA, respectively. Fig. 7b shows that with three singular values up to 64.21% of data variability is captured, while using three principal components 97.60% of data variability is achieved (Fig. 7c). However, in both cases (Fig. 7b and c) only disperse data is displayed and is not possible to identify any cluster.

To gain more insight into the cluster identification analysis, 99% of the data variability is considered to carry out cluster identification. For that, 191 singular values and 6 eigenvalues for 2W-SVD and 2W-PCA respectively are considered. Then its respective curve of validation indexes based on Davis-Bouldin (D-B) and Silhouette are outlined and an automatic search of the significant "knee" within the diagram is performed. The number of clusters at which the "knee" is observed, indicates the optimum clustering for the selected data set.

Based on the above comment, the lowest value of index D-B and the highest value of index Silhouette indicate the optimum number of clusters [43]. Both validation indexes assess the quality of the clusters in terms of compactness and separation of each cluster. For the case under study, *K*-means operates 100 times a partition of 2-6 groups. The validation indexes values obtained from the *K*-means are formulated in the form of a box-whisker plots and are presented on Fig. 8 and show the results of the D-B and Silhouette indexes, respectively. The results

Table 1
Data compression comparison

Method	Matrix decomposition						Total kb	CR
PARAFAC	$\mathbf{A} \in \mathbb{R}^{828 \times 2}$	12.937 kb	$\mathbf{B} \in \mathbb{R}^{96 \times 2}$	1.500 kb	$\mathbf{C} \in \mathbb{R}^{8 \times 2}$	0.128 kb	14.5650	333.09
2W-SVD	$\mathbf{U} \in \mathbb{R}^{828 \times 191}$	1206.60 kb	$\mathbf{\Sigma} \in \mathbb{R}^{191 \times 191}$	285.00 kb	$\mathbf{V} \in \mathbb{R}^{768 \times 191}$	1119.1 kb	2610.7	1.85
2W-PCA	$\mathbf{W} \in \mathbb{R}^{828 \times 6}$	38.812 kb	$\mathbf{E} \in \mathbb{R}^{828 \times 828}$	0.288 kb	$\tilde{V} \in \mathbb{R}^{828 \times 6}$	38.812 kb	77.9120	62.27

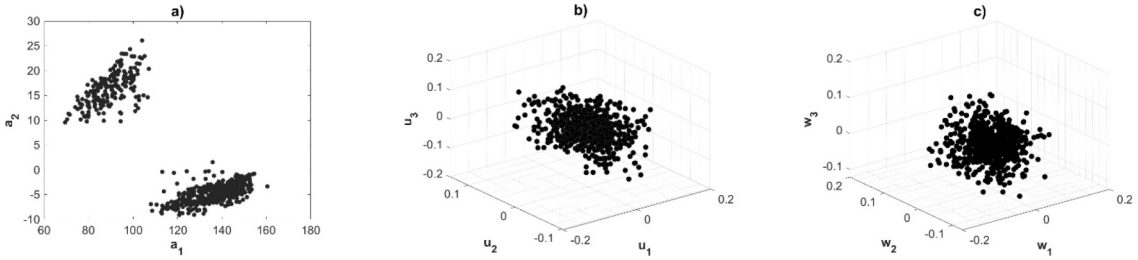

Fig. 7. . Score plot: (a) PARAFAC. (b) 2W-SVD. (c) 2W-PCA

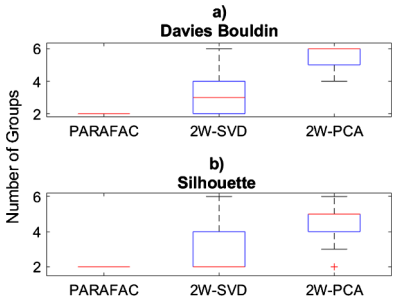

Fig. 8. Box and Whisker plots: (a) Davies Bouldin. (b)Silhouette

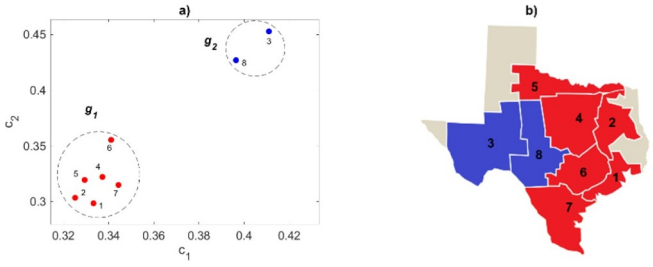

Fig. 10. (a) Score plot $\boldsymbol{c}_1$ vs $\boldsymbol{c}_2$. (b) Geographic groups.

