CONCEPTUAL COST ESTIMATION OF PUMP STATIONS PROJECTS USING FUZZY CLUSTERING

Mohamed Marzouk* and Magdy Omar

Associate Professor, CEM Program, Nile University, Egypt
** Corresponding author (mmarzouk@nileuniversity.edu.eg)*

ABSTRACT: Conceptual cost estimates, are prepared at the very early stages of a project, and generally before the construction drawings and specifications are available. At this stage, cost estimates are needed by the owner, contractor, designer, or funding agencies for determination of the feasibility of a project, financial evaluation of a number of alternative projects, or establishment of an initial budget. Traditional approaches rely heavily on experienced engineers. This paper presents a method using fuzzy clustering technique for pump station projects cost estimation. The proposed conceptual cost estimating methodology provides fast and reliable results that can be very useful in the early stages of a project. The main cost drivers are identified using stepwise regression. Collected data are utilized to build the fuzzy clustering model. A training data set and a testing data set are used to calibrate the model. Sensitivity analysis is conducted to determine the appropriate model and the corresponding number of clusters that provides minimum error.

Keywords: *Conceptual Cost Estimation, Fuzzy Clustering, Pump Stations Projects*

1. INTRODUCTION

The conceptual cost estimation during the engineering planning stage of construction projects is important process for successful execution of thoseprojects.,This is attributed to the fact main structural systems, major construction methods, and most construction materials are determined in planningstage. However, due to the lack of detail design information during the planning phase, accurate cost estimation is hard to obtain even for the professional estimators. It was found that experiencedestimators can do better in this job compared to inexperienced professionals. The emerging development of modern artificial intelligence (AI) techniques, such as fuzzy clustering systems, the aforementioned estimating experience/knowledge can be acquired by learning from historical examples, so that accurate estimation (compared with the detail estimation) could be obtained with very limited available project information. This paper presents a parametric-cost model, dedicated to pump station projects. The proposed model is considered useful for preparing early conceptual estimates when there are little technical data or engineering deliverables to provide a basis for using more detailed estimating.

2. COST FACTORS OF PUMP STATION

The sizing of pump station components in the distribution system depends upon the effective combination of the major system elements: supply source, storage, pumping, and distribution piping. Population and water consumption estimates are the basis for determining the flow demand of a water supply and distribution systems. Flow and pressure demands at any point of the system are determined by hydraulic network analysis of the supply, storage, pumping, and distribution system. Supply point locations such wells and storage reservoirs are normally known based on a given source of supply or available space for a storage facility.

The various cost drivers of pump station projects have been identified and collected from literature, instructed interviews and surveys. Fourteen cost drivers have been concluded to have the most impact on the costs of pump station projects in Egypt. These fourteen factors are used to develop the parametric cost estimating model using Fuzzy

Clustering.A survey was prepared to collect historical data records, which are used for the training and the testing in order to be ready for the prediction of future projects. A total of 44 pump station projects (cases) were collected in the survey. These projects were divided into two sets: the first set (35 projects) is used to build the fuzzy model, while the second set is used to test its performance (nine projects).

Table 1 Identifiedcost drivers.

No	*Cost Driver*
1	Project type (PT)
2	Project Location (PL)
3	Population (PO)
4	Station Capacity (SC)
5	Distance between pump station and source (D)
6	Pumps type (PT)
7	No of Pumps (NP)
8	Individual Pump Capacity (IPC)
9	Pump Head (PH)
10	Pump Arrangement (PA)
11	Pump Motor Type (PMT)
12	Pump Motor Rating (PMR)
13	Header Pipe type (PPT)
14	Pump Price (PP)

