# New Class of Simple and Efficient Clustering Algorithms for Multiscale Mathematical Programming with Demand Data Applications

**Falah Alhameli[1], Ali Elkamel[1,2] , Mohammed Alkatheri[1], Alberto Betancourt-Torcat[1], and Ali Almansoori[2]**
[1]Department of Chemical Engineering
University of Waterloo
Waterloo, ON N2L 3G1, Canada

[2]Department of Chemical Engineering
Khalifa University of Science and Technology
Abu Dhabi, P.O. Box: 127788, United Arab Emirates

falhamel@uwaterloo.ca, aelkamel@uwaterloo.ca, ali.elkamel@ku.ac.ae,
mohammed.alkatheri@uwaterloo.ca, alberto.betancourt@uwaterloo.ca, ali.almansoori@ku.ac.ae

## Abstract

Integration across a supply chain decision levels is key on improving investment returns. Integration of different time scales leads to large scale problems usually computationally intractable. Different approaches have been proposed to tackle the problem in terms of modeling and solution methods. However, most of them are problem specific or applicable only to short time horizons. Clustering has the potential to handle such problems by grouping similar input parameters together and considerably reducing the model size while not compromising solution accuracy. This work presents a new class of clustering algorithms to support the integration of planning applications of different time scales. The clustering algorithms were formulated using integer programming with integral absolute error as similarity measure. Two different clustering algorithms were developed: normal and sequence. The models were developed in the GAMS software. Two case studies are presented to assess the algorithms outputs and computational performance using utility demand data. It was found that the algorithm is capable of finding good quality solutions; and even succeed at finding optimal solutions with a small computational effort while providing clusters with high intra-cluster similarity and low inter-cluster similarity.

**Keywords**
Multiscale, Clustering, Algorithm, Modeling and Computational performance.

## 1. Introduction

Supply chain management has demonstrated ability to increase profits while upholding customer satisfaction (Papageorgiou 2009). This comprises three decision levels: strategic, tactical, and operational. Decision makers usually follow the aforementioned sequential mode (Grossmann 2012). Nevertheless, these levels are subject to each other, and comprise dissimilar time scales. Consequently, their integration is fundamental to increase efficiency and profit. The planning decisions must be carried out concurrently if one is to achieve a global optimum. Due to the difference in time scales between the three supply chain management elements, their integration frequently derives in a multiscale model which is in practice computationally intractable. Although diverse methodologies have been suggested to solve this problem from a modeling and solution perspective, most of the methodologies are specific to a problem or its applicability limited to short timeframes. Accordingly, clustering arises as a valid and suitable option to handle this type of problem by grouping similar inputs, such as price or demand, together. Input parameters typically are made up of multiple attributes like simultaneous electricity and heat demands. This allows to substantially reduce the size of the model and improve computational tractability while keeping solution accuracy.

Clustering orders the data into homogeneous groups where the within-group-objects similarity is minimized whereas the between-group-objects dissimilarity is maximized. The aim is that objects in a group shall be similar or linked to one another and dissimilar or unlinked to objects in other groups. An affective clustering is characterized by a high similarity/homogeneity within groups and high dissimilarity/heterogeneity between groups. The objects are normally represented by vectors in a multidimensional space; in which each dimension represents a specific attribute (e.g., variable, measurement) describing the object. Each attribute is considered to be represented by values. Clustering has been studied for more than 50 years in diverse disciplines (Jain 2010).

Mathematical programming is key in the development of clustering algorithms. For example, Balachandra and Chandru (P Balachandra and Chandru 1999) grouped a whole year electricity demand into 9 clusters sequentially applying discriminant analysis. Later, the clusters were used as inputs in a mathematical model of an electricity system based on supply-demand matching (Patil Balachandra and Chandru 2003). Likewise, Fazlollahi et al. (Fazlollahi, Becker, and Maréchal 2014) developed an algorithm to cluster electricity demand using $k$-means. The algorithm was extended to include attributes such as: heat demand, electricity price, and solar radiation. The clusters were used as input for the operation of fixed energy systems. However, the study does not display the solution quality nor the solution approach.

