Proceedings of the International Conference on Industrial Engineering and Operations Management
Toronto, Canada, October 23-25, 2019

Data-driven Power Generation Design and Operation Under Demand Uncertainty

Mohammed Alkatheri[1], Muhammad Rizwan[2], Falah Alhameli[1], Ali Elkamel[1,2], Ali Almansoori[2], and Peter Douglas[1]
[1]Department of Chemical Engineering, University of Waterloo, Waterloo, Ontario, N2L 3G, Canada
[2]Department of Chemical Engineering, Khalifa University of Science and Technology, SAN Campus, P. O. Box 2533, Abu Dhabi, United Arab Emirates
mohammed.alkatheri@uwaterloo.ca, muhammad.rizwan@ku.ac.ae, falhamel@uwaterloo.ca , ali.elkamel@ku.ac.ae, ali.almansoori@ku.ac.ae, peter.douglas@uwaterloo.ca

Abstract

A Data-driven stochastic optimization framework that leverages big data in design and operation of power generation units is proposed. A k-means clustering algorithm is adopted to generate uncertainty scenarios for the stochastic optimization framework. In order to do this, the power generating design and operation problem is formulated as a two-stage stochastic programming model. The first stage variables are associated with design decisions, whereas the second stage variables are associated with unit commitment operation (i.e. scheduling). The historical demand data was first collected and reprocessed. After that, the processed electrical demand (uncertain parameter) is processed and recognized using unsupervised machine learning. K-means clustering algorithm is used to produce electrical demand scenario profiles. These scenarios are used as inputs to the stochastic model. The proposed model is formulated as a mixed integer linear programming (MILP) and solved using GAMS. The stochastic data driven method enjoys the following features: it is based on information derived from real data without explicitly knowing the data distribution and it applies the recent advances of data analysis tools (e.g. machine learning) to generate a reduced size data set (i.e. clusters) integrated into mathematical model (i.e. design and planning model) that leads to a computationally tractable problem.

1. Introduction

Deterministic process design and operation models can help ensure an optimal solution for certain process parameters (e.g. demand, fuel price) where it satisfies the constraints associated with that parameter (i.e. product should be greater than or equal to demand). As most real-life problems involve some sort of uncertainty, deterministic models are incapable of resolving them. In reality, most model parameters are uncertain, such as the availability of renewable energy and power demand are not known with certainty and are hard to predict. There exist a considerable number of studies from industry and academia on optimization under uncertainty (Sahinidis 2004, Grossmann et al. 2016, Ierapetritou et al. 1996). However, these approaches do not take advantage of the recent advances in machine learning and big data analytics to leverage uncertainty data for optimization under uncertainty. Hence, the goal of this paper is to propose a data- driven stochastic decision-making framework that integrates machine learning methods to uncertainty data with the design and operation of power generation problem. Traditional models of decision-making under uncertainty assume perfect information which means either accurate values for the system parameters or specific probability distributions for the random variables. Nevertheless, such exact knowledge is rarely available, prior knowledge on uncertain parameter distribution is unknown and fitting random variables (uncertain parameter) into a popular distribution is complicated and impractical (Bertsimas and Thiele 2006). Furthermore, it is mathematically intractable to deal with erroneous inputs (all sets of uncertain data) and this could lead to infeasible solutions or exhibit poor performance when implemented (Bertsimas and Thiele 2006). Therefore, in this paper we propose a data-driven stochastic approach for power generation design and operation optimization that can efficiently utilize the available historical demand data through advances of data analytics tools such as k-means clustering (a machine learning tool) By doing this, data-driven power planning is achieved against uncertainty realization.

The rest of this paper is structured as follows: Section 2 states the methodology of the proposed method. Section 3 presents the mathematical formulations of the deterministic and the stochastic models of the design and operation of power generation problem. Section 4 covers the construction of data-driven uncertainty scenario. The results of both mathematical formulation problems are discussed in section 5. Conclusions are drawn in section 6.

2. Methodology

In this paper, the unit commitment (UC) model was reformulated in such a way that its design and operational decisions can be determined. The main objective of this paper is to formulate and solve the mathematical problem of the design and operation for power generation units. The model is expected to incorporate design and operational decisions based on uncertain electricity demand. The UC problem can be defined as finding the optimal scheduling of electric power generating units over a short-term period (i.e. typically from 24hours to one week, in order to minimize the operations costs. The unit commitment optimal solution must obey the technical constraints and must satisfy the demand. The design and operation of the power generation model can be divided into two phases, namely ***deterministic*** and ***stochastic with recourse*** formulation. In the deterministic model the hourly expected values (i.e. means) over one year for one day of the electricity demand were used as inputs. The deterministic model was solved for a one-day time horizon and the electricity demand was assumed to be certain. We have only one-day profiles for demand that represent the entire block of previous year/ years. In the stochastic approach, the problem was formulated as a two-stage stochastic programming model. The first stage variables are associated with power generating units design decisions, whereas the second stage variables are associated with unit commitment operation (i.e. scheduling). The uncertain parameters (i.e. electricity demand) were processed and recognized using unsupervised machine learning. Different scenarios were generated for this uncertain parameter. A clustering algorithm was used to produce uncertain parameters profiles (i.e. each scenario corresponds to the profile of electricity demand for 1 day, in other words each scenario is a vector with 24 dimensions and each cluster is associated with a certain occurrence/probability). Different scenarios were used as inputs to the stochastic model. The time horizon for the stochastic model was also one day, however, there were many scenarios for this day at each hour (e.g. the electrical demand at 2 hours of the day is different depending on the scenario/profile). Clustering strategies have proven their ability to aggregate cyclic data based on the concept of cyclic scheduling (e.g. electricity demand follows daily cycle). It was used extensively in process systems engineering (Pochet and Warichet 2008, Wu and Ierapetritou 2004), cyclic scheduling which requires certain demands to be processed over certain time periods repeatedly within the time horizon. However, these aggregated cyclic results, to our knowledge have only been incorporated in deterministic modelling which works only under the presence of certain information. Clustering techniques were not widely used to reduce the size of the uncertain data. Therefore, applying clustering algorithms to extract patterns from uncertain data and use its output to be fed into a stochastic optimization formulation is an interesting research area and more exploration on this approach can be performed.
This stochastic model has two level decisions (i.e. operational and design) whose objective is to minimize the capital and operating cost. The capital costs correspond to the number of generator units that need installation, while the operating costs are associated with the amount of power generated by these units while meeting electricity demands. There are several mathematical models for the unit commitment problem available in the literature (Tumuluru et al. 2014). In this study, we adopted Marcovecchio et al. (2014) the UC model formulation as the basis for our formulations of power generating model. The model is formulated as a mixed integer linear programming (MILP) (Alhameli 2017). The following sections present the model formulations and related consideration.