presented on Fig. 8 reinforce the outcomes displayed on Fig. 7. From Fig. 8a it can be observed that both indexes agree on identifying two groups with PARAFAC. Contrary to the other approaches, where in the case of the 2W-SVD method, the index D-B identifies three groups and the index Silhouette only two. However, the number of groups comprised between 25 and 75 percent of the evaluations determine two or four groups for both indexes. The results demonstrate that both methods 2W-SVD and 2W-PCA fail to identify a clear number of groups.

The proposed indexes (D-B and Silhouette) also allow obtaining the corresponding labels, which helps to evaluate the quality of the compactness and separation of each cluster as well as provide visual information. Based in this information the electric load profile (ELP) is grouped based in its associated label and is displayed on Fig. 9. The groups identified with PARAFAC are clearly shown on Fig. 9a; where $\boldsymbol{g}_1$ represents working days and $\boldsymbol{g}_2$ depicts weekends. Conversely, 2W-SVD and 2W-PCA approaches displayed on Figs. 9b and c, respectively, cannot group correctly the ELD data and mix the results.

1) ***Grouping Electrical Geographic Areas:*** As described on Eq. (4), the matrix decomposition **C** is related to the geographical location of the measurements, given that c_i elements on vector $\boldsymbol{c}_r$ weight the time series of the k point of each frontal slice. This means that r vectors $\boldsymbol{c}_r$ can be used to visualize the information related to the geographical location in the reduced-dimensional space: $R > K$. Based on this, Fig. 10a shows the score plot of vector $\boldsymbol{c}_r$, which allows to observe the data distribution on a 2D dimensional plot. Following the clustering identification procedure introduced on the previous Section, two groups are identified ($\boldsymbol{g}_1$, $\boldsymbol{g}_2$), which are conformed by the geographic zones: where $\boldsymbol{g}_1$ include the Coast (1), East (2), North-Central (4), North (5), South-Central (6) and South (7) areas, respectively. Meanwhile $\boldsymbol{g}_2$ is comprised by the Far-West (3) and the West (8) areas. These groups are displayed on Fig. 10a and b depicts these groups in the map.

It should be noted that the visualization and clustering of the geographical areas cannot be performed if the input data is stored on two-dimensional structures, because in this form, temporal and geographical information is combined in one axis. Only higher order representations (3D), allows to visualize and separate each variable in its

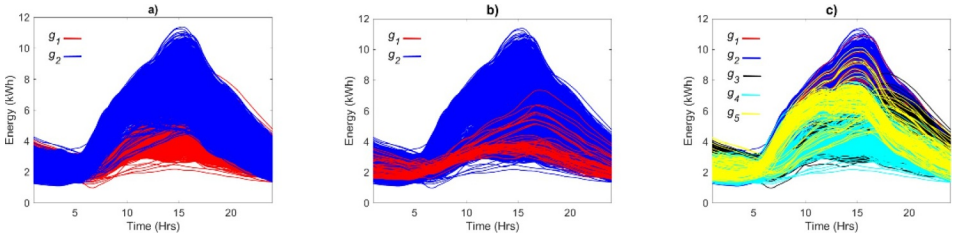

Fig. 9. ELP Clustering plot obtained by: (a) PARAFAC. (b) 2W-SVD. (c) 2W-PCA

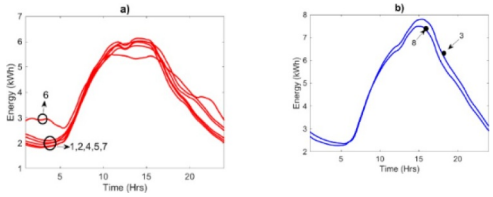

Fig. 11. Average ELP zone geographic group corresponding to: (a) $\boldsymbol{g}_1$. (b) $\boldsymbol{g}_2$

actual unit of measurement. Moreover, the presented results are not possible to replicate even if the original data are split into eight different datasets (one for each area) and then analyzed individually. This is only possible with PARAFAC because the content of the matrix C allows to build a low dimensional score plot where the geographical information can be visualized.

