Malaysian Journal of Mathematical Sciences 10(S) February: 277–297 (2016)
Special Issue: The 3^{rd} International Conference on Mathematical Applications in Engineering 2014 (ICMAE'14)

MALAYSIAN JOURNAL OF MATHEMATICAL SCIENCES
Journal homepage: http://einspem.upm.edu.my/journal

Low Cost Genetic Algorithm to Photovoltaic-Diesel Power System Design Problem

Osman, M. H.*[1], Sopian, K.[2], and Nopiah, Z. M.[1]

[1]*Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia*
[2]*Solar Energy Research Institute, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia*

*E-mail: haniff68@ukm.edu.my**

ABSTRACT

In order to find ideal design of hybrid photovoltaic-diesel power system, genetic algorithm is an efficacious technique. The optimum design gives architecture structure, which has finest selection of components and size in accordance with suitable controlled strategy to offer budget friendly and dependable energy substitute. During the search of the best solution, among potential solutions, the aforementioned algorithm tries to find out solution that has minimum total net present cost. Yet, while using complex economic models in order to calculate the value of population fitness, this hunt takes the shape of expensive optimization problem. In order to shrink the deficiencies carried by genetic algorithm, current paper proposes low cost version of cluster-based genetic algorithm grounded upon statistical approaches, which considerably reduces the cost for evaluation of fitness function and bolster the performance. Population is divided into several clusters and multiple linear regression model is obtained from each clusters. Principal component analysis takes responsibility to increase possibility of having good estimation. During the course of probing, the values of population fitness are computed from corresponding model that is comparatively cost effective than direct evaluation. Algorithm is being used to identify clusters is denoted by k-means, which operates the process at low cost. The performance of proposed method is judged on the

basis of benchmark case study. The obtained results indicate the efficacy of proposed method for the model of hybrid photovoltaic-diesel power design via genetic algorithm workflow.

Keywords: Hybrid photovoltaic-diesel energy system, genetic algorithm, regression, principal component analysis, clustering.

1. Introduction

Development of a hybrid photovoltaic-diesel renewable energy power system (PV-diesel system) might be regarded as a problem faced in arrangement of elements. For instance, taking an example of battery, PV panel, each of them have a specific, type and strategy that will define suitable size of elements and a control strategy, which will enhance the performance along with reduction of cost. The problem related to optimization of PV-diesel system design is almost impossible to be resolved in polynomial time as search space widens up and myriad types of elements are included in the consideration. This issue instigated the researchers to use heuristic techniques to obtain a satisfactory (near to optimum) solution. A genetic algorithm (GA) is one of well-known heuristic techniques used to solve not only PV-diesel system design problem but other hybrid renewable energy power system (HRES) also (Fadaee and Radzi, 2012). The GA follows a probing strategy established upon laws of natural evolution in order to find optimum solution. In addition to, demandless examination of the structure of the data and to integrate it with supporting knowledge, the capacity of GA to examine solutions in myriad directions have termed it as common selection for resolving the combinatorial optimization problems (Konak et al., 2006).

Using GA in the capacity of search engine to design PV-diesel system, it is of utmost importance to understand, the relative higher cost for the appraisal of the problem because of two reasons. First, genetic algorithm required higher costs to gauge the suitability of large number of individuals in the population. This is supposedly necessary as the genetic algorithm has ability to provide acceptable results when it is dealing with large population. Genetic drift is one of the negative outcomes of having smaller population size. Second, cost on the evaluation of individual that is associated with total net present cost (TNPC) calculation turns implementation of GA. TNPC is the cost incurred at the time of investments along with the discounted current values of all future costs that will incur throughout the useful life span of the implemented system. In fact, it embodies actual fitness value of the different potential solutions; however it

demands long simulation procedure, as depicted in HOGA (Dufo-López and Bernal-Agustín, 2005). Subsequent part will give relevant contextual information about HOGA. Generally, including knowledge into genetic algorithm can minimize calculation stress, according to which easiest method is assessing fitness of an individual after comparison with other similar individual. Clustering technique is one of tools, suggested by this approach (Wu, 2008).

According to clustering approach, population is divided into several small sub-groups. After this one individual is selected from each cluster and fitness is appraised by using original fitness function. The fitness of remaining members of group is measured against by taking proportion of fitness of representative individual, of same cluster, as benchmark. With less function evaluation, current GA is regarded as cost friendly optimization technique. Still, simplicity of GA does not provide assurance that GA will always converge to the global optimum Ref. (Santana-Quintero et al., 2010). There should be some betterment to address this concern. Ref. Wu (2008) proposed a maximum theoretical distance to scale the distance before performing fitness computation. In ref. Jong-Won and Sung-Bae (2011), through the membership function, the fitness values of other individuals are estimated from the fitness values of the representative individuals. The importance of estimation model has been surfaced, because of improperly selected representatives of the cluster can bring impediments in the integration of GA (Shi and Rasheed, 2010). The grounding of the concept is to estimate the value of individual fitness through the usage of locally developed estimation scheme for each cluster. Polynomial function, neutral network and support vector machine and Krigling model are most popular models for this purpose (Santana-Quintero et al., 2010). But, use of aforementioned models to the glitches of a higher dimension is not applicable because of the computer cost of developing and executing model in several cases are comparable to the original fitness function (Santana-Quintero et al., 2010).