3. Mathematical Formulation

3.1 Deterministic Mathematical Formulation of Power Generation Design and Operation

Consider a set of I thermal to be scheduled over a time horizon T. The goal is to minimize the overall cost (i.e. capital and operating). This goal can be achieved by optimally determining the number of power generation units (both thermal) and scheduling the thermal generating units. The mathematical programming framework ensure that these optimal solutions are meeting the electricity demands and operating within the units' capacities (i.e. technical constrains). The problem is solved for 10 thermal units and 24 hours.
Objective Function: The objective function (see equation (1)), represents the net present cost, including capital cost of the power units and their operating cost. In this study, the operating cost covers fuel consumption calculated by a linear function with fixed charges, and fixed start-up and shut-down costs (Alhameli 2017). Net present cost is the

sum of the discounted values of all the cost cash flow at the present. Assume the annual discount rate for the calculation is r and the system life span is L years. As fuel consumption considered to be the costliest component of operating expenditure in power generation, then the expected net cost value of the project over the system life-span can be minimized as:

$$\min \sum_i x_i C g_i + \sum_{L=1}^{N_{life}} \frac{N_d}{(1+r)^L} \sum_{i,t} (a_i U_{i,t} + b_i P_{i,t} + CU_{i,t} + CD_{i,t}) \quad (1)$$

Where:

x_i the binary design decision for power unit i (0 means no and 1 yes);

$U_{i,t}$ the binary operational/scheduling decision variable representing the on/off status of unit i at period t;

$CU_{i,t}$ start-up cost variable of unit i in period t;

$P_{i,t}$ power output variable of unit i in period t;

a_i, b_i coefficients of the fuel cost function of unit, their values are shown in Table 5

Cg_i the capital cost of i generating unit (3,246 $/kW, the capital cost for coal power plant)

$\sum_{y=1}^{N_{life}} \frac{N_d}{(1+r)^y}$ coefficient to convert the daily operating cost into net value, where N_d, denotes number of days per year (365 days/year), N_{life} represents the life time (i.e. system life span and it was assumed to be 25 years) of the generating units (10 Units) and r denotes the discount rate (12%)

Minimum and maximum power generation: Equation (2) ensures that the power produced by unit i at time t is within the generation power limits of that unit. (i.e. upper limit P_i^U and lower limit P_i^L). The values of the upper and lower power generating limits are shown in Table 5. These constraints fix units availability at zero when units are 'off' ($U_{i,t} = 0$) and specify the lower and upper bounds of units capacity when units are active ($U_{i,t} = 1$).

$$U_{i,t} P_i^L \le P_{i,t} \le U_{i,t} P_i^U \qquad t = 1, \dots, T;\ , i = 1, \dots, I \quad (2)$$

Electricity demand and reserve: Electricity demand should be satisfied at any t time by equation (3). In this deterministic case, the average value of Ontario demand for 2018 was used. It was assumed that we want to satisfy a portion of Ontario's demand (i.e.~ 7% of total Ontario demand in 2018 (Hourly Ontario and Market Demands 2018)). More details on demand data collection and mean calculation are shown in the section 4.

$$\sum_i P_{i,t} \ge D_t \qquad t = 1, \dots, T \quad (3)$$

Equation (4) guarantees spinning reserve by the available capacity of the active units, where, R_t represents the reserve requirements. Spinning reserve (i.e. spinning means active units that already connected to the grid) means that from the pool of available capacity, a portion is selected for a back-up role. It is assumed that the spinning reserve requirement to be met is set at 10% of the load demand for each time period.

$$\sum_i P_i^U U_{i,t} \ge D_t + R_t \qquad t = 1, \dots, T \quad (4)$$

Minimum up and down time of thermal generating units: Once a decision has been made to turn a thermal power plant on or off, it must remain in that state for a minimum amount of time. Equation (5) and equation (6) determine the online/offline status of unit i in its earliest periods of operation which are specified by its initial status (T_i^{ini}) and its minimum up (TU_i) and down (TD_i) times. T_i^{ini} denotes the number of periods that unit i has been initially offline ($T_i^{ini} < 0$) or online ($T_i^{ini} > 0$) if .The following constrains ensure that when the simulation is started if unit i is offline for T_i^{ini}, it will continue to be offline until it satisfies its minimum down requirement (TD_i) and vice versa for the online unit.

