Elsevier Editorial System(tm) for Electric Power Systems Research

Manuscript Draft

Manuscript Number: EPSR-D-15-01822R2

Title: Dynamic Clustering of Residential Electricity Consumption Time Series Data Based on Hausdorff Distance

Article Type: Research Paper

Keywords: dynamic clustering; data mining; load profiles

Corresponding Author: Dr. Ignacio Benitez,

Corresponding Author's Institution: Instituto Tecnológico de la Energía

First Author: Ignacio Benitez

Order of Authors: Ignacio Benitez; José-Luis Díez; Alfredo Quijano; Ignacio Delgado

Abstract: As the analysis of electrical loads is reaching data measured from low voltage power distribution networks, there is a need for the main agents involved in the operation and management of the power grids to segment the end users as a function of their shapes of daily energy consumption or load profiles, and to obtain patterns that allow to classify the users in groups based on how they consume the energy.

However, this analysis is usually limited to the analysis of single days. Since the smart metering data are time series formed by sequential measurements of energy through each hour or quarter of hour of the day, and also through each day, thanks to the implementation of Advanced Metering Infrastructure (AMI) and the Smart Grid technologies, it becomes clear that the analysis of the data needs to be extended to consider the dynamic evolution of the consumption patterns through days, weeks, months, seasons, and even years.

This is the objective of the present work. A new framework is presented that addresses the dynamic clustering, visualization and identification of temporal patterns in load profiles time series, fulfilling the detected gap in this area. The present development is a generic framework that allows the clustering and visualization of load profiles time series applying different classical clustering algorithms. A novel dynamic clustering algorithm is also presented, based on an initial segmentation of the energy consumption time series data in smaller surfaces, and the computation of a similarity measure among them applying the Hausdorff distance. Following, these developments are presented and tested on two dataset of energy consumption load profiles from a sample of residential users in Spain and London.

*Cover Letter

This information has been extracted from the article.

TITLE:

Dynamic Clustering of Residential Electricity Consumption Time Series Data Based on Hausdorff Distance

AUTHORS:

- Ignacio Benítez (corresponding author). Instituto Tecnológico de la Energía (Spain). ignacio.benitez@ite.es
- José-Luis Díez. Universitat Politècnica de València (Spain). jldiez@isa.upv.es
- Alfredo Quijano. Instituto Tecnológico de la Energía (Spain). alfredo.quijano@ite.es
- Ignacio Delgado. Instituto Tecnológico de la Energía (Spain). ignacio.delgado@ite.es

BRIEF INTRODUCTION AND OBJECTIVES:

The question arises on how the large amounts of smart metering data can be used in a way to be profitable to an interested party or agent. Data mining techniques can provide the tools to achieve this objective.

Clustering and classification techniques are among the descriptive objectives of the data mining, whereas prediction is included in the predictive objective. The evolution analysis is also an objective of the data mining. In the evolution analysis, the trend of the series and the temporal evolution of the data is a key factor in the objective of the analysis. Most of the objectives described for the static analysis can be extended in the evolution analysis.

The objective is to capture the evolution in time of the load profiles. It allows the obtention of patterns that evolve through time, in a time frame defined by the expert. This allows an interpretation of the results that depicts the full dynamic behavior of all the objects, therefore providing a much more complete (and also complex) information, from where conclusions can be obtained and actions can be determined, to fulfill the purposes of the data mining process. Issues such as identifying specific groups with special trends or shapes in time, or comparing clusters' differences according to their entire behavior in the time frame, can now be considered.

INNOVATION AND PREVIOUS WORKS FROM THE AUTHORS:

Concerning the analysis of load curves of energy consumption, all the works found in the literature correspond to static clustering. Realizing that no specific development had been found addressing the dynamic clustering and visualization of energy consumption load profiles time series data, the authors of the present work presented a previous paper:

- Benítez, I.; Quijano, A.; Díez, J.-L. & Delgado, I. Dynamic clustering segmentation applied to load profiles of energy consumption from Spanish customers International Journal of Electrical Power & Energy Systems , 2014, 55, 437 – 448.

where a dynamic K-means clustering algorithm was developed, by modifying the static K-means algorithm to obtain the similarity distances among objects taking into account all the Euclidean distances between each pair of objects from their coincident time stamps.

The new approach presented in this work applies the concept of Type 3 dynamic clustering. In this case the feature or dimension trajectories of the objects are clustered. The dimensions of two objects are compared as sequences of n samples, and a final distance is obtained as the average of all the comparisons of features at the 24 dimensions.

The similarity measure proposed can be seen as an ``augmented'' distance, in the sense described by Izakian et al.:

- Izakian, H.; Pedrycz, W. & Jamal, I. Clustering Spatiotemporal Data: An Augmented Fuzzy C-Means Fuzzy Systems, IEEE Transactions on, 2013, 21, 855-868.

or Izakian and Pedrycz:

- Izakian, H. & Pedrycz, W. Agreement-based fuzzy C-means for clustering data with blocks of features Neurocomputing , 2014, 127, 266 – 280.

since the similarity in static or Type 1 clustering is augmented to process time instants in n features or dimensions. It can also be seen as a development based on the description of the membership function of a time series object to a class made by Weber in the FFCM algorithm:

- de Oliveira, J. V. & Pedrycz, W. (Eds.) Advances in Fuzzy Clustering and its Applications John Wiley & Sons, Ltd., 2007.

The distance function operator d is replaced by specific distance functions able to compare two time series and yield a value of similarity. None of the previous works presented, however, have been developed to analyze and visualize the resulting clusters in the form of n or, in this case, 24 dimensions of dynamic data objects.

Moreover, a new dynamic clustering procedure is presented, as a modification of the static K-means algorithm, but applying an initial decomposition of the data object in smaller linear surfaces and comparing them applying a Hausdorff-based similarity distance.

CONCLUSIONS:

The results obtained provide a feasible and valuable analysis for the different experts and agents involved in the management of power systems, and can serve for different purposes, such as predictive maintenance, evaluation of consumption trends, detection of non-typical patterns of consumption, or identification of groups of customers with specific characteristics for the provision of energy.

Highlights

Dynamic clustering is applied on load profiles of energy consumption.

The dynamic evolution of energy at each hour of the day can be observed.

A new dynamic clustering technique based on Hausdorff distance is presented.

List of changes

The following changes have been made:

Reviewer #3: In my opinion, in this revised version of the work, the Authors have improved the paper adequately addressing the previous requests of the Reviewers. Some minor revisions:

- page 3, section 2, line 2: the comma after [4] is to be taken off;

The comma has been removed.

- page 4, line 4: Eq. 1 is to be replaced by Eq. (1);

It has been changed.

- page 4, after Eq. 1: in Vj and Vk, j and k are to be subscripted;

Subindices j and k have been subscripted.

- defining every symbol used in the paper seems to be necessary; to this purpose a list of the symbols could be advisable.

A nomenclature has been included on top of page 3, making use of the LateX package *nomencl*. A list of symbols and their corresponding definitions from Equations 1, 2, 4, 6 and 7 has been included. The descriptions of the symbols have been removed from the text, to keep the extension limit to 22 pages.

Dynamic Clustering of Residential Electricity Consumption Time Series Data Based on Hausdorff Distance

Ignacio Benítez[a], José-Luis Díez[b], Alfredo Quijano[a], Ignacio Delgado[a]

[a] *Instituto Tecnológico de la Energía*
Avda. Juan de la Cierva 24
46980 Paterna, Spain
[b] *Department of Systems Engineering and Control*
Universitat Politècnica de València
Camino de Vera, 14
46022 Valencia, Spain

Abstract

As the analysis of electrical loads is reaching data measured from low voltage power distribution networks, there is a need for the main agents involved in the operation and management of the power grids to segment the end users as a function of their shapes of daily energy consumption or load profiles, and to obtain patterns that allow to classify the users in groups based on how they consume the energy.

However, this analysis is usually limited to the analysis of single days. Since the smart metering data are time series formed by sequential measurements of energy through each hour or quarter of hour of the day, and also through each day, thanks to the implementation of Advanced Metering Infrastructure (AMI) and the Smart Grid technologies, it becomes clear that the analysis of the data needs to be extended to consider the dynamic evolution of the consumption patterns through days, weeks, months, seasons, and even years.

This is the objective of the present work. A new framework is presented that addresses the dynamic clustering, visualization and identification of temporal patterns in load profiles time series, fulfilling the detected gap in this area. The present development is a generic framework that allows the clustering and visualization of load profiles time series applying different classical clustering algorithms. A novel dynamic clustering algorithm is also

presented, based on an initial segmentation of the energy consumption time series data in smaller surfaces, and the computation of a similarity measure among them applying the Hausdorff distance. Following, these developments are presented and tested on two dataset of energy consumption load profiles from a sample of residential users in Spain and London.

Keywords:
dynamic clustering, data mining, load profiles

---

1. Introduction

The European Technology Platform on Smart Grids (ETP SG) has issued a report in 2015 [1] on the research and development needs foreseen by the platform for the EC Horizon 2020 Research and Innovation Programme [2], for the years 2016 and 2017. One of the main challenges identified by the ETP SG is the utilization of smart metering data. According to the ETP SG, "a very large amount of data is being collected whose potential has been untapped".

The question arises on how the large amounts of smart metering data can be used in a way to be profitable to an interested party or agent. Data mining techniques can provide the tools to achieve this objective. The term "data mining" gathers a number of different algorithms and techniques which have as objective the analysis and extraction of useful information from large sets of data [3].

Han and Kamber [3] define two main objectives of the data mining process, as a function of the data mined and the kind of knowledge sought. These objectives are the static analysis, or the analysis of static data, and the evolution analysis, where the trend of the series and the temporal evolution of the data is a key factor in the objective of the analysis.

The development presented in this paper allows to obtain patterns that evolve through time, in a time frame defined by the expert. This allows an interpretation of the results that depicts the full dynamic behavior of all the objects, therefore providing a much more complete (and also complex) information, from where conclusions can be obtained and actions can be determined, to fulfill the purposes of the data mining process. Issues such as identifying specific groups with special trends or shapes in time, or comparing clusters' differences according to their entire behavior in the time frame, can now be considered.

Nomenclature

$\bar{X_{i_k}}$, $\bar{V_{j_k}}$	average values of the X_{i_k} and V_{j_k} time series
μ_{ij}	membership value of object i to cluster j
ε_i	residual or model error at time or instant i
B	norm matrix
c	number of clusters or classes
$d(X_i, V_j)$	distance or similarity function between X_i and V_j
m	degree of fuzziness of the clusters (usually a value higher than 1)
n	number of features or characteristics of the data
p	total number of time samples or instants
V_j, V_k	centroids or prototypes of the classes or clusters j and k
X_i	feature vector of object i
W=$[w_0\ w_1\ w_2]^t$	vector of coefficients of a linear surface model

First, a state of the art is presented, regarding dynamic clustering algorithms found in the literature and previous clustering analyses regarding the segmentation of load profiles. Following, the development made is presented and tested on two different datasets of load profiles from a sample of residential low voltage consumers, in Spain and London. The results are described and discussed. Finally a conclusions section is included.

2. State of the art on dynamic clustering techniques applied on load profiles

With respect to the dynamic nature of the data and the cluster analysis, Weber [4] classifies the cluster analysis in four types or categories, according to the dynamic nature of the data and the clusters:

- Type 1. The data is treated as static and the clustering process is also static.
- Type 2. The data is treated as static but the number of clusters may vary at each new computation. In this system, issues such as clusters formation, collapse, split or fusion must be considered.
- Type 3. The data is treated as dynamic, evolving through time, as trajectories of the different data features or dimensions through time. The number of clusters is fixed. The resulting centroids or patterns are therefore defined by feature trajectories that evolve through time. The present work approaches this type of dynamic clustering analysis.
- Type 4. The data is also treated as dynamic, as in type 3, becoming feature trajectories that evolve through time, and the number of clusters varies dynamically at each iteration. Clusters and patterns can, therefore, as in type 2, merge or split.

Liao [5] makes a differentiation of clustering types for time series data based on three main approaches: clustering on the raw data, clustering on a feature-based transformation of the data, and clustering on a model-based transformation of the data. The clustering algorithms presented in this paper are based on the analysis of the raw time series data, therefore a brief review on the current state of the art in this field is described next.

A number of time series clustering algorithms are based on modifications of the K-means [6] or Fuzzy c-means or FCM [7] algorithms. Weber [4] describes the algorithm called Functional Fuzzy C-means or FFCM, as a time series generalization of the FCM. The FFCM algorithm presents a modified calculation of the membership value at each iteration, indicated in Eq. (1), where the distance function d is based on fuzzy inference.

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d(X_i, V_j)}{d(X_i, V_k)} \right)^{\frac{2}{m-1}}} \tag{1}$$

Regarding recent years, Izakian et al. [8] present a clustering algorithm for spatiotemporal data where the Euclidean distance is replaced by an "augmented" distance, which is the weighted sum of two Euclidean distances: the comparison of the spatial components and the comparison of the temporal features. This modification is later extended by Izakian and Pedrycz [9] to

n features or dimensions with the concept of blocks or groups of similar features, computing a weighted sum of Euclidean distances where the different weights are obtained by Particle Swarm Optimization (PSO) [10].

Concerning the analysis of load curves of energy consumption, all the works found in the literature correspond to static or Type 1 clustering. Realizing that no specific development had been found addressing the dynamic clustering and visualization of energy consumption load profiles time series data, the authors of the present work presented a previous paper [11] where a dynamic K-means clustering algorithm was developed, by modifying the static K-means algorithm to obtain the similarity distances among objects taking into account all the Euclidean distances between each pair of objects from their coincident time stamps.

3. Development of algorithms and techniques to perform dynamic clustering on load profiles time series data

The new approach presented in this work applies the concept of Type 3 dynamic clustering described. In this case the feature or dimension *trajectories* of the objects are clustered. The dimensions of two objects are compared as sequences of n samples, and a final distance is obtained as the average of all the comparisons of features at the 24 dimensions. Although the two approaches may deal similar mathematical results, they are quite different and, depending on the operators and similarity measures used, may produce very different outcomes. The first approach can be seen as a succession of Type 1 static clustering calculated for n times and clustered together. The second approach is designed as a Type 3 dynamic clustering of dynamic trajectories through time, with a fixed number of classes.

The similarity measure proposed can be seen as an "augmented" distance, in the sense described by Izakian et al. [8] or Izakian and Pedrycz [9], since the similarity in static or Type 1 clustering is augmented to process time instants in n features or dimensions. It can also be seen as a development based on the description of the membership function of a time series object to a class made by Weber in the FFCM algorithm [4]. The distance function operator d is replaced by specific distance functions able to compare two time series and yield a value of similarity. None of the previous works presented, however, have been developed to analyze and visualize the resulting clusters in the form of n or, in this case, 24 dimensions of dynamic data objects.

Moreover, a new dynamic clustering procedure is presented, as a modification of the static K-means algorithm, but applying an initial decomposition of the data object in smaller linear surfaces and comparing them applying a Hausdorff-based similarity distance. These developments are presented next.

3.1. Development of cluster validity indices for the evaluation of dynamic clustering on time series n-dimensional data

A number of cluster validity indices are used for the comparison of the results of dynamic clustering algorithms on time series data. The indices described are extensions for the dynamic analysis of common static clustering validity indices. The clustering validity indices were initially described to assess the initial selection of a number of clusters for partitional clustering algorithms [12]. Different authors, however, [13][14], have used them to compare the results of different clustering algorithms applied on the same data sets, to evaluate the performance of the different algorithms. They will also be used in this work for this purpose. The indices used are three: DB or Davies-Bouldin index [15], SD or Scatter - Distance index [14] and the XB or Xie - Beni index [16]. All of them indicate a good partition of clusters if their values are low.

The indices have been modified to evaluate partitions of dynamic objects in Type 3 clustering, by replacing the static similarity distances used (Euclidean distance) by a generic distance d between dynamic objects, which can be any of the similarity measures described for the dynamic cluster analysis.