3. IDENTIFICATION OF SIGNIFICANT COST PARAMETERS

Stepwise regression model is used to assess significance cost parameters.Stepwise regression is a systematic method for adding and removing terms from a multi-linear model based on their statistical significance in a regression.The procedure is based on generating asimple regression model for each variable and including the onewhich has the largest F statistic. The F statistic values whichindicates whether the independent variables are linearly related tothe dependent variable at a specific level of significance (correspondingto t statistic in simple regression models)[1]. Subsequently,Matlabchecks the performance of the model by adding and/orremoving independent variable(s) and comparing the resulting Fvalue against F-to-enter and F-to-remove values, respectively.The default values for F-to-enter and F-to-remove have been setto 0.05 and 0.1 level of significance, respectively.In thisprocess, a stepwise linear regression analysis was performed using MatlabStatistics Toolbox. The stepwise regression was performedin threesteps as depicted in Table 2.It should benoted that only three out of the fourteen parameters have been foundsignificant, and were used in developing the proposed method.These parameters are: (1) Population; (2) Station capacity; and (3) Pump price.

Table 2 Stepwiselinear regression analysis.

Step	*F-state*	*R-Square*	*RMSE*	*Variable*	*Coefficient*
1	6.80	0.479	8.291	Intercept	6.80314
				SC	6.77E-05
2	30.83	0.601	7.352	Intercept	2.89106
				SC	6.14E-05
				PO	0.013553
3	29.50	0.689	6.571	Intercept	0.560932
				SC	4.14E-05
				PO	0.015913
				PP	1.34E-05

4. FUZZY LOGIC MODELLLING

The concept of fuzzy logic is derived from the theory offuzzywas introduced by Zadeh[2]. Fuzzy logic provides a means forcoping with problems arising from unexpected situations. It is used to solve hard problems, by determining amathematical model that describes the system behavior,known as an unsupervised learning method.Adaptive Network Fuzzy Inference System (ANFIS) is known as fuzzy rule-based systems,fuzzy model, fuzzy expert system, and fuzzy associative memory. The ANFISis essentially an adaptive system that is able to extract rules and knowledge fromhistorical data bases and express these rulesin acomprehensive way through linguistic rules, to beeasily understood and applied.The learning process buildsa model that connects input variables with typicallyone output variable, using a set of rules. This approach allows a more efficient treatment ofthe inputs, to reduce the number of rules neededand to obtain a more clear and interpretable outputresponse surface.

4.1 Adaptive Network Fuzzy Interface System

A well known model named AdaptiveNetwork Based Fuzzy Interference System(ANFIS) proposed byJang[3]. Theprocess of fuzzification builds a certain number offuzzy sets represented by membership functions foreach

variable. The inference system uses thesefuzzy sets and the rules to build an output valuewhich is translated into real number("defuzzified").The ANFIS use a hybrid-learning algorithm inorder to identify the fuzzy sets by using thefollowing below two steps.

1. Fuzzification and Rule Identification: the sub-clustering method, as mentioned in [4], is used in order to identify the inputs of the fuzzy model. It allows building input-output functions, under the form of IF-THEN sentences.
2. Training: the training step optimizes the parameters in the ANFIS model which has been generated by the previous step.

4.2 Fuzzy Inference System

Fuzzy inference system consists of a fuzzification interface, a rule base, adatabase, a decision-making unit, and finally a defuzzification interface[4].FISconsists offive functional blockas shownin Fig. 1. The function of each blockas follows:

- Arule base containing a number of fuzzy IF–THEN rules;
- Adatabase which defines the membership functions of the fuzzy sets usedin the fuzzy rules;
- Adecision-making unit which performs the inference operations on therules;
- Afuzzification interface which transforms the crisp inputs into degrees ofmatch with linguistic values; and
- A defuzzification interface which transforms the fuzzy results of the inferenceinto a crisp output.

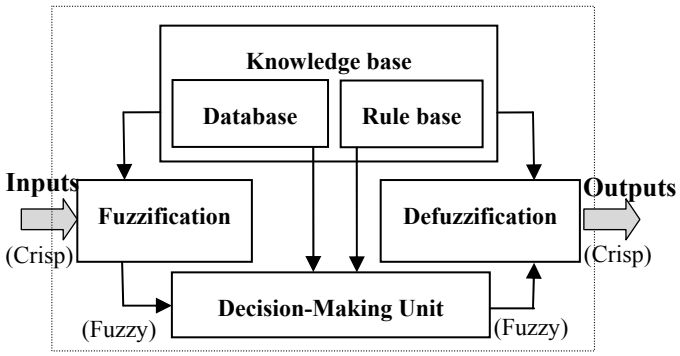

Fig. 1 Fuzzy interface system.

