Unsupervised grouping of industrial electricity demand profiles: Synthetic profiles for demand-side management applications

Javier Valdes a, Yunesky Masip Macia b, Wolfgang Dorner d, Luis Ramirez Camargo c, d, e, *

a Institute for Applied Informatics, Technologie Campus Freyung, Technische Hochschule Deggendorf, Freyung, Germany
b Escuela de Ingeniería Mecánica, Facultad de Ingeniería, Pontificia Universidad Católica de Valparaíso, Quilpué, Chile
c Institute for Sustainable Economic Development, University of Natural Resources and Life Sciences, Vienna, Austria
d Electric Vehicle and Energy Research Group, Mobility, Logistics and Automotive Technology Research Centre, Department of Electrical Engineering and Energy Technology, Vrije Universiteit Brussel, Brussels, Belgium
e Flanders Make, 3001, Heverlee, Belgium

ARTICLE INFO

Article history:
Received 30 April 2020
Received in revised form 25 September 2020
Accepted 26 September 2020
Available online 1 October 2020

Keywords:
Electricity load profiles
Demand response
Time series analysis
Clustering
Chilean energy transition

ABSTRACT

Demand side management is a promising alternative to offer flexibility to power systems with high shares of variable renewable energy sources. Numerous industries possess large demand side management potentials but accounting for them in energy system analysis and modelling is restricted by the availability of their demand data, which are usually confidential. In this study, a methodology to synthesize anonymized hourly electricity consumption profiles for industries and to calculate their flexibility potential is proposed. This combines different partitioning and hierarchical clustering analysis techniques with regression analysis. The methodology is applied to three case studies in Chile: two pulp and paper industry plants and one food industry plant. A significant hourly, daily and annual flexibility potential is found for the three cases (15%–75%). Moreover, the resulting demand profiles share the same statistical characteristics as the measured profiles but can be used in modelling exercises without confidentiality issues.

© 2020 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

One of the main challenges encountered in the transformation of fossil fuel–based energy systems into renewable energy-based systems is matching the electricity generation from variable renewable energy sources with demand. In Chile, increasing renewable energy use can reduce foreign energy dependence and the impact of its economy on the environment [1]. However, fluctuations in renewable energy threaten the stability of the electrical system and require its reconfiguration. One option to deal with imbalances in systems with high amounts of variable renewable energy sources (VRES) are demand-side management (DSM) measures and technologies [2].

The term DSM encompasses several perspectives and technologies, and its implementation in a demand response (DR) programme requires that several issues are addressed, such as data collection, consumer participation and appropriate security measures. These barriers affect different groups of end consumers differently depending on their energy use. According to Ref. [3], there is a high potential of implementing DSM in industries in comparison with the commercial and domestic sector, because: a) individual energy consumption is very high and concentrated in some sub processes; (b) DSM requires zero capital investment, because the measurement infrastructure is already installed; and c) the processes operate in isolated and mostly automatized environments inside the plant domain, which require a low or even no human resources.

In Chile, the cogeneration facilities used in industries where heat is essential for the production process or in energy-intensive industries are among the most interesting industrial resources that provide flexibility to the electricity grid [4]. The use of these cogeneration plants varies according to the needs of the production process that depend on market conditions but that are designed to meet business as usual scenario requirements, maximizing industry profits. Nevertheless, Energy cost are rarely incorporate to maximization profits functions, and are mainly considered as variable costs and even sometimes fixed costs. Therefore, through DR
programs profits be can generated and at the same time included in the profits maximization functions, as industries can provide flexibility to the electricity system either by adjusting the cycles of the production process or by modifying the electricity generation profile of the cogeneration plants according to the needs of the electricity system. Two industries that have this type of facility and flexibility potential in Chile are the food processing and the pulp and paper industries [5,6].

Studying how these industries use electricity can help to understand the specific needs of these companies and the extent to which these companies can apply DSM measures to adapt their production processes to the needs of the electricity grid. To address this question, it is necessary to identify the patterns behind the energy consumption. In the analysis of data patterns, some clustering algorithms are designed to be applied on time series. Recently, Wang et al. [7] applied a clustering algorithm to analyze the disparity of residential and industrial energy demand in different regions using mainly aggregated data. Clustering methods as the applied by Wang, have the advantage over other algorithms dedicated to the classification of data that do not require supervision. Or in other words, it is not necessary a first classification of data to train the cluster algorithm. However, its use is restricted by access to large data series.

Nevertheless, the data collection of individual final consumers is not a trivial task. It depends on many factors, ranging from meter reliability to the required protection of data and confidential information of individual users. In addition, despite applications in the industrial sector have great potential, industrial consumption profiles that are generated from specific cases involve the use of sensitive information on production and costs. Thus, a significant commitment is required from the participating organizations. Consequently, at present, the studies are few and are focused on power flows [8] or energy benchmarking [9], in which the performance of plants are compared with that of similar plants or their own past performance. Moreover, studies in other sectors are very limited due to the lack of data. Therefore, in most of the applications calculating DR potential, anonymized data normally not available to the public or aggregated consumption data is used.