To validate the clustering process shown on Fig. 10, the ELP of each region is averaged and assigned at each group using Eq. (14)

$$\overline{\mathrm{ELP}}_k = \frac{1}{I}\sum_{i=1}^{I} \hat{\underline{\mathbf{X}}}(i,:,k) \in \mathbb{R}^{1\times 96} \tag{14}$$

where k represent the *k-th* geographical area, and I the total number of days. Fig. 11 depicts the dynamic response of the averaged ELP related to groups $\boldsymbol{g}_1$ and $\boldsymbol{g}_2$, respectively. Fig. 11a shows that electricity consumption begin at 5 hrs and increases gradually to reach its maximum electricity consumption between 11 hrs and 17 hrs. Similarly, Fig. 11b displays that electricity begins around 7 hrs, and reach its maximum consumption between 15 hrs and17 hrs. Note, that $\boldsymbol{g}_2$ shows a higher energy consumption than $\boldsymbol{g}_1$.

4.3. Analysis of computational complexity

As shown on Section 4.2.1, two-way approaches such as 2W-SVD and 2W-PCA do not allow to retrieve information for a given each variable, therefore it is necessary to unfold the tensor in three different modes $\underline{\mathbf{X}}_{(1)}$, $\underline{\mathbf{X}}_{(2)}$, $\underline{\mathbf{X}}_{(3)}$ (see Section 2.5). In this form is possible to obtain the same information as it is given for the tensorial decomposition. In consequence, the computational complexity of 2W-SVD and 2W-PCA will be the sum of the three unfolding processes.

The computational cost for PARAFAC is equal to $((JK + KI + IJ) + (7R^2 + R) + 3RIJK + (I + J + K)(R^2 + R) + 11R^3)$, more information can be found in [44]. The numerical calculation of 2W-SVD is currently performed using the Golum-Reinsch algorithm because it is the most efficient, popular and numerically stable technique for computing an arbitrary matrix 2W-SVD [45]. To evaluate the components **U, Σ** and **V** [46] required in 2W-SVD, require a computational effort, which is equal to $(4m^2n + 8mn^2 + 9n^3)$. Whereas the computational cost to evaluate the covariance matrix required in 2W-PCA is (m^2n) and its eigenvalue decomposition is (m^3); therefore the computational complexity of 2W-PCA is $(m^2n + m^3)$ [47]. Where m are the number of the rows and n are the number of columns in the matrix. The values of these variables in the unfolding process are as follows: in mode-1 $m = I$ and $n = JK$; in mode-2 $m = J$ and $n = IK$; while in the unfolding mode-3 $m = K$ and $n = IJ$. If mode-2 and mode-3 are fixed and mode-1 increase, it is possible to quantify the computational complexity for the different approaches PARAFAC, 2W-SVD and 2W-PCA in function of the magnitude of mode-1. For instance, to compute the mode-1, let us assume $m = 828$ and $n = (96 \times 8)$, yields a number of operations of 3835150, 4.5232×10^{15} and 2.8827×10^{17}, for PARAFAC, 2W-SVD and 2W-PCA respectively. Fig. 12 displays the comparison of the different computational effort required for the respective approaches. It can be seen how PARAFAC is significantly less computational expensive in comparison to the other approaches.

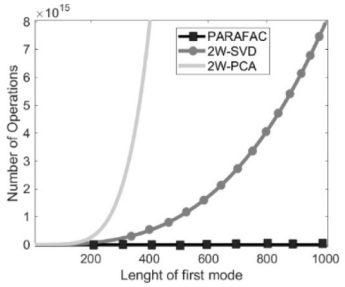

Fig. 12. Computational complexity of PARAFAC, 2W-PCA and PCA for the mode-1

4.4. Data reconstruction: missing data

Conventional statistical approaches assume that data values from variables measured are stored every sampling time and this is not always the case in real life [48]. It has been reported in [49], that power systems advance metering infrastructure (AMI) fail to record between 2.7% and 9.4% of measured data. The so called missing data could lead to computational problems and erroneous calculations. In this section, it is shown how as an indirect result of the iterative ALS algorithm, which is used to solve the PARAFAC tensor decomposition model, the iterative algorithm is also used to reconstruct missing data. The basic principle is that nonexistent elements are replaced by estimates so that the iterative procedure continue until the estimates of the missing information remains constant indicating the end of the iterative process [26].

The lack of data could be classified in two states according to the severity of the problem: 1) lack of single observations and 2) lack of variables. In the first case, no observations are stored for a limited number of samples. In the second case, an entire variable is missing, for instance: one fiber of the tensor.