3.2. Development of a common framework for the dynamic clustering and visualization of daily load profile time series

A general framework is presented for Type 3 dynamic clustering analysis, called the Equal N - Dimensional (END) Time series Clustering Framework. Within this framework, this work presents the development of partitional dynamic clustering algorithms, obtained as an extension of the classical, static techniques, where the $2D$ data and patterns are extended to $3D$ time series data and patterns. The way to perform this extension is based on the description of the FFCM by Weber, where other similarity measures have been used instead of defining fuzzy inference to compute the distances. For doing so, the END framework developed envelops the partitional clustering algorithm and modifies it in order to obtain the similarity distances among objects taking into account the distances between their feature or dimension trajectories. Applying this method, two different static clustering techniques,

K-means and Fuzzy C-means or FCM, have been extended to dynamic clustering. Two different similarity measures, based on the Euclidean distance and on the correlation measure, have been used. The resulting dynamic clustering techniques have been called: END-FCME (END FCM Euclidean-based), END-FCMC (END FCM Correlation-based), END-KME (END K-Means Euclidean-based) and END-KMC (END K-Means Correlation-based). This method, however, can be extended to most of the partitional clustering techniques found in the literature.

All the objects are assigned the same number of time samples or instants. The missing samples in this case have been filled with the average of the preceding and forthcoming values. Once all the data objects are harmonized, the distance between each cluster and the object is computed between feature trajectories. Two different techniques for obtaining this distance have been applied: Euclidean distance and correlation. In the case of Euclidean distance, the computation is shown in Eq. (2), where the identity matrix has been used as the norm B.

$$d(X_i, V_j) = \frac{1}{n}\sum_{k=1}^{n} \|X_{i_k} - V_{j_k}\|_B^2 = \frac{1}{n}\sum_{k=1}^{n}((X_{i_k} - V_{j_k})^T B(X_{i_k} - V_{j_k})) \quad (2)$$

In the case of correlation, the *Pearson* correlation coefficient between two series is computed between each pair of feature trajectories, as can be seen in Eqs. (3) and (4). The result is an index between $[-1, 0, +1]$, which yields the linear relationship degree between the two series, which is later conveniently transformed to a distance measure, applying the expression in Eq. (5). The procedure to perform the dynamic clustering with the END framework can be summarized in the following steps:

1. Initialize the C matrix of centroids with random values, or other methods.
2. Obtain all the distances of the objects to the centroids of the clusters, by the formula indicated in Eq. (2) for Euclidean distance, or Eq. (5) for correlation-based distance, or any other suitable for time series.
3. Compute membership matrix U, applying Eq. (1), in the case of END FCM, or assigning each object to the cluster with the smallest distance, in the case of END K-means.
4. Compute the cluster centroids according to the formulas for K-means or FCM in each case.

5. Repeat steps 2 to 4 until a termination condition is met, such as reaching a maximum number of iterations.

$$corr(X_i, V_j) = \frac{1}{n} \sum_{k=1}^{n} corr(X_{i_k}, V_{j_k}) \tag{3}$$

$$corr(X_{i_k}, V_{j_k}) = \frac{\sum_{m=1}^{p} (X_{i_{k_m}} - \bar{X_{i_k}})(V_{j_{k_m}} - \bar{V_{j_k}})}{\sqrt{\sum_{m=1}^{p} (X_{i_{k_m}} - \bar{X_{i_k}})^2} \sqrt{\sum_{m=1}^{p} (V_{j_{k_m}} - \bar{V_{j_k}})^2}} \tag{4}$$

$$d(X_i, V_j) = \frac{1 - corr(X_i, V_j)}{2} \tag{5}$$

3.3. Development of a two-step time series clustering algorithm with a Hausdorff-based similarity distance for the dynamic clustering of daily load profile time series

A specific development is presented in this document, being a dynamic clustering algorithm which applies a similarity measure to compare two energy consumption load profiles time series, as two dynamic surfaces, based on a two-step sequence. The energy consumption profiles from residential users are seen as 3D surfaces, defined by the 24 hours load profile, where each hour is considered as a feature or dimension of the data. Then, all the shapes are decomposed in a number of linear surfaces, by applying least squares regression. The number of surfaces and the vertices is predefined, based on the expert's knowledge of the typical behavior from residential users regarding energy consumption. Then, the resulting surfaces are compared by computing the Hausdorff distance between them, and a global similarity value is obtained, given by the average value of all the Hausdorff distances between the different surfaces. The procedure of the two-step dynamic clustering algorithm described is the following:

1. Compute centroids
2. Partition all the data objects and centroids in n x m linear surfaces
3. Compute the n x m Hausdorff distances
4. Obtain the similarity measure as the average of the n x m distances
5. Compute membership for all the objects and reassign to clusters
6. Go to first step until termination condition is met

The shapes of energy consumption are decomposed in a number of linear surfaces, applying least squares regression to model the surfaces according to the formula described in Eq. (6).

$$z_i = w_0 + w_1 x_{i1} + w_2 x_{i2} + \varepsilon_i \tag{6}$$

In order to obtain the coefficients that fit the observations to the desired function, the formula for least squares regression is used, expressed in matrix form in Eq. (7).

$$W = (X^t X)^{-1} X^t Z \tag{7}$$

The Hausdorff distance was described by Felix Hausdorff in his foundational book on Set theory [17]. The distance from a generic point x to a closed subset A, both x and A belonging to the p-dimensional subset of the closed subsets in $\Re$, is defined as the minimum of the distances of x to all the points that belong to A, as seen in Eq. (8).

$$d(x, A) = min_{\tilde{a} \in A}(d(x, \tilde{a})) \tag{8}$$

The Hausdorff metric between two non-empty closed subsets, A and B, is defined as the maximum of all possible distances $d(\tilde{a}, B)$, as seen in Eq. (9). Since this metric is not necessarily symmetric, the Hausdorff distance $d_H(A, B)$ between the two subsets is obtained as the maximum of their two Hausdorff metrics, $h(A, B)$ and $h(B, A)$, as seen in Eq. (10).

$$h(A, B) = max_{\tilde{a} \in A}(d(\tilde{a}, B)) \tag{9}$$

$$d_H(A, B) = max(h(A, B), h(B, A)) \tag{10}$$

4. Analysis and results

4.1. Description of the first dataset and the analysis performed

The first database analyzed comprises hourly energy consumption data from smart meters of a sample of 708 residential customers in Spain during two consecutive years, 2009 and 2010. The data to be clustered are comprised, therefore, of 708 objects, each one having 24 features or dimensions, and 729 daily measures. In order to validate the patterns obtained, a subset of 36 customers, representing the 5% of the sample, is extracted from the

Table 1: Type 3 dynamic clustering algorithms tested.

No.	Static clustering technique	Dynamic objects similarity measure	Dynamic clustering algorithm
1	K-means	Euclidean distance	END-KME (present work)
2	K-means	Correlation	END-KMC (present work)
3	K-means	Hausdorff distance	END-KMH (present work)
4	K-means	Euclidean distance by the same time instant	Extended static clustering (previous work by authors [11])
5	FCM	Euclidean distance	END-FCME (present work)
6	FCM	Correlation	END-FCMC (present work)
7	FCM	Hausdorff distance	END-FCMH (present work)
8	FCM	Fuzzy membership functions	FFCM (described by Weber [4])

database and will be later used for classification on the resulting patterns. The resulting set of 672 customers is used to obtain the clusters and the patterns.

The first dynamic K-means algorithm described by the authors in a previous work [11], the common framework for dynamic clustering described in the present work, including the two-step dynamic clustering algorithm with a Hausdorff-based similarity distance described also in the present work, and the FFCM dynamic clustering algorithm [4] are implemented and tested. As a result, the combination of the dynamic clustering algorithms described in Table 1 is tested on the same data set.

These 8 algorithms have been tested 10 times each, and all the resulting clustering validity indices' values have been recorded. The maximum number of iterations that each dynamic clustering algorithm is running until it stops (if no other convergence criterion is reached) has been also set to 10. These values have been chosen due to the computational effort needed to process matrices which have 729 rows and 24 columns (the complete analysis took almost four days to complete, in a workstation with a dual-core Intel processor with 12 GB of RAM). The analysis has been developed with MATLAB software.

The number of clusters to be found is set to 10, based on a previous work from Benítez et al. [18], where clustering and classification techniques are applied on a data set of energy consumption daily load profiles and the

DB index [15] is computed for a scope of cluster numbers. The value of 10 clusters is seen by the authors as a good representative value to capture the main groups of energy consumption users and also some small groups of customers with unexpected or unusual energy consumption profiles, at least for the Spanish case, since the dataset analyzed is a sample that represents the Spanish domestic energy users population. However, further studies could be made in this sense, comparing the values from different cluster validity indices. For instance 6, 8, or 12 clusters may also be a proper selection. As the number of clusters increases, there is more room for the clustering algorithm to identify new patterns that otherwise may be buried in the average pattern of bigger groups. Another factor to take into account is human perception: 10 clusters is still a reasonable number to observe global results at a glance and extract conclusions from the resulting patterns. Since one of the main objectives of the analysis is to aid in decision support, the number of clusters cannot be too big. On the other hand, if the number of clusters is too low, the capacity of the algorithm to identify small groups with uncommon load shapes is lost.

Regarding the Hausdorff distance-based algorithms, as indicated in Section 3.3, the data objects are first decomposed in a smaller number of linear surfaces and then compared by applying Hausdorff distance between them. The number of regions or surfaces to decompose the energy consumption shapes have been chosen based on previous knowledge of the behavior of the residential load profiles in Spain. The daily load profile has been divided in seven regions, according to the expected trends in a typical consumption profile of one day for the low voltage residential consumer: the first region is from 0:00 to 5:00 hours; the second from 5:00 to 8:00 hours, when the first peak of the morning is expected (getting up for going to work); the third from 8:00 to 10:00 hours, when the consumption decreases; the fourth from 10:00 to 15:00 hours, when the second peak of energy consumption is expected (due to lunchtime in Spain); the fifth is from 15:00 to 18:00 hours; when the energy consumption decreases again; the sixth is from 18:00 to 22:00 hours, when the third (and maximum) peak of energy consumption in Spanish dwellings is reached; and finally the seventh is from 22:00 to 24:00 hours. The time axis has been divided in 24 sections equal in length, which would approximately correspond to the 24 months during two consecutive years.

Table 2: Spanish data test results, DB modified index.

Cycle	END-KME	END-KMC	END-KMH	Extended K-means	END-FCME	END-FCMC	END-FCMH	FFCM
1	2.397	5.735	3.269	NaN	3.844	NaN	4.119	NaN
2	2.234	5.897	2.971	2.154	18.850	NaN	7.056	NaN
3	2.320	6.734	2.851	NaN	3.201	NaN	3.866	NaN
4	2.756	5.764	3.154	2.385	2.768	NaN	4.154	7.220
5	2.731	5.490	3.157	1.799	3.010	NaN	12.271	8.305
6	3.132	5.197	2.793	2.162	3.001	NaN	5.256	4.768
7	2.058	6.396	3.147	2.945	5.990	NaN	4.744	NaN
8	2.666	5.670	3.050	1.473	6.438	NaN	11.417	NaN
9	2.615	5.889	3.164	2.535	2.561	NaN	4.376	NaN
10	2.834	5.374	2.667	1.865	11.999	NaN	6.822	5.696

4.2. Results of the cluster analysis

The following Tables display the results obtained for the clustering validity indices DB (Table 2), SD (Table 3) and XB (Table 4). From the DB index values' Table (Table 2) it can be observed that the best results are obtained by the END-KME and the END-KMH algorithms, in all the 10 cycles. The worst result is obtained for the END-FCMC algorithm, where a NaN value in all the cycles indicates a division by zero or an error in numerical precision, probably due to the inability of the algorithm to produce well defined and separated clusters. The FCM-based algorithms, END-FCME and END-FCMH algorithms provide worse results than K-means. Finally, the FFCM and the Extended K-means yield NaN values in some cycles, therefore their reliability would be less trustworthy than the clustering algorithms with no NaN values. It can be concluded that, regarding the DB index, K-means-based dynamic clustering algorithms with Euclidean or Hausdorff-based distances provide the best results.

The analysis of the remaining Tables for the SD and XB (Table 3 and Table 4) yields similar conclusions: the best values are obtained with K-means or FCM Euclidean or Hausdorff-based distances. The FFCM and the END-KMC provide worse results, and the END-FCMC is the worst of all. These results indicate that correlation is not a good similarity measure for the dynamic clustering of load profiles time series in the way it is used in the present approach.

Table 3: Spanish data test results, SD modified index.

Cycle	END-KME	END-KMC	END-KMH	Extended K-means	END-FCME	END-FCMC	END-FCMH	FFCM
1	26.584	32.616	26.931	27.347	26.758	35.713	26.992	43.948
2	26.460	32.829	26.734	26.739	26.904	35.238	27.138	39.742
3	26.483	32.936	26.842	27.373	26.779	83.618	26.941	30.717
4	26.406	32.715	26.954	27.052	26.997	34.251	27.065	31.305
5	26.571	32.743	26.851	27.557	26.688	34.718	27.019	42.022
6	26.610	32.866	27.007	27.017	26.799	34.200	27.051	31.097
7	26.357	32.883	26.797	26.861	26.995	34.195	27.071	30.831
8	26.336	32.678	27.056	27.113	26.844	34.268	27.071	30.566
9	26.497	32.936	26.867	27.062	26.785	34.311	27.019	30.197
10	26.822	32.826	26.645	26.885	26.929	115.194	27.025	31.098

Table 4: Spanish data test results, XB modified index.

Cycle	END-KME	END-KMC	END-KMH	Extended K-means	END-FCME	END-FCMC	END-FCMH	FFCM
1	2.309	5.366	3.665	2.508	1.181	24882.341	1.591	2.960
2	2.086	4.498	3.651	1.706	17.037	19283.439	4.529	1.681
3	2.033	6.830	3.295	2.633	1.237	588970.759	2.259	3.273
4	2.274	5.160	3.304	1.497	1.339	7671.673	1.585	6.549
5	2.601	4.805	3.130	0.994	1.271	13170.104	8.718	5.259
6	2.421	4.526	3.304	1.811	1.072	7062.314	2.448	2.185
7	2.148	6.032	3.451	2.428	3.410	7005.073	2.743	2.198
8	2.455	4.832	3.845	1.034	2.750	7872.360	5.187	2.386
9	2.350	5.051	3.336	1.922	0.991	8368.214	2.173	1.696
10	1.988	5.568	3.500	1.367	7.310	960752.474	4.092	2.583

Following, the resulting patterns for each dynamic clustering algorithm are analyzed. In each case, the patterns with the best DB modified index from the 10 cycles have been chosen, however, as has been seen from the previous Tables, values from the validity indices obtained do not differ much along the 10 cycles for each algorithm, therefore any of the 10 cycles could have been used.

The resulting clusters and how the partition should be is unknown. The expected results, however, should follow previous experiences concerning the

classification of load profiles from residential or domestic electric energy users in Spain [18]. There are mainly three types of low voltage residential energy consumption users in Spain according to their load profile patterns or prototypes:

- The first type of client represents the majority of energy consumption residential users in Spain. It is represented by a daily profile of energy consumption with three ascending peaks of energy consumption: one in the morning (around 8 h.), another one at lunchtime (around 15 h.) and the highest one at night, around 22 h.

- The second type of clients represents a minority of users with a high level of energy consumption through the day. There are two different patterns in this type of clients: one with the typical shape of energy consumption, described above, but with higher energy levels (from 2500 to 7000 Wh), and another group of users that present a (more or less) flat shape of elevated energy consumption through the day, or other non-typical patterns of energy use.

- The third type comprehends a small group of clients with a higher consumption of energy at night, due to thermal energy accumulators that are used mainly at night, or in valley hours where the price of the energy is cheaper.

In all the resulting patterns for each dynamic clustering algorithm, a study is performed in order to match the clusters obtained to one of the types mentioned. To do this, first the previously described types are categorized in the following five groups or labels:

1. Common profile of residential energy consumption, with low average of daily consumption (around 500 Wh).
2. Common profile of residential energy consumption, with medium average of daily consumption (around 1500 Wh).
3. Uncommon profile, with the typical shape but with elevated average or maximum of daily energy consumption (from 2500 to 7000 Wh).
4. Uncommon profile, more or less flat through the day (or other non-typical shape) and with elevated average of daily energy consumption.
5. Peak consumption mainly at night, or shifted to other valley hours.