This work aims to tackle the integrated supply chain problem employing a clustering approach. The objective is minimizing the model size by representing the days in a year by typical days representative of the operating year. While it is true clustering has been broadly used in different applications, there is a lack of analysis in demand patterns clustering. The latter are complex to represent due to their multidimensional nature involving shape and time dependent attributes (e.g., utility demand). The present works takes a mathematical programming approach to tackle the integrated supply chain management problem and proposes a Mixed Integer Linear Programming (MILP) formulation for the clustering algorithms. Therefore, the present work aims to analyze the clustering of demand patterns with multiple attributes for multiscale models. The $L_1$-norm (least absolute value method) (Bektaş and Şişman 2010; Chelmis, Kolte, and Prasanna 2015; Green, Staffell, and Vasilakos 2014; Lyu et al. 2013; Sabo 2014) is employed as similarity measure. The paper is organized as follows: Section 2 presents the proposed clustering algorithms formulation. Section 3 presents a heuristic size-reduction algorithm. Section 4 shows two Case Studies involving electricity and heat demands data. Concluding remarks are presented at the end of this work.

## 2. Clustering Algorithm

The proposed clustering algorithm is part of the time-series data. It can cluster demand data by considering shape-similarity and trajectories-time at the same time. Thus, it can help minimizing the computational complexity of multiscale models. Input parameters typically involve multiple attributes like the simultaneous electricity and heat demands. The weighting method is used to deal with the multiplicity of the demand data attributes. This can be expressed in the following form:

$$\min\ X = \sum_a W_a\ IAE_a\ , \tag{1}$$
$$\text{s.t.} \qquad \sum_{c=1}^{C} y_{d,c} = 1 \qquad \forall\, d\ ,$$

where $X$ is the multi-objective performance criteria function to be minimized, $IAE_a$ denotes the attribute $a$'s $L_1$-norm or integral absolute error, $W_a$ attribute $a$'s weighting factor ($W_a \geq 0, \sum_a W_a = 1$), $y_{d,c}$ denotes the binary variable allocating loads for day $d$ joining cluster $c$. The integral absolute error can be defined as follows:

$$IAE_a = \frac{\Delta}{2} * \sum_{d=1}^{D} \sum_{h=1}^{H-1} D_{a,d,h} + D_{a,d,h+1} \qquad \forall\ a\ , \tag{2}$$

where $D_{a,d,h}$ represents the absolute difference between load curve $l$ and clustered curve $c$ for hour $h$ in day $d$ for attribute $a$. The absolute difference between the load and cluster curves for the performance criterion is given as:

$$D_{a,d,h} \geq \left| DEML_{a,d,h} - DEM_{a,c,h} \right|\ y_{d,c} \qquad \forall\ a,h,d,c\ , \tag{3}$$

where $DEML_{a,d,h}$ denotes the $a$'s attribute demand load for hour $h$ in day $d$, $DEM_{a,c,h}$ the demand for hour $h$ in cluster $c$ and attribute $a$. It is important to notice that other integration schemes (e.g., Simpson's 1/3 rule) could be used as well. Additionally, the use of the $L_2$-norm is also easy to implement and only requires the use of the Euclidean distance in (2).

Demand data can be clustered in a sequential way if one defines a constraint set following the string property concept (Vinod 1969). This type of clustering is important in processes with flexible operations such as processes subject to change-overs and set-ups. In order to incorporate time-dimension into the clusters and therefore sequencing, the following set of constraints can be included:

$$y_{d+1,1} \leq y_{d,1} \qquad \forall\, d < D\ , \tag{4}$$

$$y_{d+1,c} \leq y_{d,c} + y_{d,c-1} \qquad \forall\, d < D, c > 1\ , \tag{5}$$

$$y_{D,c} \leq y_{D-1,c} + y_{D-1,c-1} \qquad \forall\, c > 1\ , \tag{6}$$

Equations (4)-(6) handle the first, intermediate, and last clusters sequencing, respectively. Each subsequent equation is comparable to the previous constraints set as long as that the non-existing terms are taken out the equation. This feature can be found in many algebraic modeling systems (i.e., GAMS).

$$y_{d+1,c} \leq y_{d,c} + y_{d,c-1} \qquad \forall\, d, c\ , \tag{7}$$

The aforementioned general formulation offers a single platform for normal and sequence clustering given its equivalent algorithmic structure. It renders a mixed integer nonlinear programming (MINLP) model due to the multiplication of $DEM_{a,c,h}$ and $y_{d,c}$ variables illustrated in (3). Nevertheless, the model can be easily turned into a MILP by applying common linearization methods on the absolute function (Mangasarian 2013). The model used in this work is the linearized version of the clustering algorithm. In summary, the model for normal clustering is made up by (1)-(3); whereas sequence clustering is denoted by (1)-(3), and (7).