This research aims at redressing the technical deficiencies of cluster-based GA and suggests a substitute evaluation method. Moreover, it is trying to seek a cost effective version of GA for designing PV-diesel power system. This study believes that it can be accomplished by establishing new relationships between determining factors and TNPC. Thus, multiple linear regression (MLR) analysis is conducted to establish estimation functions for calculating fitness values of the population of the PV-diesel optimization problem. Such estimates of functions are being provided, based on size of components e.g. quantity of PV panel, diesel generator power output, etc and strategies of operation. In order to deal with the multicollinearity issue, faced in regression analysis, along with improving predictive power of the models, regression coefficients are estimated using principal component analysis (PCA). Interestingly, it should be noted

that the PCA-tuned regression analysis is conducted at initial stage of GA and every cluster will be linked to a MLR model. Although, decrease in number of fitness evaluation has not been witnessed, yet the allocation of individual fitness through usage of linear regression is simple than TNPC calculation. Moreover, this new evaluation method is more result oriented and meaningful than the indirect method applied in conventional cluster-based GA.

2. Photovoltiac-Diesel Power System Design

Hybrid PV-diesel power system is composed of solar photovoltaic panels, a battery bank, diesel generator and inverter, which function together in efficient and smart manner. Renewability resources of the energy, whenever it is placed in sunlight, are the major benefit of PV-diesel system. Moreover, diesel generator provides the back-up support facility in the absence of primary sources of energy. Having this capability, hybrid system is more advantageous, budget friendly and clean than that of conventional single source diesel system of energy (Amer et al., 2013).

The main task of developing autonomous system of power through exploiting renewable sources of energy is to determine exact selection of components along with their size and to define an appropriate strategy operate the system, which will be dependable and cost effective for a longer period of time. Fig. 1 depicts an example of PV-diesel system in the form of block diagram where related PV-diesel optimization model can be viewed in ref. Dufo-López and Bernal-Agustín (2008). The classical optimization techniques fall short to provide desired results, whenever possible combinations and variables in the model exceed a specific limit. To cope with the limitation, a computer program called Hybrid Optimization through Genetic Algorithm (HOGA) is purposefully designed to addresses the issue of size and operational control of the PV-Diesel system design (Dufo-López and Bernal-Agustín, 2005).

HOGA utilizes GA to search for an ideal structure that requires minimum investment to install the system. Optimum configuration is defined and expressed very succinctly: and the type of PV panels, the quantity of photovoltaic panels, type and numbers of batteries, power of diesel generator, the inverter power and the strategy for optimal control over the system along with its parameters. Moreover, this program reduces computational time to a great extent as it enhances the configuration of the system as well as control strategy, without rehearsing the whole simulation as the process of HYBRID2 does (Green and Manwell, 1995). Because of GA, HOGA can be extendedly used to enhance the battery state of charge set point within paltry computation

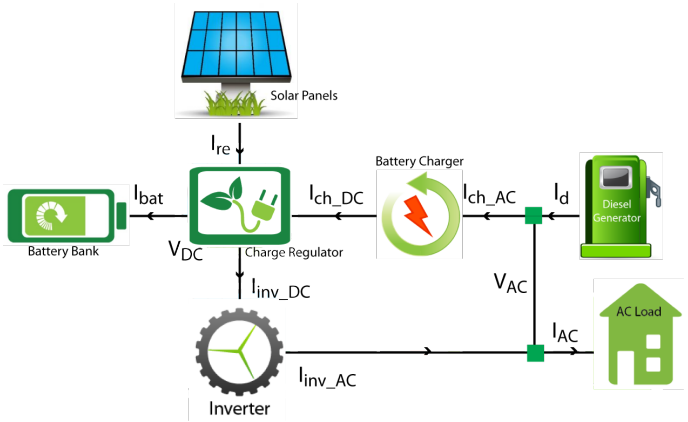


Figure 1: Block diagram of the PV-diesel power system. Every hour the following input data must be estimated: the current from the PV generator,I_{re}, AC load current I_{AC} which depends on the predicted load, and the battery state of charge SOC. The remaining current ($I_{AC} \cdot \frac{V_{AC}}{V_{DC} \cdot \eta_{inv}} - I_{re}$ where V_{AC} and V_{DC} are the DC and AC voltages and η_{inv} is the inverter efficiency) will be supplied by the batterries I_{bat} or by the diesel generator I_d or by both of them (Dufo-López and Bernal-Agustín, 2005)

time as compare to other commercial software such as HOMER (Manwell and McGowan, 1993).

In HOGA, the objective function is to minimize the TNPC of the system throughout the useful lifetime of the system. Generally, life of the system is represented by life of PV panel. The TNPC includes both fixed and variable costs as depicted below (Dufo-López and Bernal-Agustín, 2005):

$$TNPC = C_{Var} + C_{Fix}. \tag{1}$$

The variable costs C_{Var} may change in accordance with according size of the system and strategy for control and it is expressed as follows:

$$\begin{aligned} C_{Var} &= C_{ACQ\_PV} + C_{ACQ\_B} + C_{ACQ\_BCH} + C_{ACQ\_GEN} \\ &= +C_{ACQ\_INV} + C_{ACQ\_REG} + C_{REP\_B} + C_{REP\_INV} \\ &= +C_{REP\_REG} + C_{REP\_GEN} + C_{O\&M\_GEN} + C_{FUEL} \end{aligned} \tag{2}$$

where $C_{ACQ\_PV}, C_{ACQ\_B}, C_{ACQ\_BCH}, C_{ACQ\_GEN}, C_{ACQ\_INV}$ and C_{ACQ_REG}

are representing the cost incurred for acquiring PV panels, the batteries, the battery charger, the diesel generator, the inverter and the charge regulator. C_{REP_B}, C_{REP_INV}, C_{REP_REG} and C_{REP_GEN} these costs are incurred for replacement of the batteries, replacement of the inverter, replacement of regulator and generator. $C_{O\&M\_GEN}$ is the operational and maintenance costs of the diesel generator during the life span of the system. C_{FUEL} is fuel cost which is consumed by the system during the whole operational life. In the meantime, the C_Fix (Eq. (3)) has fixed opening cost and life therefore it does not dependent upon the strategy.