Table 5: Assignment of clusters from dynamic clustering algorithms to expected groups, Spanish data.

Dynamic clustering algorithm	Group 1	Group 2	Group 3	Group 4	Group 5
END-KME	3 clusters, 571 users	1 cluster, 76 users	0 clusters	3 clusters, 15 users	3 clusters, 10 users
END-KMC	5 clusters, 439 users	0 clusters	0 clusters	4 clusters, 229 users	1 cluster, 4 users
END-KMH	3 clusters, 436 users	2 clusters, 197 users	1 cluster, 5 users	2 clusters, 20 users	2 clusters, 14 users
Extended K-means	2 clusters, 572 users	2 clusters, 71 users	1 cluster, 3 users	2 clusters, 9 users	3 clusters, 17 users
END-FCME	3 clusters, 621 users	1 cluster, 28 users	0 clusters	1 cluster, 4 users	5 clusters, 19 users
END-FCMC	1 cluster, 672 users	0 clusters	0 clusters	0 clusters	0 clusters
END-FCMH	5 clusters, 574 users	2 clusters, 64 users	0 clusters	1 cluster, 14 users	2 clusters, 20 users
FFCM	3 clusters, 652 users	3 clusters, 18 users	0 clusters	2 clusters, 2 users	0 clusters

Table 5 displays this assignment. Either the results from the Table and the visualization of the obtained patterns support the conclusions obtained in the analysis of the clustering validity indices: K-means or FCM - based, with Euclidean or Hausdorff - based distances provide the best defined and well-balanced clusters and patterns. The visualization of the resulting patterns for the END-KMH (Fig. 1) algorithm is provided for visual comparison.

4.3. Classification and validation

A classification is performed of the remaining 16 users on the resulting patterns, by the computation of the Euclidean distance, as described in Eq. (2), between each user consumption data and each of the 10 patterns, and assigning each user to the cluster with the minimum distance. The clustering validity indices have been applied on the results of this classification. The values, that can be observed in Table 6, support the results obtained in the clustering process: K-means or FCM based algorithms with Euclidean or

Table 6: Validity indices on the classification results, Spanish data.

Index	END-KME	END-KMC	END-KMH	Extended K-means	END-FCME	END-FCMC	END-FCMH	FFCM
DB	1.931	6.359	2.655	1.320	2.024	5173.027	3.753	3.645
SD	1.410	1.531	1.400	1.461	1.445	3.871	1.388	1.475
XB	2.358	4.323	3.771	1.142	1.642	2670.530	4.109	2.380

Haudorff - based distances provide better quantitative results.

4.4. Description of the second dataset and the results obtained

The same analysis has been applied on a second dataset, "SmartMeter Energy Consumption Data in London Households" from the London Data Store site. The same parameters have been used for this analysis. In this case, there is no previous experience when analyzing these data and, therefore, how the resulting patterns should be is totally unknown. For this reason, only the quantitative results are displayed in Tables 7, 8 and 9, and the resulting patterns for the END- KME algorithm are displayed as an example in Fig. 2.

Although the dataset contains a sample of 5.567 users from households in London, only the first 1.000 users of the dataset, through the year 2013, have been used in the analysis. The results obtained support the main conclusions from the previous analysis. The algorithms that apply K-means or FCM with Euclidean or Hausdorff - based similarity measures appear to quantitatively provide better results. However, in this case there is also another conclusion: for this specific dataset, 10 clusters is not a good choice of clusters number. As can be seen in the Tables, there are similar values in the ten cycles for the non-fuzzy algorithms; this is due to the presence of empty clusters in the results. This can be appreciated in the visualization of the resulting patterns for the case of the END-KME algorithm in Fig. 2. There is one pattern with one user with zero value of energy consumption, i.e., a client with no data, and an empty cluster. A lower number of clusters therefore, probably 8 or 6, should have been chosen instead for the analysis of these data. These possibilities will be explored in further analyses.

5. Conclusions

The development made has been tested with two different data sets. The results show that a selection of the algorithms developed provide an appro-

Table 7: London data test results, DB modified index.

Cycle	END-KME	END-KMC	END-KMH	Extended K-means	END-FCME	END-FCMC	END-FCMH	FFCM
1	2.457	6.860	2.292	1.114	14.690	198.103	10.289	4.807
2	2.457	7.798	2.292	1.114	15.185	274.766	19.961	15.473
3	2.457	6.017	2.292	1.114	23.700	360.493	211.358	6.944
4	2.457	6.440	2.292	1.114	10.693	3303.798	17.991	11.001
5	2.457	6.419	2.292	1.114	22.007	713.202	34.459	6.812
6	2.457	8.002	2.292	1.114	263.688	291.929	13.772	9.817
7	2.457	7.552	2.292	1.114	23.579	681.747	54.445	12.027
8	2.457	7.196	2.292	1.114	11.466	7628.787	126.092	7.048
9	2.457	7.164	2.292	1.114	27.592	235.322	10.681	5.252
10	2.457	6.268	2.292	1.114	10.001	357.123	12.643	13.952

Table 8: London data test results, SD modified index.

Cycle	END-KME	END-KMC	END-KMH	Extended K-means	END-FCME	END-FCMC	END-FCMH	FFCM
1	49.467	49.705	52.216	45.551	46.456	1257.251	43.452	40.260
2	49.467	50.444	52.216	45.551	45.564	1724.442	57.374	54.206
3	49.467	50.124	52.216	45.551	49.997	2246.869	327.492	42.086
4	49.467	50.057	52.216	45.551	44.802	20183.564	50.077	45.610
5	49.467	49.887	52.216	45.551	63.059	4396.303	81.620	41.706
6	49.467	51.141	52.216	45.551	218.593	1829.037	47.082	45.207
7	49.467	50.023	52.216	45.551	60.807	4204.610	129.051	49.097
8	49.467	50.360	52.216	45.551	45.087	46540.341	377.216	41.697
9	49.467	50.505	52.216	45.551	62.200	1484.068	45.754	40.887
10	49.467	50.158	52.216	45.551	43.909	2226.333	50.608	50.736

priate segmentation in temporal patterns, either visually and numerically. The quantitative analysis, by means of specifically modified clustering validity indices, and a qualitative study, by observing the resulting patterns ad assigning them to one of five groups defined for typical profiles of residential energy consumption in Spain, are coincident in their results: the K-means or FCM - based algorithms, with Euclidean or Hausdorff - based distances, provide the best defined and well-balanced clusters and patterns.

The Extended K-means algorithm has not provided a similar performance

Table 9: London data test results, XB modified index.

Cycle	END-KME	END-KMC	END-KMH	Extended K-means	END-FCME	END-FCMC	END-FCMH	FFCM
1	13.440	4.896	16.063	0.455	4.003	990.513	3.001	1.950
2	13.440	5.996	16.063	0.455	3.900	1373.830	10.552	13.104
3	13.440	4.835	16.063	0.455	7.241	1802.465	168.239	2.324
4	13.440	4.620	16.063	0.455	3.063	16518.988	6.478	4.245
5	13.440	4.361	16.063	0.455	12.065	3566.012	24.452	2.020
6	13.440	5.692	16.063	0.455	90.884	1459.647	4.724	5.269
7	13.440	4.674	16.063	0.455	10.285	3408.734	53.559	6.797
8	13.440	5.543	16.063	0.455	4.054	38143.937	197.590	2.452
9	13.440	5.338	16.063	0.455	12.900	1176.610	4.021	2.127
10	13.440	4.384	16.063	0.455	2.581	1785.616	6.909	9.413

in all the cycles and does not provide a good segmentation of the customers. The algorithm is mixing profiles with different trends, since only the values from each dimension are compared at each instant of time. The resulting features are not dynamic trajectories, but an alignment of independently computed distances. Therefore this algorithm is not a good option for dynamic clustering.

Correlation is not a good similarity measure for the dynamic clustering of load profiles time series either, as stated previously, probably due to the non-linearity in the features or dimensions' trends. This measure, however, can be appropriate for other data sets.

The FFCM algorithm obtains worse results than Euclidean or Hausdorff-based distances. The reasons for these results could be in the definition of the membership function used to define the proximity between features.

Finally, as has been discussed, the appropriate selection of the number of clusters to be found is also a challenging issue. A low number of clusters will skip minority patterns, but a high value will produce empty or duplicate clusters.

The present analysis is intended to serve as a tool that may help in decision support, as a way to quickly identify the main patterns in consumption from a number of consumers, allowing to observe the evolution of the consumption through the day and also how this consumption evolves through the days. These results provide a new way of analyzing the load profiles of

energy consumption.

6. Acknowledgments

The data set for the Spanish case used in this work has been provided by the Spanish DSO Iberdrola Distribución Eléctrica S.A. as part of the works developed in the Spanish R&D project GAD. The GAD or "Active Demand Management" (in Spanish) project was a project financed by the INGENIO 2010 program and supported by the CDTI (Technological Development Centre of the Ministry of Science and Innovation of Spain).

References

[1] Consolidated View of the ETP SG on Research, Development & Demonstration Needs in the Horizon 2020 Work Programme 2016–2017, Tech. rep., ETP SG (European Technology Platform on Smart Grids) (2015).

[2] EC, Horizon 2020 - the framework programme for research and innovation, Communication (November 2011).

[3] J. Han, M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2006.

[4] J. V. de Oliveira, W. Pedrycz (Eds.), Advances in Fuzzy Clustering and its Applications, John Wiley & Sons, Ltd., 2007.

[5] T. W. Liao, Clustering of time series data - a survey, Pattern Recognition 38 (11) (2005) 1857 – 1874.

[6] J. B. MacQueen, Some methods for classification and analysis of multivariate observations, in: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, University of California, Berkeley, CA, 1967, pp. 281–297.

[7] J. C. Bezdek, Fuzzy mathematics in pattern classification, Ph.D. thesis, Faculty of the Gradual School of Cornell University, Ithaca, NY (1973).

[8] H. Izakian, W. Pedrycz, I. Jamal, Clustering spatiotemporal data: An augmented fuzzy c-means, Fuzzy Systems, IEEE Transactions on 21 (5) (2013) 855–868.

[9] H. Izakian, W. Pedrycz, Agreement-based fuzzy c-means for clustering data with blocks of features, Neurocomputing 127 (2014) 266 – 280, advances in Intelligent SystemsSelected papers from the 2012 Brazilian Symposium on Neural Networks (SBRN 2012).

[10] J. Kennedy, R. Eberhart, Particle swarm optimization, in: Neural Networks, 1995. Proceedings., IEEE International Conference on, Vol. 4, 1995, pp. 1942–1948 vol.4.

[11] I. Benítez, A. Quijano, J.-L. Díez, I. Delgado, Dynamic clustering segmentation applied to load profiles of energy consumption from spanish customers, International Journal of Electrical Power & Energy Systems 55 (0) (2014) 437 – 448.

[12] J. L. Díez, Técnicas de agrupamiento para identificación y control por modelos locales, Ph.D. thesis, Universidad Politécnica de Valencia, in Spanish (Julio 2003).

[13] G. Chicco, Overview and performance assessment of the clustering methods for electrical load pattern grouping, Energy 42 (1) (2012) 68 – 80, 8th World Energy System Conference, {WESC} 2010.

[14] M. Halkidi, Y. Batistakis, M. Vazirgiannis, On clustering validation techniques, J. Intell. Inf. Syst. 17 (2-3) (2001) 107–145.

[15] D. L. Davies, D. W. Bouldin, Cluster separation measure, IEEE Transactions on Pattern Analysis and Machine Intelligence 1 (1979) 95–104.

[16] X. L. Xie, G. Beni, A validity measure for fuzzy clustering, IEEE Transactions on Pattern Analysis and Machine Intelligence 13 (8) (1991) 841–847.

[17] F. Hausdorff, Grundzüge der Mengenlehre, Veit and Company, Leipzig, 1914.

[18] I. Benítez Sánchez, I. Delgado Espinós, L. Moreno Sarrión, A. Quijano López, I. Navalón Burgos, Clients segmentation according to their domestic energy consumption by the use of self-organizing maps, in: Energy Market, 2009. EEM 2009. 6th International Conference on the European, 2009, pp. 1–6.

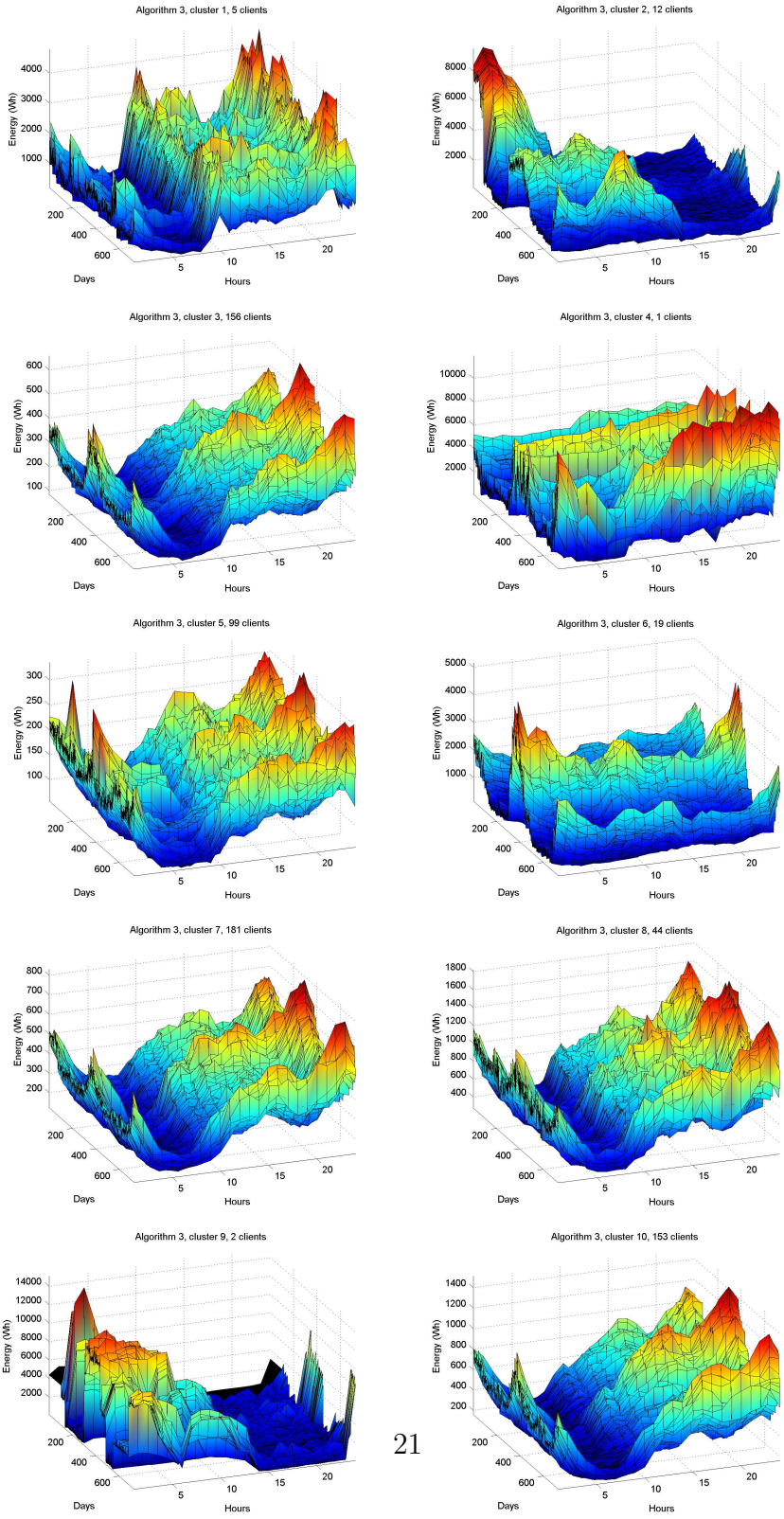


Figure 1: Results from END-KMH dynamic clustering algorithm, Spanish data.