## 3. Heuristic Algorithm for Size Reduction

Given the computational complexity of the proposed clustering algorithm described in the previous section, a simple heuristic algorithm, which compares lower and upper bound solutions in an iterative way, was developed to help reducing the problem size including single or multiple attributes. The developed size-reduction algorithm allows maintaining the linearity and programming basis of the proposed MILP model of the previous section. The developed heuristic uses the $k$-means algorithm (Xu and Wunsch 2008); however, in the present approach the clusters are arranged employing the MILP model described in Section 2. Although the $k$-means is mostly used in one-dimension time-series data, the version applied in this work is capable of dealing with higher dimensions. Figure 1 explains the developed size-reduction heuristic that can be applied to single and multiple attributes. For single attribute the weighting factor is assumed to be 1 ($W_a = 1$); which simplifies the heuristic algorithm.

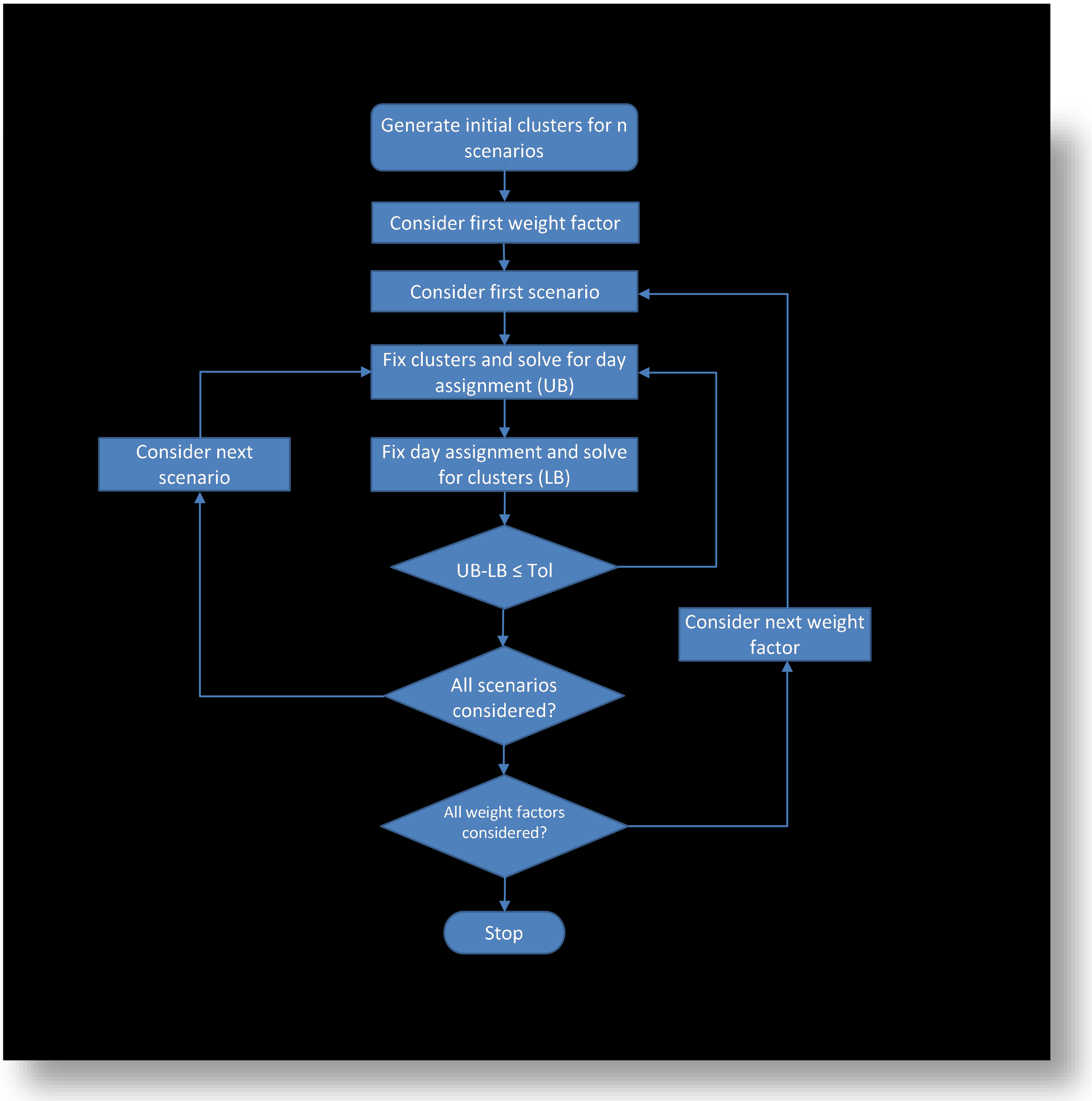
Generate initial clusters for n scenarios
Consider first weight factor
Consider first scenario
Fix clusters and solve for day assignment (UB)
Consider next scenario
Fix day assignment and solve for clusters (LB)
UB-LB ≤ Tol
Consider next weight factor
All scenarios considered?
All weight factors considered?
Stop