To demonstrate the ability of PARAFAC to handle missing data, an example using data from the ERCOT system, is now presented for both cases: lack of single observations and lack of variables. It is worth noticing that reconstruction of missing data is performed during the initial stage of any data mining algorithm, as part of the regular preprocessing phase. In this work, this process was performed in the final stage in order to have the load profiles already classified according to their daily dynamic and their geographical position. In this form, it was possible to approximate the reconstructed data with the actual data with a relatively low error.

In the first example, the effectiveness of the proposed approach against lack of single observations is demonstrated. In this case, data from one day in summer of 1998 (30.07.1998) was used, where five consecutive samples of time were artificially removed from the original data to mimic the temporary loss of data. Note that each sample of time correspond to 15 minutes and thus, in the example an absence of one hour of recorded measurements is simulated. The first step is to build a tensor for the selected date 30.07.1998, which was Friday, a working day that belongs to the coast zone. Therefore, tensor $\hat{\underline{\mathbf{X}}}$ is built only considering working days with the regions belonging to cluster $\boldsymbol{g}_1$

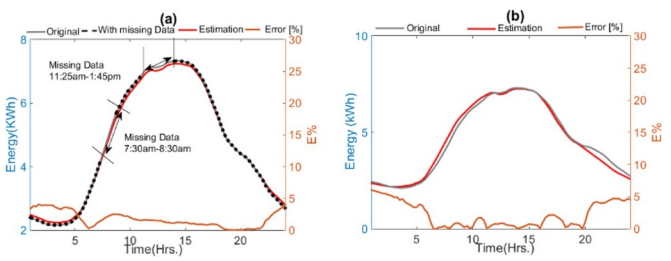

Fig. 13. (a) Lack of single observations 7: 30 – 8: 30 in the morning and 11: 25 – 13: 45 Rush hour, July 30th of 1998. (b) Approximation of a missing variable, July 31th of 1998.

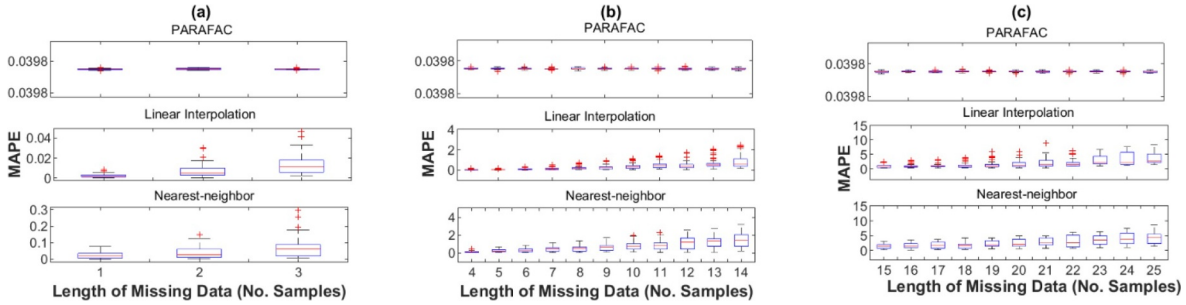

Fig. 14. MAPE for different length of missing data periods: (a) 1-3 Samples. (b) 4-50 Samples. (c) 51-96 Samples.

(1,2,4,5,6,7). In this way, the new tensor, which will be used to calculate the approximation has the following dimension $\underline{\mathbf{X}} \in {}^{65\times96\times5}$.

Fig. 13a shows the results of data reconstruction for lack of single observations twice during the same day. The first lack of observation occurred between 7:30 and 8:30 in the morning, where a linear increment on the consumer profile was missing. The second lack of observation occurred between 11:25 and 13:45, which was characterized by the effect of increase-decrease profile behavior caused by consumption during rush hour on this particular day. From the results, the effectiveness of the missing data reconstruction can be observed by the accuracy between the original data (gray) and the estimated signal (red). Moreover, the error in percent of the estimation is also displayed (orange).

In the second example, the procedure to reconstruct the lack of a variable, a fiber in this case, is presented. After the tensor $\hat{\underline{\mathbf{X}}}$ has been reconstructed and the load matrices **A, B, C** have been calculated, any missing variable can be retrieved using the equations of fibers or slices (7)–(9) and (10)-(12), respectively. Carrying on with data from example one and using the load matrices, is possible to reconstruct any day of a particular region or any region of a particular day, represented as: $\hat{\underline{\mathbf{X}}}(:,:,k)$ or $\hat{\underline{\mathbf{X}}}(i,:,:)$ using Eqs. (7) and (9), respectively.