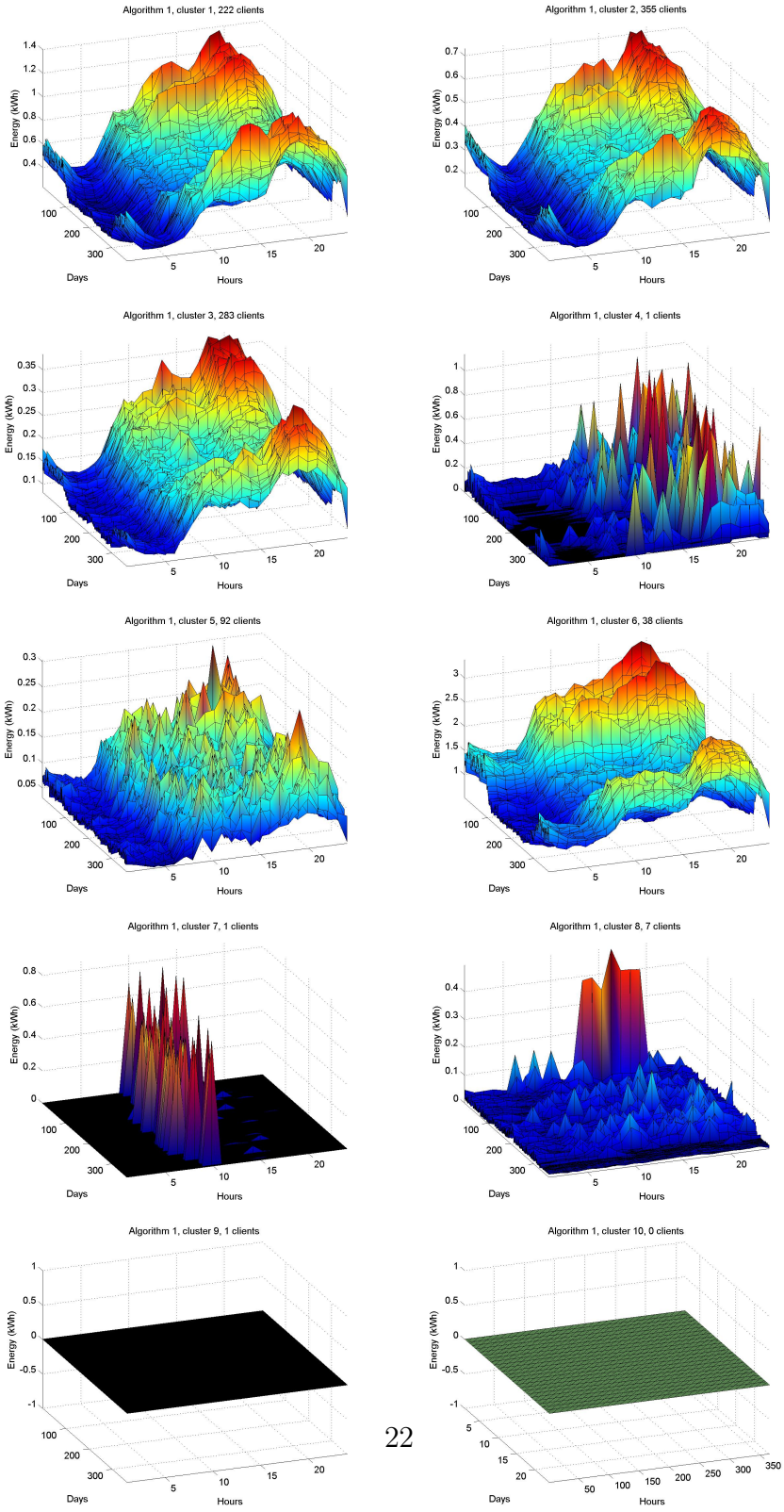
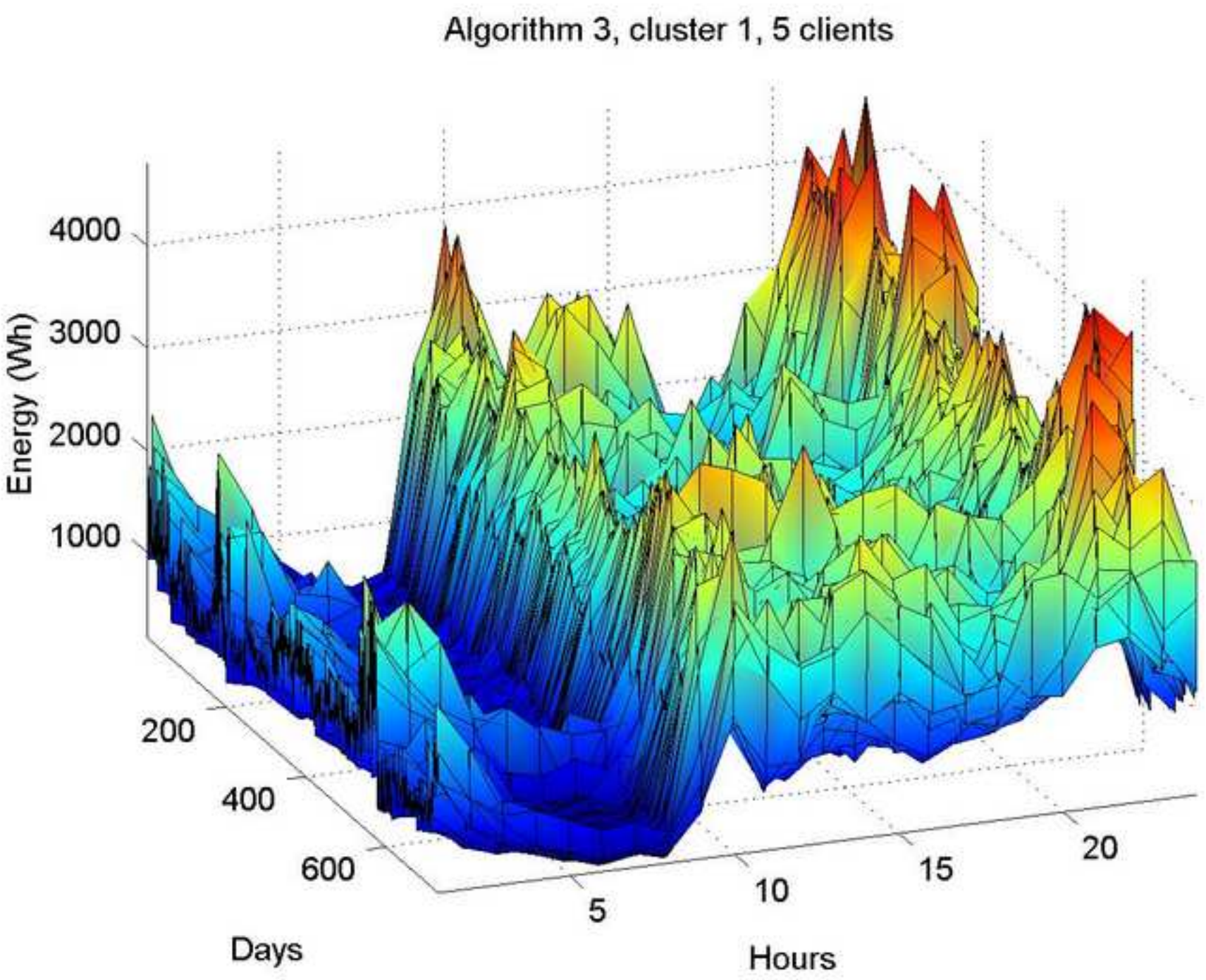


Figure 2: Results from END-KME dynamic clustering algorithm, London data.

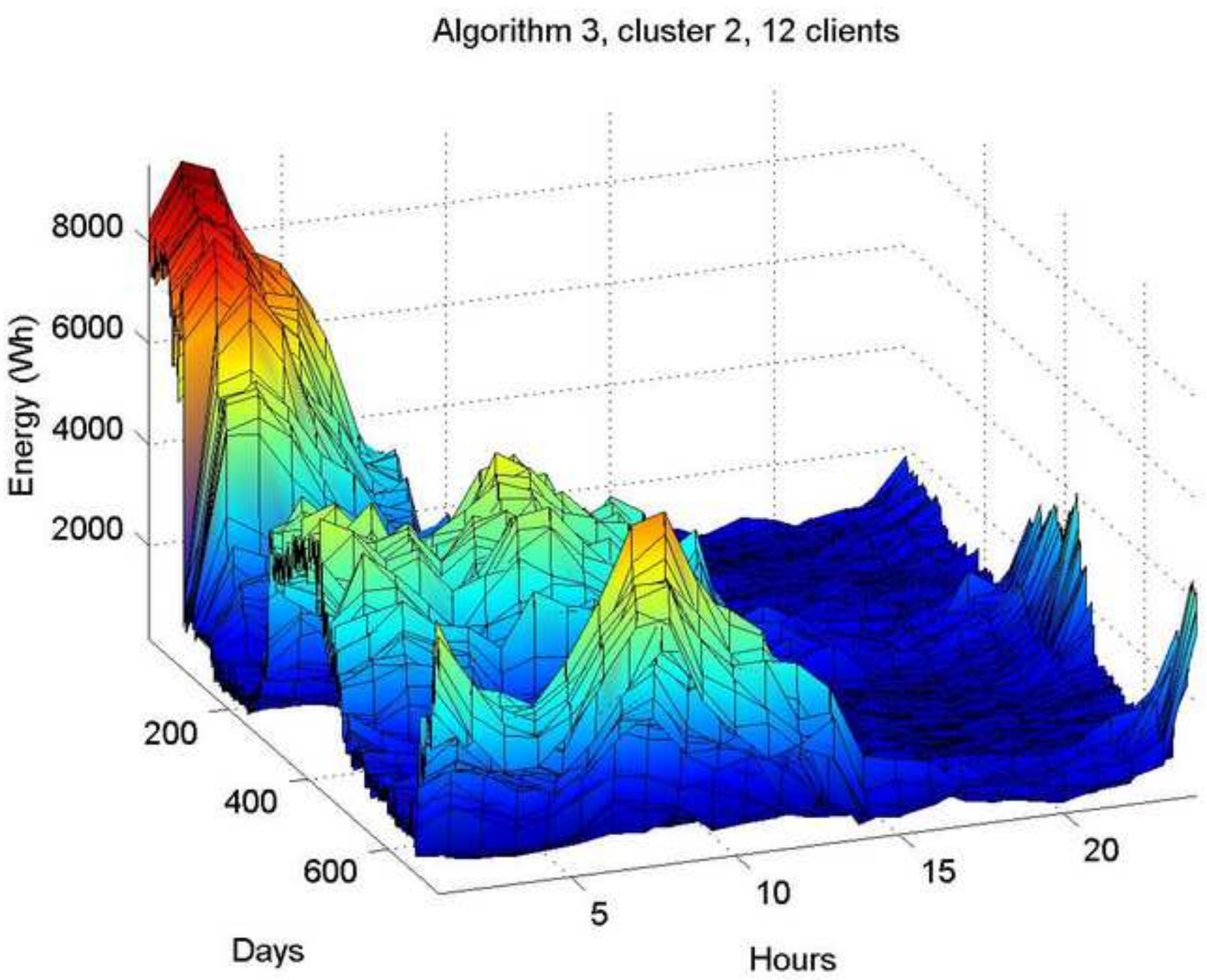
Figure(s)

Click here to download high resolution image



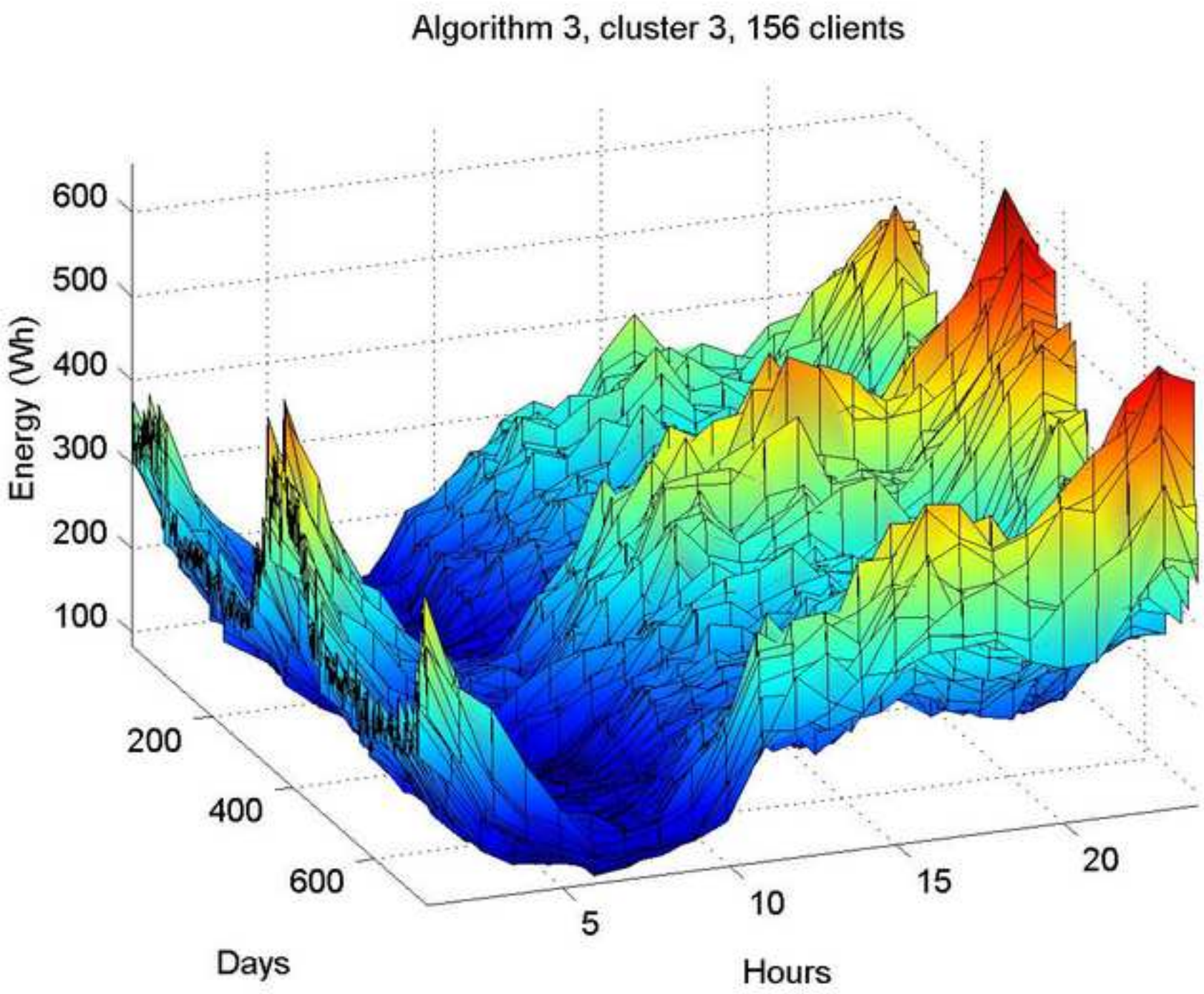
Figure(s)