$$C_{Fix} = C_{REP\_BCH} + C_{O\&M\_PV} + C_{O\&M\_B} \tag{3}$$

where C_{REP_BCH} is the costs associated with of replacement of batteries and charger battery of the during the whole life span of the system. $C_{O\&M\_PV}$ and $C_{O\&M\_B}$ are the correspondent costs of maintenance of the PV panel and the batteries. Individuals are advised to refer the original paper about HOGA (Dufo-López and Bernal-Agustín, 2005) for further details about mathematical equations underlying each cost.

Nevertheless, it is important to highlight how the functions in Eqn. (2) and Eqn. (3) could cause high complexity in evaluation of individual fitness in HOGA. Only three costs, from a total of fifteen aforementioned costs, have fix values, while rest of the costs computation demands simulation results as inputs. For instance, replacement cycle during the year is demanded for calculation of costs of acquisition of components and consumption of diesel by the generator is determined by annual hours. Information about these factors can be only obtained after conducting simulation on PV-diesel system with specific configuration, depending upon solar irradiation and load demand. Thus, there will be complexity in computation of fitness value in HOGA since the number of simulation conducted in HOGA is according to the size of population and repetition of GA.

3. Proposed Method

Through combined the application of regression analysis, principal component analysis and clustering technique, we will purpose a cost effective and promising method to evaluate fitness value for GA. Interestingly, the proposed method will occur at the initial stage of GA process, soon after performing individual population initialization. As it is appeared once in GA, this method can also be considered a pre-processing tool for optimization technique based

on population.

This process is developed in two parts. During the first part, individuals comes with an actual fitness value i.e. TNPC, are clustered into a group according to their similarities by using k-means technique. Only component sizes are used when performing clustering. The reason for which we have used partial system information is that impact of control strategy on regression model can be examined separately for each cluster. The factor could be avoided from the model if adding it adds nothing to the explanation of TNPC. Apart from that, a specific cluster validity index is analysed to resolve the problem of ideal number of clusters.

In the second part, a multiple regression method is applied to regress TNPC value on the system configuration. Principal component analysis estimates regression coefficients due to its capability of tackling a multicollinearity problem i.e. high correlation between the explanatory variables. The analysis is independently done for each resultant clusters. The proposed method along with its flow chart is depicted in Fig. 2.

Throughout the evolution process of GA, a new individual is assigned to an appropriate cluster before the implementation of associate MLR model for the estimation of fitness value. Further changes are not required in conventional GA-based PV-diesel system design.

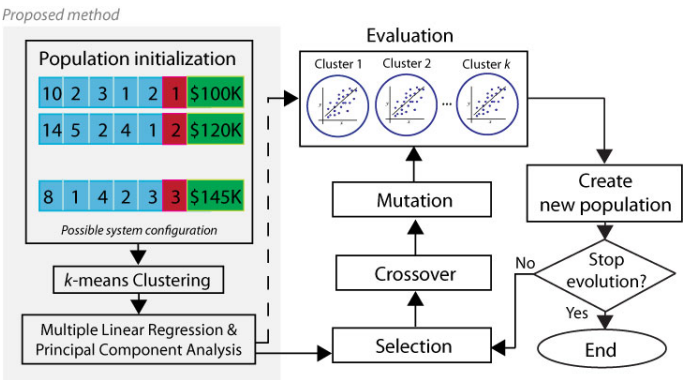


Figure 2: Proposed method takes place at an initial stage of GA

3.1 Designing GA for PV-Diesel System Design

Through using GA, finding an optimum system configuration for PV-diesel system, is the basic objective of this study. Therefore, it is essential to design GA for the problem. GA performs by engendering a population of numerical vectors referred as individuals and all of the presents a potential solution to a specific problem. In GA terminology, an individual is a string of genes, which has effective value for parameter of interest. In this study, each individual comprises over three parts; size, strategy for control and fitness value. The first part consist information regarding size such as quantity of PV panels, total number of battery, solar generator, type code of battery and type code of generator. The second part is a sole gene with three possible values and these values are indicated by numbers (1, 2 and 3). Every number identifies if the system rub in strategy of load following (1), cycle of charging (2), otherwise combination of those two strategies (3). Final part, preliminary individuals holds a TNPC value, in the meantime fitness values of later arrived individuals is assessed by their particular regression model.

New individuals can be developed by three operators; crossover, mutation and reproduction. Reproduction operator duplicates individuals into subsequent breeding pool along with their probable fitness value. In the meantime, crossover picks two individuals and then mates them, producing two new individuals. Sometimes, new individual might be produced through mutation instead of crossover operation. A new individual is gained through mutating by keeping a value at single gene location. As the PV-diesel system design optimization problem employs fundamental individual representative therefore fundamental operators like elitism for reproduction, uniform mutation and single-point crossover are relatively satisfactory.

