



# Implementing Data Reduction Strategies for the Optimal Design of Renewable Energy Systems

Jorge Luis Tena-García<sup>1</sup> · Luis Miguel García-Alcala<sup>2</sup> · Dulce Celeste López-Díaz<sup>3</sup> · Luis Fabián Fuentes-Cortés<sup>1</sup> 

Received: 28 April 2021 / Revised: 20 August 2021 / Accepted: 30 August 2021 / Published online: 28 September 2021  
© The Author(s), under exclusive licence to Springer Nature Singapore Pte Ltd. 2021

## Abstract

Phenomena such as the growing use of environments associated with Big Data have been conducive to the access and generation of databases of energy consumption and environmental conditions. Due to data-driven analysis and the availability of large data sets, the current issue of handling “a lot of information” has arisen in the Renewable Energy Systems (RES) design. The problem of a high amount of data to feed a model, either control or optimal design, is the increase in computational costs. Therefore, it is necessary to define strategies to discriminate data without significant loss of information. This paper presents a comparison of three different strategies of data reduction, random selection of real days, and two approaches to construct typical days (TDs) profiles of solar irradiance  $\alpha$ , temperature ( $T$ ), power demand ( $W^D$ ) and wind speed ( $v$ ) for the optimal design of RES. The addressed strategies to obtain TD are based on principal component analysis (PCA) and k-means clustering for pattern recognition, respectively, which are compared with the use of random real days (RRD) from a database of 1 year of measurements of the four mentioned variables in the northwest region of Mexico. The  $W^D$  data corresponds to a residential building. The optimal design algorithm minimizes the total annual cost of the system; additionally, results include the size of wind turbine, photovoltaic system and battery. The analysis performed allows identifying advantages in the use of the different strategies for catching the behavior of all variables as well as their influence in the design of RES and computational issues associated with the optimization process as time of computation (ToC). Results show that using a conservative number of typical days to feed the model may be sufficient to obtain a similar design to the one obtained using the full data. On the other hand, a wind turbine design gap caused by wind speed variability is analyzed.

**Keywords** Data analysis · k-means · Principal component analysis · Energy consumption profiles · Renewable energy systems

## Introduction

The inclusion of renewable energies in the main electricity grids is one of the objectives of the United Nations 2030 agenda to achieve sustainable development worldwide (U Nations 2015). In energy systems, this depends directly on the efficiency with which the technologies to be implemented are selected and on how their equipment and backup systems are dimensioned, as well as the operation of the system (Kakran and Chanana 2018). In the design of renewable energy systems (RES), the main obstacle

that arises when using clean energies such as solar or wind is their intermittence and randomness (Fuentes-Cortés and Flores-Tlacuahuac 2018). If, in addition, the dynamic characteristics of electricity demand are considered, the analysis becomes even more difficult (García et al. 2019). A solution that is used to guarantee an adequate design of energy systems is to review extensive databases, which usually cover at least one year to visualize the behavior of variables in all seasons of the year (Kakran and Chanana 2018).

Traditionally, variations and uncertainties in input data for optimization models are addressed by two main strategies: stochastic programming and the use of forecasting models. Regarding stochastic programming, one of the most widely used methods is multi-scenario analysis (Yan et al. 2021). This can use the totality of available data, or generating scenarios based on a probabilistic approach

✉ Luis Fabián Fuentes-Cortés  
luis\_fuentes@iqcelaya.itc.mx

Extended author information available on the last page of the article.

of the phenomena, to address uncertainties and possible variations and find solutions under expected conditions or reduce impacts of scenarios with extreme conditions or minimize the chances of a worst-case scenario (Zakaria et al. 2020). Strategies such as chance-constrained (Odetayo et al. 2018), conditional value at risk (Cao et al. 2017) and Monte Carlo analysis (Meschede et al. 2017) have been used to achieve these objectives. In terms of forecasting models, there have been developed models for planning with stability analysis of operating conditions (Ahmad et al. 2020), data distribution analysis (Hernández-Romero et al. 2019; Wang et al. 2018) or models based on chaos theory (Cui et al. 2019). However, these types of strategies involve a high level of complexity in terms of modeling and a high computational cost, which increases when aspects such as the non-linearity of the phenomenon addressed are considered (Rudin et al. 2021).

Particularly, for systems based on renewable sources, addressing the variation in the availability of wind or solar radiation is a fundamental problem to define an optimal sizing of the system and an adequate operation policy of interconnection with local grids or the use of storage systems. Variations on availability of resources leads to instability and low quality of energy supply (Ciupageanu et al. 2020). In this sense, for distributed generation systems based on wind turbines, considering the variability of wind behavior, significant contributions have been made based on forecasting and scenario generation (Li et al. 2020). For photovoltaic systems, both approaches have been used for addressing variations and abnormal behavior associated with weather changes (Ahmed et al. 2020).

Derived from the trend of data-driven analysis and the availability of large data sets, a current issue has emerged in the design of RES: the handling of “a lot of information” (Calvillo et al. 2016). The difficulty that this presents, especially when calculating the optimal design (OD) of the RES, is that the computational capacity and resources required by optimization models (MO) to find a solution to the multivariate problem can be very demanding (Azuatalam et al. 2019). In general, to lighten the burden of handling so much information, one approach is to identify patterns of behavior of the variables of interest for different purposes as: the sizing of the RES, the forecast of the operation, and the generation of scenarios to perform sensitivity analysis of the system (Abdmouleh et al. 2017).

To identify the characteristics of the data sets there are different approaches, those that are simply statistical, probabilistic or those that use Statistical Learning. In Kettaneh et al. (2005) the importance of managing large data sets and how their size increases the difficulty and modifies the results of various problems of general interest is reviewed. Thus, due to the fact that large data sets continue to grow, it is concluded that their management and

synthesis is a necessary subject of study (Kettaneh et al. 2005).

In the statistical branch, the principal component analysis (PCA) (Jolliffe 1986) leads to data reduction based on its dimensionality reduction (Shlens 2014). This technique allows comparing several variables or dimensions, then identifies statistical measures to form the matrix of principal components, in which the variance is recorded (Shlens 2014). The first principal component retains most of the variance and is then used as a model of the actual data set, which can be mapped to the initial data to return to the corresponding size or range (Shlens 2014). In Gordillo-Orquera et al. (2018), PCA was used to model a couple of large electrical load data sets in a first approach, by reducing the dimensionality of the original data, the authors were able to forecast electricity load consumption a year in advance. You can review other case studies where solar irradiance ( $\alpha$ ) (Azadeh et al. 2008), wind speed ( $v$ ) (Zhang et al. 2018), energy demand ( $W^D$ ) (Ribeiro et al. 2016) and other RES variables are reduced by methods involving PCA.

In the pattern recognition category, the k-means clustering method is used to group the initial data (Hamerly and Elkan 2004). This method consists of an optimization where random values are tested to identify centers or means of the cluster  $k$ , each cluster is formed by finding all the values close to the center of the cluster, until the distance of the closest data to each center is similar for each group (Hamerly and Elkan 2004). The classification of typical days (TD) of the solar irradiance ( $\alpha$ ) identifying its frequency and probability through a k-means-Markov chain coupled algorithm to construct a typical year is developed in Li et al. (2017), wind speed can be modeled using a k-means clustering approach if there are gaps in the data sets as demonstrated in Yesilbudak (2016) and electricity demand is modeled using k-means in Azad et al. (2014) to understand its usage behavior.

This article presents a comparison of three data reduction approaches. In the first approach, a random sample of actual days is evaluated with results indicating clear variability and dependence of the input data, then typical days are modeled using a PCA reduction ( $TD_{PCA}$ ) and a stochastic simulation based on a pattern recognition by k-means ( $TD_{km}$ ). The optimization results are quantitatively compared using performance analysis (iterations required to achieve OD) and computation time (ToC) with data reduction using all three methods versus a no reduction approach as a benchmark.

Main contributions of the analysis performed are:

- Developing a strategy based on statistical learning algorithms (PCA and k-means) for finding a balance between data reduction and keeping performance of OD.

- Providing a strategy for implementing PCA and k-means, with a minimal loss of information over the original data, and obtaining suitable profiles of energy consumption and ambient conditions for the OD of RES.
- A contrast and comparison is presented between the use of the total data and the patterns obtained through statistical learning strategies by analyzing deviations and gaps associated with data reduction.
- The impacts of implementing statistical learning algorithms on the computational cost associated with solving a NLP model are considered.

This article is divided in seven sections: “[Introduction](#)”; “[Problem Statement](#)”, where the problem to be addressed is presented; “[Model Definition](#)”, where the optimization model is presented; “[Data Reduction](#)”, where the three approaches for data reduction are explained; “[Case Study and Computational Issues](#)”, where main features of data used are introduced; “[Results and Discussion](#)”; and, finally, “[Conclusions](#)”.

## Problem Statement

The configuration of the addressed energy system is shown in Fig. 1. The set of wind turbines dynamically produces electrical energy, like the set of photovoltaic panels, the battery bank can be charged, or it can power the load while the electricity grid is used as a backup. In the inverter, depending on the battery charging conditions and the available energy, clean energies are directed to power the

load or recharge the batteries, while the grid can only power the load.

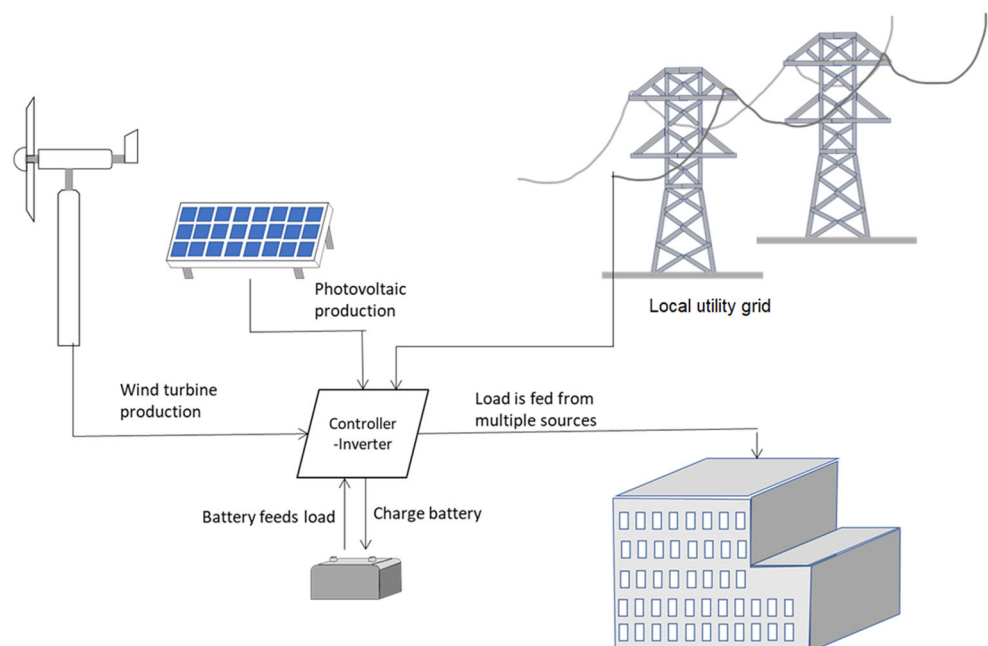
When the algorithm for calculating the optimal design is fed with large databases, the computation time and iterations for computing an optimal result increase in most cases. As can be seen in the “[Results and Discussion](#)” section, the system is also susceptible to the characteristics of the input parameters, it being possible that the number of iterations does not always increase as the input data increases, but sometimes fluctuates. This shows that removing unnecessary or outlier data, in addition to improving computational performance, can also ensure that the results are representative of the problem. Thus, data reduction becomes relevant and improves computational performance, achieving results that are similar to having large data sets involved in the calculation (see Fig. 2).

The process to have OD is shown in Fig. 3. The impact of performance of data reduction strategies can only be analyzed after the OD is obtained by the OM, however, some features can be revised to assure that the data inserted in the OM is representative of the original registers.