Click here to download high resolution image

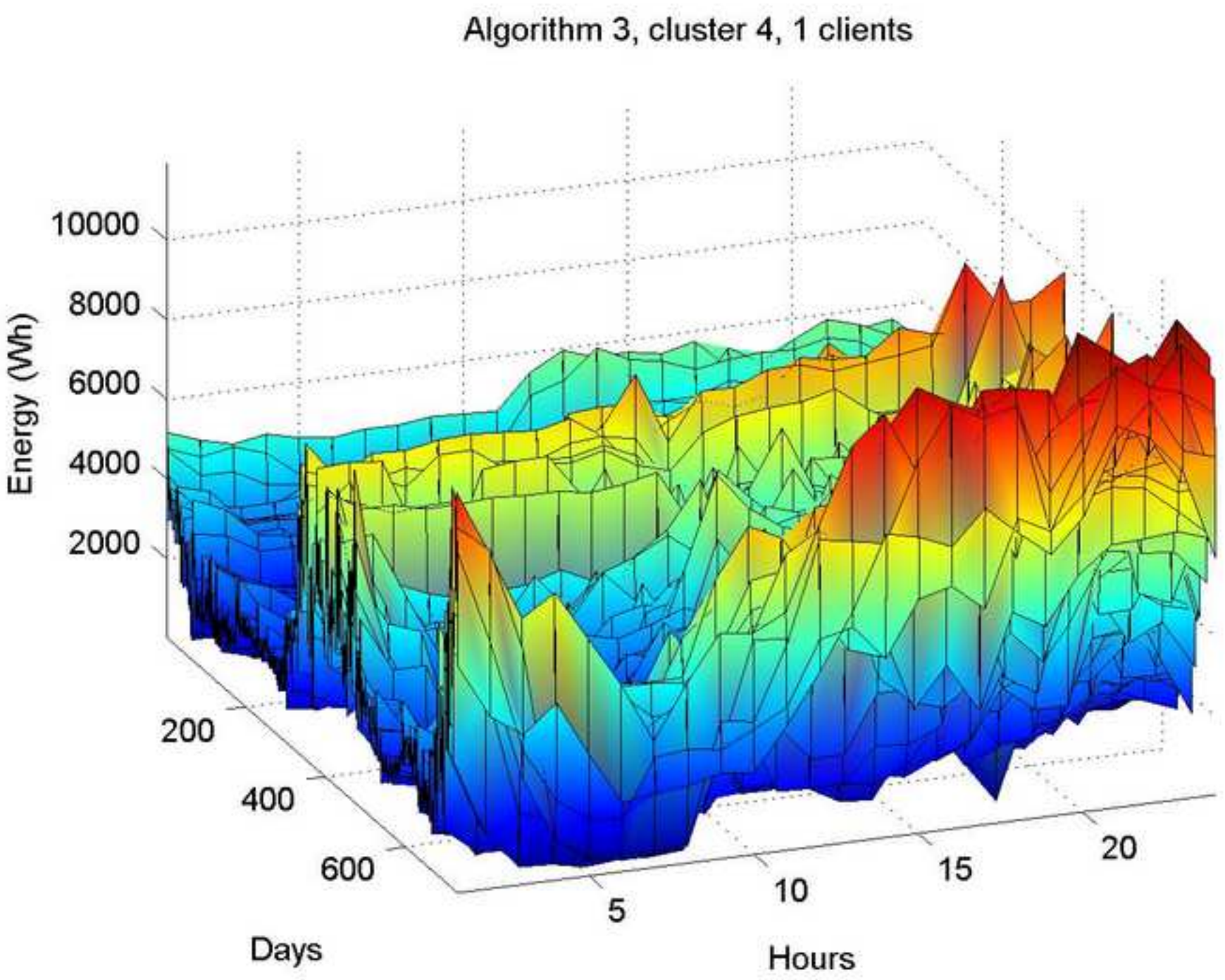


Figure(s)

Click here to download high resolution image

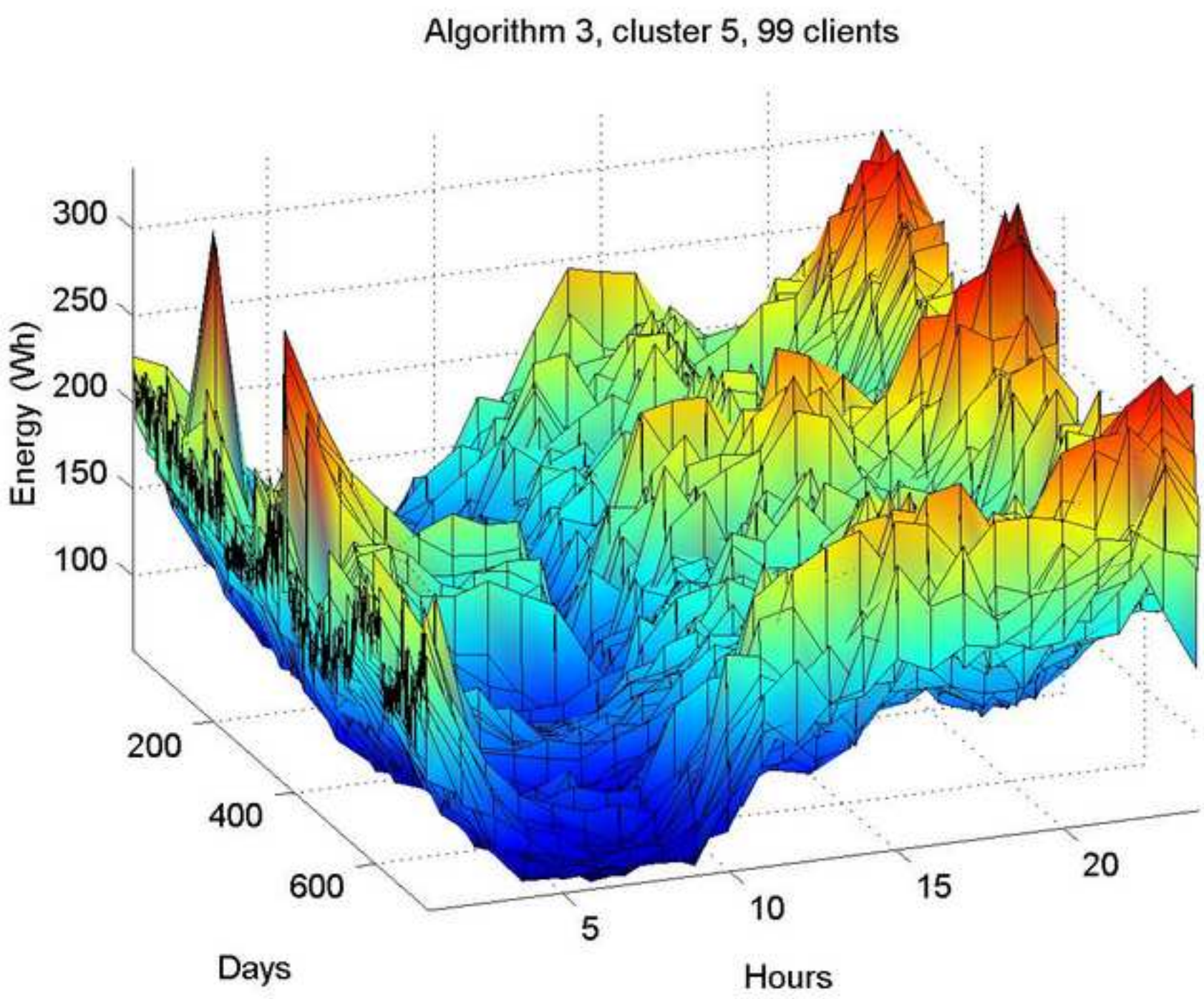


Figure(s)
Click here to download high resolution image



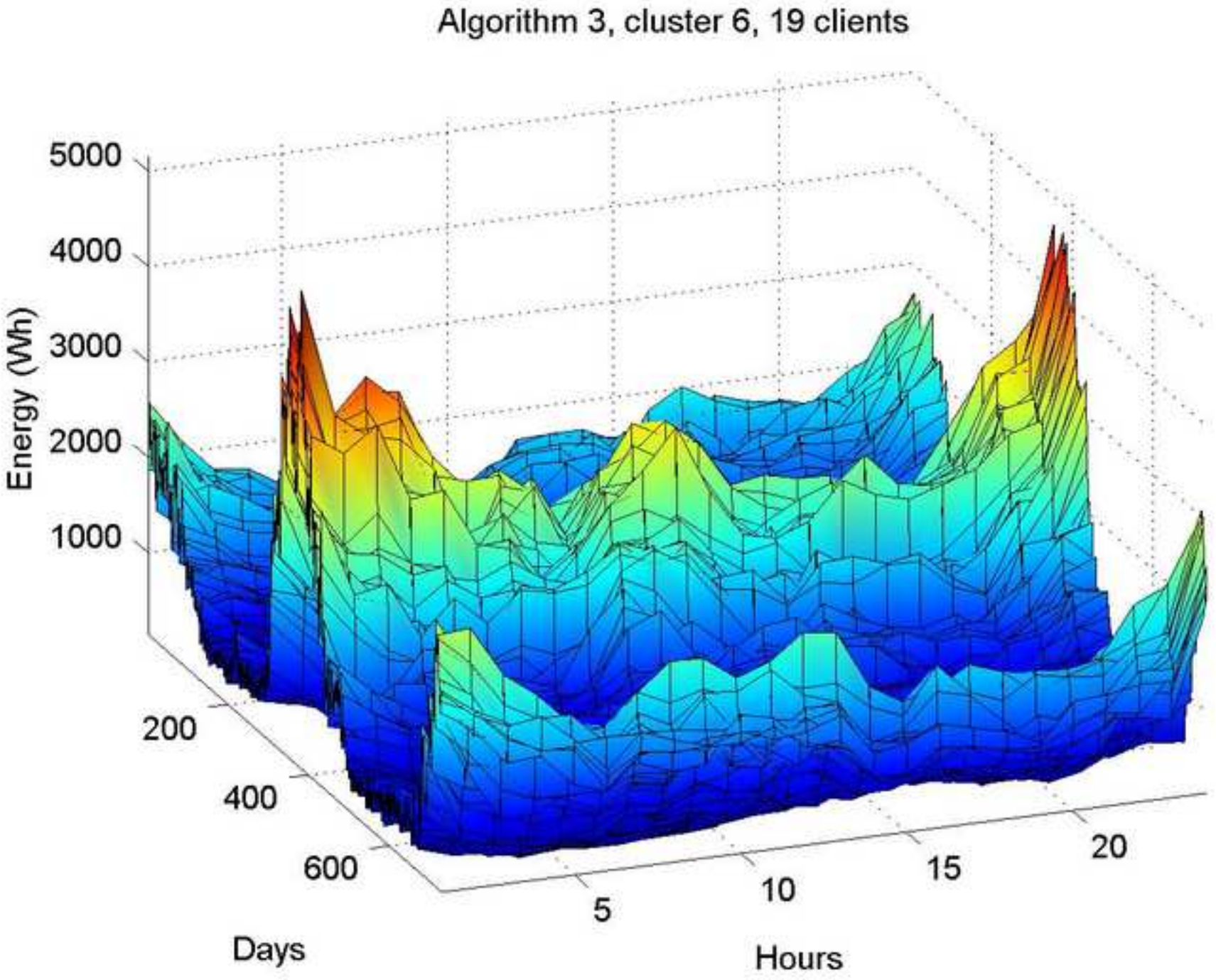
Figure(s)

Click here to download high resolution image



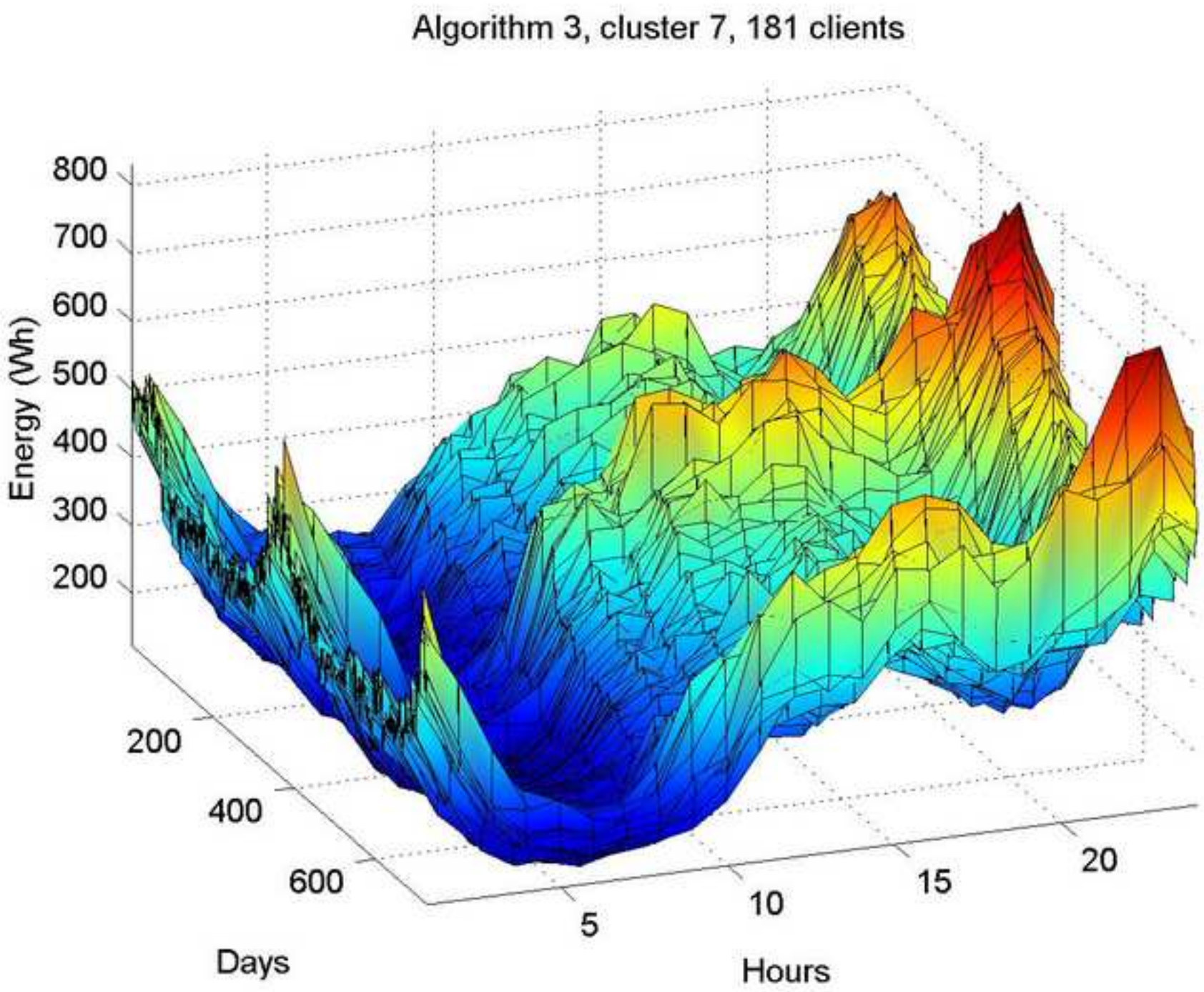
Figure(s)