Figure 1. Proposed heuristics algorithm for single and multiple attributes.

## 4. Case Studies

The proposed clustering algorithms were applied in two different case studies: 1) single attribute, and 2) multiple attributes. The first case study involves the Unit Commitment (UC) model (Marcovecchio, Novais, and Grossmann 2014; Padhy 2004); whereas the second includes a full energy hub model (Maroufmashat et al. 2015).

### 4.1 Case Study 1: Single Attribute Problem

The present case study analyzes the impact on solution accuracy when clustered demand in a normal and sequence mode are applied to a planning model (Marcovecchio, Novais, and Grossmann 2014; Padhy 2004). The UC model was selected for this analysis given its wide application. The objective is to minimize the operating cost of existing power generators while meeting the demand. The analyzed UC problem is modeled as a MILP (Marcovecchio, Novais, and Grossmann 2014). The analysis was conducted on 10 thermal units. Ontario-Canada's (Hourly Ontario and Market Demands, 2002-2014 n.d.) power demand of the first 30 days of 2014 was used to illustrate the capabilities of the proposed algorithms. The model size effect is tested by doubling and tripling the initial number of thermal units. The full-scale model has a time horizon of 8760 hours while the clustered ones feature 96, 120, 144, and 168 hours for 4, 5, 6, and 7 clusters; respectively.

Table 1. Summary of results for normal clustering with different number of units.

| N° of units | Statistic | Optimal | Number of Clusters - Normal | | | |
|---|---|---|---|---|---|---|
| | | | 4 | 5 | 6 | 7 |
| 10 | CPU time (s) | 2228 | 3 | 5 | 5 | 6 |
| | Objective function ($) | 1.37 E8 | 1.37 E8 | 1.37 E8 | 1.37 E8 | 1.37 E8 |
| 20 | CPU time (s) | 33580 | 9 | 14 | 12 | 22 |
| | Objective function ($) | 2.73 E8 | 2.72 E8 | 2.73 E8 | 2.73 E8 | 2.73 E8 |
| 30 | CPU time (s) | 99280 | 63 | 28 | 129 | 37 |
| | Objective function ($) | 4.09 E8 | 4.08 E8 | 4.08 E8 | 4.08 E8 | 4.08 E8 |

Table 2. Summary of results for sequence clustering with different number of units.

| N° of units | Statistic | Optimal | Number of Clusters - Sequence | | | |
|---|---|---|---|---|---|---|
| | | | 4 | 5 | 6 | 7 |
| 10 | CPU time (s) | 2228 | 3 | 5 | 6 | 9 |
| | Objective function ($) | 1.37 E8 | 1.37 E8 | 1.37 E8 | 1.38 E8 | 1.38 E8 |
| 20 | CPU time (s) | 33580 | 11 | 13 | 15 | 26 |
| | Objective function ($) | 2.73 E8 | 2.73 E8 | 2.74 E8 | 2.74 E8 | 2.74 E8 |
| 30 | CPU time (s) | 99280 | 177 | 1830 | 1720 | 979 |
| | Objective function ($) | 4.09 E8 | 4.09 E8 | 4.10 E8 | 4.11 E8 | 4.11 E8 |

Tables 1 and 2 present the results of normal and sequence clustering for different number of units. The application of both normal and sequence clustering shows a great advantage in terms of solution time compared to the full-scale model. The solution times of normal and sequence clustering for 10 and 20 units are very similar. However, it takes much less time to solve the normal clustering model compared to the sequence one for the 30 units' case. Figures 2 and 3 illustrate the values of the objective function in error percentage compared with the optimal non-clustered solution for the 10 and 30 units of the normal and sequence clustering, respectively.