3.2 *k*-means Clustering Algorithm

An unlabelled data set can go toward a treasure when certain concealed structure or specific grouping is identified. Cluster or group entails samples of data that may express similarities to a certain degree within the group. Due to the availability of data mining techniques, it can be unravel through clustering techniques. The most famous and frequently used technique is k-means, as it is easy to implement and use.

k-means technique for clustering, divides dataset into smaller and similar sets. This separation is based on distance measure in which data sample is incorporated to the cluster related to the closest centroid. Cluster centroid is a center of cluster. For crisp clustering like k-means, each cluster is completely

separate from the other clusters and no overlapping occurs. There are several methods of distance measures such as city block distance ($m = 1$), Euclidean distance ($m = 2$) and Minkowski distance ($m \geq 3$) (Jong-Won and Sung-Bae, 2011). These methods compute the distance from the notation:

$$d_{ij} = d(\vec{x}_i, \vec{x}_j) = \sqrt[m]{\sum_{t=1}^{|\vec{x}|} |x_{it} - x_{jt}|^m} \tag{4}$$

Probably the toughest task of k-means is to indicate or pick the number of clusters to be made, k in data set. It is very critical as the quality of separation can be hampered by the selected value of the k. A set of k values could be opted instead of selecting a single value of k. As suggested in ref. Maulik and Bandyopadhyay (2002), several runs of the k-means algorithm are performed for a fixed value of k, and the clustering corresponding to the run that provides the maximum value of the Dunn′s index (cluster validity index for crisp clustering) is assumed to be appropriate (Maulik and Bandyopadhyay, 2002).

4. Multiple Linear Regression Model

For each observed cluster, our aim is to construct a multiple linear regression model of the form:

$$\hat{Y} = \vec{X}\vec{\beta} + \lambda_{load} Load + \lambda_{cycle} Cycle + \lambda_{cmb} Cmb \tag{5}$$

where $\hat{Y}$ is the projected cost of PV-diesel system throughout lifetime of the system and $\vec{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_5)$ represents the vectors of regression coefficients. $\vec{X}$ is a 5×1 vector, which gauge the distinguishable variables that are related to size of system components. Description of every X_i can be noticed in the Table 1. *Load*, *Cycle* and *Cmb* are three dummy variables and λ_{load}, λ_{cycle} and λ_{cmb} are their corresponding coefficients. To demonstrate, when the system employed load following in the capacity of control strategy, here *Load* is set to $'1'$ allocating $'0'$ toward both *Cycle* and *Cmb*. Therefore, fitness value can be estimated by using the upcoming model $\hat{Y} = \vec{X}\vec{\beta} + \lambda_{load}$. The aforementioned rule can be implemented for the selecting other strategies for control.

Assuming there is multicollinearity among the explanatory variables, X_i thus PCA has been used to estimate the value of $\vec{\beta}$. Multicollinearity is an

Table 1: Explanatory variables in the multiple linear regression model

Variable	Component
X_1	Number of PV panel in parallel
X_2	PV panel peak power (Wp)
X_3	Number of battery in parallel
X_4	Battery nominal capacity (Ah)
X_5	Diesel generator output power

issue that may arise in multiple regressions through higher covariance and variance of coefficients when predictor variables have witnessed higher correlation. Concisely, PCA transforms the original variables into a new set of orthogonal or uncorrelated variables called principal components of the correlation matrix. This transformation ranks the new orthogonal variables in order of their importance. An ordinary least square estimation is employed to a set of principal components to obtain a multiple regression model of response variable. Henceforth, we call it as principal component regression (PCR). Although good results and predictions are achieved through PCR, but there is problem with interpretability of new variables of data set. Because of this reason it is suggested that principal component should be transformed into original variables in regression analysis.

Let suppose $\vec{b}_{pc} = (b_{1,pc}, b_{2,pc}, \ldots, b_{5,pc})$ is the vector of estimated coefficients of the parameters of vector $\vec{\beta}$. It should be noted, subscript pc is merely used to denote the estimators are principal component estimators instead of ordinary least squares estimators. Note that;

$$\begin{pmatrix} b_{1,pc} \\ b_{2,pc} \\ \vdots \\ b_{5,pc} \end{pmatrix} = \left(\vec{v}_1 \vec{v}_2 \ldots \vec{v}_k\right) \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_k \end{pmatrix} \tag{6}$$

where $(\vec{v}_1 \vec{v}_2 \ldots \vec{v}_k)$ is termed as $k \times k$ matrix belongs to eigenvector and has association with the principal components or eigenvalues and $(\alpha_1 \alpha_2 \ldots \alpha_k)^T$ is a $k \times 1$ vector belongs to coefficients associated with PCR. Parameter k represents a number of original variables hence the problem has k principal components. It is important to note that not all principal components are important for regression model. Removing less informative principal components will reduce the total variation in the model; therefore producing significantly improved prediction or diagnostic model. Since the principal components are ordered

according to its eigenvalues, it is sufficient to select the first r principal components ($r<k$) containing an amount of variation larger than a pre-defined percentage threshold (e.g. 85%) (Fekedulegn et al., 2002).

Assume r components are retained, therefore in Eq. (6) will be written as

$$\begin{pmatrix} \beta_{1,pc} \\ \beta_{2,pc} \\ \vdots \\ \beta_{k,pc} \end{pmatrix} = \left(\vec{v}_1 \vec{v}_2 \dots \vec{v}_{k-r}\right) \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{k-r} \end{pmatrix} \tag{7}$$

where $(\vec{v}_1 \vec{v}_2 \dots \vec{v}_{k-r})$ is the $k\times(k\text{-}r)$ matrix of eigenvectors associated with the retained principal component. The $(\alpha_1 \alpha_2 \dots \alpha_{k-r})^T$ refers a reduced vector of coefficient (α). After completing the findings $\vec{\beta}$, Eq.(6) will take form that is presented below:

$$\hat{Y} - \vec{X}\vec{\beta} = \lambda_{load} Load + \lambda_{cycle} Cycle + \lambda_{cmb} Cmb. \tag{8}$$