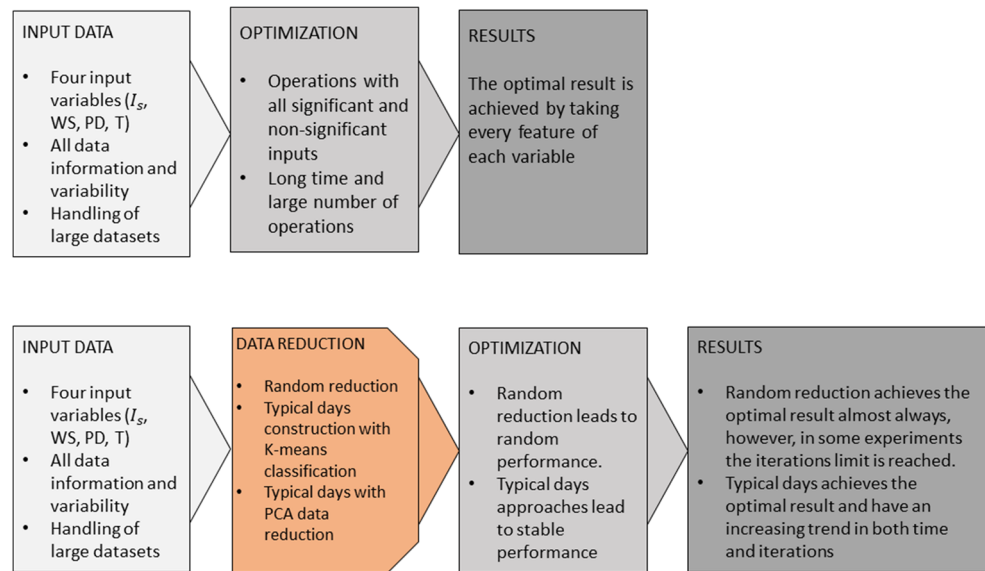
The most important parameters of OM are solar irradiance  $\alpha$ ,  $T$ ,  $W^D$  and  $\nu$ , as well as the efficiency parameter and other technical factors, as well as the energy selling price to main grid. As variables of the OM, the sizing of PV, WT and BS, many technical factors and considerations, also some of the balances that are described in the “[Model Definition](#)” section. The objective is solely the total annual cost.

In sum, the data reduction problem is addressed by considering the following points:

**Fig. 1** RES with batteries and grid backup



**Fig. 2** Data reduction implementation



- The aim is finding a balance between data reduction and keeping performance of OD. A first approach is to reduce data by sampling random real days (RRD) of a four-parameter-database, which represents no data handling nor pre-processing.
- PCA is applied to four large data sets to identify the principal component (PC) characteristics that enable to reduce their dimensionality after each PCA, while retaining the most relevant values related to variance of the analyzed data set after reconstruct the PC over the sample.
- $T D_{kms}$  of four variables are built based on the patterns identified by k-means clustering. First, each hour of the day is classified in a k-cluster, then the characteristic cluster (CC) of each hour of the day is selected, and then stochastic scenarios are constructed selecting values within standard deviation of CC.
- A comparison among the three approaches of data reduction is presented, analyzing the gaps and differences on OD results.

Therefore, the optimization problem can be stated as follows: Given a large amount of raw data sets of energy demand, ambient temperature, wind speed and solar radiation, statistical learning techniques are applied to obtain different typical days which are used to feed an optimal design model of an energy supply system based on renewable energies: photovoltaic units, wind turbines and battery storage systems. The optimal design includes the computation of the system sizing to minimize the total annual cost of the system. The analysis to determine the effectiveness of the data reduction includes a comparison with the solution obtained using the total data sets.

## Model Definition

The objective function for the NLP multi-scenario multi-period model is the total annual cost (TAC). The model inputs are: solar irradiance ( $\alpha$ ), wind speed ( $v$ ), temperature ( $T$ ) and power demand ( $W^D$ ). Among these and some other characteristics of these energy sources, the power obtained by each equipment is calculated. In both cases, the catchment area is the common parameter and is proportional to the size of the installation for each of the two: photovoltaic panels (PV) and wind turbines (WT).

## NLP Model

NLP modeling is defined by the sets  $\zeta = \{1, \dots, S\}$  for all the scenarios and  $\tau = \{1, \dots, T\}$  for all the operational periods. This approach allows considering different variations in ambient conditions and energy demand, as well as defining the optimal operational policy for each period and scenario.

## PV System Modeling

The efficiency of PV system is defined by the ambient temperature conditions (Skoplaki and Palyvos 2009).

$$\eta_{t,s}^{PV} = \eta_0^{PV} \left[ 1 - \beta_{Ref} \cdot (T_{t,s}^{amb} - T_{Ref}) \right], \forall t \in \tau, \forall s \in \zeta \quad (1)$$

Where  $\eta_{t,s}^{PV}$  is the efficiency at the period  $t$  and scenario  $s$ .  $\eta_0^{PV}$  is the design efficiency,  $\beta$  is the temperature coefficient associated with the material of the PV panel.

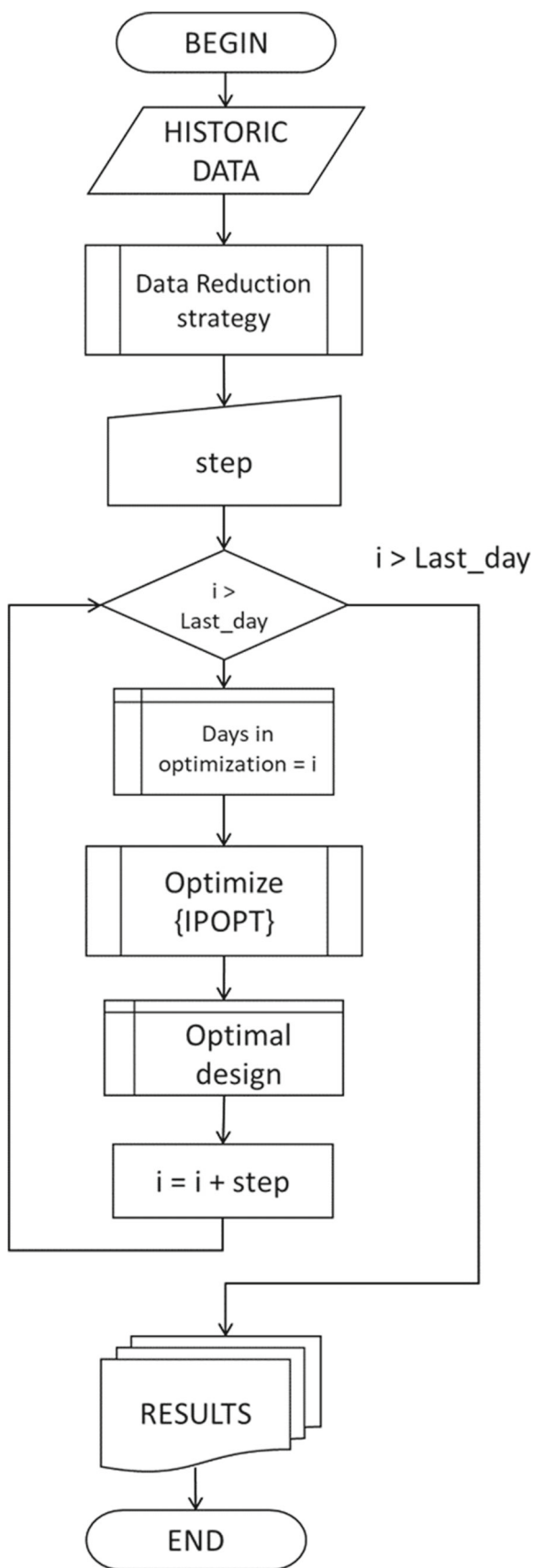


Fig. 3 Data reduction flowchart

$T^{amb}$  is the ambient temperature and  $T_{Ref}$  is the reference temperature associated with  $\eta_0^{PV}$ . A common value for  $T_{Ref}$  is 25 °C. As shown, efficiency is dependable on known data. As a consequence, it results in a known value.

Power generation ( $W^{PV}$ ) results from multiplying  $\alpha$ , efficiency ( $\eta^{PV}$ ) and the area of the PV system ( $A^{PV}$ ).

$$W_{t,s}^{PV} = \alpha_{t,s} \cdot \eta_{t,s}^{PV} \cdot A^{PV}, \forall t \in \tau, \forall s \in \zeta \quad (2)$$

The power generation can be used for meeting the energy demand of the end user, in this case an apartment building ( $W^{PVB}$ ), sent to the battery system ( $W^{PVBBS}$ ) or sold to the local utility grid ( $W^{PVG}$ ).

$$W_{t,s}^{PV} = W_{t,s}^{PVB} + W_{t,s}^{PVBBS} + W_{t,s}^{PVG}, \forall t \in \tau, \forall s \in \zeta \quad (3)$$

As shown,  $A^{PV}$  is the area for collecting solar energy and defines the size of the PV system. It is constrained by the available area in the apartment building for installing the energy system,  $A^B$ .

$$A^{PV} \leq A^B \quad (4)$$

### WT System Modeling

Power generation of WT system ( $W^{WT}$ ) is defined by  $v$  presence and the limit that is known of the flow of air bring by Betz ( $\beta$ ) coefficient. In addition,  $v$  is classified into: cut-in ( $v_{in}$ , rated  $v_r$ ) and cut-out ( $v_{out}$ ). The  $v_{in}$  is the speed at which power begins to be generated, around 4 meters per second. The  $v_r$  is that at which the nominal power of the electric generator ( $W_{nom}^{WT}$ ) is reached, around 12 meters per second.  $v_{out}$  is at which the electric generator is at risk of being damaged and therefore disconnects from the wind turbine at around 30 meters per second (García et al. 2019). Thus, there are three values of  $v$  that indicate how  $W^{WT}$  performs:

$$W_{t,s}^{WT} = \begin{cases} 0 & \text{if } v_{t,s} < v_{in} \\ \beta_{t,s} \cdot \rho^{air} \cdot A^{WT} \cdot v^3 & \text{if } v_{in} \leq v_{t,s} < v_r \\ W_{nom}^{WT} & v_r \leq v_{t,s} < v_{out} \\ 0 & \text{if } v_{t,s} \geq v_{out} \end{cases}, \forall t \in \tau, \forall s \in \zeta \quad (5)$$

Similarly to PV system, energy generated by the WT is sent to the apartment building ( $W^{WTB}$ ), the battery system ( $W^{WTBS}$ ) and the local utility grid ( $W^{WTG}$ ).

$$W_{t,s}^{WT} = W_{t,s}^{WTB} + W_{t,s}^{WTBS} + W_{t,s}^{WTG}, \forall t \in \tau, \forall s \in \zeta \quad (6)$$

### BS System Model

The energy stored ( $E^{BS}$ ) is defined by the inlets from the PV and WT systems ( $W^{BS}$ ), affected by the charge efficiency ( $\eta^{BS}$ ), and the outlets, determined by the energy sent to



the apartment building ( $W^{BSB}$ ), and the local utility grid ( $W^{BSG}$ ).  $\eta^{BS}$  is a function of the status of charge ( $0 \leq SoC \leq 1$ ).  $SoC$  is a relationship between the energy stored at the operational period  $t$  and the size of the battery system ( $E^\mu$ ). As shown, the input and output efficiencies of the battery are defined by different functions dependable on the  $SoC$ . These functions are determined by coefficients  $a$ ,  $b$  and  $c$  which are associated with the nature of the technology (Fuentes-Cortés and Flores-Tlacuahuac 2018; Ranaweera and Midtgård 2016; Yu et al. 2018).