In this study, we present a new method to estimate flexibility potential by generating simulated and anonymized demand profiles for industrial electricity consumers in the food processing and the pulp and paper industries. The characteristics of the daily load profiles are studied, and clustered into representative categories. Together with the estimation of DR potential for specific plants, the objective of the analysis is to generate public data sets that maintain the statistical characteristics and descriptive value of the data obtained in the different plants, while being sufficiently simple to be easily integrated into optimization models without compromising the confidentiality of the companies. Thus, twenty-two alternative time series clustering methods are applied to time series of electricity demand collected from these industries in Chile. Different alternative methods are used, which vary in the type of clustering (hierarchical and partitioning), centroid, distance matrices and linkage methods applied.

The methodology proposed here goes beyond the previous contributions in Ref. [10,11] by applying a thorough inter-comparison of unsupervised clustering methods to generate anonymized load profiles of industries and determine DR potential based on these. The proposed approach to anonymize demand data of industries should help researchers to be able to share data that is very scarce. Moreover, it allows to quantify demand response potential of industries and use it to assess the flexibility potential of the studied industries. These results are useful to analyse DSM and DR potentials and design policies to exploit them. Potential further applications of our approach include electrical system planning and operation, load demand forecasting, development of electricity tariffs and any other measure that can introduce flexibility into the system through demand side management [10,12].

This work presents all the necessary steps to apply the methodology used to other companies in order to calculate the potential for flexibility and generate representative consumption profiles. To this end, the following section presents a review of the literature on the use of clustering techniques and the calculation of flexibility potentials. Section three describes the different steps of the applied methodology. Section four shows the data for three energy-intensive companies. Finally, the conclusions are presented.

2. State of the art

Unlike efficiency measures that aim to reduce total energy consumption, DSM measures aim to shift consumption based on fluctuations in electricity prices. An example of a DSM measure is the cancellation or rescheduling of consumption. Valdes et al. [2] have classified both DSM programmes and technologies. Most industries that are involved in DR programmes, which is a DSM measure, participate through ancillary services programmes that are essential for supporting a reliable, safe and high-quality power system. These services, which are currently under development in numerous power systems [13], are based on industrial production and electricity consumption data and periodic load profiles that estimate potential reductions and increases in the load [14].

The pulp and paper industry are among the energy-intensive industries and therefore has a high DSM potential. Current studies identify a wide range of energy efficiency technologies that have already been marketed to this industry for pulp production and paper-making machines; however, information on the ability of these technologies to provide flexibility to the grid is scarcer. In the food processing industry, the heat generated by cogeneration plants is used to dry, cool and preserve the product [15,16]. These systems are usually costly and energy-intensive.

The level of power consumption therefore depends on production capacity and the programming of various intermediate products. The literature on the short-term flexibility of industrial processes is often presented from a top-down perspective, with little focus on the short term. Gils [5] evaluated the theoretical potential of DR in Europe using the technical capabilities of different processes from the literature rather than empirical data. Shiljkut and Rajakovic [17] estimated the DR potential by comparing daily consumption profiles for different years, albeit at a very low level of detail and by using a typical annual profile. Paulus and Boggrefe [18] used literature and interviews with company representatives to study DSM potential in Germany, including the paper and pulp sector. Lastly [19], provides information on the DR potential in industry, which is obtained by collecting consumption data from participants in DR programmes. Despite the wealth of data collected, the authors only provide information on the probability distribution of the consumption of certain industries and exclude the paper industry from their analysis. Therefore, the current literature on modelling existing production processes, in both food processing and pulp and paper sectors, is typically aimed at improving plant operations and not at assessing DSM potential [20,21].

Analyses of electricity consumption patterns are focused primarily on categorising the load pattern for tariff purposes. These studies are mainly focused on aggregate residential or non-residential individual load data [10]. For non-residential customers, clustering the electrical load patterns is carried out primarily to identify an appropriate partition of load patterns in subsets of customers depending on the shape of the load pattern. Two types of clustering can be identified from the literature:
longitudinal clustering, which is designed to establish the periods of the year that are consistent with each other in terms of load pattern shape, and ii) transversal clustering, which groups the load patterns of a series of customers whose data have been collected under similar conditions [22].

Outstanding among the studies on the use of clustering techniques to analyse the behaviour of electricity demand is [10]. This paper provides an overview and assessment of the performance of clustering methods for electrical load patterns and shows that most cluster validity indices generally represent the ability of the clustering method to isolate outliers. The dissemination of smart meters has increased data availability to a satisfactory level, and a large portion of the recent literature focuses on utilising this data source in the residential sector [12]. An analysis of the literature on these methods and the application of their results to DR programmes can be found in Ref. [23]. Among the most notable applications is the ability to segment the market using the daily consumption profile and the power load of the different resulting clusters. These applications focus primarily on the aggregation of large industrial and commercial customers, on medium and low voltage feeders or on a combination of small customers with much more regular load profiles [24]. With regard to the characterization of consumption profiles, some works have recently investigated the use of clustering algorithms in household consumption data [11,24], which shows the interest of the modelling community in knowing the applicability of these methods to generate load pattern categorization.