The working of FIS is as follows. The crisp input is converted in to fuzzyby using fuzzification method. After fuzzification the rule base is formed.The rule base and the database are jointly referred to as the knowledge base.Defuzzification is used to convert fuzzy value to the real world value which isthe output.The steps of fuzzy reasoning (inference operations upon fuzzy IF–THENrules) performed by FISs are:

1. Compare the input variables with the membership functions on the antecedentpart to obtain the membership values of each linguistic label(this step is often called fuzzification).
2. Combine (through a specific t-norm operator, e.g.,multiplication) the membership values on the premise part to get firing strength (weight) of each rule.
3. Generate the qualified consequents (either fuzzy or crisp) for each rule depending on the firing strength.
4. Aggregate the qualified consequents to produce a crisp output (this step is called defuzzification).

4.3 Determination of Fuzzy Clusters

Cluster algorithms(self-organizing map) are used to group data into subsetsor clusters that contain data having similar feature(s). Differentcluster algorithms have been developed in various applications using fuzzy logic. In those algorithms, data is expressed in a form offuzzy rules, each representing a cluster. Those algorithms include:(1) The fuzzy C-Means (FCM) clustering method (Bezdek andPal 1992);(2) mountain clustering method [5]; and (3) subtractive clustering method [4]. FCM clustering methodis an iterative technique that starts with a set of cluster centers and generates membership grades, used to induce new cluster centers [6]. The number of iterations depends on the choice of the initial values of the clusters' centers. Mountain clustering method is based on creating a grid of data space and computing the potential value (mountain function) for each point on the grid, based on its distance to the actual data point [7]. The greatest potential point (one of the grid vertices) represents the first cluster (highest point on the mountain). Subsequently, the potential for each grid point is adjusted, allowing for the

determination of all remaining clusters. Subtractive clustering method [4]is an extension of the mountain clustering method, where the potential is calculated for the data rather than the grid points. As a result, clusters are elected from the system training data according to their potential. Subtractive clustering has an advantage over mountain clustering in that there is no need for estimating a resolution for the grid. This method was adapted in the developmentmade in this research.

The cluster radius indicates the range of influence of a cluster when considering the data space as a unit hypercube. Specifying a small cluster radius yields many small clusters in the data, (resulting in many rules). While, specifyinga large cluster radius yields a few large clusters in the data, (resulting in fewer rules). The radius defining the neighborhood for each cluster has been determined using asensitivity analysis. Different values have been assigned to R using a trial and error procedure. An initial guess was made by trying values for R in the 0.5 to 0.65 range, with an increment of 0.05.Errors during training and testing were recorded for each R as perTable 3. In the developed model, R value was set to be 0.65 so asto minimize training the error. The two clusters associated with a model developed for estimating the cost of pump station projects are listed in Table 4 and shown in Fig 2.

Table 3 Identified two clusters centers.

Cluster No.	*Population (thousands)*	*Total Capacity (m^3/day)*	*Pump Price (L.E)*	*Project Cost (Million L.E)*
1	65,000	47,692	135,000	10.51
2	750,000	95,000	67,550	13.50

Table 4 Modelsensitivity analysis.