$$\eta_{t,s}^{BSinput} \cdot W_{t,s}^{BS} - \eta^{BSoutput} \cdot (W_{t,s}^{BSB} - W_{t,s}^{BSG}), \quad \forall t \in \tau, \forall s \in \zeta \tag{7}$$

$$W_{t,s}^{BS} = W_{t,s}^{WTBS} + W_{t,s}^{PVB}, \quad \forall t \in \tau, \forall s \in \zeta \tag{8}$$

$$SoC_{t,s} = \frac{E_{t,s}^{BS}}{E^\mu}, \quad \forall t \in \tau, \forall s \in \zeta \tag{9}$$

$$\eta_{t,s}^{BSinput} = a_1 \cdot SoC_{t,s}^2 + b_1 \cdot SoC_{t,s} + c_1, \quad \forall t \in \tau, \forall s \in \zeta \tag{10}$$

$$\eta_{t,s}^{BSoutput} = \frac{a_2 \cdot SoC_{t,s}}{b_2 \cdot SoC_{t,s} + c_2}, \quad \forall t \in \tau, \forall s \in \zeta \tag{11}$$

$$E^\mu \geq E_{t,s}^{BS}, \quad \forall t \in \tau, \forall s \in \zeta \tag{12}$$

**Power Supply**

Energy demand ( $W^D$ ) is met using energy from the PV, WT, BS and the utility grid.

$$W_{t,s}^D = W_{t,s}^{PVB} + W_{t,s}^{WTB} + W_{t,s}^{BSB} + W_{t,s}^{GB}, \quad \forall t \in \tau, \forall s \in \zeta \tag{13}$$

**Total Annual Cost (TAC)**

The TAC is the minimized objective function. It is determined by the Capital cost of equipment ( $CCost$ ), operation and maintenance cost ( $OMCost$ ) and the cost of energy from the utility grid ( $PCost$ ). In addition, incomes from energy sales ( $PInc$ ) are included as a negative term in the expression.

$$TAC = CCost + OMCost + PCost - PInc \tag{14}$$

Capital cost ( $CCost$ ) is defined by variable cost ( $v$ ) associated with the size of the equipment as well as the fixed cost of each of the technologies ( $\phi$ ) and the annualization factor  $\psi$ .

$$CCost = \psi \cdot (\phi^{BS} + \phi^{PV} + \phi^{WT} + v^{PV} \cdot A^{PV} + v^{WT} \cdot A^{WT} + v^{BS} \cdot E^\mu) \tag{15}$$

$OMCost$  results from multiplying the unit O&M cost for each equipment by the total energy produced by the PV system and the total energy stored in the BS during all the operational periods ( $\Theta$ ) associated with the annual scenarios and operational periods.

$$OMCost = \Theta \cdot \sum_{s=1}^S \sum_{t=1}^T (\Upsilon^{OMPV} \cdot W_{t,s}^{PV} + \Upsilon^{OMWT} \cdot W_{t,s}^{WT} + \Upsilon^{OMBS} \cdot W_{t,s}^{BS}) \tag{16}$$

Similarly, the cost of external energy from the utility grid ( $PCost$ ) is computed using the unit cost of energy from the grid based on a scheduling tariff  $\Upsilon^G$ .

$$PCost = \Theta \cdot \sum_{s=1}^S \sum_{t=1}^T \Upsilon_t^G \cdot W_{t,s}^{GB} \tag{17}$$

The incomes ( $PInc$ ) are computed considering the unit price  $\varpi^P$  of energy sent to the end user and the local utility grid.

$$PInc = \Theta \cdot \varpi^P \cdot \sum_{s=1}^S \sum_{t=1}^T (W_{t,s}^{GS} + W_{t,s}^{BSG}) \tag{18}$$

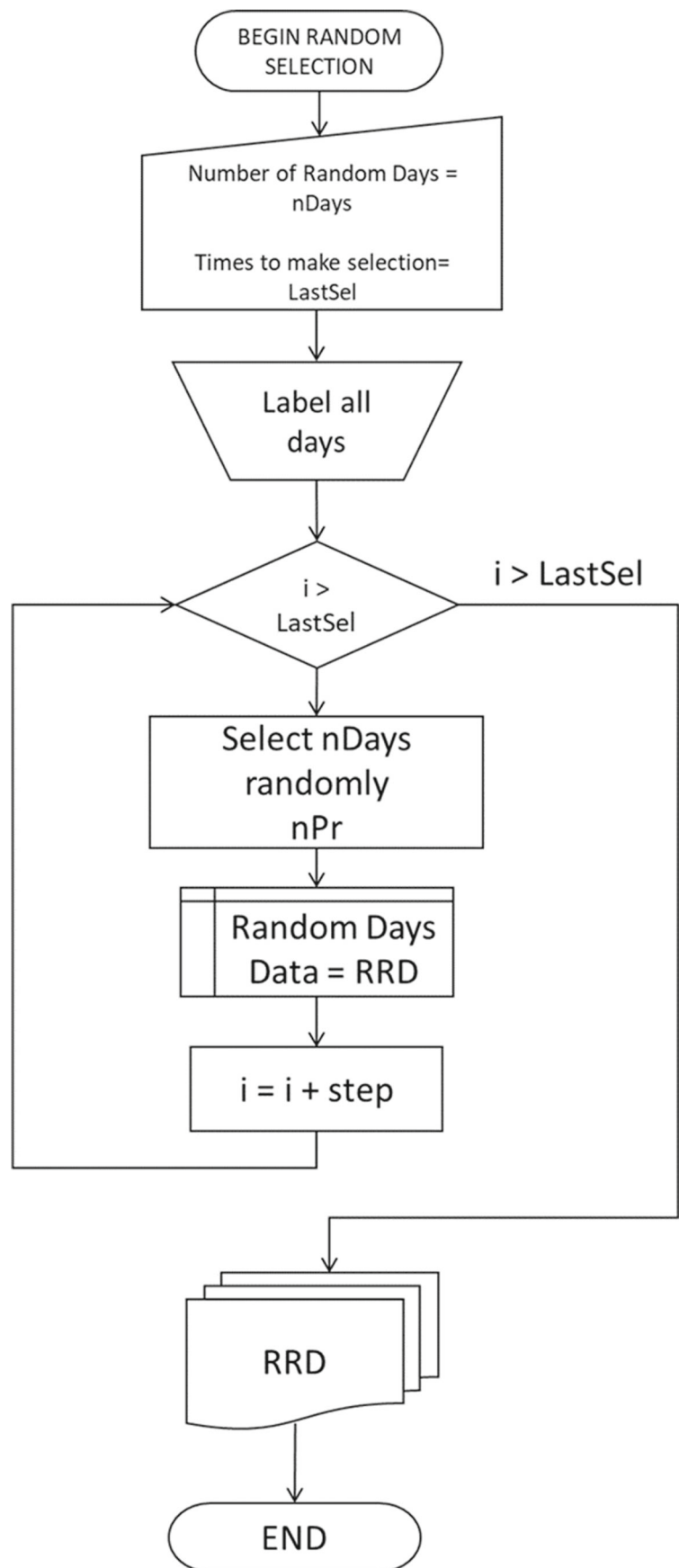
**Data Reduction**

Seeking to obtain an OD similar to that obtained when all the data is fed into the OM, three approaches to reduce the sample size were used. First, real days were randomly selected from the real sample, for which it was decided to select 60 days for two different seasons. Second, PCA was used to reduce the dimensionality of the sample that was arrayed in a certain way to be able to train, test and rebuild the principal components. Finally, k-means was used to find the general pattern of the behavior of each variable and typical days were generated from the characteristics of the clusters found. Each technique is described in detail below.

**Selecting Random Real Days (RRD)**

The first approach to reduce the input data and testing the performance and results obtained in the OD of RES was to randomly take  $r$  real days (RRD) as a sample from the database. All variables correspond to the same RRD. In a first selection, the RRD were taken from the total days of the year (365 days), later it was decided, based on the atmospheric conditions of different seasons of the year, to divide the population into two seasons: spring-summer and autumn-winter (see Fig. 4).

Fig. 4 RRD selection process



RRD were selected using a free selection and the number of permutations corresponds to the known formula:

$${}_n P_r = \frac{n!}{(n-r)!} \quad (19)$$

where  $r$  is the number of days to be selected and  $n$  is the total data or days from which the selection is made. Five different selections were made to have a benchmark, it should be noted that in each RRD selection days that have been selected on other occasions may appear. So, there are  $r$  RRD for two seasons, where each selected day had the same probability of being selected as any other day within the season. With this, it is expected to observe a performance of the model such that the results will begin to stabilize as there is a greater number of input data.

In the “[Results and Discussion](#)” section, the comments about this topic will be expanded; however, it can be anticipated that the results are highly dependent on the RRD used and as they were selected by a simple random sample, days with atypical conditions may be included in the calculation, which can lead to a variable behavior of the OM, even when there is “enough” data.

### Typical Days with PCA Data Reduction

The PCA process consists on identifying little relevant information in a database of  $n$  dimensions (Jolliffe 1986). Using the specific programming package in Julia called PCA, which belongs to MultivariateStats package (Statistics 2014), three functions that are found within PCA are used to correctly model the input data for the OD (see Fig. 5).

Previously to initiate PCA, the data of the four variables are standardized by  $X = \frac{(x_i - \bar{x})}{s}$  where  $\bar{x}$  is the mean of sample,  $s$  is the standard deviation and  $x_i$  is the  $i^{th}$  observation. This is a usual standardization use in statistical analysis (Forkman et al. 2019). Afterwards, the option to model and reconstruct approximately the original data is presented in Julia documentation as follows (Blaom et al. 2020): Given a PCA model  $M$ , one can use it to transform observations into principal components, as

$$\mathbf{y} = \mathbf{P}^T(\mathbf{x} - \boldsymbol{\mu}) \quad (20)$$

or use it to reconstruct (approximately) the observations from principal components, as

$$\tilde{\mathbf{x}} = \mathbf{P}\mathbf{y} + \boldsymbol{\mu} \quad (21)$$

here,  $\mathbf{P}$  is the projection matrix.

To complete the transform-reconstruction process, some functions of the PCA package must be used. First, a training step for the PCA model must be performed using part of the original data. When traversing this resulting matrix, each subsequent principal component has less representative of the original variance than the previous one, so that PC2 is

much less relevant than PC1 and PC3 is much less relevant than PC2

The next step is to transform the results of the training step by a matrix operation with part of the original data to generate a matrix of  $n$  number of PCs, the first of these being the one that preserves the most information from the original data, this stage is called the testing stage.

Finally, in the reconstruction step the results obtained in the second process is projected to some original data reserved to perform test to the model to calculate an approximation in the same units and range (Statistics 2014).

### Typical Days with k-means Pattern Recognition

When k-means is run on a database, each variable is treated as a vector, since this process examines each vector separately. The k-means technique is an optimization in which the data sample is separated into clusters according to some random values called cluster centers that change until approximately the same distance of data near to the center of cluster do not change (Hamerly and Elkan 2004).