We can suppose, without losing generality let $\hat{Y} = Y$ and Eq.(8) is defined as following

$$Y - \vec{X}\vec{\beta} = \lambda_{load} Load + \lambda_{cycle} Cycle + \lambda_{cmb} Cmb. \tag{9}$$

$$E = \lambda_{load} Load + \lambda_{cycle} Cycle + \lambda_{cmb} Cmb. \tag{10}$$

Here E is the residual term where cluster density, N determines size of E. The coefficients of dummy variables were estimated using an ordinary least square estimator, so that

$$\begin{pmatrix} \lambda_{load} \\ \lambda_{cycle} \\ \lambda_{cmb} \end{pmatrix} = (X_S^T X_S)^{-1} X_S^T E \tag{11}$$

where $X_S = [Load\ Cycle\ Cmb]$ and each *Load*, *Cycle* and *Cmb* is a $N\times1$ vector with binary elements.

5. Experimental Results

We have conducted experimentation to confirm and prove the effectiveness of proposed method to accelerate GA to solve PV-diesel system optimization problem. With the aim to show exceptional performance of the method, we have made comparison of HOGA and conventional cluster-based GA. To provide fair comparison of different GA-based methods, only the first algorithm of HOGA was considered in the experiments.

5.1 Experimental Settings

The paper of author of original research paper of HOGA was settled as benchmark by us. A PV-diesel system nourishing the peoples of Zaragoza, Spain, has been developed and augmented. With 4.37 Wh/m^2 solar radiation and bright sunshine for the most part of the year, indicate that the region is favourable for the hybrid renewable sources of energy. The average radiation received per day is depicted in Table 2. Sequence of the power represents the daily demand for power of the community, which is supposed constant for a time unit of 1 hour, as depicted in Fig. 3. Variation in demand for different seasons is ignored in current study; daily load demand remains constant throughout the year. The demand for AC load is fulfilled by hybrid PV-diesel system, which has 48 DC voltages and 230 AC voltages as shown in Fig. 1.

Table 2: Average daily irradation

Month	Wh/m^2	Month	Wh/m^2
January	2138	July	6644
February	2688	August	5593
March	4150	September	4830
April	4931	October	3456
May	6318	November	2555
June	6941	December	2138

As adumbrated in previous section, GA will be utilized to identify a suitable control strategy along with value for five size parameters of PV-diesel system. Number of possible different PV panel types is 9. Maximum number of panels in parallel is 25. Number of different battery types is 12. Maximum number of batteries in parallel is 15. Number of possible diesel generator types is 7 (commercial diesel generators from 3 to 13 kW, with prices according to ref. Schmid and Hoffmann (2004): 0.55 €/W). The costs of the different PV panels and batteries are shown in Tables 3 and 4. Hence, the total number of ways

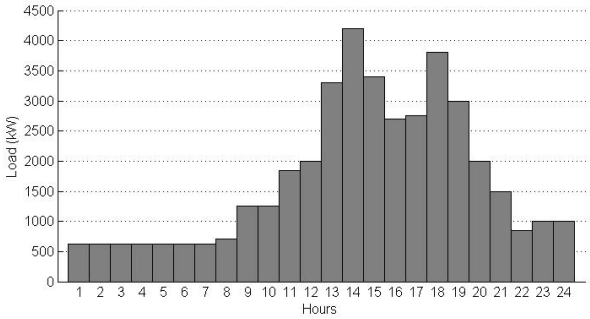


Figure 3: Daily load profile

the system could be configured is 9×25×12×15×7×3=850,500.

When we want to assess all available combinations, we will consume around 2.65 hours, at the rate of hundred calculations per second. Through GA, an acceptable solution will be found by examining not more than 30% of total combinations. Hence, it is indicated that ideal solutions provided by PV-diesel system can be found in fewer than 255,000 assessments irrespective of GA version. With a population of 1000 individuals, the GA requires maximum 250 generations to converge. On reaching stagnation stage i.e. cumulative change in the fitness function value over stall generations is less than or equal 10^{-3}, GA process would be instantly ended. Roulette wheel method for selection was adopted by us, for reproduction cycle, crossover rate was 90

Table 3: Investment costs of 12V PV panels

Peak power (Wp)	20	36	50	55	75	90	100	110	125
Cost (€)	278	297	385	413	525	676	744	812	884

Table 4: Investment costs of 12V batteries

Nominal capacity (Ah)	Cost (€)	Nominal capacity (Ah)	Cost (€)
43	155	187	433
64	202	200	565
69	207	308	843
96	258	385	971
144	288	462	1017
160	357	524	1054

Besides HOGA, there are two types of cluster-based GA were utilized to assess and compare the performance k means algorithm along with Euclidean distance was applied with k ranging from 2 to the nearest integer of $\sqrt{G}$ where G is GA population size. Appropriate number of clusters from the population that was seeded is provided with the highest average value of Dunn′s Index over 20 random runs.

In conventional cluster-based method which evaluates individuals partially, cluster centroids are appraised as cluster representatives. The fitness of remaining of the individuals was estimated in proportion to the distance from the cluster representative.

For the proposed method, two statistics were used in MLR to evaluate model fit: R-squared and the F-test. The F-test was used to evaluate the null hypothesis that adding dummy variables adds nothing to explanation of the model. An equivalent null hypothesis is that all regression coefficients of dummy variables equals zero. We reject null hypothesis if the p-value is smaller than the significance level α=0.01. Following this, a plot of the standardized residuals against the predicted values was studied to examine whether the MLR models are adequate for the data or the addition of terms involving the predictor should be involved. A random scatter of points signals that the current model is adequate.