$$U_{i,t} = 1 \qquad \forall\, i : T_i^{ini} > 0;\ \ t = 1, .., (TU_i - T_i^{ini}) \quad (5)$$

$$U_{i,t} = 0 \qquad \forall\, i : T_i^{ini} < 0;\ \ t = 1, .., (TD_i + T_i^{ini}) \quad (6)$$

Equation (7) and equation (8) are expressing the constraints on minimum uptime and downtime unit as follows:

$$U_{i,t} - U_{i,t-1} \le U_{i,t+j} \qquad i = 1, \dots, I;\ t = 2, \dots, T\ ;\ j = 1, \dots, (TU_i - 1) \quad (7)$$

$$U_{i,t+j} \le U_{i,t} - U_{i,t-1} \qquad i = 1, \dots, I,;\ t = 2, \dots, T;\ \ j = 1, \dots, (TD_i - 1) \quad (8)$$

In the first time period equation (7) and equation (8) reduce to equation (9) and equation (10) respectively.

$$U_{i,1} \le U_{i,1+j} \qquad T_i^{ini} < 0;\ \ i = 1, \dots, I;\ \ j = 1, \dots, (TU_i - 1) \quad (9)$$

$$U_{i,1+j} \le U_{i,1} \qquad T_i^{ini} > 0;\ \ i = 1, \dots, I;\ \ j = 1, \dots, (TD_i - 1) \quad (10)$$

Unit ramp rates: Thermal generating units are limited with respect to how quickly they can change their power output and also this limit is known as a unit's ramp rate (RU_i ramp-up rate, RD_iramp-down rate, SD_i shutdown ramp rate and SU_i is start-up ramp rate per unit of time period) . The ramp-up and ramp-down rates of each unit are set to be at 20% of the unit maximum power output per time period. Whereas the start-up and shutdown ramp rates of each

unit are chosen to be at its maximum generation output (Simopoulos et al. 2006). The ramp rate limits are modelled by equation (11) and equation (12)

$$P_{i,t} - P_{i,t-1} \leq RU_i U_{i,t-1} + SU_i(1 - U_{i,t-1}) \qquad i = 1, \dots, I; t = 2, \dots, T \qquad (11)$$

$$P_{i,t-1} - P_{i,t} \leq RD_i U_{i,t-1} - DU_i(1 - U_{i,t}) \qquad i = 1, \dots, I; t = 2, \dots, T \qquad (12)$$

Start-up and shut-down unit costs: The costs involved in turning on and off generating units are essential and considered to be an important element of the operation cost of power thermal unit. In this study, it is assumed that there are two fixed start-up costs per unit (hot start and cold start), depend on the time periods that the unit was off (Marcovecchio et al. 2014). The start-up cost function is defined as a hot start cost ($CU_{i,t} = Hsc_i$) if downtime ≤ ($TD_i + T_i^{COLD}$) and a cold start cost ($CU_{i,t} = Csc_i$) otherwise. Where Hsc_i, Csc_i and T_i^{COLD} are parameters that represent the hot start cost of unit i, the cold start cost of unit i, and the cold start hour of unit i, respectively. The values of these parameters are reported in Table 5. This start-up cost function can be modelled by equations (13 -16)

$$(U_{i,t} - U_{i,t-1})Hsc_i \leq CU_{i,t} \qquad i = 1, \dots, I;\ t = 2, \dots, T \qquad (13)$$

$$U_{i,1} Hsc_i \leq CU_{i,t} \qquad i = 1, \dots, I; T_i^{ini} < 0 \qquad (14)$$

$$\left(U_{i,t} - \sum_{j<TD_i+T_i^{COLD}} U_{i,t-j}\right) Csc_i \leq CU_{i,t} \qquad i = 1, \dots, I;\ t > TD_i + T_i^{COLD} \qquad (15)$$

$$(U_{i,t} - \sum_{j<t} U_{i,t-j}) Csc_i \leq CU_{i,t} \qquad i = 1, \dots, I;\ T_i^{ini} < 0\ ; TD_i + T_i^{COLD} \leq t < TD_i + T_i^{COLD} + 1 \qquad (16)$$

Equation (17) ensures that variable $\boldsymbol{CU_{i,t}}$, takes value 0 when the unit is not turned on at the time period

$$0 \leq CU_{i,t} \qquad i = 1, \dots, I; t = 1, \dots, T \qquad (17)$$

We assumed that there is no shut-down cost when units are turned off (Alhameli 2017 and Marcovecchio et al. 2014).

Design constraints: The following constraints (equations (18 & 19)) ensure that the thermal unit will not be develop if it is not used (not needed). x_i denotes a binary decision variable to determine whether unit i should be installed or not.

$$U_{i,t} \leq x_i \qquad i = 1, \dots, I;\ t = 2, \dots, T \qquad (18)$$

$$x_i \leq \sum_t U_{i,t} \qquad i = 1, \dots, I \qquad (19)$$

Variable specification: Finally, the specification on the variables is as follows:

$$U_{i,t} \in [0,1] \qquad i = 1, \dots, I;\ t = 1, \dots, T \qquad (20)$$

$$x_i \in [0,1] \qquad i = 1, \dots, I \qquad (21)$$

After that, these equations are solved, and results obtained are shown in section 6.