The k-means technique is a classical method for clustering or vector quantization. It produces a fixed number of clusters, each associated with a center (also known as a prototype), and each data point is assigned to a cluster with the nearest center. From a mathematical standpoint, k-means is a coordinate descent algorithm that solves the following optimization problem (Statistics 2012):

$$\text{minimize} \sum_{i=1}^n |x_i - \mu_{z_i}|^2 \text{ w.r.t. } (\boldsymbol{\mu}, \mathbf{z}) \quad (22)$$

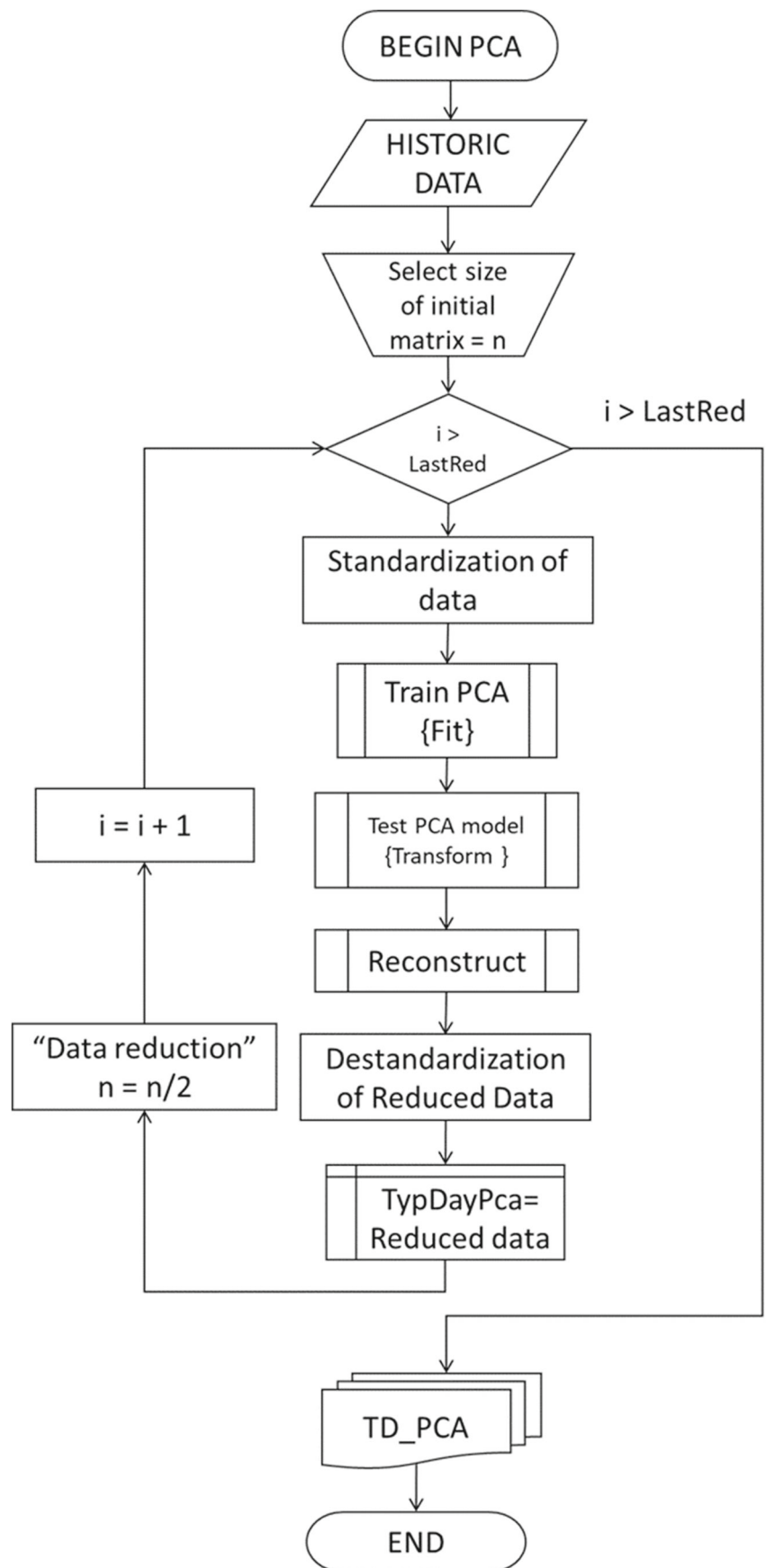
Here,  $\mu_k$  is the center of the  $k$ th cluster, and  $z_i$  is an index of the cluster for  $i^{th}$  point  $x_i$ .

Once the data has been separated into clusters each real datum is labeled with the cluster to which it belongs. After classifying the data, a probabilistic analysis is made to determine which is the most common cluster for each hour of the day, in this way it is assumed that the cluster with the highest frequency for a specific hour of the typical day ( $TD_{km}$ ) is its characteristic cluster (CC).

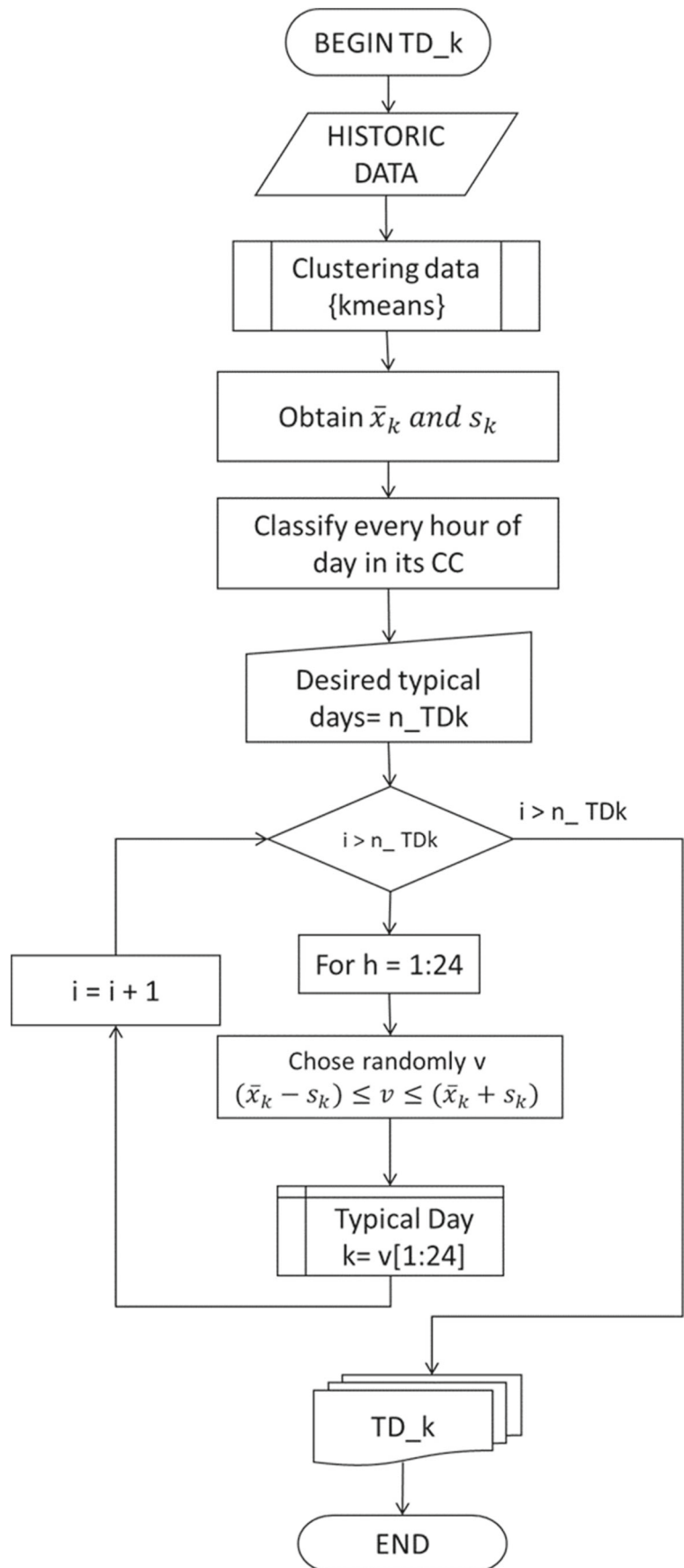
In order to define the  $TD_{km}$ 's, it is assumed that the most common values for each hour of  $TD_{km}$  are those that are within the CC, therefore any value within the common values of the CC would be an acceptable value of a typical day. Thus, the  $TD_{km}$  indicators are the center or mean of the CC and its standard deviation. Therefore, to construct the typical days using this methodology, a scenario simulation can be performed within these parameters for each hour of a  $TD_{km}$  for any of the four variables, obtaining  $n$   $TD_{km}$  with the most common characteristics of the analyzed data. Each  $TD_{km}$  is a probabilistic typical scenario within the parameters of corresponding CCs.



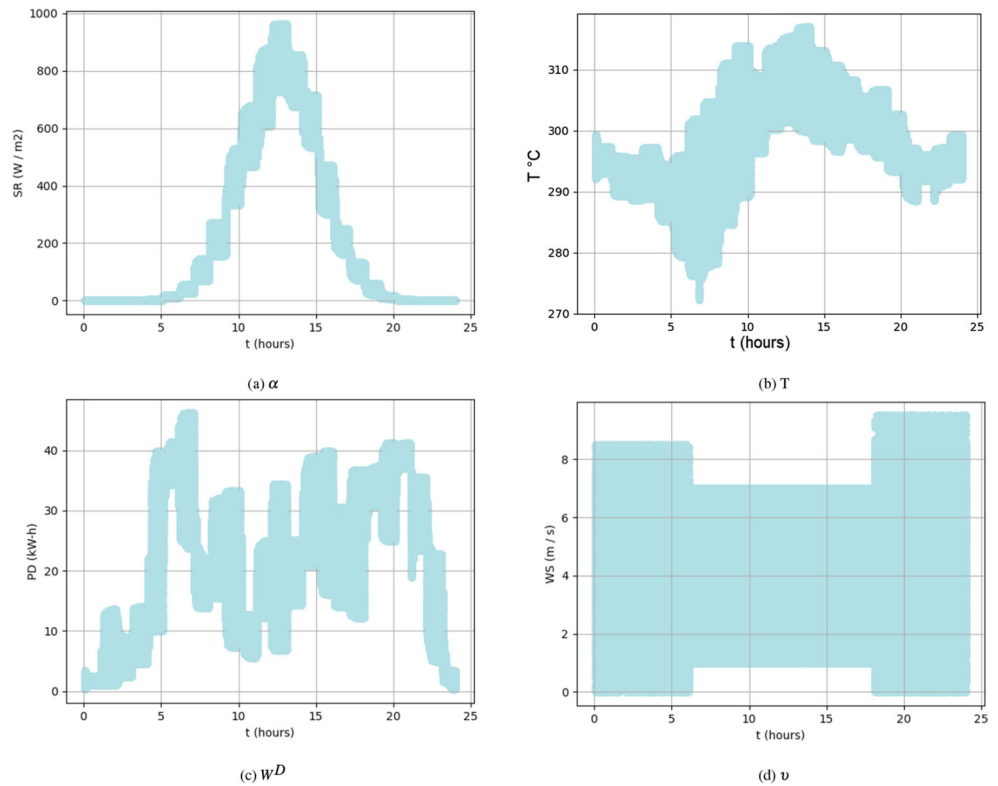
**Fig. 5** Flowchart of PCA reduction process, selecting  $TD_{PCA}$



**Fig. 6** Flowchart of typical days process with k-means pattern recognition, obtaining  $TD_k$



**Fig. 7** Real data, four variables



The scenario simulation consists in take a random value that is in the interval:

$$(\mu_k - s_k) < v_k < (\mu_k + s_k) \tag{23}$$

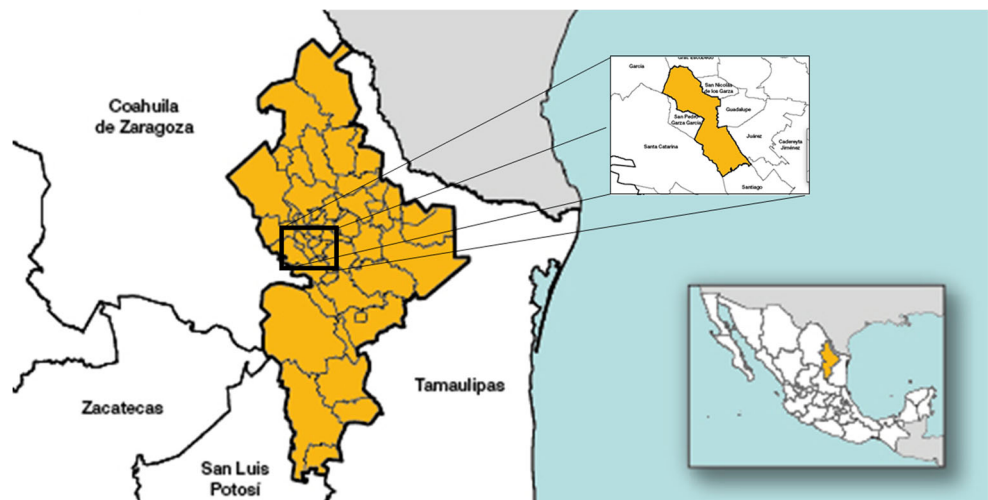
For each variable, where  $\mu_k$  and  $s_k$  are features of the CC of that hour and  $v_k$  is a typical value for an hour of the  $TD_{km}$ . This process can be observed in Fig. 6.