The numerous recent studies with large data applications in the residential sector contrasts with the few applications in the industrial sector. This disparity is due in part to the difficulties of finding companies willing to collaborate over the timeframe of the study. One of the few exceptions is [25], in which clustering was used to explore the DSM potential of the manufacturing industry using a fuzzy self k-means clustering algorithm. Other studies have analysed the consumption of industries and consumers together [26], which does not allow differentiation among various industries. There are also studies that have analysed demand curves for different commercial and industrial buildings to create automatic classification processes of electricity users, such as [27]. The authors of this study apply the k-means clustering algorithm to the meat processing, foundry and plastic industry sectors, using data at intervals of 15 min. The results show different consumption profiles spread throughout the year, but the authors have not ventured to use these profiles to calculate any DR potentials. More recently, two research groups have also published studies on comparing clustering approaches for domestic electricity load profiles [11,24]; their results show that the use of one general pattern is not sufficient to correctly characterize the electricity demand of households.

In this paper, several contributions to the literature are made involving consumer profile generation and the calculation of the DSM potential.

- First, a method is presented to calculate a flexibility potential based on cluster calculations. The flexibility potential calculation has two origins: the DSM potential derived from the reorganisation of production processes over more than a day and the intra-day DSM potential.
- Unlike studies that solely focus on generating annual consumption profiles, a longitudinal and DSM potential analysis is performed to classify consumers according to their consumption and flexibility potentials throughout the year.
- The flexibility potential calculation is based on the analysis and clusters of the hourly electricity consumption per year. This method is easier to implement and provides a more realistic estimate of the electrical potential of the DSM than merely technical top-down estimates. It also enables the dissemination of the profiles produced by the anonymization process that is performed using cluster techniques.

3. Data and methodology

The approach followed here is based on [10] and complemented with statistical techniques and a new phase: the DSM potential calculation. There are six phases in the procedure: i) data gathering and processing; ii) pre-clustering to prepare the data for statistical analysis; iii) time series analysis; iv) clustering analysis, which is carried out in parallel with the previous phase; v) assessment, in which the results of the previous two phases are compared and vi) potential calculation and profiles development. Fig. 1 describes these phases graphically.

The methodology proposed by Ref. [10] has served as a basis and been extended by the introduction of a potential estimation phase and a data analysis phase, wherein regression techniques are used to verify the consistency of the selected clusters. Clustering is an unsupervised learning technique: the time series analysis is used to establish a benchmark for the clustering analysis. The clustering technique has been chosen over other classification algorithms for its ability to process a large amount of data without an a priori hypothesis for the time series (unsupervised learning). Time series clustering is an active research area with applications in a wide range of disciplines and usually has one or more of the following objectives: data reduction, hypothesis generation, hypothesis testing or cluster-based prediction [28].

The objective in both the time series analysis and clustering phases is to reduce the complexity of the data by creating representative profiles of electricity consumption throughout the year. Formally, given a dataset D of i time series data S, D = {S1, S2, S3, ..., Sj}, the unsupervised partitioning process of dividing D into a K number of cluster K = {K1, K2, K3, ..., Kj} is carried out such that the time series are grouped together. Therefore, the clustering and time series analysis enables a population to be classified into a certain number of groups in terms of the similarities and discrepancies in the existing profiles among the different elements of the population. Furthermore, the assessment phase is modified to include alternative measures of the central tendency, because the calculated centroids may be optimal for cluster selection, but not for the creation of representative profiles. The last step of this phase, the calculation of the flexibility potential based on the cluster results, is also new. In this phase, the clusters are sorted by their mean values to identify the daily consumption profile with the highest power demand, which is used as a benchmark for calculating the annual potential. This daily consumption profile is nothing more than the difference between the representative or simulated maximum consumption profile and the simulated daily consumption profile.

3.1. Data gathering, processing and pre-clustering phase

Data from three plants are used to represent three different scenarios of power consumption management and its subsequent role in the electrical system. All of these plants use large amounts of heat and electricity in the production process. The first company is dedicated to the food processing and has a system of natural gas boilers with different levels of efficiency for steam production. This company does not have a combined heat and power (CHP) installation but does have high heating demand from gas boilers. The electricity used in the production process is provided by the power grid. The second company is a paper manufacturer and has a CHP installation that provides a portion of the electricity needed in the
production process and the necessary steam for paper production. The third company is a paper manufacturer and uses CHP equipment that generates a significant portion of the electricity needed and all of the steam necessary for the production process.

The three companies released available data on consumption generation and electricity consumption. However, these relatively rich data sets include negative values (0.0005% of all of the data), missing values for some days (0.02% of all of the data) and differences in the data frequency. Table 1 shows the main features of the obtained data.

The company with a medium to large CHP system and DSM presents the most significant challenges to data processing stage because the data for this company has the most missing values of all of the obtained data. Data preparation, including data clean-up, is not the objective of this paper and is only described briefly here. As the objective of this case study is to generate time series as input to a model that optimises the energy consumption and production of the plant, the data were complemented as follows. First, the amount of energy supplied to the system (as reported by the system operator in Ref. [29]), the generator production and the total consumption were compared to the monthly consumption provided by the company. These consumption figures were used to determine the maximum consumption of the plant based on technical characteristics, and the missing values in the data were filled in based on the final consumption. It was assumed that the company only supplies power to the system when consumption falls below generation, because the cost of energy generated by the company is less than that obtainable from the grid. Second, a hypothesis was proposed for the periods when the generator does not operate, during which all of the power consumed is supplied by the grid. Third, in the absence of data, consumption was considered to equal the maximum capacity of the plant. Last, the average values were calculated as hourly values using a frequency of 2 h. This methodology was used to determine the monthly consumption values, which were compared to those recorded by the company. The average error was found to be 15%, with a maximum error of 35% and a minimum error of 2%.