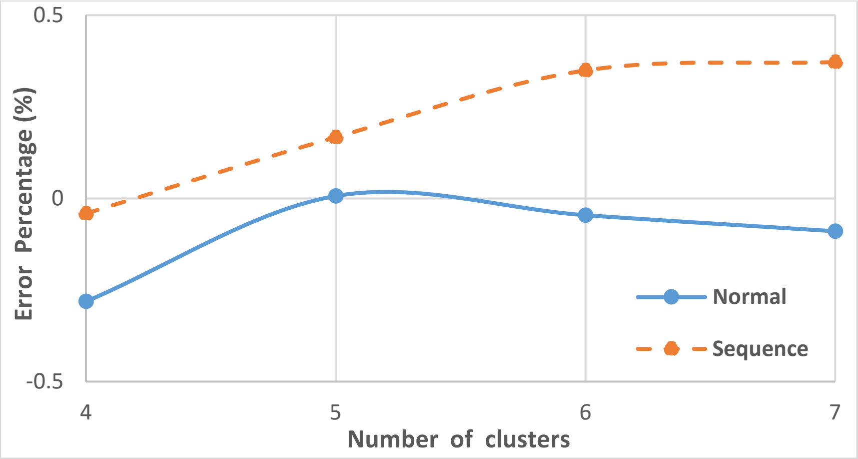

Figure 2. 10 units' objective function values in error percentage for normal and sequence clustering.

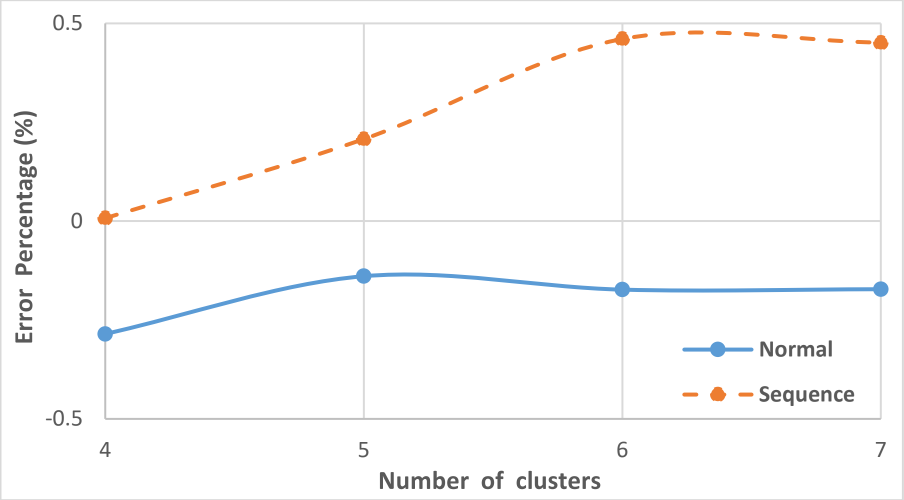

Figure 3. 30 units' objective function values in error percentage for normal and sequence clustering.

The error range is within ± 0.5 % for all cases. The four sequence clusters case is the closest to the optimal showing that four clusters is the optimal representation of the electricity demand curve. This trend could be expected since typically electricity demand behaves seasonally, which is often clustered into four well-known seasons. This further validates the proposed clustering algorithm.

### 4.2 Case Study 2: Multiple Attributes Problem

This case study evaluates the clustering algorithms (normal and sequence) compared with a full energy hub model with multiple demand attributes. The objective is the minimization of the energy hub's operating cost while meeting the power and heat demands. The energy hub problem is formulated as a linear programming (LP) model (Maroufmashat et al. 2015). The energy hub includes: one boiler, one combined heat and power (CHP) unit, and option to purchase power from the grid. The power demand is met by the CHP and grid; whereas heat by the boiler and CHP.