Cluster Radius	*Clusters No*	*% AAE*	*RMSE*
0.45	6	130	43.04
0.50	6	72	10.19
0.55	5	32	5.90
0.60	4	37	9.29
0.65	2	28	5.31

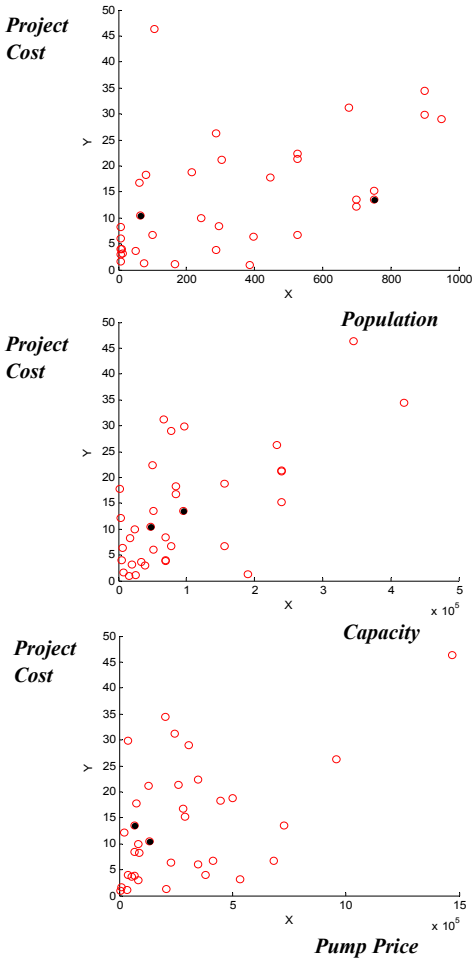

Fig. 4 Generated fuzzy rules.

4.4 FUZZY MODEL STRUCTURE

The Sugeno fuzzy model was implemented to estimate the pump station project cost. The proposed Sugeno fuzzy model is an effort to formalize a system approach to generating fuzzy rules from an input–output data set. The membership functions of the actual data are obtained by the projection of the clusters by the expectation maximization algorithm are shown graphically in Fig 3.Gaussianmembership functions were used for this model.The fuzzy rule base contains a set of fuzzy decision rules. Every fuzzy decision rule consists of a set of fuzzy linguistic terms for expressing values of every attribute in the precondition part; it also contains a set of fuzzy linguistic terms for the single output in the consequence part. Every fuzzy linguistic term is coupled with a fuzzy

membership function. A typical fuzzy rule in a Sugeno fuzzy model hasthe following: format

IF x is A and y is B THEN z = f(x, y),

Where AB are fuzzy sets in the antecedent; Z = f(x, y) is a crisp function in the consequent [8].

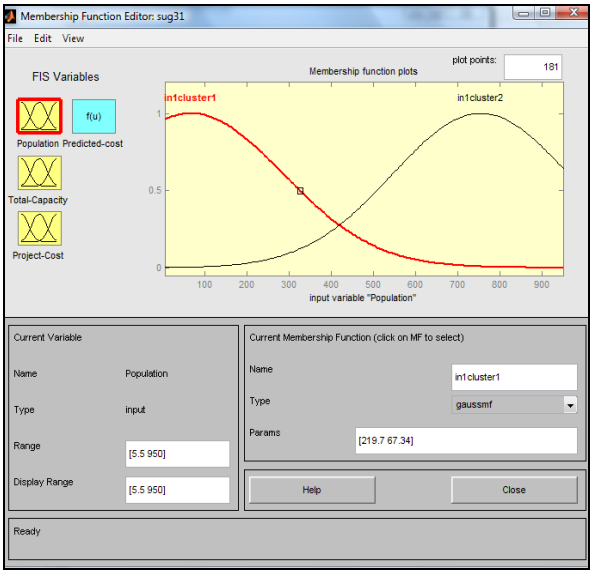

Fig. 3 Membership functions for the population variable.

The Sugeno output membership functionsare either linear or constant. A typical rule in a Sugeno fuzzy model hasthe form

IF Input 1 = x AND Input 2 = y, THEN Output is z = ax + by + c.