All the data were provided directly by the companies and have the quality requirements of an energy management certification programme associated with ISO 50001, which was launched by the

### Table 1

| Industry       | Size     | CHP | DSM | Periods | Frequency |
|----------------|----------|-----|-----|---------|-----------|
| Pulp and paper | Med-large| Yes | No  | 2017    | Hourly    |
| Pulp and paper | Small-med| Yes | Yes | 2017    | Bi-hourly |
| Food           | Small    | No  | No  | 2017    | Hourly    |

**Fig. 1.** Profile generation procedure (based on [10]).
government of Chile [30]. This paper does not provide any private information, and data are presented in graphic and standardised form to comply with a confidentiality agreement with the companies. An example of the original data is presented in normalized form in Fig. 2.

3.2. Cluster analysis

Data grouping or clustering analysis creates homogeneous groups of objects. Ideally, a cluster consists of time series that are similar to each other, while being as different as possible from the time series in a different cluster. Different available algorithms can be used to perform clustering in time series, because there is no single definition of a cluster, and the time series to be grouped has variable characteristics. Moreover, the methods for defining a cluster are specific to the algorithm applied [31]. As each application can have different objectives, a clustering algorithm is chosen based on the type of clusters preferred. Additionally, as no single clustering algorithm outperforms other algorithms, different clustering methods must be tried and compared. Two broad clustering methods are considered here: partitioning clustering and hierarchical clustering.

On the one side, partitioning clustering is the simple division of a set of data objects into non-overlapping subsets (clusters), such that each data object is precisely located in a subset. Three different partitioning algorithms (Partition Around Medoids — PAM, k-shape algorithm and the dynamic time warping barycentre averaging - DBA) and three different distance measures (dynamic time warping distance - DTW, shape-based distance — SBD and Euclidean distance) are used in this study. On the other side, hierarchical clustering is a set of nested clusters that are organised into a tree. In hierarchical clustering, seven different linkage methods as well as the same distance measures as in partitioning clustering are used. The structure is summarized in Fig. 3 and each part is explained in the supplementary material.1

3.3. Time series analysis

Time series analysis is used in a wide range of areas of study from market behaviour and national accounts to the analysis of specific industries. Ongoing efforts have resulted in extensive improvements in time series analysis. Numerous approaches are available depending on the type of phenomena studied and the objectives and the hypothesis used in the research study. In the field of energy economics, time series approaches have been widely used to model electricity demand.

McLoughlin et al. [32] have provided a brief review of the use of statistical techniques (in which descriptive statistics and probability are used to generate clusters) [34] and regression [35] to describe electricity consumption. These methods produce highly diversified load profiles and vary in complexity. In addition, a combination of descriptive analysis and regression techniques is used to identify key data aspects. In this study, descriptive analysis is used to establish the main characteristics of the time series and the hypothesis as well as to subsequently develop a regression analysis to generate groups with similar characteristics.

A regression analysis is carried out to identify common consumption patterns over time. A cycle analysis is carried out using binary variables that represent different time cycles. These variables are valued at 0 or 1 to indicate the absence or presence of a categorical effect that can be expected to change the outcome. Fictitious (binary or dummy) variables can be used to measure the effect of a qualitative factor and determine the relevance of this effect. These variables are used in several linear regression models for months, days of the month, days of the week and hours. These models are calculated for both the entire time series and monthly. In a regression model, the effect of different days of the week and hours of the day on electricity consumption are considered:

$$ED_i = \alpha + \theta h + \varphi w,$$

Eq. 1

where $ED_i$ represents the electricity demand of industry $i$; $\alpha$ is a constant; $\theta$ is a vector that contains a dummy variable for each day of the week except for Monday (w); and $\varphi$ is a vector that contains a dummy variable for each hour (h) of the day except for the first and does not include the intercept or all of the seasonal dummies to prevent co-linearity issues. This model corresponds to a seasonal dummy model, wherein the deterministic seasonality is expressed as a function of seasonal dummy variables. This relatively simple

1 The aim of this study is not to analyse different clustering methods, but to produce load profiles to calculate the flexibility potential of various industries. A detailed analysis of the methods used in this study is given in Ref. [32] and the different studies presented in section 2. Based on these studies, a combination of internal validity indices and clustering methods were selected to represent a set of algorithms that are widely used in the cluster analysis of time series with proven efficiency [33].
model can provide a first approximation to the data to identify common characteristics in consumption patterns and an approximate maximum cluster number for use in the next phase.

4. Results

4.1. Time series analysis results

The results for the pulp and paper industries show that the total plant electricity consumption is reduced in the winter months, especially between 18 and 23 h, although this effect is less pronounced on Fridays and Saturdays. Using May as a representative month, the adjusted R-squared value of the regression model for the medium size pulp and paper industry plant is 0.71. Significant maximum cluster number for use in the next phase.