The full scale model includes hourly heat and power demands loads for 365 days; whereas the clustered cases hourly loads considered 4, 5, and 6 clusters (clusters are considered as days). Given that the energy hub is a LP, it only takes a few seconds to solve the full scale model. Nevertheless, computational time reduction using clustering has been proven in the previous case study. In this case study the focus is the quality of the solution.

Figure 4 shows the values of the objective function alongside with the relative error compared with the optimal case. All clustered cases underestimate the value of the objective function. Normal clustering is closer to the optimal compared with sequence. The objective function's error average is -1.7 %, and 4.2% for normal and sequence clustering, respectively. It was found that increasing the clusters number improves solution quality in both types of clustering as it closes the gap between the clustered cases' and optimal solution. Also, changing the weight factors does not cause a strong effect on the objective function values. This might be the result of a similar symmetry between the heat and electricity demands.

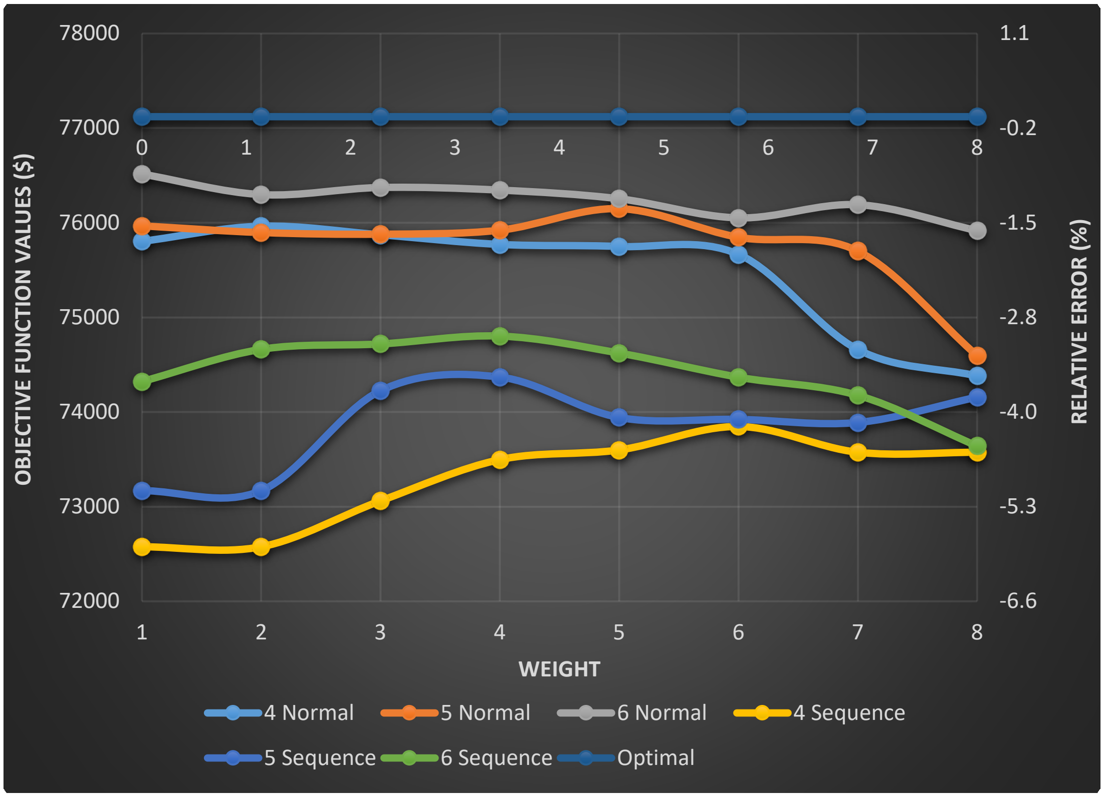

Figure 4. Energy hub's objective function values combining all runs and weight factors.

## 5. Conclusions

This study targets the integrated supply chain problem employing a clustering methodology. Since the use of shorter time periods lead to large and intractable models, this study main objective is to decrease the model size by denoting the days of the year by typical days during the operating year while maintaining accuracy in the results. Accordingly, a mathematical programming methodology was used to model the clustering problem with single or multiple attributes. There are clustering advantages in terms of solution time compared with the full scale model, while increasing size had a minor effect on solution accuracy. It was found that normal clustering yields better objective function, average, and standard deviation error values compared with sequence. For the one attribute case study, the error range is within ± 0.5 % for all studied cases. The error grows as the number of cluster increases suggesting there is an optimal number of normal or sequence clusters regardless of cluster quality. The computational burden associated with solving the MILP model even with the $L_1$-metric still denotes a drawback for large planning horizons. Nonetheless, the application of a heuristic algorithm helps reaching optimal solutions in shorter times.