5.2 Experimental Results

Referring to a recommendation in ref. Pakhira et al. (2004), the maximum possible number of clusters an initial GA population (dataset) could have is $\sqrt{1000}$ ≈31. Ranging from 2 to 31, the non-hierarchal k-means is executed on the normalized dataset. Box plots for 20 runs in Fig. 4 shows that the average value of Dunn′s Index peaks at cluster number 3, indicating that this is the optimal number of clusters for the dataset. For the following paragraphs, the three clusters are denoted as *Cluster1*, *Cluster2* and *Cluster3*.

Table 5 gives computed eigenvalues and percentages of variance associated with each principal component for each of the identified clusters. Obeying the ground theory of PCA, it can be seen that the first principal component accounts for the maximum proportion of variance from the original dataset in each cluster. For the first two clusters, the total percentage variance explained by the four components is greater than 85%. This indicates the fifth component will be excluded from regression analysis. Based on the same judgement, it seems that the first three components are the most important and have to be included in the regression model for the third cluster.

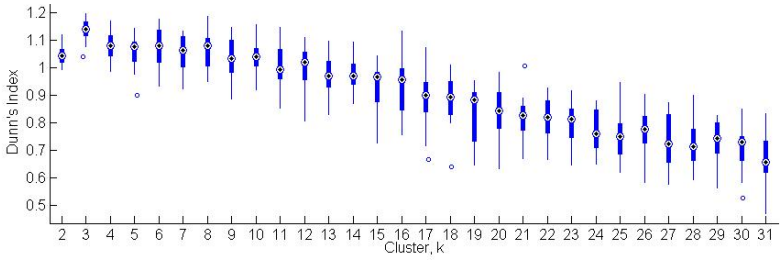


Figure 4: Values of Dunn′s index in the range of k=2,...,31 over 20 initial GA population using k-means

A first multiple regression analysis includes five elements of sizing components as explanatory variables to predict TNPC. For all clusters, the linear combination of the five variables was significantly related to the TNPC, for example in the *Cluster1*, $R^2 = 0.819, F(5, 366) = 325.79, p < 0.01$. As can be seen in Table 6, PV-related variables had negative regression weights, indicating that a system with a higher number of PV panels and/or high voltage panels was expected to be more economic by lowering the TNPC. These relationships might reflect key factors that could maximize energy utilization from the solar PV panel Lal et al. (2011). However, the effect of using a different type of PV panel in the MLR model of *Cluster2* was relatively small compared to other two models. On the other hand a system installed with a large number of batteries and/or high battery capacity was expected to have a higher TNPC. This is clearly demonstrated since both variables have a positive regression weight. Based on the results, an optimum system in this study is expected to be incorporated either with a high number of small capacity batteries or with fewer batteries but which have high storage capacity. Positive correlation between the fifth variable and TNPC was an unexpected occurrence due to the fact that both the acquisition and O&M costs of a diesel generator are calculated based on kW/€.

Next, the second analysis was conducted to evaluate the effect of dummy variables in the previous models. With respect to p-value$<$0.01 which indicates the rejection of null hypothesis at the 1% significance level, all the regression models with additional variables were statistically significant. The null hypothesis is when all the coefficients of dummy variables are zero. Table 6 demonstrates that the use of a *Load Following* strategy has a positive effect (denoted by a positive coefficient) on the total cost for systems in all clusters except *Cluster1*. However, systems in *Cluster2* and *Cluster3* that imposed either *Cycle Charging* or *Combined Strategy* as a control strategy can reduce

Table 5: Summary of PCA for the identified clusters

Cluster	Variable	Eigenvalue	Variance	Cumulative Eigenvalue (%)	Cumulative Variance (%)
Cluster1	**PC1**	23.73	40.72	23.73	40.72
	PC2	10.44	23.06	34.16	63.78
	PC3	10.47	17.97	44.63	81.75
	PC4	8.90	15.28	53.54	97.03
	PC5	1.73	2.97	55.27	100.00
Cluster2	**PC1**	15.11	29.15	15.11	29.15
	PC2	12.89	24.87	28.00	54.02
	PC3	11.86	22.89	39.86	76.92
	PC4	7.78	15.00	47.64	91.92
	PC5	4.19	8.08	51.82	100.00
Cluster3	**PC1**	19.10	38.85	19.10	38.85
	PC2	11.84	24.08	30.93	62.92
	PC3	9.96	22.25	40.89	85.18
	PC	46.17	10.56	47.06	95.73
	PC	52.10	4.27	49.16	100.00

Table 6: Regression summary of MLR model for investment cost (€) of PV-diesel power system

Item	Variable/Type	*Cluster1*	*Cluster2*	*Cluster3*
Cluster size	N	366	278	356
Coefficient	β_1	-8,263	-9,322	-3,774
	β_2	22,237	27,506	20,257
	β_3	-7,526	-648	-2,230
	β_4	31,336	38,525	29,691
	β_5	46,327	7,763	29,140
	λ_{load}	-20,557	69,501	66,962
	λ_{cycle}	16,018	-20,444	-44,883
	λ_{cmb}	22,602	-518	-10,060
R^2	5 variables	0.819	0.828	0.821
	8 variables	0.835	0.846	0.836
F-value	F(5,8)	325.79	261.88	321.06
	F(8-5,N-8)	11.262	7.697	10.732
p-value	5 variables	0.0	0.0	0.0
	8 variables	0.0	0.0	0.0

cost.