3.2 Stochastic Mathematical Formulation of Power Generation Design and Operation

This section discusses the model for the stochastic problem for the design and operation of a power generation under uncertain demand. The mathematical model was formulated as a two-stage stochastic with recourse, where the first-stage decisions decide the existence of the thermal generating unit while the second-stage decisions plan the operation of the system (i.e. power scheduling). One main difference of stochastic model compared to the deterministic one, is that the optimal power scheduling can be different for each different realization of the uncertainty (in each cluster of demand in the system). The two-stage stochastic recourse formulation is basically a bi-level optimization formulation whose inner optimization problems mimic the second-stage planning process. Due to its special structure, two-stage stochastic programs can be naturally reformulated into an equivalent single-level optimization problem. Therefore, the single-level optimization formulation of two-stage recourse of power generation design and operation can be directly written as follows:

Objective Function: The objective function in equation (22) represents the net present cost of the stochastic power generation design and operation. The second part of the equation denotes the annual net cost from operating the power unit (i.e. basically fuel consumption because of power generation), which depends on the scenario of uncertainty realization s with probability $Prob_s$:

$$\min \sum_i x_i Cg_i + \sum_{L=1}^{N_{life}} \frac{N_d}{(1+r)^L} \sum_{i,t,s} Prob_s(a_i U_{i,t,s} + b_i P_{i,s} + CU_{i,t,s}) \qquad (22)$$

The remaining equations are almost the same as the deterministic one, except the new subscript s. where the new subscript s [$\{1, \dots S\}$] is used in the stochastic model for all the variables and parameters whose values may be different in the S different uncertainty scenarios (the uppercase s denotes the total number of scenarios).

Minimum and maximum Power generation

$$U_{i,t,s} P_i^L \leq P_{i,t} \leq U_{i,t,s} P_i^U \qquad t = 1, \dots, T;\ , i = 1, \dots, I; s = 1, \dots, S \qquad (23)$$

Electricity demand and reserve

$$\sum_i P_{i,t,s} \geq D_{t,s} \qquad t = 1, \dots, T;\ s = 1, \dots, S \tag{24}$$

$$\sum_i P_i^U U_{i,t} \geq D_{t,s} + R_{t,s} \qquad t = 1, \dots, T;\ s = 1, \dots, S \tag{25}$$

Minimum up and down time of thermal generating units

$$U_{i,t,s} = 1 \qquad \forall\, i : T_i^{ini} > 0;\ t = 1, .., \left(TU_i - T_i^{ini}\right);\ s = 1, \dots, S \tag{26}$$

$$U_{i,t,s} = 0 \qquad \forall\, i : T_i^{ini} < 0;\ t = 1, .., \left(TD_i + T_i^{ini}\right);\ s = 1, \dots, S \tag{27}$$

$$U_{i,t,s} - U_{i,t-1,s} \leq U_{i,t+j,s} \qquad i = 1, \dots, I;\ t = 2, \dots, T\ ;\ j = 1, \dots, (TU_i - 1);\ s = 1, \dots, S \tag{28}$$

$$U_{i,t+j,s} \leq U_{i,t,s} - U_{i,t-1,s} \qquad i = 1, \dots, I, ;\ t = 2, \dots, T;\ \ j = 1, \dots, (TD_i - 1);\ s = 1, \dots, S \tag{29}$$

$$U_{i,1,s} \leq U_{i,1+j,s} \qquad T_i^{ini} < 0;\ \ i = 1, \dots, I;\ \ j = 1, \dots, (TU_i - 1);\ s = 1, \dots, S \tag{30}$$

$$U_{i,1+j,s} \leq U_{i,1,s} \qquad T_i^{ini} > 0;\ \ i = 1, \dots, I;\ \ \ j = 1, \dots, (TD_i - 1);\ s = 1, \dots, S \tag{31}$$

Unit ramp rates

$$P_{i,t,s} - P_{i,t-1,s} \leq RU_i U_{i,t-1,s} + SU_i\left(1 - U_{i,t-1,s}\right) \qquad i = 1, \dots, I; t = 2, \dots, T;\ s = 1, \dots, S \tag{32}$$

$$P_{i,t-1,s} - P_{i,t,s} \leq RD_i U_{i,t-1,s} - DU_i\left(1 - U_{i,t,s}\right) \qquad i = 1, \dots, I; t = 2, \dots, T;\ s = 1, \dots, S \tag{33}$$

Start-up and shut-down unit costs

$$\left(U_{i,t,s} - U_{i,t-1,s}\right) Hsc_i \leq CU_{i,t,s} \qquad i = 1, \dots, I;\ t = 2, \dots, T;\ s = 1, \dots, S \tag{34}$$

$$U_{i,1,s} Hsc_i \leq CU_{i,t,s} \qquad i = 1, \dots, I; T_i^{ini} < 0; s = 1, \dots, S \tag{35}$$

$$\left(U_{i,t,s} - \sum_{j < TD_i + T_i^{COLD}} U_{i,t-j,s}\right) Csc_i \leq CU_{i,t,s} \qquad i = 1, \dots, I;\ \ t > TD_i + T_i^{COLD};\ s = 1, \dots, S \tag{36}$$

$$\left(U_{i,t,s} - \sum_{j<t} U_{i,t-j,s}\right) Csc_i \leq CU_{i,t,s} \quad \forall\, i;\ T_i^{ini} < 0\ ; TD_i + T_i^{COLD} \leq t < TD_i + T_i^{COLD} + 1;\ s = 1, \dots, S \tag{37}$$

$$0 \leq CU_{i,t,s} \qquad i = 1, \dots, I; t = 1, \dots, T;\ s = 1, \dots, S \tag{38}$$