For the multiple attribute case study, it was found that all clustered cases underestimate the values of the objective function. Normal clustering is closer to the optimal case compared with sequence. The objective function error average is -1.7 % and -4.2 % for normal and sequence clustering, respectively. Additionally, varying the weight factors does not have a major effect on the value of the objective function. This could be the result of a similar symmetry in heat and electricity demands.

## References

Balachandra, P, and Vijay Chandru. 1999. “Modelling Electricity Demand with Representative Load Curves.” *Energy* 24(3): 219–30.

Balachandra, Patil, and Vijay Chandru. 2003. “Supply Demand Matching in Resource Constrained Electricity Systems.” *Energy Conversion and Management* 44(3): 411–37.

Bektaş, S, and Y Şişman. 2010. “The Comparison of L11 and L22-Norm Minimization Methods.” *International Journal of Physical* 5(11): 1721–27.

Chelmis, Charalampos, Jahanvi Kolte, and Viktor K. Prasanna. 2015. “Big Data Analytics for Demand Response: Clustering over Space and Time.” In *2015 IEEE International Conference on Big Data (Big Data)*, IEEE, 2223–32.

Fazlollahi, Samira, Gwenaelle Becker, and François Maréchal. 2014. “Multi-Objectives, Multi-Period Optimization of District Energy Systems: III. Distribution Networks.” *Computers and Chemical Engineering* 66: 82–97.

Green, Richard, Iain Staffell, and Nicholas Vasilakos. 2014. “Divide and Conquer&#x003F; <formula Formulatype="inline"><tex Notation="TeX">${k}$</Tex></Formula>-Means Clustering of Demand Data Allows Rapid and Accurate Simulations of the British Electricity System.” *IEEE Transactions on Engineering Management* 61(2): 251–60.

Grossmann, Ignacio E. 2012. “Advances in Mathematical Programming Models for Enterprise-Wide Optimization.” *Computers and Chemical Engineering* 47: 2–18.

“Hourly Ontario and Market Demands, 2002-2014.”

Jain, Anil K. 2010. “Data Clustering: 50 Years beyond K-Means.” *Pattern Recognition Letters* 31(8): 651–66.

Lyu, Qin, Zhouchen Lin, Yiyuan She, and Chao Zhang. 2013. “A Comparison of Typical ?P Minimization Algorithms.” *Neurocomputing* 119: 413–24.

Mangasarian, Olvi L. 2013. “Absolute Value Equation Solution via Dual Complementarity.” *Optimization Letters* 7(4): 625–30.

Marcovecchio, Marian G., Augusto Q. Novais, and Ignacio E. Grossmann. 2014. “Deterministic Optimization of the Thermal Unit Commitment Problem: A Branch and Cut Search.” *Computers & Chemical Engineering* 67: 53–68.

Maroufmashat, Azadeh et al. 2015. “Modeling and Optimization of a Network of Energy Hubs to Improve Economic and Emission Considerations.” *Energy* 93: 2546–58.

Padhy, N.P. 2004. “Unit Commitment—A Bibliographical Survey.” *IEEE Transactions on Power Systems* 19(2): 1196–1205.

Papageorgiou, Lazaros G. 2009. “Supply Chain Optimisation for the Process Industries: Advances and Opportunities.” *Computers and Chemical Engineering* 33(12): 1931–38.

Sabo, Kristian. 2014. “Center-Based L1?Clustering Method.” *International Journal of Applied Mathematics and Computer Science* 24(1): 151–63.

Vinod, Hrishikesh D. 1969. “Integer Programming and the Theory of Grouping.” *Journal of the American Statistical Association* 64(326): 506–19.

Xu, Rui, and Donald C. Wunsch. 2008. 1 *Clustering*. Hoboken, NJ, USA: John Wiley & Sons, Inc.