4.2. Clustering results

The application of the clustering methods produced a large number of clusters due to the numerous possible combinations of methods, summarized in Table 2. In Refs. [11] the optimal number of clusters for each approach was determined based on the highest silhouette score. In our application, the optimal results of the multiple iterations were determined by calculating the following five cluster validity indices (cvi) using the dtwclust package v3.1.1 for R [36]: the Silhouette (Sil) index [37], the Dunn index (Dunn) [38], the COP index (COP) [39], the Davies-Bouldin (DB) index [40], the Calinski-Harabasz (CH) index [41] and the score function (SF) index [42]. These indices are explained in detail in the supplementary material. In addition, because the partitioning clustering methods may be affected by the initial centroid locations selected, each of the methods was executed 100 times. Finally, the maximum number of clusters proposed was between three and twenty-five, and the clusters that maximised or minimised cvi were selected. Table 3 shows the hierarchical clustering results, including the linkage method used to maximise or minimise the cvi for each algorithm.

The results of the partitioning and hierarchical clustering show that there is a tendency to generate three or 25 clusters. Note that in these cases, most cluster groups contain only one time series, which enables the method to detect outliers successfully. To address the lack of agreement among algorithms, it is possible to generate voting rules to select the optimal number of clusters. The application of such rules would probably produce three or 25 clusters as the optimal values for these data. However, the time series analysis indicated that the structure of the data for our research purposes could be better classified using an intermediate number of clusters. Therefore, the results for a number of clusters equal to the minimum or maximum number of groups are automatically discarded, which reduces the task complexity of selecting an optimal number of clusters considerably.

4.3. Assessment, profile generation and potential estimation

4.3.1. Results for small-to-medium-sized food industry cogeneration plant

The best-performing algorithms in our study for calculating the flexibility potential in this industry is the partitioning algorithm with PAM centroids and the DTW distance and the DBA centroid

---

2 These results are available on request.

3 The number of clusters was restricted to a maximum of 5 in the first phase, to a maximum of 15 in the second phase and lastly to a maximum of 25. No significant variation was found among the results obtained from choosing between a maximum number of clusters of 15 and 25. A 25-cluster-maximum was chosen arbitrarily, although using knowledge of the time series. A maximum of 15 clusters can provide valid results and reduce computational needs, especially for methods using DTW and SBD distances.
with the DTW distance, where the Sil or DB index are used to calculate the optimal number of clusters to obtain the best approximation of the results of the time series analysis. If the number of clusters need to be increased for a more detailed classification using a method that is not based on identifying outliers, index D could be used with algorithm PAM and the Euclidean distance, but less heterogeneous clusters may be generated (Figure S5 in the supplementary material). To resolve the differences among the optimal number of clusters with different indices, a composite index could be created, or a majority vote rule could be used after the results with clusters of three or 25 are discarded. The effectiveness of the type of methods proposed in Ref. [43] could thus be increased.

The Sil, Dunn and DB indices also select an optimum level of 4 clusters, which provide satisfactory results in some cases (see Figures S2, S3 and S4 in the supplementary material); however, less homogeneous clusters are obtained after a descriptive analysis. This observation is especially true for the k-shape algorithm results, because the calculation of the best alignment between the time series includes the generation of a set of clusters that are irrelevant from a potential calculation standpoint (Figure S1 in the supplementary material).

Fig. 5 shows the results obtained using the combination of PAM centroids and Euclidean distance. The general consumption profiles were obtained by using simple statistical calculations of the centroids. Three different averages were analysed: the conditional

### Table 2
Cvi results for partitioning clustering.

| Distance | Centroid | no CHP | Small-medium CHP | Medium-large CHP |
|----------|----------|--------|------------------|------------------|
|          | Sil      | Dunn   | COP              | DB               | CH    | SF    | Sil   | Dunn   | COP | DB   | CH  | SF |
| Euclidean| PAM      | 25     | 14               | 24               | 4     | 25    | 25    | 8     | 17   | 24  | 3    | 25  | 25 |
| DTW      | PAM      | 4      | 17               | 25               | 3     | 3     | 3     | 3     | 23   | 21  | 3    | 3   | 3   |
| DTW      | DBA      | 4      | 21               | 25               | 3     | 3     | 3     | 3     | 24   | 21  | 3    | 3   | 3   |
| SBD      | Shape    | 8      | 4                | 25               | 3     | 3     | 3     | 3     | 24   | 3   | 3    | 3   |

### Table 3
Cvi results for hierarchical clustering.

| Case            | Distance | Sil   | Dunn   | COP   | DB     | CH     | SF    |
|-----------------|----------|-------|--------|-------|--------|--------|-------|
| no CHP          |          |       |        |       |        |        |       |
| Small-medium CHP| Euclidean| 4     | median | 16    | average| 25     | ward  | 11    | centroid| 3  | complete| 5   | single|
| DTW             | 4        | complete| 20    | complete| 25     | ward  | 3     | ward  | 3     | complete| 3   | ward  |
| SBD             | 3        | single | 3     | single | 25     | ward  | 4     | single| 5     | ward  | 8   | centroid|
| Medium-large CHP| Euclidean| 3     | average| 3     | single | 25     | ward  | 3     | centroid| 3  | ward  | 5   | single|
| DTW             | 3        | centroid| 24    | ward  | 25     | ward  | 3     | centroid| 3  | ward  | 5   | single|
| SBD             | 3        | single | 3     | single | 25     | ward  | 4     | median| 3     | ward  | 7   | single|
| SBD             | 3        | average| 3     | average| 25     | ward  | 13    | median| 3     | ward  | 25  | single|
| SBD             | 3        | single | 3     | single | 25     | ward  | 18    | single| 3     | complete| 22 | single|
| SBD             | 3        | single | 3     | single | 25     | ward  | 13    | single| 3     | ward  | 25  | single|