Design Constraints

$$U_{i,t,s} \leq x_i \qquad i = 1, \dots, I;\ t = 2, \dots, T\ ;\ s = 1, \dots, S \tag{39}$$

$$x_i \leq \sum_t U_{i,t,s} \qquad i = 1, \dots, I\ ;\ s = 1, \dots, S \tag{40}$$

4. Data-Driven Uncertainty Scenario Construction Using Clustering

In this section, we propose a method on how to generate scenarios from demand data using a clustering algorithm. These scenarios with its corresponding occurrence can be used as input parameters for the stochastic model. This method begins by collecting historical data of the attribute or uncertain parameter under study (in this study it was electrical demand). Following which, the raw input data must be pre-processed into the right format. The raw data time-series (i.e. electricity demand) are first processed (arranged) into the candidate periods (considered to be 365 days for 1 year, with each day consisting of 24 hours). This reordering process is shown in the following matrix (see Figure 1) in which the number of columns is defined by the multiple of the number of time steps (i.e. 24 hour) and number of rows corresponding to the number of periods (i.e. 365 days). A single row represents a candidate period (i.e. one day). Raw data of electricity demand are reshaped into new matrix where the number of rows represent the number of days in one year (i.e. 365 days) and the number of columns represent the number of hours in one day (i.e. 24 hours).

$$parameter_{8764} = \begin{pmatrix} parameter_1 \\ parameter_2 \\ \vdots \\ parameter_{8764} \end{pmatrix} \xrightarrow{rearrange} \begin{pmatrix} parameter_{1,1} & \cdots & parameter_{1,24} \\ \vdots & \ddots & \vdots \\ parameter_{366,1} & \cdots & parameter_{366,24} \end{pmatrix}$$

Figure 1. Process of rearranging the dimension electricity demand

As it can be noticed in the demand data in Figure 2, there are some daily oscillations. The benefit of clustering the demand data is that, instead of using the whole set of data only the centroids (i.e. centres of each cluster) of each uncertain parameter (in this study we have one uncertain parameter; electrical demand) will be used. These centroids can represent the whole set of data. Based on the matrix introduced in Figure1.,k-mean (Pedregosa et al. 2011) clustering algorithm was applied to group the independent candidate periods (i.e. each row/day of the processed data matrix) into cluster. Accordingly, representative periods are derived. Each cluster/group of each uncertain parameter (i.e. demand) is represented by a representative profile (i.e. curve). These representative profiles are the centres of each cluster. Each uncertain parameter (e.g. demand, wind speed, solar intensity, fuel supply) will be represented by several clusters with each cluster corresponding to one scenario with a certain probability of occurrence. The probability of occurrence corresponds to the weight of each cluster. Figure 4 shows the process of scenario

construction for stochastic optimization using clustering machine learning algorithm. In order to determine the most applicable cluster number needed to divide electricity demand data. The k-mean clustering algorithm was applied to the processed data (i.e. electricity demand) using different number of clusters. The clustering algorithm was applied to the processed electricity demand using a different number of clusters. Figures 5 and 6 show the error average and standard deviation as a function of the cluster number for electricity demand. The following relative error function (see equation (41)) was used as a validation measure between the representative cluster profile (centre) and the actual processed data (i.e. candidate period or row/day) which correspond to that cluster.

$$error_{c,d,h}(\%) = \frac{Cl_{c,h} - paramater_{d,h}}{paramater_{d,h}} * 100 \quad (41)$$

An ideal cluster would be compact and isolated (Alhameli 2017). In other words, an ideal cluster would have a minimum error average with a minimum standard deviation. The error and standard deviation average in clustering appears to drop as the number of clusters increases until they reach a certain value and after that it starts to oscillate as it can be seen in Figure 5 and 6. This reveals that there is some sort of optimal number of representative curves (i.e. cluster) for the electricity demand. Therefore, the number of clusters (i.e. cluster centre and its corresponding weight/ probability) with the minimum error average were selected as scenarios that represent the uncertain parameter of the stochastic power generation model. Figure 7 shows actual data and cluster centres used as representatives of the data for electricity demand. As it can be noticed in Figure 7, clustered results are in good agreement with demand data.

It can be seen from Figures 5 and 6 that the minimum error happens when the number of clusters was 9 for electricity and Figure 3 shows the clustering result that will be used for stochastic programming. By comparing the demand clustering profiles with its actual daily profile, we can say that the clustering results are following the actual demand profile.

5. Results and discussions

The deterministic power generation model (Equations (1-21)) was implemented in GAMS. The average electrical demand profile was used to solve the deterministic problem (the representative average day of one year is calculated by averaging each column (i.e. each time step; hour) of the processed matrix for the whole time period (365 days)). The model was solved using the MILP (Mixed Integer-Linear Programming) solver CPLEX which is based on the branch and cut algorithm (Elf et al. 2001). The MILP problem contains 250 discrete variables and 483 continuous variables. The number of constraints are 2923. The GAMS program executes successfully in 0.2 seconds on an Intel Core i7 commodity personal computer.

It was discussed in the previous section that 9 clusters (i.e. scenarios) were chosen to represent the uncertain demand (see Figure 4). The stochastic model (Equation (22-40)) was also implemented in GAMS. The model was solved using the MILP (Mixed Integer-Linear Programming) solver CPLEX. The stochastic problem contains 2170 binary variables and 4323 continuous variables. The GAMS program executed successfully in 145.687 seconds.