Click here to download high resolution image



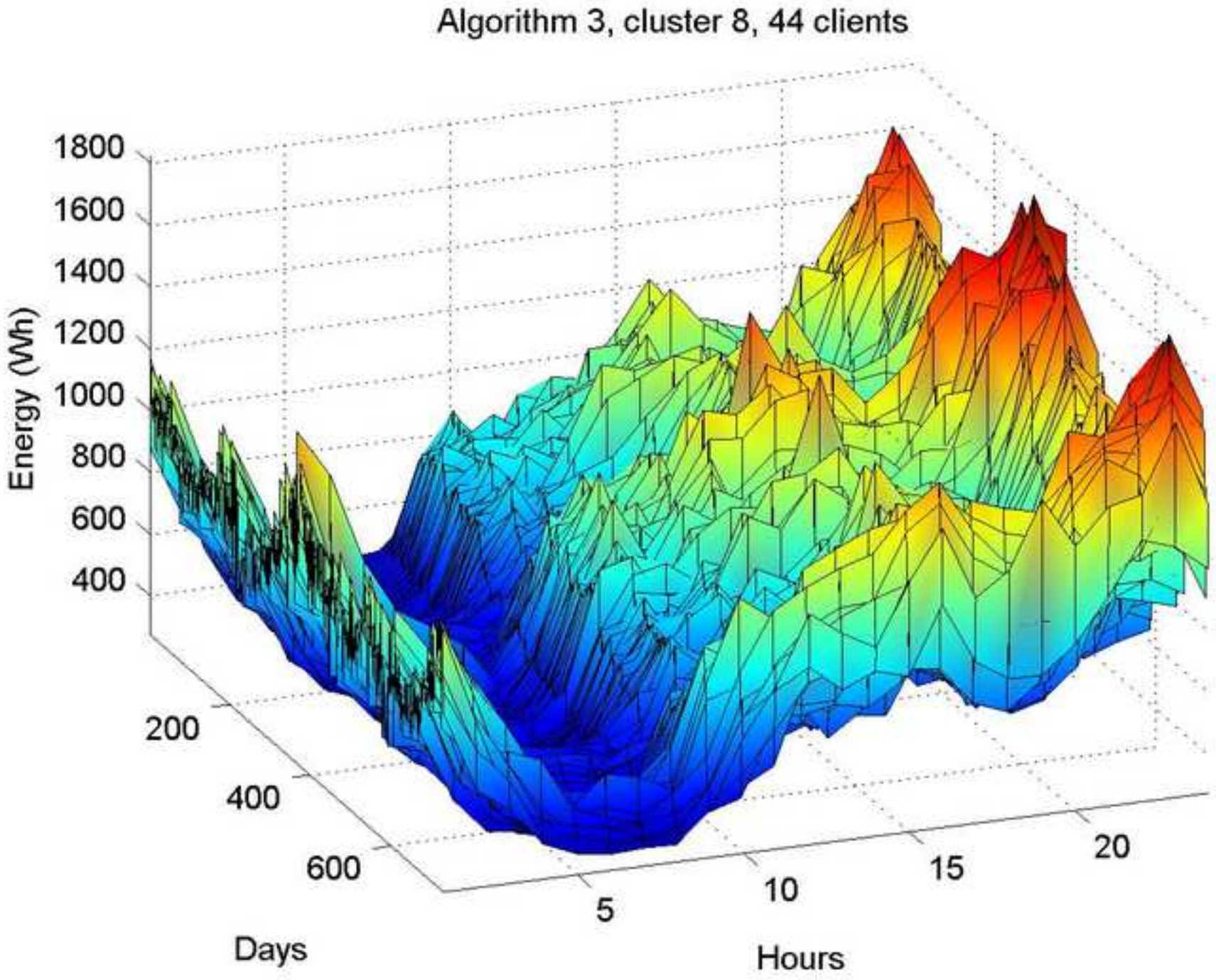
Figure(s)

Click here to download high resolution image



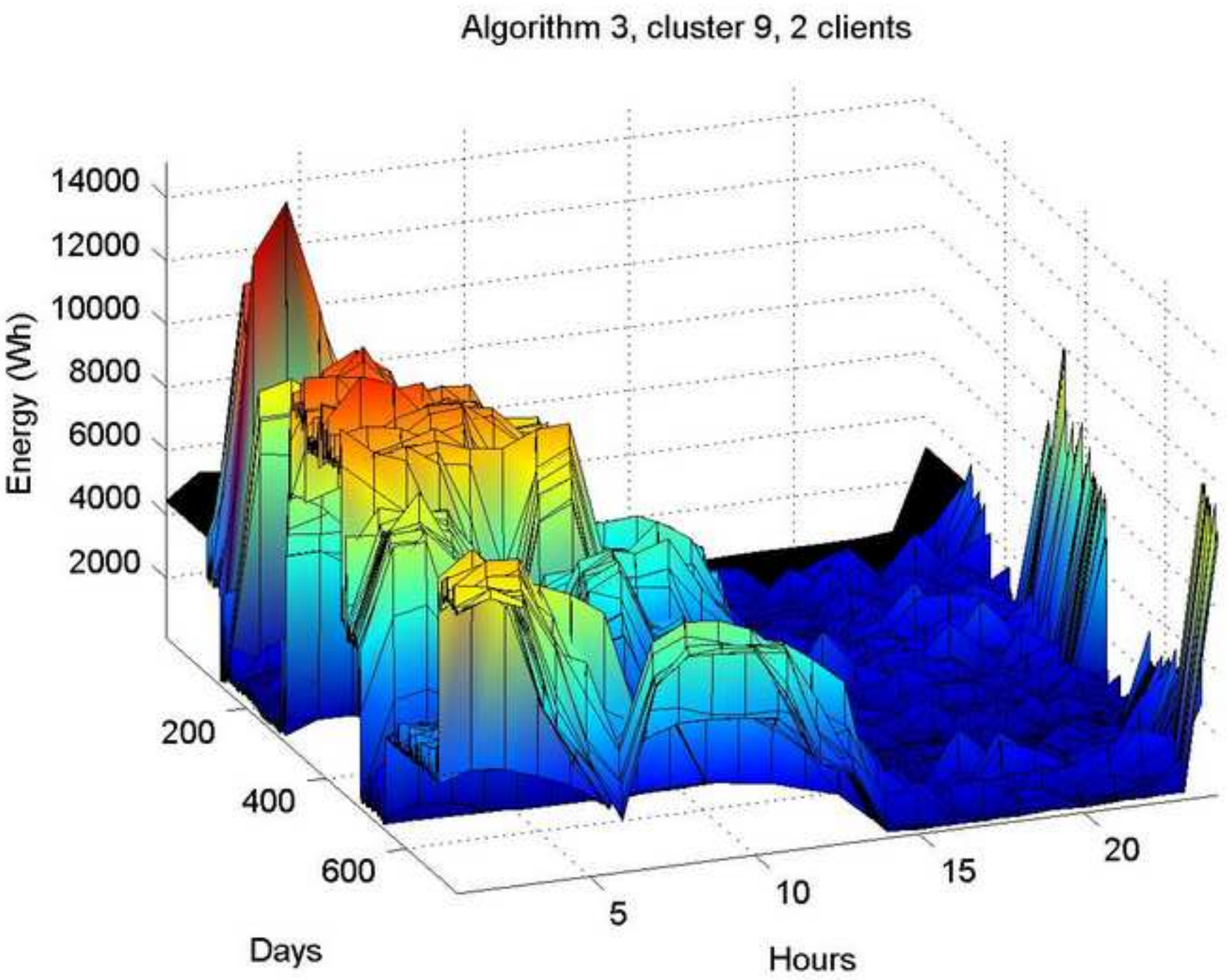
Figure(s)

Click here to download high resolution image



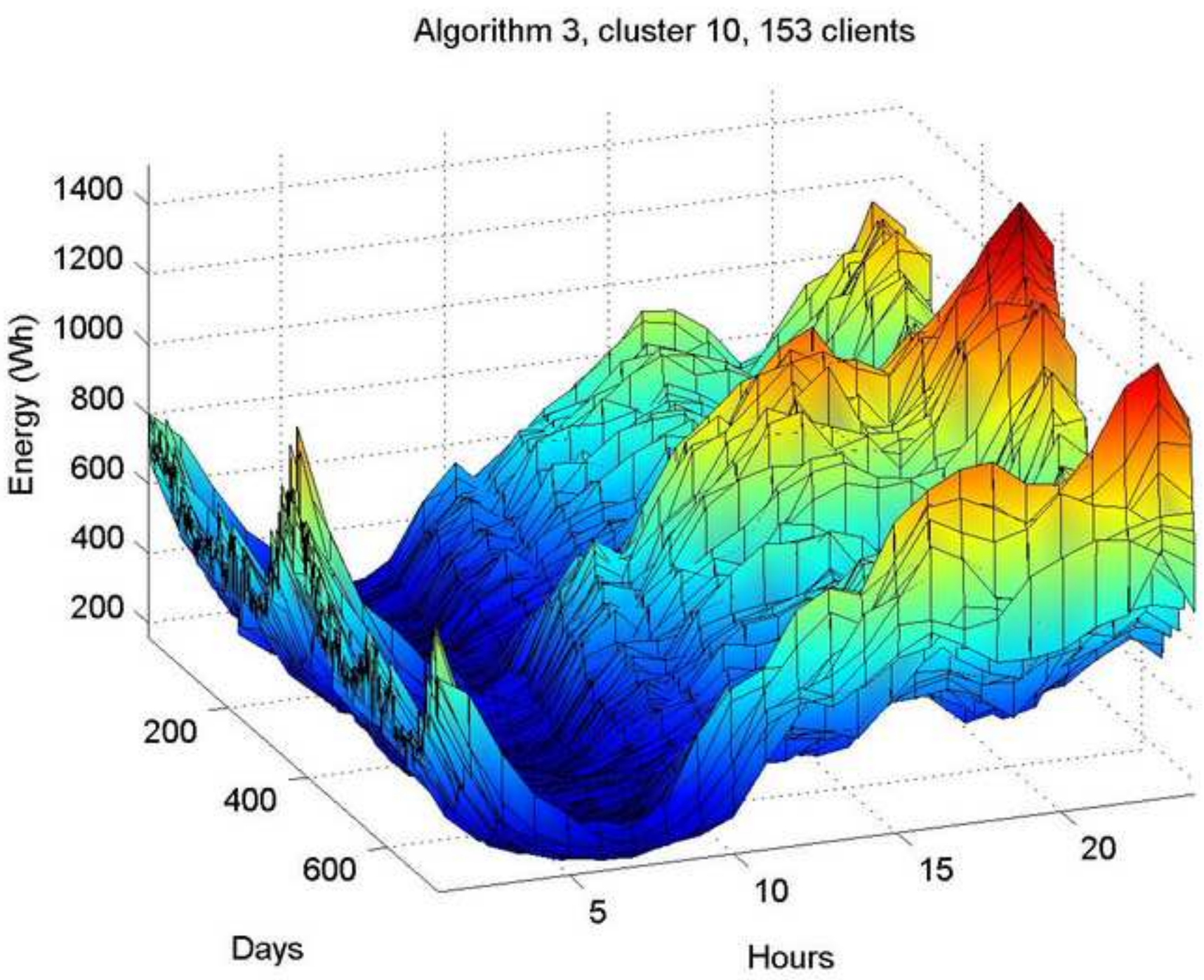
Figure(s)

Click here to download high resolution image



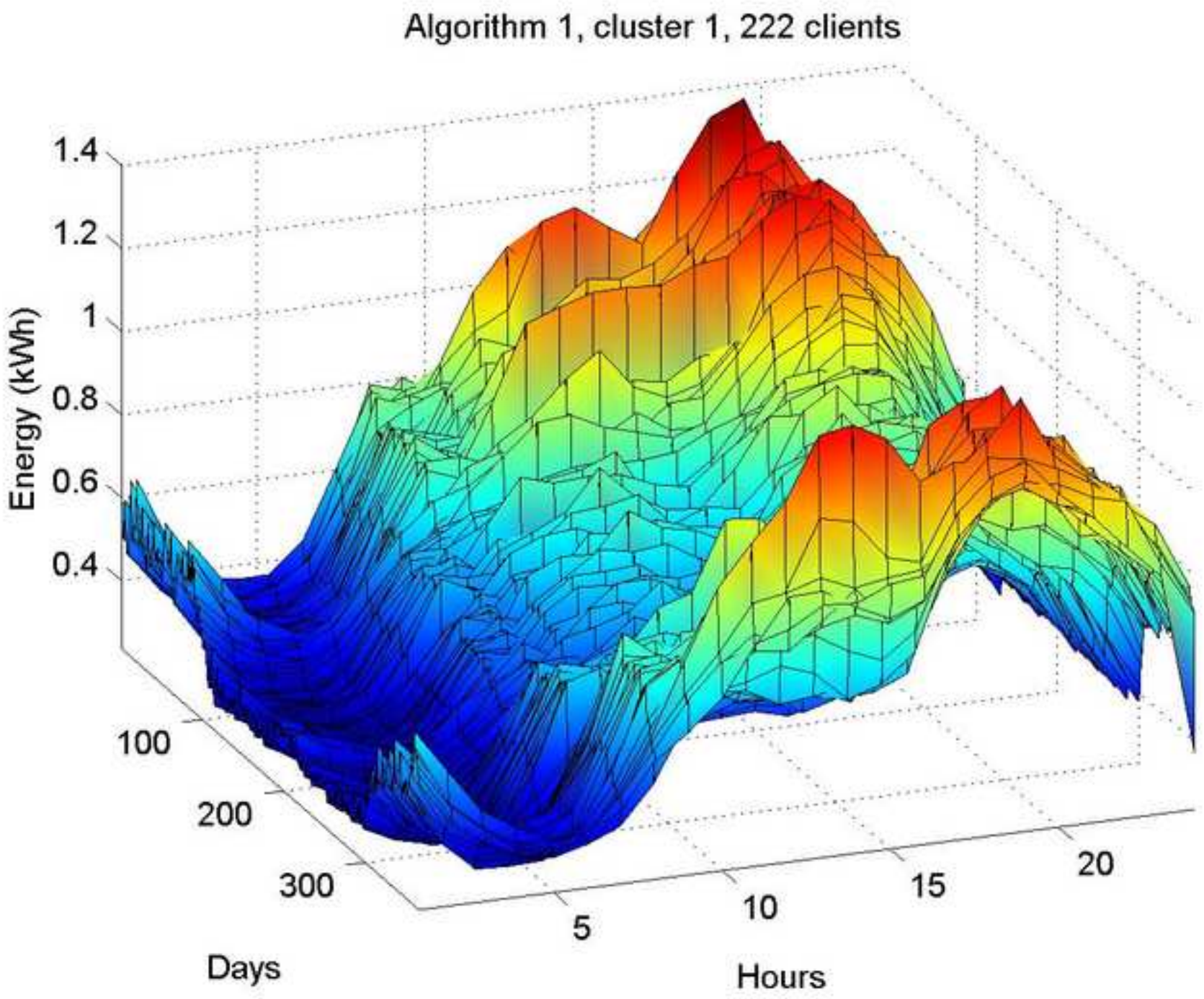
Figure(s)

Click here to download high resolution image



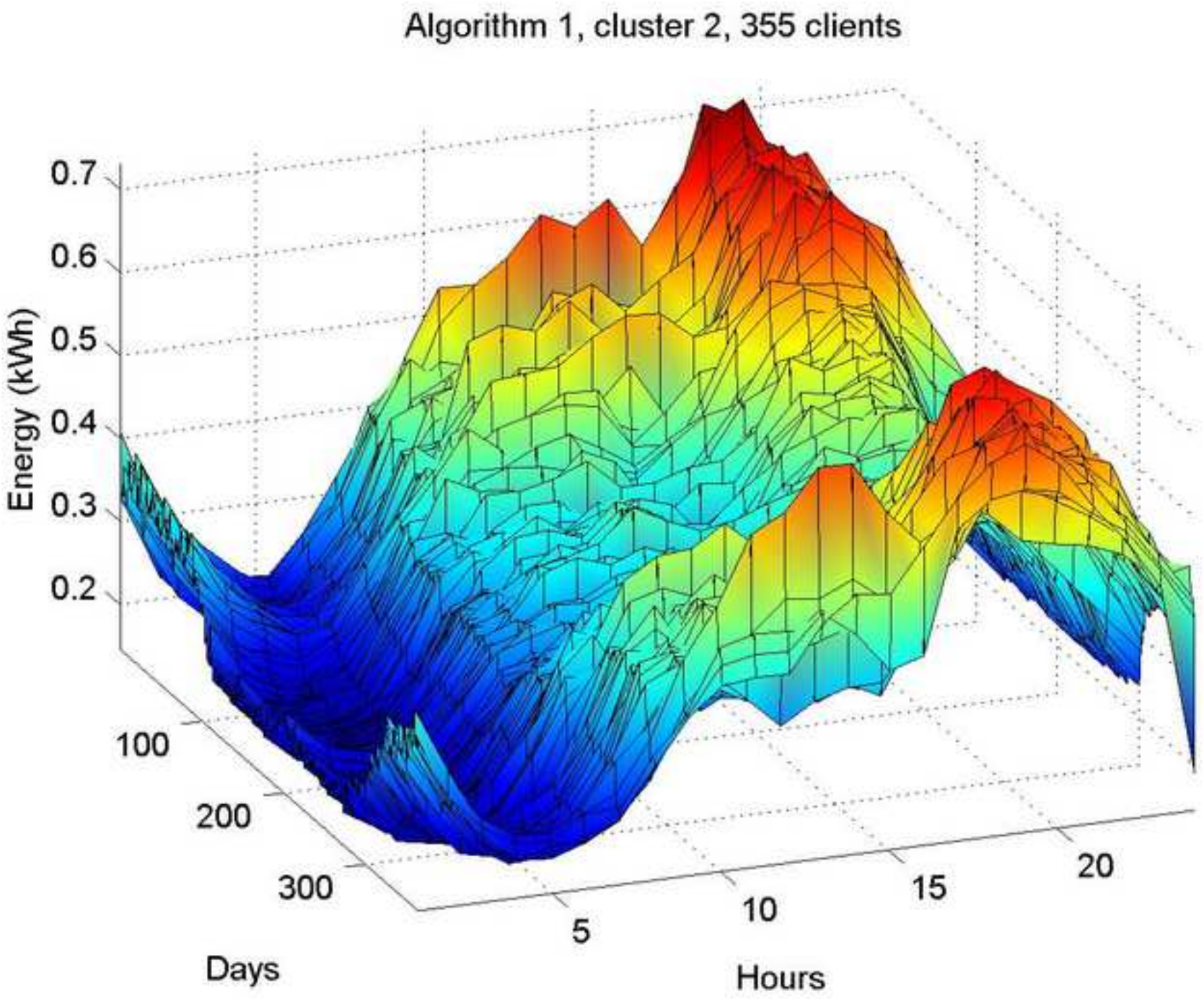
Figure(s)