Fig4illustrates the fuzzy rules obtained for the developed model. The Sugeno system is a compact and computationally efficient representation which lends itself to the use of adaptivetechniques for constructing fuzzy models. These adaptive techniques can beused to customize the membership functions so that the fuzzy system bestmodels the data. As referred to earlier, 35 data sets are used for training and the remaining 9 data sets are used for testing (see Table 4). The training aids in minimizing the errors depicted in Fig 5. It is worth to note that the error ranges from -38% to 55% for test data. The absolute average error (AAE) is estimated to be 26%, whereas, the root mean square error (RMSE) is estimated to be 5.12%.

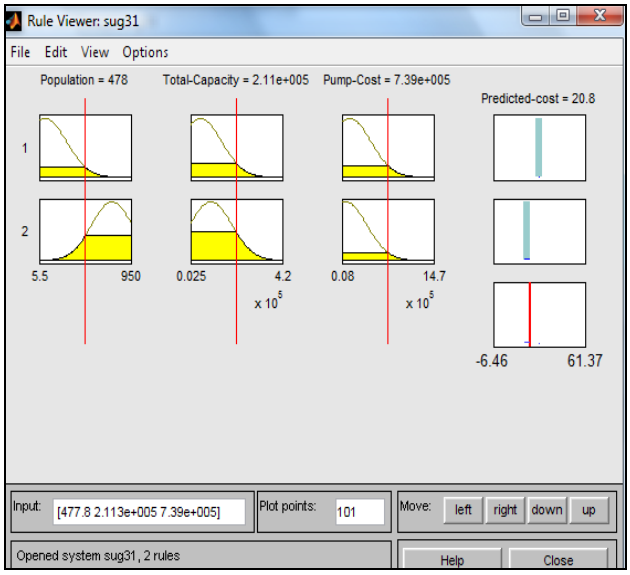

Fig. 4 Generatedfuzzy rules.

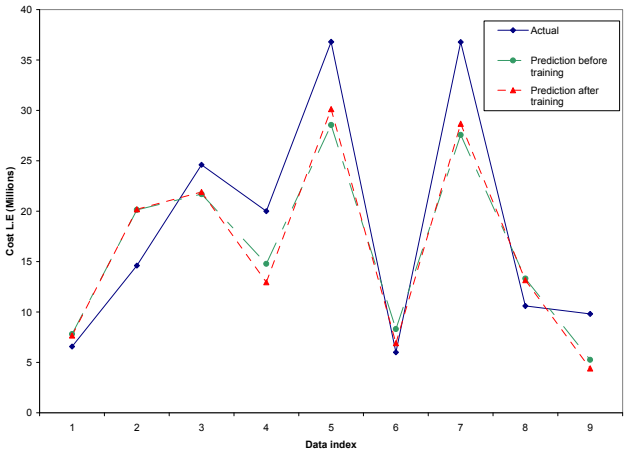

Fig. 5 Actualcost versus predicted cost.

5. SENSITIVITY ANALYSIS

One of the most valuable features for datamining is the graphical presentation of the mindedpatterns or knowledge. For cost estimation, it is convenient to know the most sensitive factorsaffecting the construction cost. In this regards,sensitivity analysis of various influential attributes onoverall construction cost is very useful for valueengineering and best alternative selection. Figure 6 show the sensitivity analyses for the different cost parameters.

Table 3 Estimated error of test data sets.

Population	*Total Capacity (m^3/day)*	*Pump Price (L.E)*	*Cost (Million L.E)*		*Error (%)*
			Actual	*Predicted*	
65	129,600	105,764	6.57	7.67	-17
179	207,360	585,000	14.60	20.20	-38
303	360,000	210,000	24.60	21.91	11
525	77,760	97,750	20.0	12.96	35
750	360,000	782,883	36.80	30.12	18
8	52,000	350,000	6.00	6.91	-15
303	360,000	782,883	36.78	28.68	22
304	181,440	90,000	10.60	13.16	-24
242	13,824	85,000	9.80	4.40	55