Table 1 shows the design results and the objective function values of three different cases, namely, deterministic, stochastic and the worst-case scenario. The worst-case scenario was generated by tackling the maximum representative demand day for the whole year (extreme demand). It was calculated by taking the maximum demand value of each column (i.e. each time step; hour) of the processed matrix (see Figure 1) for the whole period (365 days). As it can be noticed in Table 1 that the deterministic approach gives the lowest expensive solution whereas, the worst-case is the most expensive. However, the deterministic approach has the lowest reliability because it was designed for certain parameters (e.g. demand) which made it incapable of resolving most real-case scenario problems. The stochastic approach is more expensive than the deterministic because there is a price for the uncertainty. The stochastic approach is designed for the most likely scenarios while the worst-case is designed for the extreme demand which rarely occurs. The development of the system under the worst-case scenario will cause the system to be over-designed and therefore, the full capacity of the power generation will not be of use or rarely used. Table 2 shows the difference if we are using the stochastic design solution with external electricity supply when the demand is extreme in the worst-case scenario. We assumed that the extra electricity required by the extreme demand will be supplied by an external power provider with very expensive price (the levelized cost was assumed to be 300 $/MWh, which is double the levelized electricity cost reported by EIA for coal in 2018). If we assume that 20% of the year will be subjected to extreme demand, the extra cost that will be added to the stochastic solution is 0.293 billion$ as it can be seen in Table 2. Therefore, the total cost of the stochastic solution with external electricity needed for extreme demand is less than if we design the power generation plants for worst-case.

We can say from this analysis that the design and operation under stochastic approach is more practical than designing under extreme case (i.e. worst-case scenario) that rarely happens.
Table 3 shows comparison between the size of the problem if use the whole year demand data and the stochastic data-driven approach. It can be said that, machine learning method (clustering) was applied to the data set to get a reduced size data set that represent the whole data set. Moreover, the stochastic data-driven approach does not require to fit the data into known distribution as the case for the regular stochastic optimization where the uncertain parameter should be fit into know distribution and only samples from the distribution are used in the stochastic optimization. Furthermore, sometimes fitting uncertain parameter into a common distribution is complicated and impractical.

Table 3 comparison between the size of the problem if use the whole year demand data and the stochastic data-driven approach

	whole year demand data	stochastic data-driven
Binary Variables	87610	250
Continuous Variables	175203	4323

6. Conclusion

A deterministic and stochastic data-driven power system design and operation models were developed. The deterministic approach was based on a single population parameter (i.e. mean) from data which does not perfectly cover the data behavior and consequently its solution is unreliable. On the other hand, the stochastic data-driven approach was based on more detailed information from the data without explicitly knowing its distribution and therefore its solution was more reliable. In the stochastic data-driven approach, instead using the whole set of available data for the design and planning model which will be computationally expensive, machine learning method (clustering) was applied to the data set to get a reduced order data set (i.e. clusters) that was used in the mathematical model (i.e. design and planning model). We can say that the proposed data-driven stochastic method is a trade-off between computational effort and data accuracy.
In the future, this model can be further expanded to include the integration of renewable energies such as solar and wind. The interconnected power units that powered with different type of fuel will be studied and the optimal scheduling of these units will be determined. Moreover, a carbon capture unit can be added to the power generating unit and the system behavior can be demonstrated under real data behavior. The carbon capturing unit could be modeled using a surrogate data-driven model, the developed model will be added to the stochastic data-driven power generating formulation.