Click here to download high resolution image



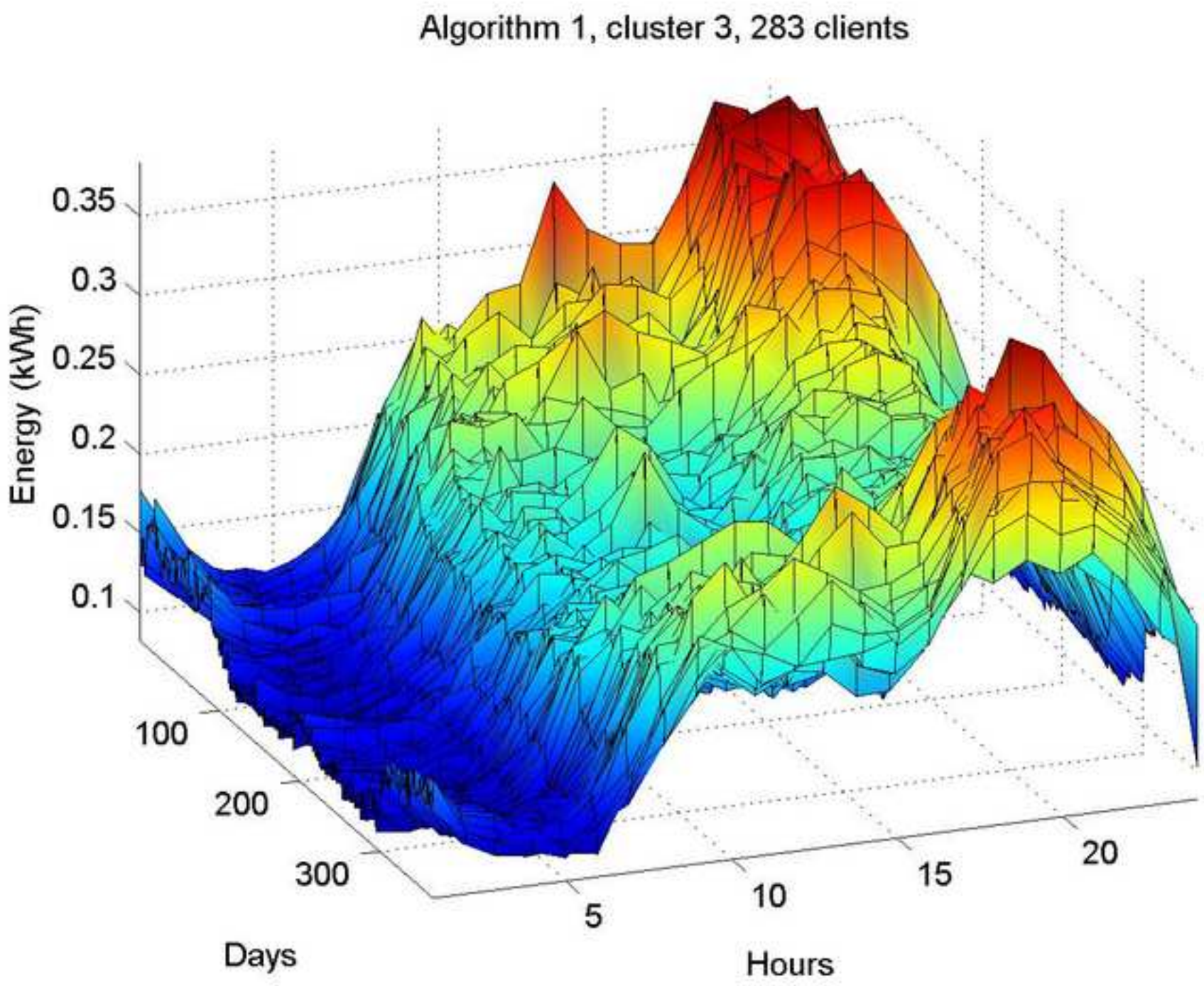
Figure(s)

Click here to download high resolution image



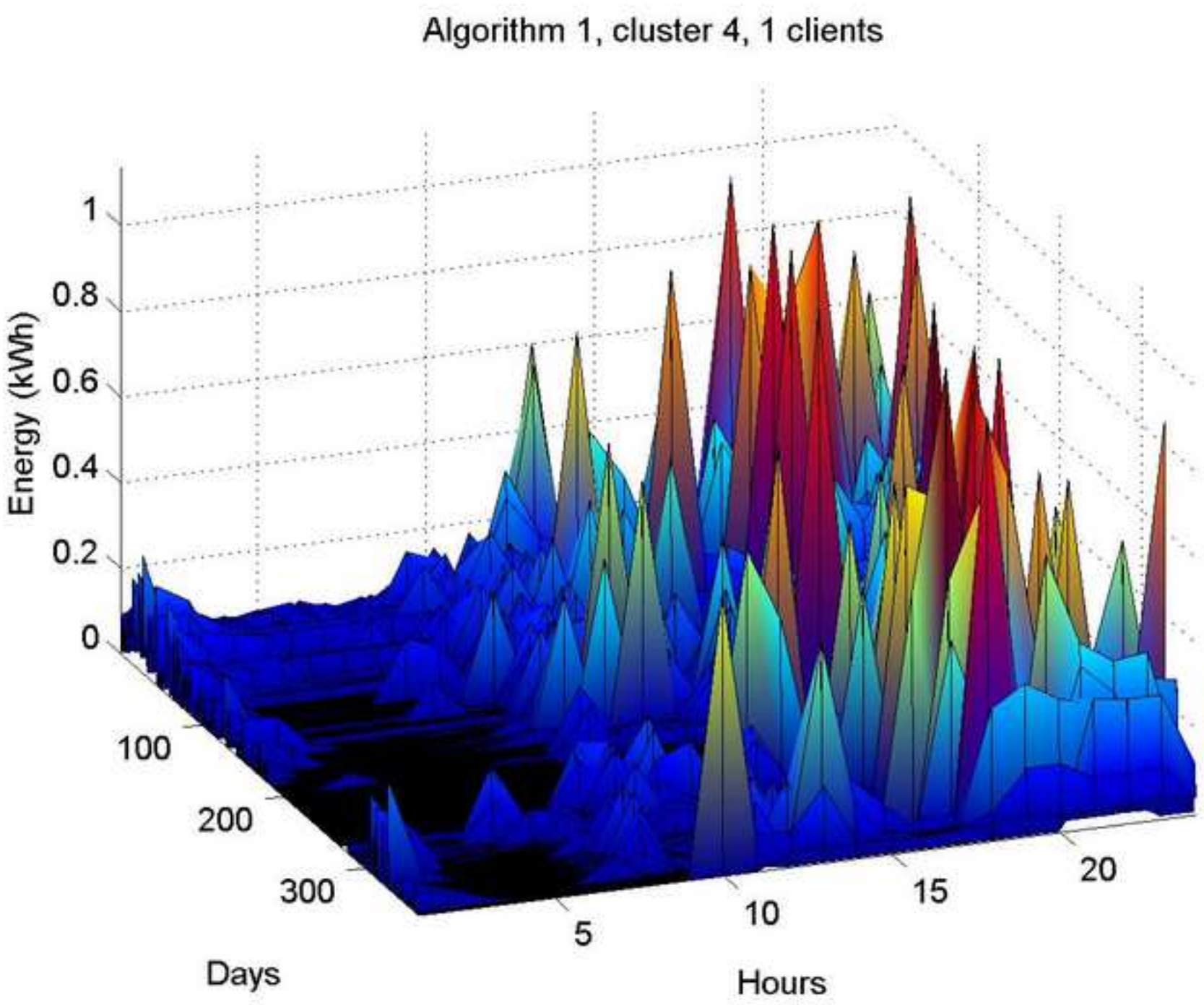
Figure(s)

Click here to download high resolution image



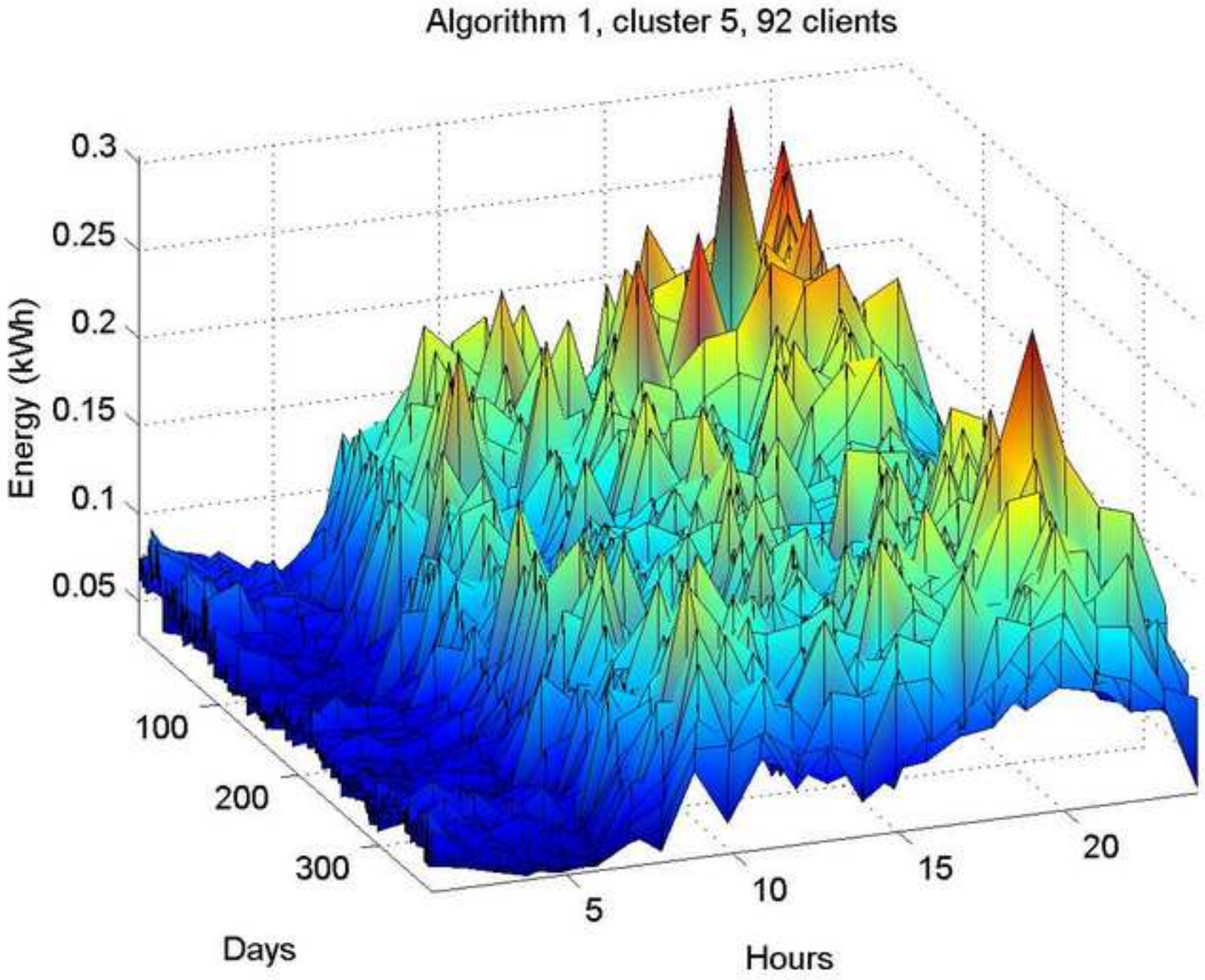
Figure(s)

Click here to download high resolution image



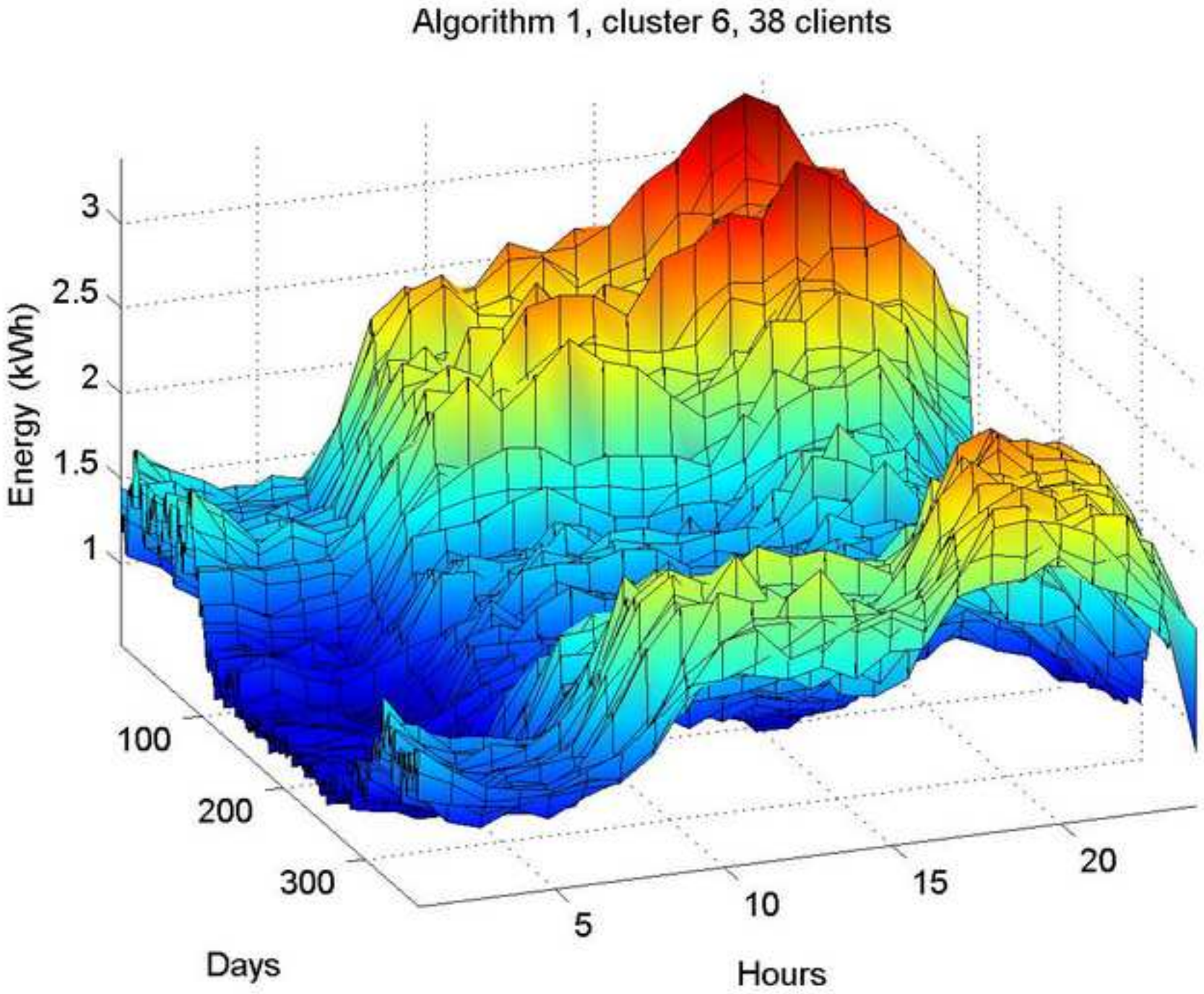
Figure(s)