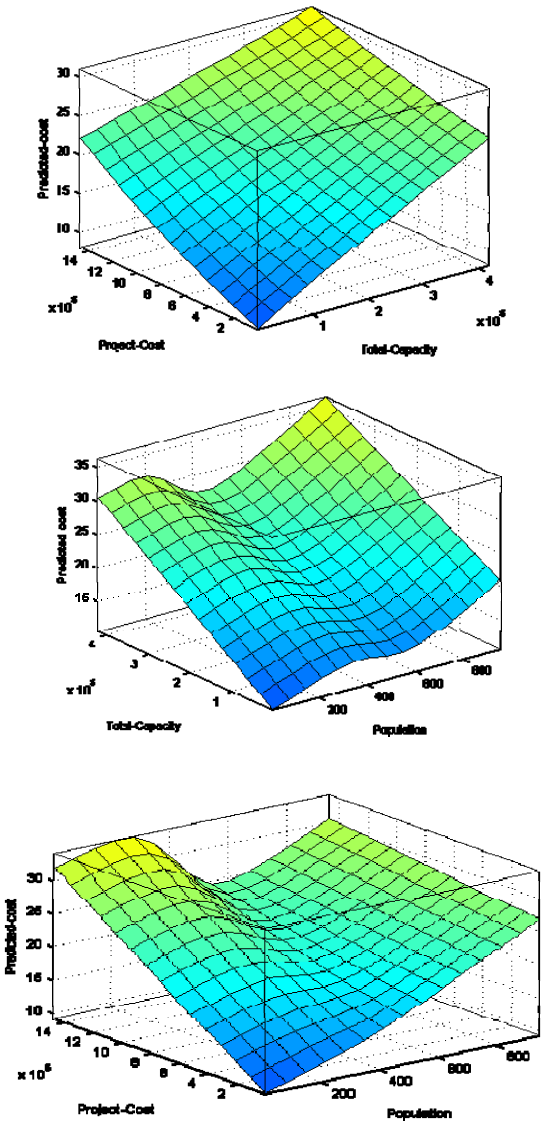

Fig 5 Selectivity analysis of cost parameters.

6. Conclusions

In the preliminary cost estimate of a pump station project, the intent is not to determine the pump type or details of the station structural design, but rather to estimate the cost of the pump station project that is capable of provide the required discharge at given head conditions. The various cost drivers of pump station projects have been identified. The paper provided an overview of a newly developed fuzzy clustering model that can be used as a parametric cost model for pump station projects.The performance of the fuzzy clustering model was tested and the error was found to be within the acceptable limits of parametric cost estimates at early stages. Although the proposed parametric cost model is limited to pump station projects, which are classified as infrastructure projects, the approach can be extended to include other types of construction projects such as residential and industrial buildings.

REFERENCES

[1] Marzouk, M. and Moselhi, O.,"Fuzzy Clustering Model for Estimating Haulers' Travel Time",*Journal of Construction Engineering and Management*, Vol. 130(6),pp. 878-886, 2004.

[2] Zadeh. L.,"Fuzzy Sets",*Information and Control*, Vol. 8, pp. 338–353, 1965.

[3] Jang, J.S.R.,"ANFIS: Adaptive Network Based Fuzzy Inference System". *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 23(3), pp. 665–684, 1993.

[4] Chiu, S.L. 1994. "Fuzzy Model Identification Based on Cluster Estimation".*Journal of Intelligent and Fuzzy Systems*, Vol. 2,pp. 267-278, 1994.

[5] Matlab, *Fuzzy Logic Toolbox User's Guide*. Natick: The Math Works Inc., 2008.

[6] Bezdek, J.C, and Pal, S.K., *Fuzzy Models for Pattern Recognition*. IEEE Publication, New York, NY, 1992.

[7] Yager, R.R., and Filev, D.P.,"Approximate Clustering Via the Mountain Method". *IEEE Transitionson Systems, Man, and Cybernetics*, Vol. 24(8), pp. 1279-1284, 1994.

[8] Sivanandam, S.N., Sumathi, S. and Deepa, S.N., *Introduction to Fuzzy Logic using MATLAB*, Springer-Verlag Berlin Hidelberg, 2010.