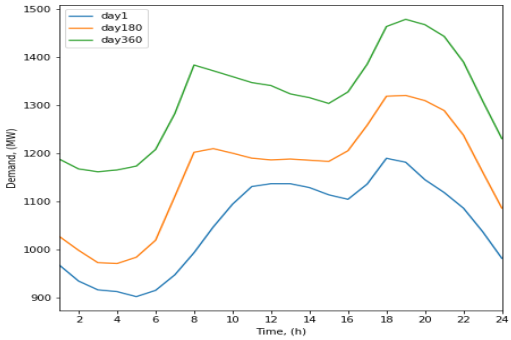

Figure 2. Demand profile for selected days

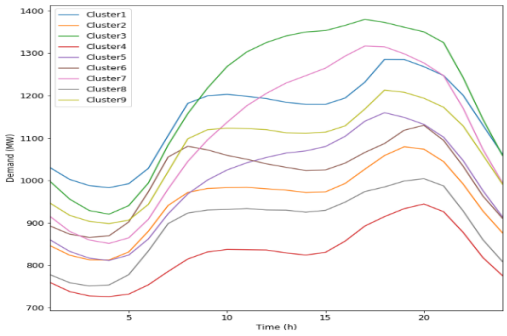

Figure 3. Electricity demand clusters

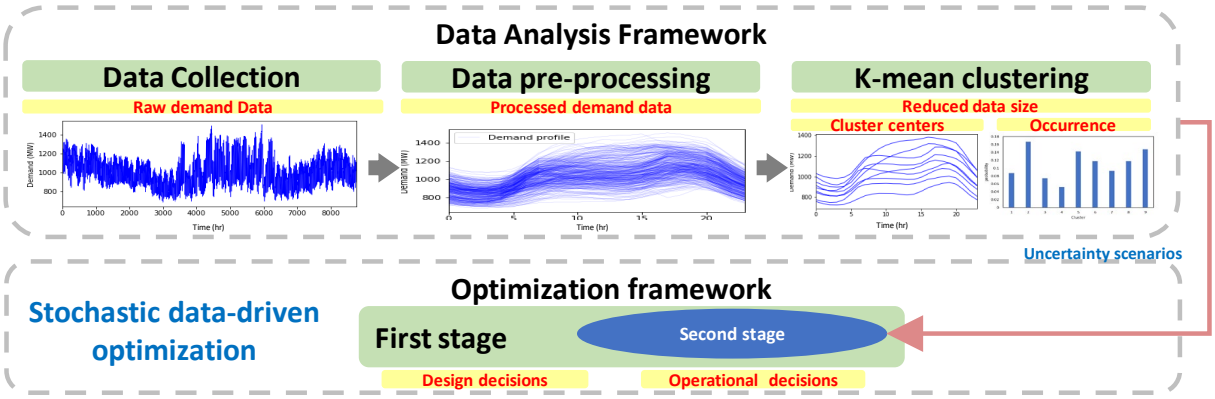

Figure 4. Stochastic data-driven design and operation of power system

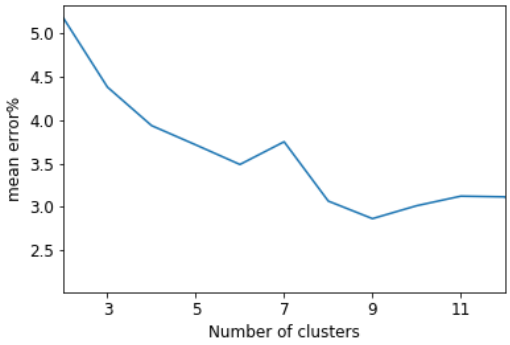

Figure 5. Effect of cluster number on the average error for electricity demand

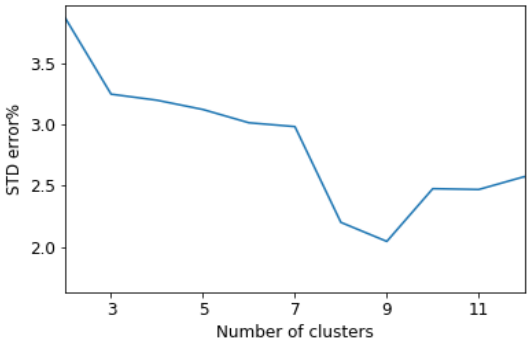

Figure 6. Effect of clusters number on the average standard deviation for electricity demand

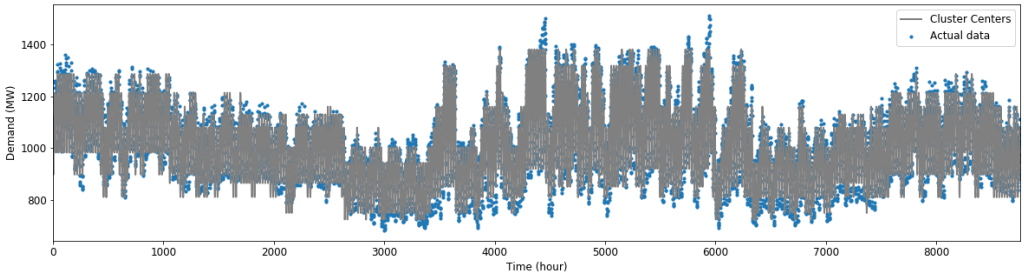

Figure 7. Actual electricity demand and its computed cluster centres for 1-year time horizon

Table 1. Comparison between deterministic, stochastic and worst-case design and objective function solution of power generating model

Deterministic		Stochastic		Worst-case scenario	
Number of generating units	Capacity (MW)	Number of generating units	Capacity (MW)	Number of generating units	Capacity (MW)
2	455	2	455	2	455

1	130	2	130	2	130
1	162	1	162	1	162
1	80	1	80	1	85
		2	55	1	80
				3	55
Total thermal generating units	5 with total capacity of 1282 (MW)	8 with total capacity of 1522 (MW)		10 with total capacity of 1662 (MW)	
Objective function (net present cost)	Total cost: 5.56 billion $ Capital: 4.16 billion $ Operating: 1.40 billion $	Total cost: 6.35 billion $ Capital: 4.94 billion $ Operating: 1.41 billion $		Total cost: 7.37 billion $ Capital: 5.40 billion $ Operating: 1.97 billion $	

Table 2. Comparison between stochastic solution with external electricity supply and worst-case design objective function solution of power generation model

	Stochastic solution with external electricity supply	**Worse-case scenario**
Total cost	6.64 billion $ 6.35 billion $ (stochastic solution) 0.29 billion $ (extra needed when 20% of year demand is extreme)	7.37 billion $

Table 5. Data for the thermal generating unit

Unit	PL	PU	a	b	TU	TD	Hsc	Csc	Tcold	Tini	RD	RU
	MW	MW	$/h	$/MWh	H	H	$/h	$/h	h	h	MW/h	MW/h
1	150	455	960.61	16.479	8	8	4500	9000	5	8	91	91
2	150	455	944.56	17.447	8	8	5000	10000	5	8	91	91
3	20	130	691.13	16.9	5	5	550	1100	4	-5	26	26
4	20	130	670.65	16.817	5	5	560	1120	4	-5	26	26
5	25	162	423.06	20.447	6	6	900	1800	4	-6	32.4	32.4
6	20	80	355.05	22.972	3	3	170	340	2	-3	16	16
7	25	85	477.93	27.827	3	3	260	520	2	-3	17	17
8	10	55	656.49	26.188	1	1	30	60	0	-1	11	11
9	10	55	663.11	27.414	1	1	30	60	0	-1	11	11
10	10	55	668.53	27.902	1	1	30	60	0	-1	11	11

Acknowledgements

References

Bertsimas D. and A. Thiele, "Robust and Data-Driven Optimization: Modern Decision MakingUnder Uncertainty," in *Models, Methods, and Applications for Innovative Decision Making*, ed: INFORMS, 2006, pp. 95-122.