In the second example it was assumed that the measurements of an entire day (31.07.1998) were missing, which correspond to variable $\hat{\underline{\mathbf{X}}}(31,:,1)$ and to the frontal slice $\hat{\underline{\mathbf{X}}}(:,:,1)$, computed with Eq. (9). Fig 13b depicts the original and the estimated data of the entirely day (31.07.1998) and it can be observed the accuracy in the approximation. Note that this day can also be reconstructed using Eq. (11).

In order to demonstrate the higher performance of PARAFAC, the former approach is compared against two of the most popular algorithms used to estimate intervals of missing data: the Linear Interpolation (LI) and the Nearest-neighbor approach [50]. LI estimates a missing value x_i from the nearest preceding and succeeding available value using linear interpolation. Meanwhile, Nearest-neighbor technique requires the value of the nearest available observation. To compare the performance of PARAFAC against the other two estimation approaches, data from the 30.07.1998 is used. 25 samples were withdrawn in increasing order, and this process was repeated 50 times randomly. For the length of each missing period, the Mean Absolute Percentage Error was calculated: $MAPE = 1/N \sum_{i=1}^{N} |x_o - x_e/x_o|$, where x_o is the original value, and x_e is the estimate value. Fig. 14 shows the box-whisker plot, done based on the MAPE values for different lengths of missing data.

From Fig. 14 it can be observed that for PARAFAC the error (MAPE) remains constant, regardless of the length of missing data. Contrary to LI and Nearest-neighbor approach, where the error is proportional to the length of missing data e.g. marginal for small number of missing samples and large in the opposite case.

5. Conclusion

In this paper an innovative application of unsupervised data mining algorithm for Electrical Load Profile using tensor decomposition is proposed. For the ERCOT data base, and example of using multi-dimensional data and how to process these information has been shown; in order to obtain: data compression, data visualization and data clustering. The results demonstrate the ability of PARAFAC to obtain an individual reduction for every variable; this was proved by obtaining a score plot to visualize and cluster the time variable, and a second score plot for the spatial variable. The results of PARAFAC against traditional techniques show that common methodologies fail to manage multivariate data.

The comparison of the computational complexity between PARAFAC, 2W-PCA and 2W-SVD confirm the advantage of working with tensor decompositions over traditional 2D arrangements. The results illustrate that the computational complexity with 2W-PCA and 2W-SVD grows exponentially in relation to PARAFAC and that the rate of grow is equal to the square and cubic exponents that weigh the dimension of the matrix that stores the data.

Finally, since PARAFAC is solved using an iterative procedure, indirectly is also possible to provide a solution to the missing data problem. The additional attribute of the proposed approach was demonstrated comparing the performance of PARAFAC following the artificial loss of 1 to 25 samples measured, in random position. The comparison was carried out against two of the most popular techniques available in the literature, namely Linear Interpolation and the Nearest-neighbor. The results show that the error when using PARAFAC is small and remains constant even for the case with more loss of missing data, while the error grows as the missing data increases when using more traditional techniques.

CRediT authorship contribution statement

Betsy Sandoval: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing - original draft, Visualization. **Emilio Barocio:** Conceptualization, Methodology, Validation, Resources, Writing - original draft, Visualization, Supervision. **Petr Korba:** Writing - review & editing, Supervision, Project administration, Funding acquisition. **Felix Rafael Segundo Sevilla:** Writing - review & editing, Supervision, Project administration.

Acknowledgment

This work was supported by the Swiss National Science Foundation (SNSF) under the project number PZENP2 173628 of the program Ambizione Energy Grant (AEG) and also is part of the activities of the Swiss Centre for Competence in Energy Research on the Future Swiss Electrical Infrastructure (SCCER-FURIES) - Digitalization program, which is financially supported by the Swiss Innovation Agency (Innosuisse - SCCER program). The authors also thank to the Federal Commission for Scholarships for Foreign Students (FCS) and the Federal Department of Economic Affairs, Education and Research for the Swiss Government Excellence Scholarship to conduct this research.