## Biographies

**Falah Alhameli** is currently a research and development engineer at the Abu Dhabi National Oil Company (ADNOC). He earned a BS and an MSc in Chemical Engineering from the Petroleum Institute (now part of Khalifa University of Science & Technology) and a PhD from the University of Waterloo. He has published journal and conference papers. Dr. Alhameli has completed research projects related to gas processing, planning of power production, and renewable energy integration in the gas and oil industry. His current research interests focus on big data analytics and integration in multiscale decision making in oil and gas operations.

**Ali Elkamel** is a Professor of Chemical Engineering. He holds a BS in Chemical Engineering and BS in Mathematics from Colorado School of Mines, MSc in Chemical Engineering from the University of Colorado-Boulder, and PhD in Chemical Engineering from Purdue University – West Lafayette, Indiana. His specific research interests are in computer-aided modelling, optimization and simulation with applications to energy production planning, carbon management, sustainable operations and product design. Professor Elkamel is currently focusing on research projects related to energy systems, integration of renewable energy in process operations and energy production systems, and the utilization of data analytics (Digitalization), machine learning, and Artificial Intelligence (AI) to improve process and enterprise-wide efficiency and profitability. Prof. Elkamel has supervised over 90 graduate students and more than 30 post-doctoral fellows/research associates. Among his accomplishments are the Research Excellence Award, the Excellence in Graduate Supervision Award, the Outstanding Faculty Award, the Best teacher award, and the IEOM (Industrial Engineering and Operations Management) Outstanding Service and Distinguished Educator Award. He has more than 280 journal articles, 141 proceedings, and 33 book chapters. He is also a co-author of four books; two recent books were published by Wiley and entitled Planning of Refinery and Petrochemical Operations and Environmentally Conscious Fossil Energy Production.

**Mohammed Alkatheri** holds a BS degree in Chemical Engineering from United Arab Emirates University, and MSc degree in Chemical Engineering from the Petroleum Institute in Abu Dhabi. During his MSc, he developed research on modelling and simulation of kinetics and single particle growth for the heterogeneous polymerization of Ziegler-Natta catalyst. From 2015 – 2017, he worked as a research assistant at the Petroleum Institute where he studied the economics of different ultra-sour natural gas sweetening processes, assessed sweeting of ultra-sour natural gas using hybrid processes and carried out green-house gases life cycle assessment for the United Arab Emirates electricity sector. In May 2017, he joined the PhD program in Chemical Engineering at University of Waterloo. His PhD research is focused on the application and integration of big-data tools (i.e. Artificial Intelligence and Machine Learning) in chemical process optimization and process system engineering. The scope of his PhD project is to address the challenges associated with chemical engineering process design and operation, namely, uncertainty handling, parameter estimation and unit process equation complexity. Therefore, high-level optimization tasks such as planning and scheduling will highly benefit from information mined from massive data, since optimization has always been based on the interchange between models and data.

**Alberto Betancourt-Torcat** is a researcher at the University of Waterloo. He holds a BS in Chemical Engineering from University Simon Bolivar in Venezuela, and MSc in Chemical Engineering from the University of Waterloo. He was a Research Associate at the University of Waterloo from September 2011 to June 2012. From August 2012 to November 2018, he worked at the Petroleum Institute (currently Khalifa University of Science & Technology) in Abu Dhabi, as a Research Engineer and Lecturer in the Department of Chemical Engineering. He has published numerous articles in renowned refereed journals, book chapters, and conference proceedings. He has also delivered several presentations in international conferences and seminars. Aditionally, he serves as a reviewer for various reputable international journals in the area of energy systems and energy policy.

**Ali Almansoori** is a Professor of Chemical Engineering at Khalifa University of Science & Technology in Abu Dhabi. He earned a Ph.D. in Chemical Engineering from the Imperial College in London, an Executive MBA from London Business School, and a BS in Chemical Engineering from Florida Institute of Technology. During his profession, he has held several administrative positions including: the Coordinator of President's Duties, Dean of Engineering, and Chair and Deputy Chair of the Chemical Engineering Department. He also was the Interim Senior Vice President for Academic Affairs during the merge between PI, Masdar Institute, and Khalifa University of Science, Technology, and Research. He has published numerous articles, book chapters, and conference proceedings. Dr. Almansoori was also a research fellow at the Organization of the Petroleum Exporting Countries (OPEC) in Vienna, Austria during the summer of 2012. He was recently awarded the Mohammed Bin Rashid medal for scientific excellence on January 2019.