### Case Study and Computational Issues

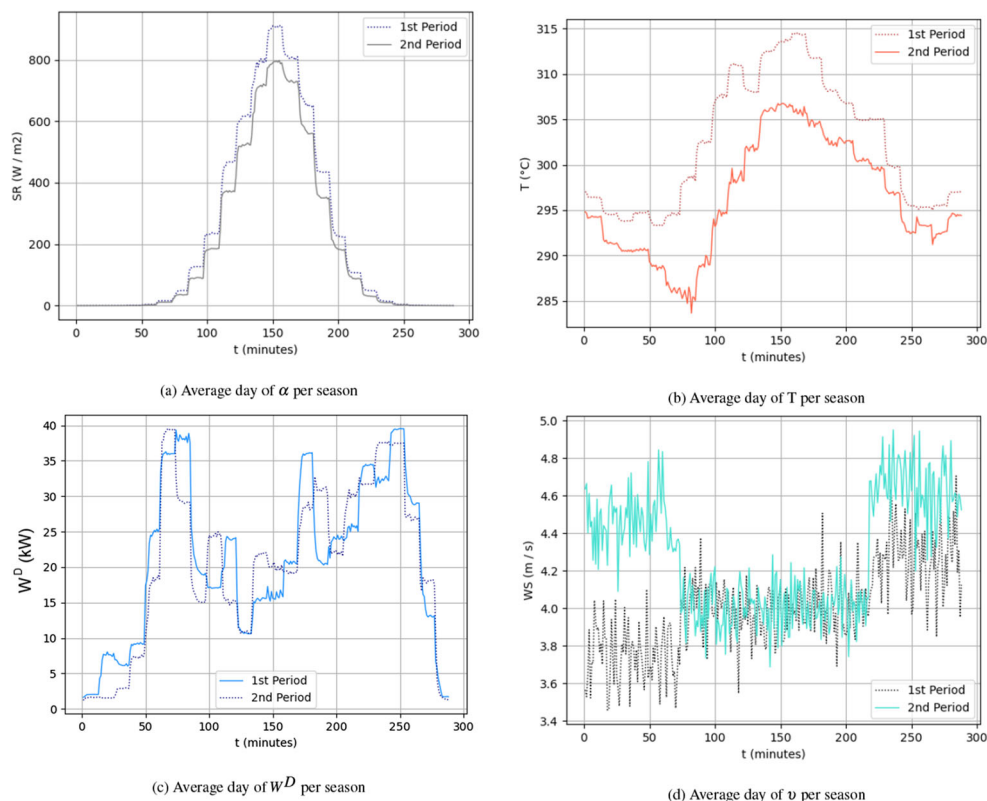
In the case study presented, the electrical load of a residential building in the northwestern region of Mexico is analyzed (see Fig. 8), which location is between the coordinates: longitude  $100^\circ 26' 31.20''$  W and  $100^\circ 10' 01.20''$  W, latitude  $25^\circ 28' 55.56''$  N and  $25^\circ 47' 50.28''$  N.

The main environmental conditions in the site are that the available wind energy is relatively scarce, while solar radiation and temperature have a relatively consistent

**Fig. 8** Location of Monterrey, Nuevo León, México. Obtained from Inegi (2017)



**Fig. 9** Real data average day for two seasons: spring-summer and autumn-winter



behavior throughout of the year. In this study, four variables are analyzed: temperature, solar radiation, wind speed and power demand. There are databases of the four parameters for a full year with measurements every 5 minutes, however, they have been condensed into hourly averages to feed the OM. In Fig. 7 these data are shown (See Fig. 8 for geographic allocation).

In Fig. 9 the average by hour of all variables is shown for two seasons of the year, spring-summer and autumn-winter. Where can be observed that for  $T$ ,  $\alpha$  and  $W^D$  patterns are obvious and there is a clear gap or change between these two seasons, however,  $v$  has a high randomness in both seasons and does not present a “smooth behavior”. In Table 1 the costs and parameters that are included in the calculation of OD are listed.

$\Upsilon^G$  does not change during the year (see Fig. 10). No other changes in time are considered for these variables.

Since in the optimization problem the objective function to minimize is TAC, the results include the values of TAC, the size of photovoltaic panels ( $W^{PV}$ ), wind turbines ( $W^{WT}$ ) and battery system, as well as the time of computation (ToC) and the number of iterations (Iter) required in each optimization process.

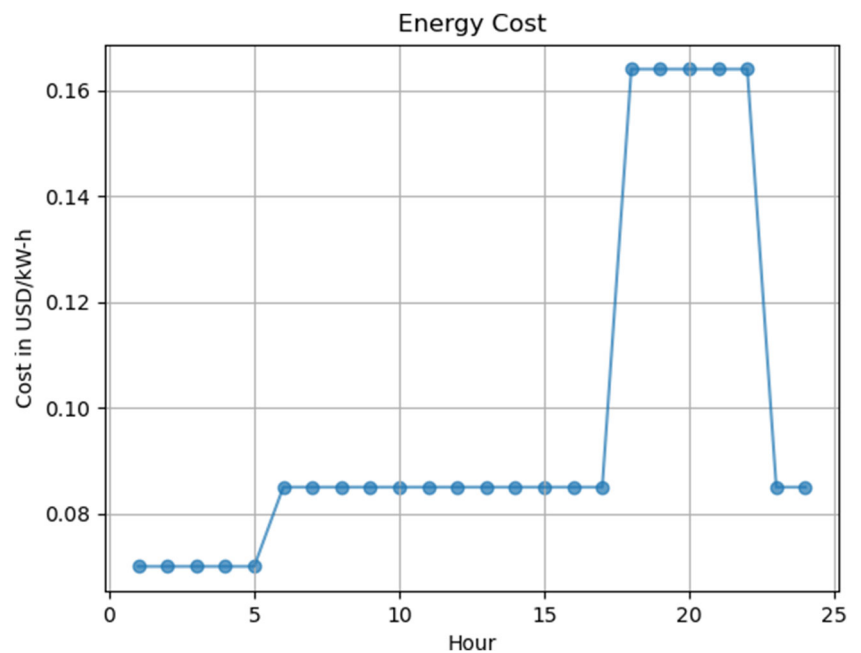
The NLP model was implemented in the mathematical language Julia, using the optimization environment JuMP and the solver Ipopt (Dunning et al. 2017; Wächter and Biegler 2006) which is commonly used to solve nonlinear

optimization problems using the interior point filter with a line-search algorithm (Breuer et al. 2018). As Ipopt is a local solver, for determining a suitable initial value for all the experiments, an algorithm for seeking feasibility, based on bootstrapping, was implemented (Chinneck 2008). In addition, the Multivariate Julia package was used to perform the principal component analysis (PCA) (Statistics 2014). Meanwhile, the package clustering was employed to carry out the CC classification (Statistics 2012). All OD were carried out in a core i5 2nd generation processor, with 4 GB of RAM, with Windows 10® operative system.

**Table 1** Values of cost variables

Parameter	Value
Fixed cost PV cells	80 (\$)
Fixed cost Battery system	30 (\$)
Variable cost PV cells	1400 (\$/kW)
Variable cost Battery system	25 (\$/A-hr)
Variable cost Wind Turbine	1700 (\$/kW)
Fixed cost Wind Turbine	100 (\$)
Operation and maintenance unit cost for PV cells	0.0003 (\$/kW)
Operation and maintenance unit cost for WT	0.03 (\$/kW)
Operation and maintenance unit cost for BS	0.00012 (\$/A-hr)
Unit cost of power	0.08 (\$/kW-hr)
Annualization factor	0.24 (\$/kW-hr)

**Fig. 10** Costs of energy from grid in USD/kW-h

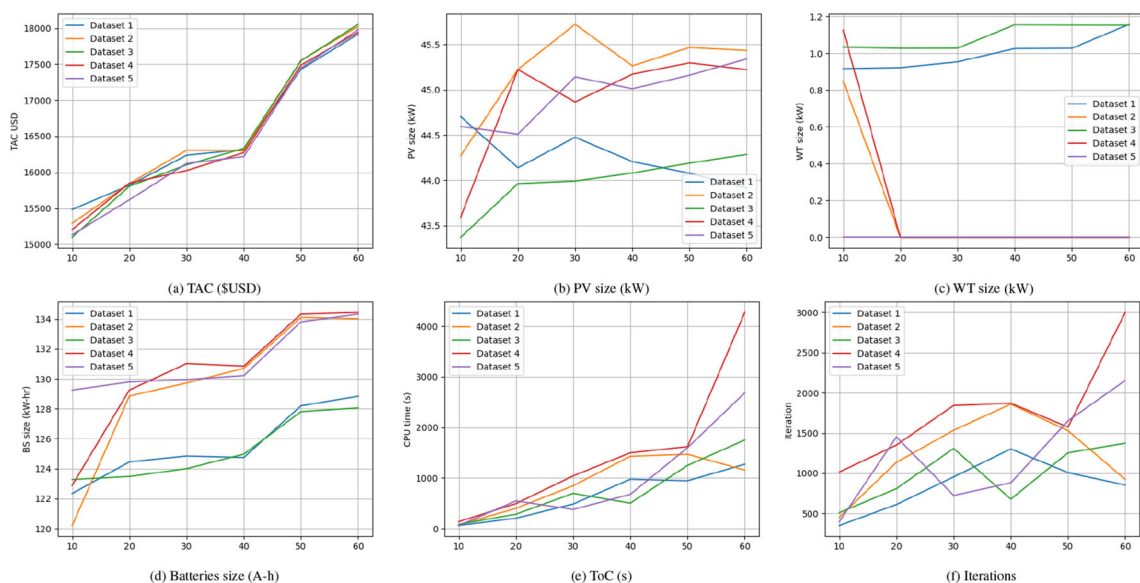


### Results and Discussion

Results demonstrated that the OD for RES are highly dependent of the input data. Although the model can calculate solutions from as little information as 1 day of input data, it is advisable to include as much information as possible to achieve results that are closer to the actual operation. However, including all the available data produces long computation times, in addition to using infrequent data that could skew the results. Thus, three strategies to reduce input data were performed to original data and the OD resultant are presented below.

**Random Real Days (RRD)** For the first approach (random selection of real days) 60 RRD for two seasons of the year are included in 5 different samples. With these sets, several optimizations are performed where the input RRD increases, and their results are recorded (see Fig. 11 where are plotted: (a) TAC, (b) PV, (c) WT, (d) BS, (e) ToC and (f) Iterations).