Fig. 4. Regression analysis results: medium size pulp and paper industry.
mean function, the centroids of the clusters and the hourly averages of each cluster. The blue line represents the conditional mean function, the black line represent the hourly averages, and the red line represents the centroids. Although no large differences can be observed for this case, the slight differences among the centroid values in the figure may become much larger depending on the clustering method (see supplementary material). Based on these results, the hourly average is chosen as the measure for generating a standard profile over the respective time period and to calculate the DR potential. The annual potential calculation is based on the calculation of the hourly averages per cluster in terms of the difference between the electricity demand of the cluster with the highest hourly average and the other clusters:

$$p_{i,t} = d_{\text{max},t} - d_{i,t},$$  

where $$p_{i,t}$$ is the potential of cluster i at hour t; $$d_{\text{max},t}$$ is the electricity demand d in the cluster with the greatest demand at hour t; and $$d_{i,t}$$ is the electricity demand in cluster i at hour i. Fig. 6 shows the calculated potential. Cluster 2 exhibits a significant potential that is concentrated on Saturdays and holidays (Fig. 5), followed by clusters 1 and 4, which coincide with the days before and after the days in cluster 2. Fig. 6 also shows the number of days in each cluster, and Fig. 7 shows the distribution of these clusters over a year. A high degree of weekly seasonality can be observed.

4.3.2. Results of small-to-medium-sized paper and pulp industry cogeneration plant

Similar results were obtained for the small-to-medium-sized paper and pulp industry plant as for the food industry case study: either a very low number of three clusters or a very high number of clusters, between 21 and 25, were obtained. Most of these clusters were composed of a time series, as indicated by the tendency of these algorithms to find outliers. The only algorithm that provided an intermediate result was the partitional algorithm with Euclidean distance using PAM centroids. Fig. 8 shows eight different clusters with a less concentrated distribution of days. Clusters 3, 7 and 8 are clearly different from the other clusters, showing a significant consumption reduction between 17 and 23 h. Clusters 4 and 2 exhibit high levels of consumption with hourly drops on some days. Clusters 1, 5 and 6 exhibit more erratic reductions in consumption levels. These results do not appear optimal at first because of the similarities that can be discerned among the different profiles. However, these results exhibit more intra- and inter-group
homogeneity than the results of the time series analysis.

As in the previous case, the difference between the cluster with the highest average and the other clusters was calculated based on the hourly averages. Cluster 4 had the highest average hourly consumption and was used as the reference: the results are shown in Fig. 9. It is very interesting that this company is using its high DSM potential: this potential can be visually identified in clusters 3, 5, 6, 7 and 8. This potential, which represents up to 50% of the maximum daily consumption of the plant, is temporally concentrated in only few hours per day. Although concentrated, the potential has a very short ramp up and ramp down of approximately an hour, which may facilitate its contribution to DR programmes on short notice.

The annual distribution of the clusters in Fig. 10 shows a marked flexibility concentration in the winter months. During these months, the company adjusts its electricity consumption according to the energy price. In the winter in Chile, the price of electricity increases between 17 h and 23 h. However, cluster 6 can be distinguished, even though its hourly distribution should not be affected by this rate, which indicates flexibility potential. Furthermore, clusters 2 and 8 are obtained for the summer months, which also represents a flexibility potential to the system, though it is lower, at approximately 10%.

In 2018, the paper industry had a CHP-installed capacity of 1109 MW [44]. The 10% flexibility mentioned above is equivalent to 110 MW, which is around 1% of the total peak demand of Chile in 2018 [45]. Considering that pulp and paper represent 20% of the final electricity consumption in Chile, the potential is even greater. However, this plant is small-to-medium size, and the results cannot be directly extrapolated to larger plants.

4.3.3. Results for medium-to-large-sized cogeneration plant

The various algorithms mostly produced similar results. Several algorithms produced the minimal number of clusters, three, or the maximum, 25. The hierarchical clustering data for the second case study was also promising. However, 12 of the 13 clusters only had one-time series. Based on these results, only the k-shape algorithm with SBD distance, provided an intermediate number of clusters.
5. Discussion

The results show that DSM potentials can be calculated for industrial plants on an annual basis. This DSM potential is based on the hypothesis that the plants considered in this study can pull ahead or delay daily consumption by reorganising production activities throughout the year. To this end, the organisation of consumption and the corresponding production processes could be aligned with the needs of the system. Thus, the consumption clusters identified throughout the year would need to be re-distributed. This re-distribution could be complemented by reorganising the production processes throughout the day to create a second source of flexibility.