A normal probability plot of the residuals from each cluster is shown in Fig. 5. Apart from a general linear trend that can be observed we do not have any severe outliers in each subplot in Fig. 5. In addition for each cluster, the normality test indicates that the coefficient of correlation between the ordered residual and their expected values under normality is greater than the critical value at 1% significance level. These observations suggest that the assumption that the residuals follow a normal distribution is reasonably satisfied.

A graph of average population fitness versus generation number, as shown in Fig. 6, is referred for methods comparison for a PV-diesel system design problem. The methods are the HOGA, the conventional cluster-based GA (CCGA) and the proposed method (MLR-aid). This study is more interested in the *average values* instead of *best result* thus we can observe how the new fitness estimation function reacts during the evolution process. Fig. 6 is also a representation of typical behaviour for 20 runs that we observed. It is clearly shown that the CCGA has failed to converge less than 50 generations and the corresponding plot (Fig. 6(b)) suggests that it might not happen even if it performs more evolution processes. However, contrary to this, both MLR-aid and HOGA reached their respective optimum solution and surprisingly it was completed before one-tenth of maximum generation. Speaking in terms of number of iterations both methods performed, not more than 20,000 fitness evaluations, which is a relatively small number compared to 255,00 iterations if running a full simulation. In addition, it seems that the convergence rates of both methods are quite close for the first five evolution process before the proposed method got stuck in a local optimum for some generations before it managed to converge.

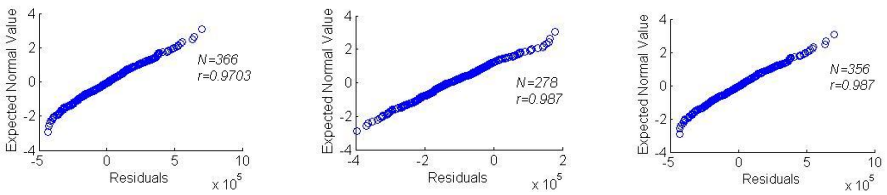
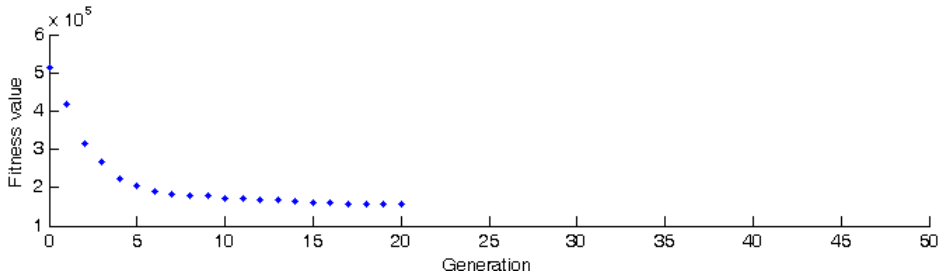
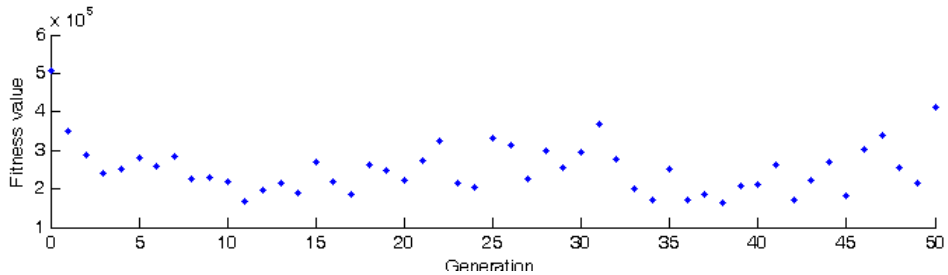


Figure 5: Normal probability plot of the residual for the three identified clusters, (a) *Cluster1*, (b) *Cluster2* and (c) *Cluster3*

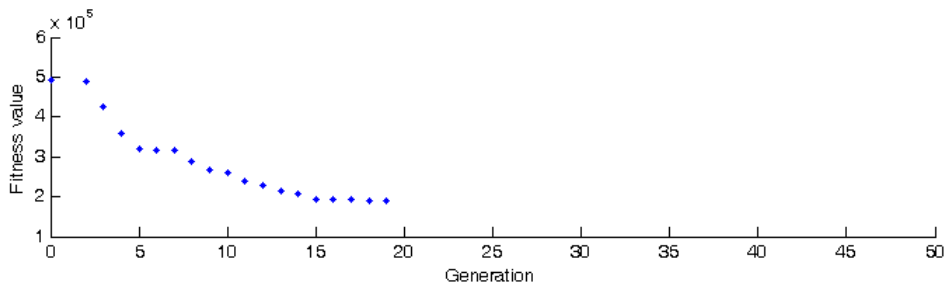
Table 7 shows the best solution for a PV-diesel system design problem from HOGA, CCGA and MLR-aid. HOGA′s solution is considered as the optimal for the design problem since it is only method in this study which used TNPC as fitness function. The optimum configuration costs the system were €158,527. According to MLR-aid, the best solution was estimated to cost



(a)



(b)



(c)

Figure 6: Evolution processes for the PV-diesel system optimization problem. Results are displayed for three methods, (a) *HOGA*, (b) *CCGA* and (c) *MLR-aid*

€168,750 but the corresponding TNPC value is €169,100 which is less than 7% above the global optimum. Different to this, the CCGA found the best combination of components of the system should cost €157,051 but in reality the cost was €194,484. In terms of component selection with respect to the optimum solution, CCGA′s solution is obviously flawed compared to MLR-aid since it has fewer similarity points. As a comparison, MLR-aid suggests the same type of PV panel, even if the system has to use four more panels. It also uses the same number of batteries but with a different power selection. Nonetheless, all methods agreed that *Cycle Charging* is the suitable dispatch strategy for the PV-diesel system.