In Fig. 11 results for RRD indicate that in each RRD sample, as the number of days included in the optimization increases, this does not ensure that results of OD converge to a unique value, nor the (a) TAC, nor the (b)  $W^{PV}$ , nor the (c)  $W^{WT}$  nor the (d)  $W^{BS}$ . It is noticed that the green



**Fig. 11** OD results performance with 5 sets of RRD



**Table 2** RRD comparison with benchmark (60 RRD per season)

Result	Benchmark	RRD 1	RRD 2	RRD 3	RRD 4	RRD 5
TAC (USD)	17998.48	− 0.45% ↓	0.14% ↑	0.31% ↑	− 0.30% ↓	− 0.10% ↓
PV size (kW)	45.36	− 3.20% ↓	0.16% ↑	− 2.38% ↓	− 0.31% ↓	− 0.05% ↓
WT size	WT off	<b>WT on</b>	WT off	<b>WT on</b>	WT off	WT off
Battery size	132.53	− 2.79% ↓	1.13% ↑	− 3.38% ↓	1.44% ↑	1.36% ↑
ToC (s)	2915.25	− 1643.15 ↓	− 1758.90 ↓	− 1164.44 ↓	1356.50 ↑	− 233.69 ↓
Iter	901	− 47 ↓	23 ↑	473 ↑	2099 ↑	1253 ↑

and blue lines retain the WT reduce the size of PV and BS. In addition, (e) ToC has a peaking trend as increasing the number of RRD, however some fluctuations can be observed, meanwhile (f) Iter have a variant behavior with each change in RRD and not a clear trend is appreciated. However, the TAC increases and reaches values close to the benchmark (less than 0.5% of difference),  $W^{PV}$  reaches 45 kW of main power,  $W^{WT}$  oscillates around 1 kW until it is discarded in some cases, although in some optimizations it remains,  $W^{BS}$  ranges between 130 and 140 A-hr, which is close to those of the benchmark's. In Table 2 the results of 5 samples with 60 RRD per seasons are compared to all-data optimization.

**Typical Days with PCA Data Reduction ( $TD_{PCA}$ )** The original information is presented in matrix form and then processed by the PCA, with two columns, in which the first represents half of the data and the second one is the remaining data. Previously, the data were separated into two seasons, spring-summer and autumn-winter. These two use 4608 data corresponding to 192 days, with an overlap between the end of the first and the beginning of the second season. This 192 actual days sample enhance the data reduction process due to it can be divided by two and the result is an integer for each new process. In this way, in the first iteration the matrix of the real data has two columns and 2304 rows, 50% of the data is used to train the PCA model obtaining a certain number of principal components (PC), in our case only  $PC_1$  and  $PC_2$  are computed. Then, the

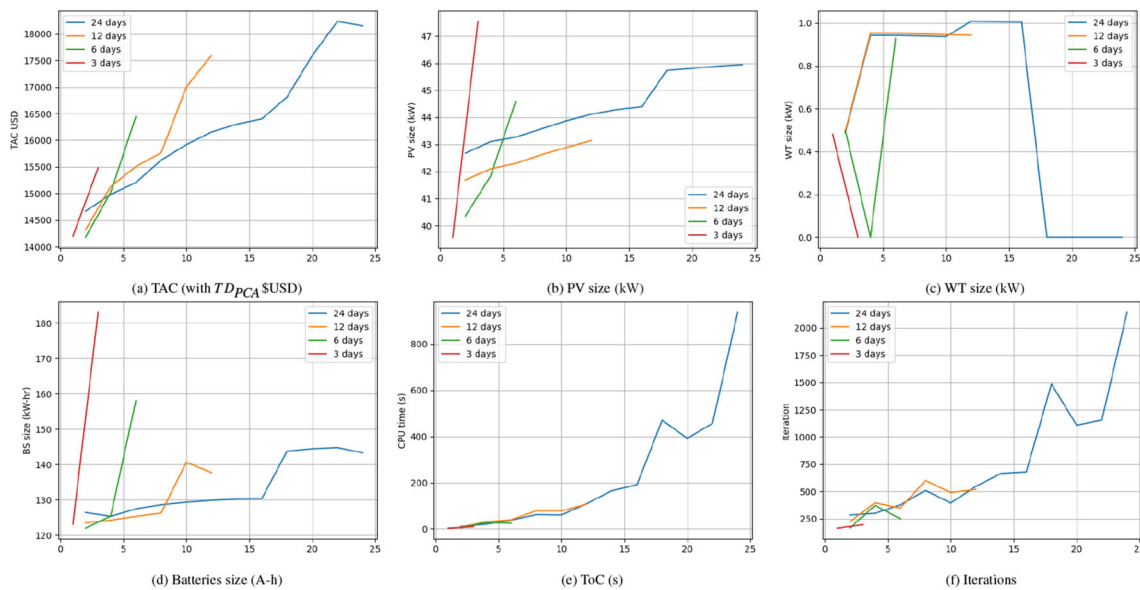
model is tested on the remaining data and, finally, function PC are re-size the to original range. At this point, as  $PC_1$  is solely considered, only half of initial number of data is conserved, now reconstructed, however, this data retains a high percentage of the variance of each variable data as found in the “PCA ratio” of  $PC_1$  (see Table 3). The results of this reduction are treated as a new sample and are arranged in a  $2 \times 1152$  matrix, then, they are used as the new training set and are reduced once more in half on the next iteration and so on until a set of only three  $TD_{PCA}$  of each variable is reached. To achieve this, six reductions were necessary. Please observe that  $\nu$  retain 100% of the variance after the whole process in both seasons, as total PCA ratio ( $ratio_{total}$ ) is calculated by multiplying PCA ratio of every reduction, while the lowest  $ratio_{total}$  is 97.52% for the  $W^D$  in season 1. These values indicate that a high variance of the variables are preserved in the reduction process, even when only three  $TD_{PCA}$  remain.

The results of the reduction have shown that when there is little data even, when they are significant, the OD results are not very reliable; thus, the more the number of  $TD_{PCA}$  increases, a consistent performance is reached (ToC and Iter) and with a certain proximity to benchmark results.

In Fig. 12 the smallest  $TD_{PCA}$  sets are presented, with 3, 6, 12 and 24  $TD_{PCA}$ . Given the characteristics of PCA, sometimes the reconstructed model can give results outside the physics of the problem, so a restriction to reconstructed  $\alpha$  was included: there cannot be negative data.

**Table 3** PCA ratio for each reduction

Reduction	Days conserved	$\alpha$ season1	$\alpha$ season2	T season1	T season2	$W^D$ season1	$W^D$ season2	$\nu$ season1	$\nu$ season2
1	96	99.88%	99.82%	100%	100%	100%	100%	100%	100%
2	48	99.95%	99.86%	99.67%	100%	99.14%	99.060%	100%	100%
3	24	99.99%	99.77%	99.79%	100%	99.61%	100%	100%	100%
4	12	99.99%	99.74%	99.67%	100%	99.48%	100%	100%	100%
5	6	99.99%	99.78%	99.43%	100%	99.27%	100%	100%	100%
6	3	99.99%	99.95%	99.18%	99.96%	100%	99.77%	100%	100%
	<b>Total PCA ratio</b>	<b>99.79%</b>	<b>98.92%</b>	<b>97.75%</b>	<b>99.96%</b>	<b>97.52%</b>	<b>98.83%</b>	<b>100%</b>	<b>100%</b>



**Fig. 12** Results of  $TD_{PCA}$  reduction, four reduction sets

Data reduction with PCA retain more than 97% of the variance of the original data sample for each reduction, according to the internal evaluation and the PCA ratio Fig. 3. However, if the most reduced model is used (3  $TD_{PCA}$  for each season), results of OD are highly variant, this can be observed in Fig. 12 with the red line, (a) TAC, (b) PV, (c) WT, (d) BS are “volatile”, as the Reduction is more conservative (blue line is the most conservative reduction), all values start to reach those of the benchmark’s see the 24  $TD_{PCA}$  reduction. The behavior of the model demonstrates some trends the more  $TD_{PCA}$  it receives: more Iter and ToC are required to calculate OD; TAC reaches a benchmark-like value, as well as all the other variables,  $W^{PV}$  approaches to 45 kW,  $W^{WT}$  is finally discarded, and  $W^{BS}$  lies around to 240 A-hr. These results are consistent to those of the benchmark, as presented in Table 4, where the reductions are labeled according to  $TD_{PCA}$ .

**Typical Days with Pattern Recognition by k-Means ( $TD_{km}$ )**

For the third approach, the number of clusters to be considered for each variable was empirically tested, and

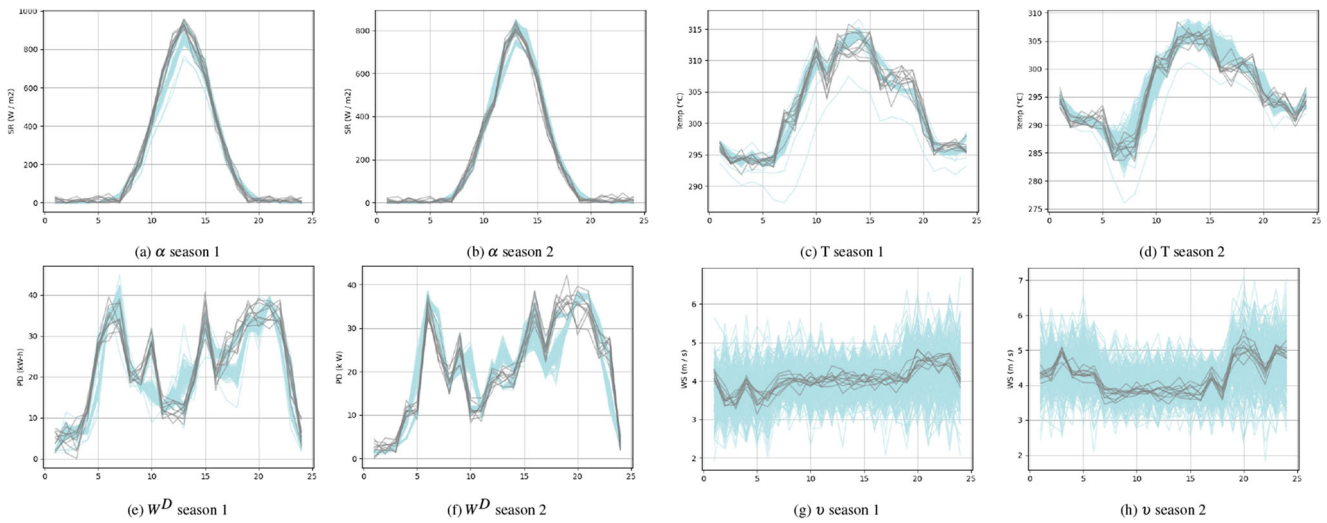
heuristically,  $\alpha$  considers 8 clusters, while  $W^D$ , T and  $v$  only 5 clusters. The data was not treated previously to the clustering process. As describe in the “Typical Days with k-means Pattern Recognition” section,  $TD_{km}$  are constructed by using the center and the standard deviation of each CC, thus, sets of 60  $TD_{km}$  were simulated for two seasons, in Fig. 13  $TD_{km}$  over total data can be observed, note that  $TD_{km}$  of T,  $\alpha$  and  $W^D$  are close to the area that actual data occupies, meanwhile,  $TD_{km}$  of  $v$  are concentrated around the mean and do not show a recognizable pattern, actual  $v$  appear to have a “white-noise-like” behavior.

In Fig. 14 the ODs obtained with this strategy are presented as well as their performance. Three sets of  $TD_{km}$  with 60 days are presented. Their behavior is consistent, notice that in all cases the WT remains, although it does not significantly changes the other results.