Appendix

Unfolding 2-Way methods: 2W-SVD and 2W-PCA

Alternative approaches to PARAFAC based on tensor decomposition of the response matrix (1) have been developed in [30,33]. Therefore, two standard approaches to deal with matrix unfolding of $\underline{\mathbf{X}} \in {}^{I\times J\times K}$, which is denoted as $\underline{\mathbf{X}}_{(1)}{}^{I\times J*K}$ were implemented in this paper: 2W-SVD and 2W-PCA.

The modal decomposition of 2W-SVD applied to tensor model $\underline{\mathbf{X}}$ yields:

$$\underline{\mathbf{X}}_{(1)} = \mathbf{U}\Sigma\mathbf{V}^{\mathrm{T}} = [\mathbf{U}][\boldsymbol{\Sigma}_{\mathrm{I}} \quad 0]\begin{bmatrix}\mathbf{V}_{\mathrm{I}}^{\mathrm{T}}\\ \mathbf{V}_{\mathrm{s}}^{\mathrm{T}}\end{bmatrix} \tag{15}$$

where $\mathbf{U}\in{}^{I\times I}$ is an orthonormal matrix containing the left singular vectors, $\boldsymbol{\Sigma}\in{}^{I\times I}$ is a matrix containing the singular values, σ_i,and $\mathbf{V}\in{}^{J*K\times J*K}$ is a matrix containing the right singular vectors.

Similarly, 2W-PCA decomposes the tensor of data using the following expression:

$$\underline{\mathbf{X}}_{(1)} = 1^{*}\bar{\boldsymbol{x}}' + \mathbf{W}\bar{\boldsymbol{V}}^{\mathbf{T}} + \boldsymbol{\varepsilon} \tag{16}$$

From [51], $1^{*}\bar{\boldsymbol{x}}'$ represents the averages value of the original variables, which results from a pre-processing step. The matrix product $\mathbf{W}\bar{\boldsymbol{V}}^{\mathbf{T}}$ is the model of the structure, where $\mathbf{W}\in{}^{I\times I}$ denotes the scores, $\mathbf{V}\in{}^{J*K\times I}$ loadings and $\boldsymbol{\varepsilon}$ the residual. The solution to the problem to PCA using eigenvector decomposition, the covariance matrix of $\underline{\mathbf{X}}_{(1)}$: is denoted as:

$$\mathbf{C}_{\mathbf{X}}\mathbf{W} = \mathbf{E}\mathbf{W} \tag{17}$$

where the right eigenvectors are the score $\mathbf{W}$, $\mathbf{E}$ return the eigenvalues of $\mathbf{C}_{\mathbf{X}}$ sorted in descending order, and the loadings $\bar{\boldsymbol{V}}$ are computing as $\bar{\boldsymbol{V}} = \mathbf{W}\mathbf{E}^{\frac{1}{2}}$. The first few columns of $\mathbf{W}^{\mathbf{t}}$ and $\bar{\boldsymbol{V}}$ explain most of the variance in $\underline{\mathbf{X}}_{(1)}$.

References

[1] E. Hossain, I. Khan, F. Un-Noor, S.S. Sikander, M.S.H. Sunny, Application of big data and machine learning in smart grid, and associated security concerns: a review, IEEE Access 7 (2019) 13960–13988.

[2] Y. Wang, Q. Chen, T. Hong, C. Kang, Review of smart meter data analytics: applications, methodologies, and challenges, IEEE Trans. Smart Grid 10 (3) (2019) 3125–3148.

[3] C. Institude, D.J. Sailor, J.R. Muñoz, Sensitivity of electricity and natural gas consumption to climate in the U.S.A. - methodology and results for eight states, Energy 22 (10) (1997) 987–998.

[4] C. Institude, *Munich personal RePEc archive the dynamic link between energy consumption, economic growth, financial development and trade in China : Fresh Evidence from Multivariate framework analysis Muhammad, Shahbaz and Saleheen, Khan and Mohammad*, no. 42974. 2012.

[5] N. Yu, S. Shah, R. Johnson, R. Sherick, M. Hong, K. Loparo, Big data analytics in power distribution systems, *2015 IEEE Power and Energy Society Innovative Smart Grid Technologies Conference ISGT 2015*, 2015, pp. 1–5.

[6] B.R.J. Hyndman, Visualizing big energy data: solutions for this crucial component of data analysis, IEEE Power Energy Mag. 16 (April) (2018) 18–25.

[7] G. Chicco, R. Napoli, F. Piglione, Comparisons among clustering techniques for electricity customer classification, IEEE Trans. Power Syst. 21 (2) (2006) 933–940.