To determine this second source of flexibility, the distance between the centroid of each cluster and the maximum and minimum values of the data set should be considered or the first and third quartiles could be considered, for example, to provide a safer estimate. Cluster 3 in Fig. 5 shows that the electricity consumptions of the company from Monday through Friday average slightly over 75% of maximum consumption, although most consumption is around 60% and 90%. The results show that a estimation of potential based on installed capacity, as in Ref. [5,18], and the one presented in this study, show a similar flexibility potential. These studies, based on literature and interviews [18] and technical capabilities [5] set the maximum potential on 80% for the pulp machines and 70% for the paper machines. The same observation can be made for cluster 4 in Fig. 8 for the pulp industry. These differences also represent a flexibility potential of approximately ± 0.15 MWh (For a load normalized to 1 MWh). If companies can modify their production process to reduce consumption during certain time slots from Monday to Friday and increase consumption in other time slots, the companies could participate in DSM programmes, given the provision of sufficient incentives and technical conditions [46].

The estimated flexibility potential depends on the clustering method and the cluster validity index applied. Applications resulting on a large number of clusters should imply less within cluster variability in some of the clusters reducing the within cluster flexibility potential. In the other hand, the higher number of clusters increases the flexibility potential associated to the new clusters. The distribution of the new flexibility potential will be reflected in an increased heterogeneity of the annual potential distribution shown in Figs. 7, 10 and 12 (Figures S3 and S5 in the supplementary material show the extent to which the clusters are related and how the selection of a larger number of clusters affects these flexibility calculations). The heterogeneity of the clustering results, which might be viewed as a source of uncertainty, is actually advantageous. It allows to adjust the allocation of the flexibility potential within days and over the year. Thus, the time series analysis is concluded to be a good benchmark in selecting the best balance.

Ramp up and ramp down capabilities are among the most important technical features of the production process, in addition to the necessary mechanisms for information transmission and reprogramming capabilities. The data collected show that in the paper and pulp industry, ramp up and ramp down could be used to respond to short term changes in consumption. Cyclical reductions and increases exceeding 0.4 MW in the hourly consumption can sometimes be observed in Figs. 11 and 12, which appear to be unrelated to system failures. In fact, these variations in the demand curves are associated with demand management techniques launched by the company to cope with the higher electricity rate from 17 to 23 h. This peculiarity is not extendible to other industries; for example, there is no evidence of these variations in the food industry case study. Thus, a deeper technical analysis is required. These results show in what degree that the paper industry is sensitive to financial incentives, something that is also true for households, but which had not been studied using industries profiles yet [47].

The results obtained using the different clustering techniques for the time series have significant potential and versatility. However, significant barriers are posed to the application of these methods to highly heterogeneous real data, as in the case of the paper and pulp industry. The results show the advantages of the DTW and the Euclidean distances over the SBD. This difference may
be attributed to the modifications of the start and end points and the rotation of the time series that are allowed by the k-shape algorithm. Such modifications in the analysis of the consumption series can lead to the generation of clusters with time series that include clusters 1 and 4 in Fig. 5, for example. Partitioning clustering methods to calculate the potential appears to be superior to hierarchical methods.

The index results show that some of the applied internal validity indices are not suitable for the task at hand in this paper. By instance, the COP index appears to have a remarkable ability to find outliers. Groupings with large quantities of clusters containing small numbers of elements are selected as the optimal cluster number, because the COP index is severely affected by the existence of outliers (Supplementary material). Similar number of clusters are not obtained using the considered indices, which complicates the combined use of these indices or the generation of composite metrics with the aggregation of results from several clusters. Thus, research should be conducted on the use of indices considering these properties, including or excluding indices based on the number of time series included in each cluster.

6. Conclusions

A detailed statistical study using both time series analysis and demand profiles clustering for industrial consumers has been used to generate simulated profiles and consumption ranges to determine the flexibility ability of particular industries. The simulated profiles properly reproduce the original data from a statistical perspective and can be used without compromising confidential information of the companies. The results provide an improved understanding of the possibilities of reducing or increasing electricity consumption in the studied industries and thereby encourage the use of new management mechanisms. Specifically, applications of these methods are noteworthy in the generation sector, for example, increasing consumption during renewable energy surpluses or decreasing consumption when marginal production costs are high. Thus, action could be taken on the demand curves of the entire electrical system.

The results show that there is a flexibility potential that is being used by the pulp and paper companies to cope with peak rates in the wintertime. However, this potential is not used for the rest of the year because there are no price incentives to carry out DSM. This potential differs between companies of the sector and is estimated at up to 50% of the maximum average consumption for the company with a medium size installation and 17.5% for a company with a larger size installation. However, this potential is far greater for the food industry, reaching up to 70% of the maximum average consumption on holidays and Sundays. Moreover, an additional flexibility potential associated with the consumption distribution within the identified clusters has been identified. This potential is approximately 15% of the maximum average consumption and essentially corresponds to the capability of the plant to modify its intra-day consumption even at peak hours in the day, with little ability to redistribute its medium-term consumption. These results provide valuable information for modelling energy demand in these industries.

The results also show the need to combine different types of analysis in order to reduce the sensitivity of the model and methods with the change of computational parameters. Results of the flexibility potential are highly dependent on the clustering method under the existence of outliers. Most of the clustering algorithms do provide a very high or very low number of clusters, mostly populated with only one daily demand profile. Therefore, we propose to include statistical analysis as a complement of the clustering method. Future research should continue with the analysis of different clustering methods as well as other methods to group data. In this sense, it is imperative to analyse to what extent the resulting anonymized consumption profiles from the different methods can be used in energy system modelling.