In Fig. 14 (a), (b), (c) and (d) have some interesting features, additionally, for one of the sets, while the number of days increases, suddenly there is an unexpected peak of (e) ToC and (f) Iter with few  $TD_{km}$ , nevertheless, all the other OD have a rampant ToC as  $TD_{km}$  increase, and the

**Table 4**  $TD_{PCA}$  comparison with benchmark (lowest number of TDs)

Result	Benchmark	3 $TD_{PCA}$	6 $TD_{PCA}$	12 $TD_{PCA}$ TD	24 $TD_{PCA}$
TAC (USD)	17998.48	-13.98% ↓	-8.63% ↓	-2.23% ↓	0.85% ↑
PV size (kW)	45.36	4.82% ↑	-1.71% ↓	-4.89% ↓	1.28% ↑
WT size	WT off	WT off	<b>WT on</b>	<b>WT on</b>	WT off
Battery size	132.53	38.09% ↑	19.26% ↑	3.91% ↑	8.13% ↑
ToC (s)	2915.246	-2905.06 ↓	-2889.885 ↓	-2797.475 ↓	-1902.162 ↓
Iter	901	-706 ↓	-630 ↓	-334 ↓	1376 ↑



**Fig. 13** Contrasting 10  $TD_{km}$  over original data per hour

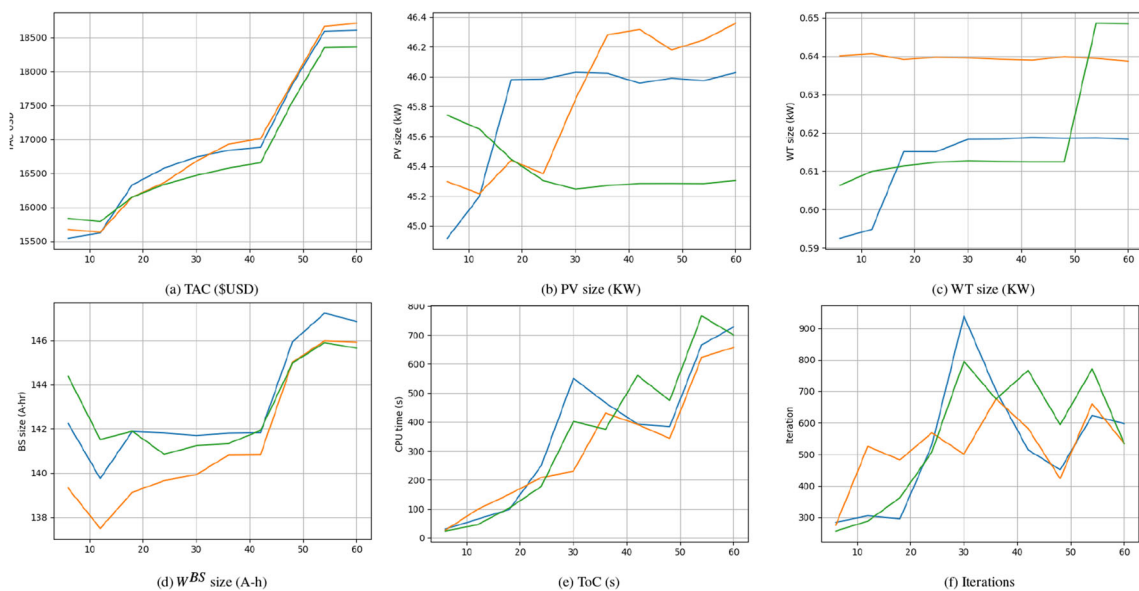
results rapidly stabilize. This could be due to the stochastic simulation used to generate  $TD_{km}$  and the presence of the WT in every optimization.

While for a greater number of typical days the results of TAC,  $W^{PV}$ ,  $W^{WT}$  and  $W^{BS}$  stabilize quickly, as expected, it is striking that the number of iterations required and the ToC have a somewhat volatile behavior that is directly proportional between them. This fact is worth commenting because given the origin of the  $TD_{km}$ , it could be expected that the ToC and Iter would simply increase with each increment in the number of days, however, this may be due to the model obtained for the wind speed that does not

seem to represent the real behavior of the wind and the randomness added by the stochastic simulation.

Some inconsistencies are found, when comparing these results to the benchmark, despite the results proximity, shown in Table 5. The criteria to select the CC to trace a general pattern has reduced significantly the variability in all cases of the modeled parameters, however,  $v$  has a behavior more “uncertain” than the model. The  $TD_{skm}$  of  $v$  favor the  $W^{WT}$  result that are close to 0.6 kW, which differs from all previous results, where WT is discarded.

The modeling of  $v$  has the main obstacle of the CC frequency, due to the fact that the CC for each hour of  $TD_{km}$



**Fig. 14** Results and performance of OD with  $TD_{km}$

**Table 5**  $TD_{km}$  comparison with benchmark 60 TD per season

Result	Benchmark	$TD_{km}$ 1	$TD_{km}$ 2	$TD_{km}$ 3
TAC (USD)	17998.48	3.40% ↑	3.96% ↑	2.01% ↑
PV size (kW)	45.36	1.46% ↑	2.20% ↑	− 0.13% ↓
WT size	WT off	<b>WT on</b>	<b>WT on</b>	<b>WT on</b>
Battery size	132.53	10.81% ↑	10.11% ↑	9.90% ↑
ToC (s)	2915.25	− 2187.08 ↓	− 2257.48 ↓	− 2215.31 ↓
Iter	901	− 304 ↓	− 367 ↓	− 368 ↓

has a frequency about 20%, which indicates that any hour may be classify in any cluster. It is evident that the pattern recognition of the  $v$  needs to be enhanced, or to be modeled by another approach. However, TAC,  $W^{PV}$  and  $W^{BS}$  results are similar to the benchmark's.

In addition, there is the fact that the optimized size of the wind turbine oscillates around 0.5 and 0.6 kW for all sets of  $TD_{km}$ . This fact is surely due to the modeling of the wind in which when determining the CC for different hours of the typical day, the production capacity of this variable has been overestimated on occasions (see Fig. 13 where the  $TD_{km}$  and the actual data are shown).

As shown in Tables 2, 4 and 5, it is possible to reach a solution near to the benchmark reducing TD without a significant variation on the results mitigating computational costs. However, it is important to consider the scale and limits of the case study used. In addition, using different statistical learning algorithms, in this case PCA and k-means have a different effect in the optimal design of the energy system. In addition, the use of limited computational resources leads to explore different strategies of data reduction for minimizing misinformation and design gaps.

### Conclusions

The handling of large databases is a problem that produces a high computational cost and high calculation times for the optimal RES design. In this work, a comparison between three methods to reduce the input data in an optimization model is presented.

To perform the data reduction to feed the OM, the first approach is the simplest and consists of randomly sample real days from the database, this strategy allows reducing the amount of input data, at the same time that having enough data the RRD achieve results similar to those of the all-data optimization. This method turned out to be quite dependent on the selected RRD and presented rather random results. Mainly, the biggest difference with the result of the benchmark was that the sizing of the wind turbines remains around 1 kW, while in other cases WT is discarded.

The second approach is to reduce information using the PCA, which proved to be quite reliable for modeling the four variables involved in the calculation, since a total PCA ratio above 97% for the four variables is obtained. Also, a quite stable behavior of the results is observed, although the maximum reduction reach three  $TD_{PCA}$ , it may be necessary to be more conservative and reduce sample down to 24  $TD_{PCA}$  to assure a smooth behavior of the results as TD in the OD increase. If the smallest samples are used, some error in the calculation of the OD may occur, although when the fewest information was used (three  $TD_{PCA}$ ) similar results to the all-data optimization are achieved, while when using 24  $TD_{PCA}$ . The third approach is to use a k-means pattern recognition technique as a basis for obtaining typical days within the most usual values. The  $TD_{km}$  use the CCs for each hour, which are determined according to the frequency with which they appear, and then the means and standard deviation of those clusters are used to obtain the typical values for that hour of  $TD_{km}$ . The results showed limitations because, e.g., despite having a stable behavior, wind energy seems to be overestimated because the wind turbine always remains as part of the optimal solution with values ranging between 0.5 and 0.6 kw, differing from results of benchmark, surely, this is due to the selection of the CCs for  $v$ . All CC of  $v$  had low frequencies, approximately 20%, this indicates that the five clusters used appear throughout the year in a similar proportion, so although some of them are selected as the CC for a certain time on the  $TD_{km}$  pattern, about 78% of the remaining data is in the other clusters evenly distributed. For this reason, it is difficult to ensure that the behavior of wind in  $TD_{km}$  will represent accurately actual  $v$  at the studied site.

From the three approaches used, the one that achieves the most adequate results is  $TD_{PCA}$  since similar results to those of benchmark optimization are achieved and also presents a low variation of the results. Meanwhile, RRDs produce highly variant ODs, where WT is relevant in some solutions while in others does not appear, which affects other variables and objective function. Finally,  $TD_{km}$  has issues since it overestimates the availability of wind energy,

this is because when simulating the  $TD_{km}$  for  $\nu$ , the CCs are almost as frequent as the clusters that are discarded.

As future work, some enhancing action to this work are anticipated, as a better selection of the CC to trace a better pattern of  $\nu$  in the  $TD_{km}$  approach, this may need a probabilistic selection among some CC to each hour. To do this, Markov Chains can be used. Another projects may concentrate to analyze the BS usage or operation to identify patterns of interest.

**Symbol**  ${}_n P_r$ , Permutation formula;  $\alpha$ , Solar irradiance;  $\beta$ , Temperature coefficient of PV panel;  $\eta^{BS}$ , Batteries efficiency;  $\eta^{PV}$ , PV efficiency;  $\mu$ , The mean vector;  $\mu_k$ , Center of the  $k$ th cluster;  $\nu$ , Variable cost of equipment;  $\phi$ , Fix cost of each technology;  $\tau$ , time periods (hours);  $\Theta$ , Annual periods;  $\tilde{x}$ , Reconstructed data from PC;  $\Upsilon$ , Unit cost;  $v$ , Wind Speed;  $\zeta$ , Scenarios or timesets;  $A^{PV}$ , Area of PV;  $E^\mu$ , Energy store size;  $E^{BS}$ , Energy stored in battery;  $ratio_{total}$ , Total PCA ratio;  $s_k$ , Standard deviation of the  $k$  cluster;  $T^{amb}$ , Ambient temperature;  $T^{Ref}$ , Reference temperature;  $TD_{km}$ , Typical day with based on k-means pattern recognition;  $TD_{PCA}$ , Typical day with PCA dimension reduction;  $v_k$ , Random value inside a  $k$  cluster;  $W^{BS}$ , Power inlet to BS;  $W^{BSB}$ , Power sent to load from batteries;  $W^{BSG}$ , Power sent to utility grid;  $W^{PV}$ , Power produced by PV;  $W^{PVB}$ , Power demand;  $W^{WT}$ , Power produced by WT;  $z_i$ , Index of the cluster.

**Abbreviations** CC, Characteristic Cluster; CCost, Capital cost of equipment; Ipopt, Interior point optimization algorithm; Julia, Programming language; JuMP, Domain-specific modeling language for mathematical optimization embedded in Julia; k-means, Clustering technique to label data; NLP, Non Linear Programming; OD, Optimal design; OMCost, Operation and maintenance cost; **P**, Projection matrix; PC, Principal Component; PCA, Principal component analysis; PCA ratio, Variance preserved of data in each PC; PCost, Cost of energy from the utility grid;  $W^D$ , Power Demand; PInc, Incomes from energy sales; PV, Photovoltaic system, panel; r, Number of random days; RES, Renewable energy system; RES, Renewable energy system; RRD, Random real days; SR, Solar Radiation; T, Temperature; TAC, Total annual cost; TD, Typical day; ToC, Time of computation; WT, Wind Turbine; **X**, Centered data matrix; **y**, PC matrix.

**Funding** This work was supported by the Consejo Nacional de Ciencia y Tecnología through the program *Posdoctorados por México*, the Departments of Chemical Engineering and Graduate Studies and Research from TecNM - Instituto Tecnológico de Celaya

**Data Availability** The data that support the findings of this study are available on request from the corresponding author LFFC. The data are not publicly available due to privacy and security concerns of the owners of the apartment building.

## Declarations

**Conflict of Interest** The authors declare no competing interests.

## References

Abdmouleh Z, Gastli A, Ben-Brahim L, Haouari M, Al-Emadi NA (2017) Review of optimization techniques applied for the integration of distributed generation from renewable energy sources. *Renew Energ* 113:266.