[8] A. Arechiga, E. Barocio, J.J. Ayon, H.A. Garcia-Baleon, Comparison of dimensionality reduction techniques for clustering and visualization of load profiles, *2016 IEEE PES Transmission & Distribution Conference & Exposition America PES T D-LA 2016*, 2 2017, pp. 1–6.

[9] S. Ryu, H. Choi, H. Lee, H. Kim, V.W.S. Wong, Residential load profile clustering via deep convolutional autoencoder, *2018 IEEE International Conference on Communications, Control, and Computing Technologies Smart Grids, SmartGridComm 2018*, 2018.

[10] G. Chicco, O.M. Ionel, R. Porumb, Electrical load pattern grouping based on centroid model with ant colony clustering, IEEE Trans. Power Syst. 28 (2) (2013) 1706–1715.

[11] E. Cuevas, E. Barocio Espejo, A. Conde Enríquez, Clustering representative electricity load data using a particle swarm optimization algorithm, Stud. Comput. Intell. 822 (2019) 187–210.

[12] I. Benítez, A. Quijano, J.L. Díez, I. Delgado, Dynamic clustering segmentation applied to load profiles of energy consumption from Spanish customers, Int. J. Electr. Power Energy Syst. 55 (2014) 437–448.

[13] J.K. Lin, S.K. Tso, H.K. Ho, C.M. Mak, K.M. Yung, Y.K. Ho, Study of climatic effects on peak load and regional similarity of load profiles following disturbances based on data mining, Int. J. Electr. Power Energy Syst. 28 (3) (2006) 177–185.

[14] R. Li, Z. Wang, C. Gu, F. Li, H. Wu, A novel time-of-use tariff design based on Gaussian mixture model, Appl. Energy 162 (2016) 1530–1536.

[15] R. Granell, C.J. Axon, D.C.H. Wallom, Clustering disaggregated load profiles using a Dirichlet process mixture model, Energy Convers. Manag. 92 (2015) 507–516.

[16] W. Labeeuw, G. Deconinck, Residential electrical load model based on mixture model clustering and markov models, IEEE Trans. Ind. Inform. 9 (3) (2013) 1561–1569.

[17] M. Sun, I. Konstantelos, G. Strbac, C-Vine Copula mixture model for clustering of residential electrical load pattern data, IEEE Trans. Power Syst. 32 (3) (2017) 2382–2393.

[18] R. Granell, C.J. Axon, D.C.H. Wallom, Impacts of raw data temporal resolution using selected clustering methods on residential electricity load profiles, IEEE Trans. Power Syst. 30 (6) (2015) 3217–3224.

[19] A. Notaristefano, G. Chicco, F. Piglione, Data size reduction with symbolic aggregate approximation for electrical load pattern grouping, IET Gener. Transm. Distrib. 7 (2) (2013) 108–117.

[20] S. Rabanser, O. Shchur, and S. Günnemann, "Introduction to tensor decompositions and their applications in machine learning," pp. 1–13, 2017.

[21] E.E. Papalexakis, C. Faloutsos, N.D. Sidiropoulos, Tensors for data mining and data fusion: models, applications, and scalable algorithms, ACM Trans. Intell. Syst. Technol. 8 (2) (2016).

[22] T.G. Kolda, B.W. Bader, Tensor decompositions and applications, SIAM Rev 51 (3) (2009) 455–500.

[23] R. Bro, PARAFAC. Tutorial and applications, Chemom. Intell. Lab. Syst. 38 (2) (1997) 149–171.

[24] C.J. Appellof, E.R. Davidson, Strategies for analyzing data from video fluorometric monitoring of liquid chromatographic effluents, Anal. Chem. 53 (13) (1981) 2053–2056.

[25] R.A. Harshman, Determination and proof of minimum uniqueness conditions for PARAFAC1, UCLA Work. Pap. Phon. 22 (1972) 111–117.

[26] R.B. Cattell, "Parallel proportional Profiles and other principles for determining the choice of factors by rotation, Psychometrika 9 (4) (1944) 267–283.

[27] C.J. Appellof, E.R. Davidson, Strategies for analyzing data from video fluorometric monitoring of liquid chromatographic effluents, Anal. Chem. 53 (13) (1981) 2053–2056.

[28] W. Deburchgraeve, et al., Neonatal seizure localization using PARAFAC decomposition, Clin. Neurophysiol. 120 (10) (2009) 1787–1796.