Click here to download high resolution image



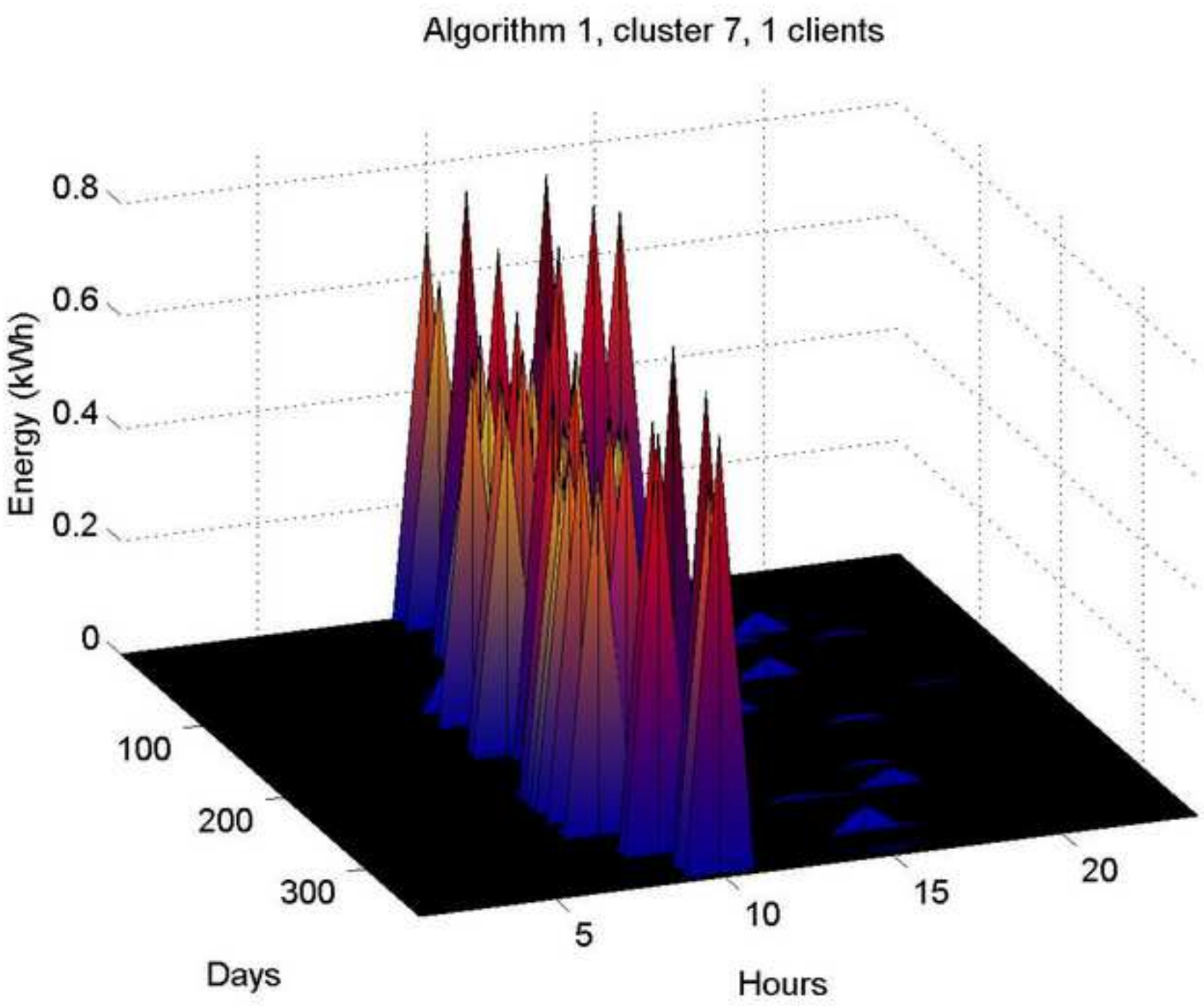
Figure(s)

Click here to download high resolution image



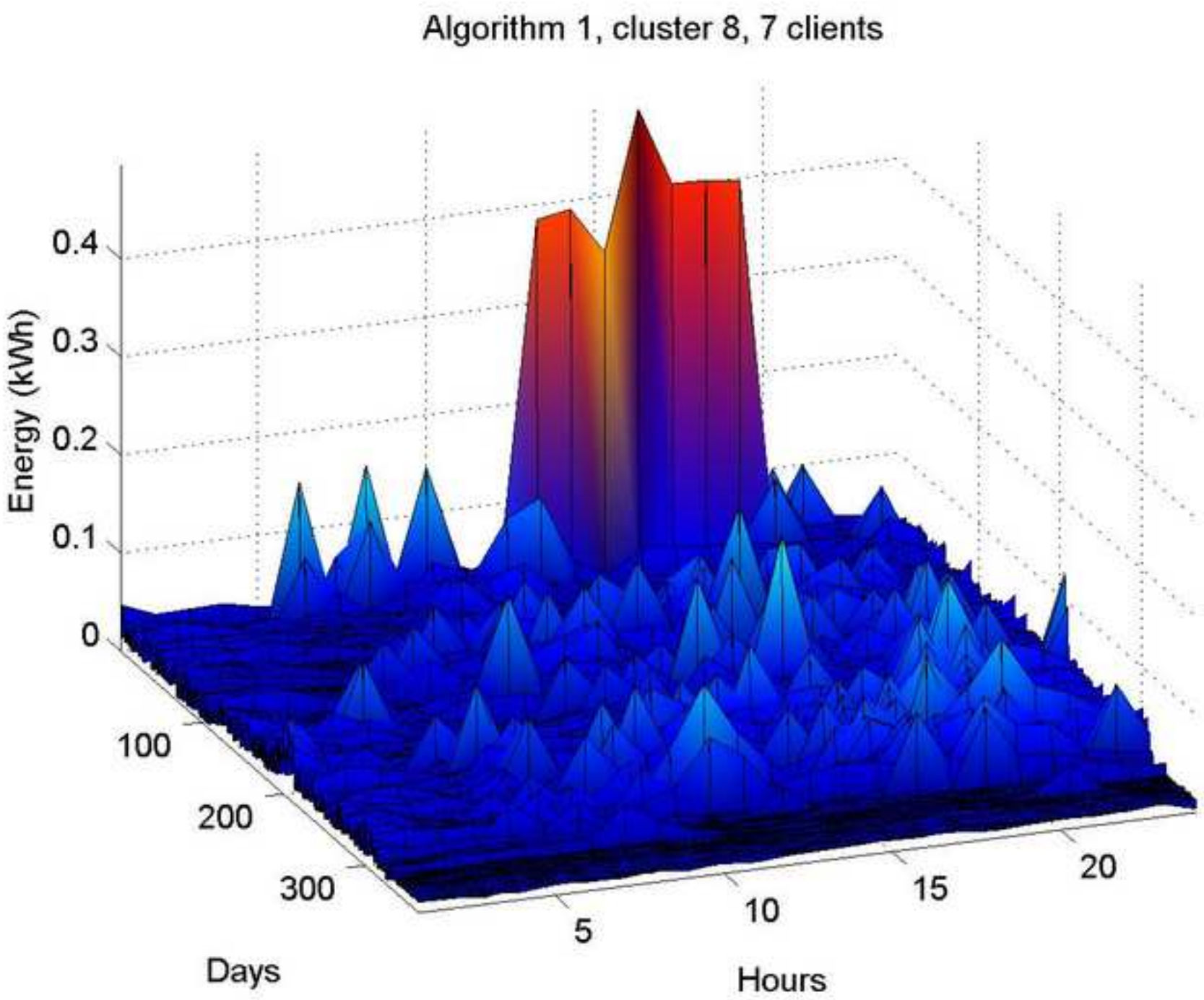
Figure(s)

Click here to download high resolution image

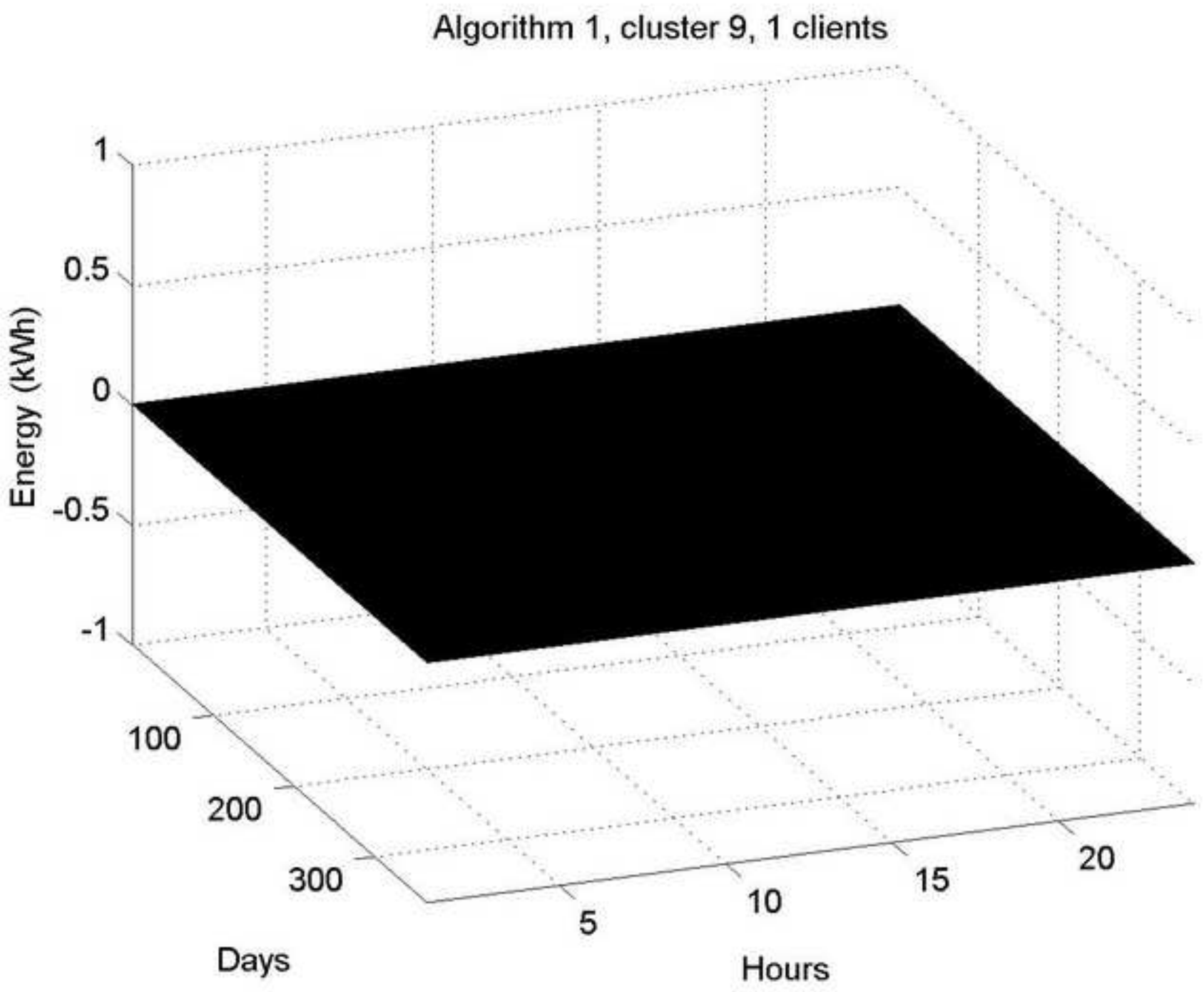


Figure(s)

Click here to download high resolution image

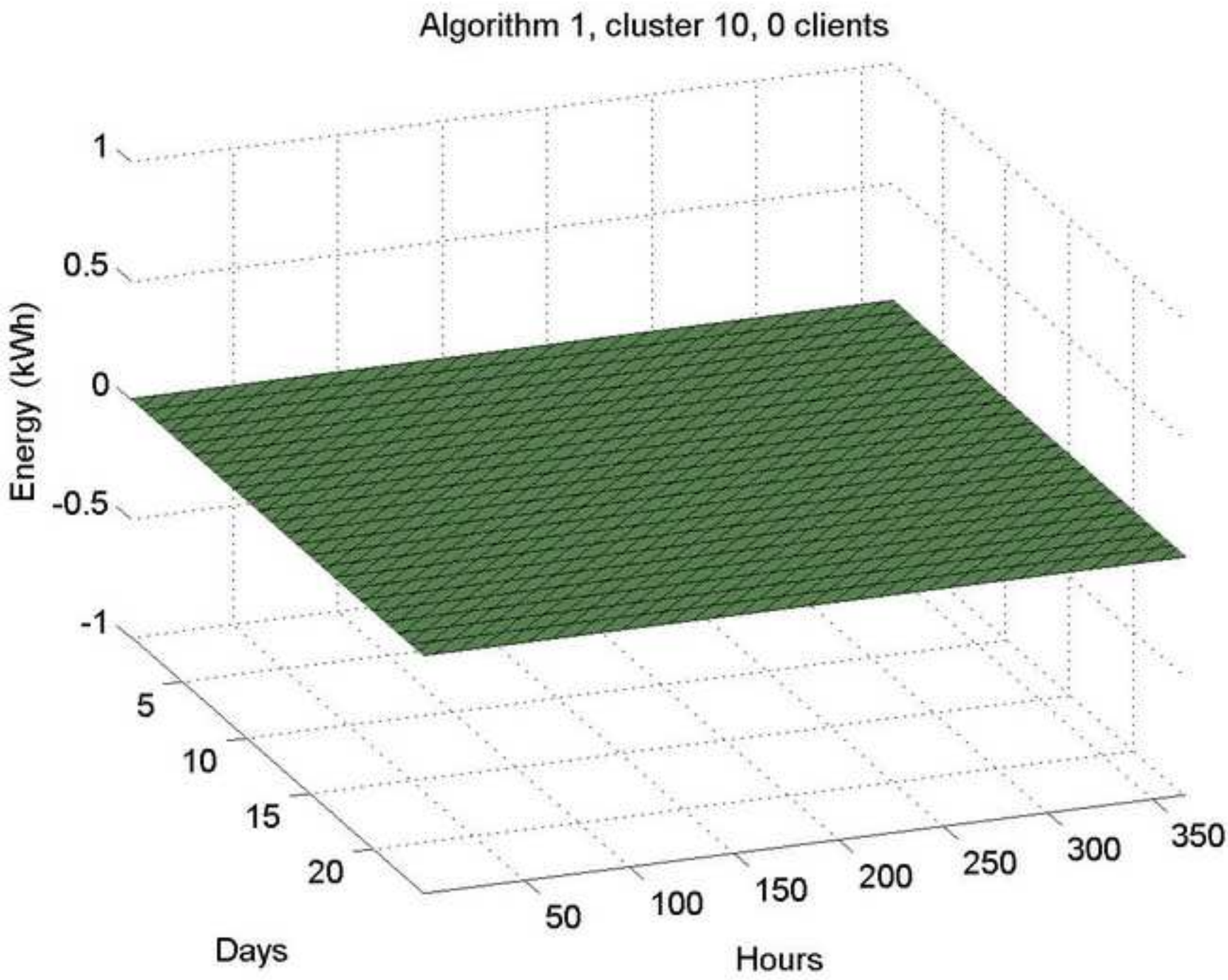


Figure(s)
Click here to download high resolution image



Figure(s)

Click here to download high resolution image



LaTeX Source Files

Click here to download LaTeX Source Files: EPSR_dynamic_clustering_hausdorff_distance_v3.tex