- <https://doi.org/10.1016/j.renene.2017.05.087>. <https://www.sciencedirect.com/science/article/pii/S0960148117304822>
- Ahmad T, Zhang H, Yan B (2020) A review on renewable energy and electricity requirement forecasting models for smart grid and buildings. *Sustainable Cities and Society* 55:102052. <https://doi.org/10.1016/j.scs.2020.102052>. <https://www.sciencedirect.com/science/article/pii/S2210670720300391>
- Ahmed R, Sreeram V, Mishra Y, Arif M (2020) A review and evaluation of the state-of-the-art in PV solar power forecasting: Techniques and optimization. *Renew Sust Energ Rev* 124:109792. <https://doi.org/10.1016/j.rser.2020.109792>. <https://www.sciencedirect.com/science/article/pii/S1364032120300885>
- Azad SA, Ali ABMS, Wolfs P (2014) Identification of typical load profiles using K-means clustering algorithm. In: *Asia-Pacific World congress on computer science and engineering*, pp 1–6. <https://doi.org/10.1109/APWCCSE.2014.7053855>
- Azadeh A, Ghaderi S, Maghsoudi A (2008) Location optimization of solar plants by an integrated hierarchical DEA PCA approach. *Energy Policy* 36(10):3993
- Azuatalam D, Paridari K, Ma Y, Förstl M, Chapman AC, Verbič G (2019) Energy management of small-scale PV-battery systems: A systematic review considering practical implementation, computational requirements, quality of input data and battery degradation. *Renew Sust Energ Rev* 112:555. <https://doi.org/10.1016/j.rser.2019.06.007>. <http://www.sciencedirect.com/science/article/pii/S1364032119303983>
- Blaom AD, Kiraly F, Lienart T, Simillides Y, Arenas D, Vollmer SJ (2020) Mlj: A julia package for composable machine learning
- Breuer T, Bussieck M, Cao KK, Cebulla F, Fiand F, Gils HC, Gleixner A, Khabi D, Koch T, Rehfeldt D, Wetzel M (2018) Optimizing large-scale linear energy system problems with block diagonal structure by using parallel interior-point methods. In: Kliewer N, Ehmke JF, Borndörfer R (eds) *Oper Res Proceedings*, vol 2017. Springer International Publishing, Cham, pp 641–647
- Cao Y, Fuentes-Cortes LF, Chen S, Zavala VM (2017) Scalable modeling and solution of stochastic multiobjective optimization problems. *Comput Chem Eng* 99:185. <https://doi.org/10.1016/j.compchemeng.2017.01.021>. <https://www.sciencedirect.com/science/article/pii/S0098135417300212>
- Calvillo C, Sánchez-Mirallas A, Villar J (2016) Energy management and planning in smart cities. *Renew Sust Energ Rev* 55:273. <https://doi.org/10.1016/j.rser.2015.10.133>. <https://www.sciencedirect.com/science/article/pii/S1364032115012125>
- Chinneck JW (2008) Seeking feasibility in nonlinear programs. In: *Feasibility and infeasibility in optimization: algorithms and computational methods*. Springer, Boston, pp 51–88. [https://doi.org/10.1007/978-0-387-74932-7\\_5](https://doi.org/10.1007/978-0-387-74932-7_5)
- Ciupageanu DA, Barelli L, Lazaroiu G (2020) Real-time stochastic power management strategies in hybrid renewable energy systems: A review of key applications and perspectives. *Electr Power Syst Res* 187:106497. <https://doi.org/10.1016/j.epsr.2020.106497>. <https://www.sciencedirect.com/science/article/pii/S037877962030300X>
- Cui Y, Yan S, Zhang H, Huang S (2019) Ultra-short-term prediction of wind power based on chaos theory and ABC optimized RBF neural network. In: *2019 IEEE 3rd International Electrical and Energy Conference (CIEEC)*, pp 1422–1427. <https://doi.org/10.1109/CIEEC47146.2019.CIEEC-2019517>
- Dunning I, Huchette J, Lubin M (2017) JuMP: a modeling language for mathematical optimization. *SIAM Rev* 59(2):295–320. <https://doi.org/10.1137/15m1020575>
- Forkman J, Josse J, Piepho HP (2019) Hypothesis tests for principal component analysis when variables are standardized. *J Agric Biol Environ Stat* 24(2):289



- Fuentes-Cortés L. F., Flores-Tlacuahuac A (2018) Integration of distributed generation technologies on sustainable buildings. *Appl Energy* 224:582. <https://doi.org/10.1016/j.apenergy.2018.04.110>. <http://www.sciencedirect.com/science/article/pii/S0306261918306779>
- García J. L. T., Calderón EC, Heras ER, Ontiveros CM (2019) Generating electrical demand time series applying SRA technique to complement NAR and sARIMA models. *Energy Effic* 12(7):1751. <https://doi.org/10.1007/s12053-019-09774-2>
- García JLT, Calderón EC, Ávalos GG, Heras ER, Tshikala AM (2019) Forecast of daily output energy of wind turbine using sARIMA and nonlinear autoregressive models. *Adv Mech Eng* 11(2):1687814018813464. <https://doi.org/10.1177/1687814018813464>
- Gordillo-Orquera R, Lopez-Ramos LM, Muñoz-Romero S, Iglesias-Casarrubios P, Arcos-Avilés D, Marques AG, Rojo-Álvarez JL (2018) Analyzing and forecasting electrical load consumption in healthcare buildings. *Energies* 11(3):493
- Hamerly G, Elkan C (2004) Learning the k in k-means. *Advances in neural information processing systems* 16:281
- Hernández-Romero IM, Fuentes-Cortés LF, Mukherjee R, El-Halwagi MM, Serna-González M, Nápoles-Rivera F (2019) Multi-scenario model for optimal design of seawater air-conditioning systems under demand uncertainty. *J Clean Prod* 238:117863. <https://doi.org/10.1016/j.jclepro.2019.117863>. <https://www.sciencedirect.com/science/article/pii/S0959652619327337>
- Jolliffe IT (1986) *Principal component analysis*. Springer, Berlin, pp 129–155
- Kakran S, Chanana S (2018) Smart operations of smart grids integrated with distributed generation: A review. *Renew Sust Energ Rev* 81:524. <https://doi.org/10.1016/j.rser.2017.07.045>. <https://www.sciencedirect.com/science/article/pii/S1364032117311188>
- Kettaneh N, Berglund A, Wold S (2005) PCA and PLS with very large data sets. *Comput Stat Data Anal* 48(1):69. <https://doi.org/10.1016/j.csda.2003.11.027>. <http://www.sciencedirect.com/science/article/pii/S0167947303002949>
- Partial Least Squares
- Inegi (2017) Mexico en cifras <https://www.inegi.org.mx/app/areasgeograficas/>
- Li J, Zhou J, Chen B (2020) Review of wind power scenario generation methods for optimal operation of renewable energy systems. *Appl Energy* 280:115992. <https://doi.org/10.1016/j.apenergy.2020.115992>. <https://www.sciencedirect.com/science/article/pii/S0306261920314380>
- Li S, Ma H, Li W (2017) Typical solar radiation year construction using k-means clustering and discrete-time Markov chain. *Appl Energy* 205:720. <https://doi.org/10.1016/j.apenergy.2017.08.067>. <https://www.sciencedirect.com/science/article/pii/S0306261917310851>
- Meschede H, Dunkelberg H, Stöhr F, Peesel RH, Hesselbach J (2017) Assessment of probabilistic distributed factors influencing renewable energy supply for hotels using Monte-Carlo methods. *Energy* 128:86. <https://doi.org/10.1016/j.energy.2017.03.166>. <https://www.sciencedirect.com/science/article/pii/S0360544217305741>
- Odetayo B, Kazemi M, MacCormack J, Rosehart WD, Zareipour H, Seifi AR (2018) A chance constrained programming approach to the integrated planning of electric power generation, natural gas network and storage. *IEEE Trans Power Syst* 33(6):6883. <https://doi.org/10.1109/TPWRS.2018.2833465>
- Ranaweera I, Midtgård OM (2016) Optimization of operational cost for a grid-supporting PV system with battery storage. *Renew Energy* 88:262. <https://doi.org/10.1016/j.renene.2015.11.044>. <https://www.sciencedirect.com/science/article/pii/S0960148115304651>
- Ribeiro LD, Milanezi J, da Costa JPC, Giozza WF, Miranda RK, Vieira MV (2016) PCA-Kalman based load forecasting of electric power demand. In: 2016 IEEE international symposium on signal processing and information technology (ISSPIT). IEEE, pp 63–68
- Rudin C, Chen C, Chen Z, Huang H, Semenova L, Zhong C (2021) *Interpretable machine learning: Fundamental principles and 10 grand challenges*
- Shlens J (2014) arXiv:1404.1100
- Skoplaki E, Palyvos J (2009) On the temperature dependence of photovoltaic module electrical performance: A review of efficiency/power correlations. *Solar Energy* 83(5):614. <https://doi.org/10.1016/j.solener.2008.10.008>. <http://www.sciencedirect.com/science/article/pii/S0038092X08002788>
- Statistics J (2014) *Multivariatestats.jl*. <https://github.com/JuliaStats/MultivariateStats.jl>
- Statistics J (2012) *Clustering.jl*. <https://github.com/JuliaStats/Clustering.jl>
- U Nations (2015) The 2030 agenda for sustainable development – goal 7: Ensure access to affordable, reliable, sustainable and modern energy for all <http://www.un.org/sustainabledevelopment/energy/>
- Wächter A, Biegler L (2006) On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Math Program* 106(1):25. <https://doi.org/10.1007/s10107-004-0559-y>
- Wang J, Niu T, Lu H, Guo Z, Yang W, Du P (2018) An analysis-forecast system for uncertainty modeling of wind speed: A case study of large-scale wind farms. *Appl Energy* 211:492. <https://doi.org/10.1016/j.apenergy.2017.11.071>. <https://www.sciencedirect.com/science/article/pii/S0306261917316719>
- Yan R, Lu Z, Wang J, Chen H, Wang J, Yang Y, Huang D (2021) Stochastic multi-scenario optimization for a hybrid combined cooling, heating and power system considering multi-criteria. *Energy Convers Manag* 233:113911. <https://doi.org/10.1016/j.enconman.2021.113911>. <https://www.sciencedirect.com/science/article/pii/S0196890421000881>
- Yu Y, Narayan N, Vega-Garita V, Popovic-Gerber J, Qin Z, Wagemaker M, Bauer P, Zeman M (2018) Constructing accurate equivalent electrical circuit models of lithium iron phosphate and lead-acid battery cells for solar home system applications. *Energies* 11(9):2305. <https://doi.org/10.3390/en11092305>
- Yesilbudak M (2016) Clustering analysis of multidimensional wind speed data using k-means approach. In: 2016 IEEE International Conference on Renewable Energy Research and Applications (ICRERA), pp 961–965. <https://doi.org/10.1109/ICRERA.2016.7884477>
- Zakaria A, Ismail FB, Lipu MH, Hannan M (2020) Uncertainty models for stochastic optimization in renewable energy applications. *Renew Energy* 145:1543. <https://doi.org/10.1016/j.renene.2019.07.081>. <https://www.sciencedirect.com/science/article/pii/S0960148119311012>
- Zhang Y, Zhang C, Zhao Y, Gao S (2018) Wind speed prediction with RBF neural network based on PCA and ICA. *J Electr Eng* 69(2):148

## Affiliations

Jorge Luis Tena-García<sup>1</sup> · Luis Miguel García-Alcala<sup>2</sup> · Dulce Celeste López-Díaz<sup>3</sup> · Luis Fabián Fuentes-Cortés<sup>1</sup> 

<sup>1</sup> Departamento de Ingeniería Química, Tecnológico Nacional de México - Instituto Tecnológico de Celaya, Antonio García Cubas Pte 600, Celaya, Guanajuato, 38610, Mexico

<sup>2</sup> Departamento de Ingeniería en Sistemas Computacionales, Tecnológico Nacional de México - Instituto Tecnológico de Estudios Superiores de Zamora, Zamora, Michoacán, 59727, Mexico

<sup>3</sup> Departamento de Ingeniería Química, Alimentos y Ambiental, Universidad de las Americas Puebla, Sta. Catarina Mártir, Cholula, Puebla, 72810, Mexico