[29] F. Cong, Q.H. Lin, L.D. Kuang, X.F. Gong, P. Astikainen, T. Ristaniemi, Tensor decomposition of EEG signals: a brief review, J. Neurosci. Methods 248 (2015) 59–69.

[30] A. Cichocki, et al., Tensor decompositions for signal processing applications: From two-way to multiway component analysis, IEEE Signal Process. Mag. 32 (2) (2015) 145–163.

[31] H.A. Song, B. Hooi, M. Jereminov, A. Pandey, L. Pileggi, C. Faloutsos, PowerCast: mining and forecasting power grid sequences, Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinform.) 10535 (LNAI) (2017) 606–621.

[32] M. Figueiredo, B. Ribeiro, A. De Almeida, Electrical signal source separation via nonnegative tensor factorization using on site measurements in a smart home, IEEE Trans. Instrum. Meas. 63 (2) (2014) 364–373.

[33] L. Burgas, J. Melendez, J. Colomer, J. Massana, C. Pous, N-dimensional extension of unfold-PCA for granular systems monitoring, Eng. Appl. Artif. Intell. 71 (February) (2018) 113–124.

[34] T.G. Kolda, B.W. Bader, Tensor decompositions and applications, SIAM Rev 51 (3) (2009) 455–500.

[35] H.A.L. Kiers, Towards a standardized notation and terminology in multiway analysis, J. Chemom. 14 (3) (2000) 105–122.

[36] N.S.E. Papalexakis, C. Faloutsos, Tensors for data mining and data fusion: models, applications, and scalable algorithms, ACM Trans. Intell. Syst. Technol. 8 (2) (2016) 16:1-16:44.

[37] R. Bro, H.A.L. Kiers, A new efficient method for determining the number of components in PARAFAC models, J. Chemom. 17 (5) (2003) 274–286.

[38] C.A. Andersson, R. Bro, The N-way toolbox for MATLAB, Chemom. Intell. Lab. Syst.

(2000).
[39] "ERCOT Repository of Load Profile Data.".
[40] Y. Wang, Q. Chen, C. Kang, Q. Xia, M. Luo, Sparse and redundant representation-based smart meter data compression and pattern extraction, IEEE Trans. Power Syst. 32 (3) (2017) 2142–2151.
[41] L. Wen, K. Zhou, S. Yang, L. Li, Compression of smart meter big data: a survey, Renew. Sustain. Energy Rev. 91 (2018) 59–69 January 2017.
[42] H. Hakkak, M. Azarnoosh, Analysis of lossless compression techniques time-frequency-based in ECG signal compression, Asian J. Biomed. Pharm. Sci. 9 (66) (2019) 16–25.
[43] G. Chicco, Overview and performance assessment of the clustering methods for electrical load pattern grouping, Energy 42 (1) (2012) 68–80.
[44] P. Comon, X. Luciani, A.L.F. de Almeida, Tensor decompositions, alternating least squares and other tales, J. Chemom. 23 (7–8) (2009) 393–405.
[45] J. Saira Banu, R. Babu, R. Pandey, Parallel implementation of Singular Value Decomposition (SVD) in image compression using OpenMp and sparse matrix representation, Indian J. Sci. Technol. 8 (13) (2015) 1–10.
[46] F. Liang, R. Shi, Q. Mo, A split-and-merge approach for singular value decomposition of large-scale matrices, Stat. Interface 9 (4) (2016) 453–459.
[47] C. Syms, Principal Components Analysis, Encycl. Ecol. Five-Vol. Set (2008) 2940–2949.
[48] A. Paul. D, "Missing data," 2001.
[49] D. Kodaira, S. Han, Topology-based estimation of missing smart meter readings, Energies 11 (1) (2018).
[50] J. Peppanen, X. Zhang, S. Grijalva, M.J. Reno, Handling bad or missing smart meter data through advanced data imputation, *2016 IEEE Power and Energy Society Innovative Smart Grid Technologies Conference ISGT 2016*, 2016, pp. 0–4.
[51] J. T., C.V.L. Eriksson, T. Byrne, E. Johansson, Multi- and megavariate data analysis, Metabonomics in Toxicity Assessment, 2005, pp. 323–355.
[52] S. Bahrami, M. Hadi Amini, M. Shafie-Khah, J.P.S. Catalao, A decentralized renewable generation management and demand response in power distribution networks, IEEE Trans. Sustain. Energy 9 (4) (2018) 1783–1797.