Alhameli F., "Multiscale modeling in mathematical programming: Application of Clustering," *Ph.D. Dissertation Chemical Engineering*, University of Waterloo, Waterloo, Ontario, Canada, 2017.

Pedregosa, F., G. Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., *et al.*, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.,* vol. 12, pp. 2825-2830, 2011.

Y. Pochet and F. Warichet, "A tighter continuous time formulation for the cyclic scheduling of a mixed plant," *Computers & Chemical Engineering,* vol. 32, pp. 2723-2744, 2008/11/24/ 2008.

Wu, D. and Ierapetritou, M., "Cyclic short-term scheduling of multiproduct batch plants using continuous-time representation," *Computers & Chemical Engineering,* vol. 28, pp. 2271-2286, 2004/10/15/ 2004.

Tumuluru, V. K., Huang, Z., and Tsang, D. H. K., "Unit commitment problem: A new formulation and solution method," *International Journal of Electrical Power & Energy Systems,* vol. 57, pp. 222-231, 2014/05/01/ 2014.

Marcovecchio, M. G., Novais, A. Q., and Grossmann, I. E., "Deterministic optimization of the thermal Unit Commitment problem: A Branch and Cut search," *Computers & Chemical Engineering,* vol. 67, pp. 53-68, 2014/08/04/ 2014.

Hourly Ontario and Market Demands *2018*. Available: http://www.ieso.ca/Pages/Power-Data/Data-Directory.aspx January 15, 2019.

Simopoulos, D. N., Kavatza, S. D., and Vournas, C. D.,"Unit Commitment by an Enhanced Simulated Annealing Algorithm," in *2006 IEEE PES Power Systems Conference and Exposition*, 2006, pp. 193-201.

Elf M. , Gutwenger, C. , M. J,, nger, and Rinaldi, G., "Branch-and-Cut Algorithms for Combinatorial Optimization and Their Implementation in ABACUS," presented at the Computational Combinatorial Optimization, Optimal or Provably Near-Optimal Solutions [based on a Spring School], 2001

Sahinidis, N.V., "Optimization Under Uncertainty: State-of-The-Art And Opportunities",*Computers & Chemical Engineering*. 28, 971–983 .

Grossmann, I.E. , Apap, R.M. , Calfa, B.A. , García-Herreros, P. , Zhang, Q., "Recent Advances In Mathematical Programming Techniques For The Optimization of Process Systems Under Uncertainty", *Computers & Chemical Engineering*, vol. 91, pp. 3–14, 2016.

Ierapetritou, M.G. , Pistikopoulos, E.N. , Floudas, C.A., "Operational planning un- der uncertainty". *Computers & Chemical Engineering*, vol. 20, pp. 1499–1516, 1996

Biographies

Mohammed Alkatheri holds a BSc 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 process and carried out green-house gases life cycle assessment for the United Arab Emirates electricity sector. In May 2017, he joined PhD program in Chemical Engineering at University of Waterloo. His PhD research is focusing 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.

Muhammad Rizwan is a Postdoctoral Fellow at the Department of Chemical Engineering, Khalifa University of Science and Technology, Abu Dhabi, United Arab Emirates. He holds a PhD degree in Chemical Engineering from Korea Advanced Institute of Science and Technology (KAIST), South Korea. His main research interests are in the area of Process Systems Engineering (PSE) with the focus on energy systems design, modeling, simulation and optimization. He developed methodological frameworks for the superstructure based optimization of biorefinery networks as well as waste-to energy networks. He is currently focusing on big data analytics and the application of machine learning in gas operations to improve the process performance and profitability.

Falah Alhameli is currently a research and development engineering at Abu Dhabi national Oil Company (ADNOC). He earned a B.S. and an M.S. 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 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 BSc in Chemical Engineering and BSc 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 supervised over 90 graduate students (of which 35 are PhDs) 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.

Ali Almansoori is Professor of Chemical Engineering at Khalifa University in Abu Dhabi. During his profession, Dr. Almansoori 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. His main research interest is in the area of Process Systems Engineering with the focus on energy systems design, simulation, modelling and optimization. He also conducts general research in the area of renewable energy and fuel cell technology with applications to the oil and gas industry. He has published numerous articles in renowned refereed journals and conference proceedings. He also delivered several presentations in international conferences and is the author of a few book chapters. Furthermore, he serves as a reviewer for reputable international journals in the area of energy and process systems.

Peter Douglas is the Associate Dean of Engineering (Undergraduate Studies) and a Professor of Chemical Engineering at the University of Waterloo. He was previously the Director of the University of Waterloo United Arab Emirates Campus in Dubai from 2009 to 2013, the Associate Dean of Engineering (Computing), and the Associate Dean of Engineering (Graduate Studies). Professor Douglas was a founding member of WISE the Waterloo Institute for Sustainable Energy at UWaterloo. His primary research area of interest is in the development and application of PSE technology to industrial processes including process modelling, simulation, control and optimization. He is currently working on simulation and optimization issues related to the mitigation and capture of carbon dioxide from large scale emitters. Professor Douglas has consulted on a world-wide basis for many clients and has worked in Canada, Australia, Malaysia, Thailand, the UAE. Additionally, he is a co-inventor of the Dryer Master online measurement and control systems for the food processing industry; such systems are finding widespread use in Canada, USA, Europe and Asia. In addition to his research work, Professor Douglas has co-authored more than 200 related research publications and has supervised more than 80 postgraduate students

.