Credit author statement

Javier Valdes: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, data curation, Writing - Original Draft, Writing - Review & Editing, Visualization; Yuneksy Masip Macia: Conceptualization, Resources, Writing - Review & Editing, Project administration, Funding acquisition; Wolfgang Dorner: Conceptualization, Project administration, Funding acquisition; Luis Ramirez Camargo: Conceptualization, Validation, Investigation, data curation, Writing - Original Draft, Writing - Review & Editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The study was conducted under the auspices of the project INCREASE “Increasing renewable energy penetration in industrial production and grid integration through optimised CHP energy dispatch scheduling and demand-side management” (grant number BMBF150075) funded by the German Federal Ministry of Education and Research (BMBF) and the Chilean National Commission for Scientific Research and Technology (CONICYT). The study was also supported by the Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH through the Energy Program in Chile. We also gratefully acknowledge support from the European Research Council (“refUEL” ERC-2017-STG 758149).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.energy.2020.118962.

References

[1] Valdes J, Escribano Francés G, San Martín González E. Energy security and renewable energy deployment in the EU: liaisons dangereuses or virtuous circle? Renew Sustain Energy Rev 2016;62:1032–46. https://doi.org/10.1016/j.rser.2016.04.069.
[2] Valdes J, Poque González AB, Ramirez Camargo L, Valin Fernández M, Masip Macia Y, Dorner W. Industry, flexibility, and demand response: applying German energy transition lessons in Chile. Energy Res Soc Sci 2019;54:12–25. https://doi.org/10.1016/j.erss.2019.03.003.
[3] Zhang Q, Grossmann IE. Enterprise-wide optimization for industrial demand side management: fundamentals, advances, and perspectives. Chem Eng Res Des 2016;116:114–31. https://doi.org/10.1016/j.cherd.2016.10.006.
[4] International Energy Agency. Energy policies beyond IEA countries: Chile. 2018, 2018.
[5] Gils HC. Assessment of the theoretical demand response potential in Europe. Energy 2014;67:1–18. https://doi.org/10.1016/j.energy.2014.02.015.
[6] Lindberg C-F, Zahedian K, Solgi M, Lindkvist R. Potential and limitations for industrial demand side management. Energy Proc 2014;61:415–8. https://doi.org/10.1016/j.egypro.2014.11.1138.
[7] Wang S, Liu H, Pu H, Yang H. Spatial disparity and hierarchical cluster analysis of final energy consumption in China. Energy 2020;197:117195. https://doi.org/10.1016/j.energy.2020.117195.
[8] Hong C-T, Ma C-M, Chen H-W, Chuang K-J, Chang C-T, Su T-L. Energy flow analysis in pulp and paper industry. Energy 2011;36:3063–8. https://doi.org/10.1016/j.energy.2011.02.051.
[9] Rogers JG, Cooper SJ, Norman JB. Uses of industrial energy benchmarking with reference to the pulp and paper industries. Renew Sustain Energy Rev
Kong L, Hasanbeigi A, Price L. Assessment of emerging energy-efficiency technologies for electrical load pattern grouping. Appl Energy 2012;42:68–80. https://doi.org/10.1016/j.apenergy.2011.12.031.

Yilmaz S, Chambers J, Patel MK. Comparison of clustering approaches for domestic electricity load profile characterization - implications for demand side management. Energy 2019;180:665–77. https://doi.org/10.1016/j.energy.2019.05.124.

Yildiz B, Bilbao JL, Dore J, Sproul AB. Recent advances in the analysis of residential electricity consumption and applications of smart meter data. Appl Energy 2017;208:402–27. https://doi.org/10.1016/j.apenergy.2017.10.014.

Torriti J, Hassan MG, Leach M. Demand response experience in Europe: policies, programmes and implementation. Energy 2010;35:1575–83. https://doi.org/10.1016/j.energy.2009.05.021.

Paulus M, Borggreve F, The potential of demand-side management in energy-saving strategies for the pulp and paper industry: a technical review. J Clean Prod 2018;95:23-37. https://doi.org/10.1016/j.jclepro.2018.06.019.

Alasseri R, Tripathi A, J. Energy. 2019;180:665–77. https://doi.org/10.1016/j.apenergy.2019.04.089.

Luo X, Zhu X, Lim EG. A parametric bootstrap algorithm for cluster number determination of load pattern categorization. Energy 2019;180:50–60. https://doi.org/10.1016/j.energy.2019.04.089.

Zhang C, Xia Y, Wang J, Sun L. A new clustering algorithm for demand side management. Energy 2019;180:665–77. https://doi.org/10.1016/j.energy.2019.04.089.

Valdes J, Y. Masip Macia, W. Dorner et al. Energy 215 (2021) 118962

Yilmaz S, Weber S, Patel MK. Who is sensitive to DSM? Understanding the determinants of the shape of electricity load curves and demand shifting: socio-demographic characteristics, appliance use and attitudes. Energy Pol 2019;133:110909. https://doi.org/10.1016/j.enpol.2019.110909.