In the last column of Table 7, CPUTime required for the three methods is represented. The time was reduced by almost 25% and 1% respectively using CCGE and MLR-aid. This result shows the effectiveness of the fitness approximation approach by means of a clustering technique that can reduce computational time in GA when solving combinatorial-type optimization problems.

Table 7: An optimum system configuration for PV-diesel power system

Item/Method	HOGA	CCGA	MLR-aid
X_1	14	15	13
X_2	2	1	2
X_3	125Wp	110Wp	125Wp
X_4	69Ah	64Ah	144Ah
X_5	3kW	4kW	6kW
Strategy	*Cycle Charging*	*Cycle Charging*	*Cycle Charging*
Est. Cost (€)	-	157,502	168,750
TNPC (€)	158,527	195, 483	169,100
Difference (%)	-	19.7	6.67
CPU Time	2776.25±5.38	651.74±7.24	11.51±2.21

6. Conclusion

A fitness approximation approach has gone through specific improvements to speed up GA to find an optimum design for a PV-diesel system. Substituting the use of TNPC during fitness evaluation, a number of multiple linear regression models were constructed before the GA starts searching. Prior to the model construction, an initial population of GA was partitioned into several clusters using k-means. Each cluster was then associated with a linear model. Regression coefficients of the models were estimated using principal component analysis and validity of the models were statistically tested. The

model enables GA to estimate individual fitness values from the model which cluster the individual belongs to.

The experiment reveals that computational time can be reduced by the proposed method and there will be no compromising of the GA where convergence is concerned. However, it would seem that gaps still do exist between conventional GA and any proposed methods as there exists alternate optimal solutions for design problems in a PV-diesel design. Future and further investigation should, however, help to minimize these gaps when focused on robust estimation procedures and a non-linear regression model. Fitness evaluation employs different economic models, so where renewable energy power system design problems exist, this model will be appropriate.

Acknowledgments

The authors would like to acknowledge financial support from the Ministry of Higher Education of Malaysia MOHE under Grant No.FRGS/2/2013/TK06//1.

References

Amer, M., Namaane, A., and M'Sirdi, N. (2013). Optimization of hybrid renewable energy systems (hres) using {PSO} for cost reduction. *Energy Procedia*, 42(0):318 – 327. Mediterranean Green Energy Forum 2013: Proceedings of an International Conference MGEF-13.

Dufo-López, R. and Bernal-Agustín, J. L. (2005). Design and control strategies of pv-diesel systems using genetic algorithms. *Solar Energy*, 79(1):33–46.

Dufo-López, R. and Bernal-Agustín, J. L. (2008). Influence of mathematical models in design of pv-diesel systems. *Energy Conversion and Management*, 49(4):820 – 831.

Fadaee, M. and Radzi, M. A. M. (2012). Multi-objective optimization of a stand-alone hybrid renewable energy system by using evolutionary algorithms: A review. *Renewable and Sustainable Energy Reviews*, 16(5):3364–3369.

Fekedulegn, B. D., Colbert, J. J., Hicks Jr., R. R., and Schukers, M. E. (2002). *Coping with multicollinearity : an example on application of principal components regression in dendroecology*. U.S. Dept. of Agriculture, Forest Service, Northeastern Research Station.

Green, H. J. and Manwell, J. (1995). Hybrid2-a versatile model of the performance of hybrid power systems. In *Proceedings of WIndPower'95*.

Jong-Won, Y. and Sung-Bae, C. (2011). An efficient genetic algorithm with fuzzy c-means clustering for traveling salesman problem. In *2011 IEEE Congress on Evolutionary Computation (CEC)*, pages 1452–1456.

Konak, A., Colt, D. W., and Smith, A. E. (2006). Multi-objective optimization genetic algorithms: A tutorial. *Reliab. Eng. Syst. Safe*, 91:992–1007.

Lal, D. K., Dash, B. B., and Akella, A. K. (2011). Optimization of pv/wind/micro-hydro/diesel hybrid power system in homer for the study area. *International Journal on Electrical Engineering and Informatics*, 3:1.

Manwell, J. F. and McGowan, J. G. (1993). Lead acid battery storage model for hybrid energy systems. *Solar Energy*, 50(5):399 – 405.

Maulik, U. and Bandyopadhyay, S. (2002). Performance evaluation of some clustering algorithms and validity indices. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(12):1650–1654.

Pakhira, M. K., Bandyopadhyay, S., and Maulik, U. (2004). Validity index for crisp and fuzzy clusters. *Pattern Recognition*, 37(3):487–501.

Santana-Quintero, L., Montaťildeo, A., and Coello, C. (2010). A review of techniques for handling expensive functions in evolutionary multi-objective optimization. In Tenne, Y. and Goh, C.-K., editors, *Computational Intelligence in Expensive Optimization Problems*, volume 2 of *Adaptation Learning and Optimization*, pages 29–59. Springer Berlin Heidelberg.

Schmid, A. L. and Hoffmann, C. A. A. (2004). Replacing diesel by solar in the amazon: short-term economic feasibility of pv-diesel hybrid systems. *Energy Policy*, 32(7):881 – 898.

Shi, L. and Rasheed, K. (2010). *Computational Intelligence in Expensive Optimization Problems*, chapter A Survey of Fitness Approximation Methods Applied in Evolutionary Algorithms, pages 3–28. Springer Berlin Heidelberg.

Wu, X. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1):1–37.