4th Construction Specialty Conference
4e Conférence spécialisée sur la construction

Montréal, Québec
May 29 to June 1, 2013 / *29 mai au 1 juin 2013*

An Innovative Technique for Improving Productivity Forecasting Models

Farid Mirahadi[1], Emad Elwakil[2], Tarek Zayed[3]
[1,3]Department of Building, Civil & Environmental Engineering, Concordia University
[2]Department of Civil Engineering and Construction Management, California State University, Northridge

Abstract: Productivity forecasting of construction operations has gained tremendous momentum in both the construction industry and academia. Many of the developed models utilize clustering methods in order to recognize existing hidden patterns among the historical data and improve the modelling performance by mimicking these patterns. It is shown that the improper determination of the number of clusters in such models can noticeably distort their fitness, which is not treated well in the literature. This paper scrutinizes the impacts and benefits of optimizing the number of clusters in productivity forecasting models. To this end, *Subtractive Clustering* is applied to optimize the clustering performed by *K-Means* method. A set of internal indices that consider *Separation* and *Compactness* of the resulted clusters are used to validate the method. The proposed technique is further investigated for a Neural-Network-Driven Fuzzy Reasoning (NNDFR) model developed to simulate a construction operation, in which several qualitative and quantitative factors are considered. Empirical results show that the model performance, in terms of Mean Squared Error (MSE), improves by up to 60 percent when the optimal number of clusters is determined using the presented technique. The developed technique benefits researchers and practitioners to improve the accuracy of modelling in productivity estimation based on a set of construction historical data.

1 Introduction

Fuzzy models and neural network systems have provided an effective tool for addressing uncertainty in decision-making. Uncertainty, as the ineradicable nature of construction projects convinced researchers to approach such intelligent systems. In the past few years, these systems were dramatically applied to develop forecasting models in the construction management area (Kim, An, & Kang, 2004) (Li, 1995) (Boussabaine, 1996) (Bowena & Edwardsa, 1985) (Moselhi, Hegazy, & Fazio, 1992) (Martin Skitmore & Thomas Ng, 2003) (Leu, Chen, & Yang, 2001) (Tah & Carr, 2000) (Chan, Chan, & Yeung, 2009) (Cheng & Ko., 2003). Productivity estimation of construction operations, as a decision criterion in project planning and control, has become an interesting target for forecasting models.

With the increasing volume of historical data provided to these kinds of models, an urgent need for data analysis techniques became apparent. Data clustering can be regarded as the most well-known and prevalent technique in exploratory data analysis. It provides a requisite data-preprocessing step to identify homogeneous patterns among data on which consequent supervised models are built. Furthermore, the wide appeal and usefulness of data clustering techniques have pushed researchers to combine them with other technologies, such as artificial neural network (ANN) and fuzzy reasoning. The recent trend has resulted in a diversity of cluster-aided models. Adaptive Neuro-Fuzzy Inference System (ANFIS) is a well-known example of such models which decomposes the input data space to subspaces and builds linear (or single value) rule consequences in each subspace. In this system, fuzzy clustering can accomplish the partitioning in which each cluster represents one particular behavior of the system.

Although the main purpose of these cluster-aided systems is to provide an accurate and interpretable model, any unconsidered decision about the involved parameters can deteriorate the performance of the model. It is shown that the improper determination of the number of clusters as one of these parameters, can noticeably distort the fitness of such models. This issue, besides emerging tendencies to the application of cluster-based fuzzy and neural network systems led us to scrutinize the impacts and benefits of optimizing the number of clusters in productivity forecasting models. The objectives of the current research can be summarized as follows:

1. Determine the optimum number of clusters to set the initial value for robust clustering techniques
2. Validate the selected number of clusters based on the performance of a cluster-based model besides statistical indices of validation

2 Background

Clustering is considered as one of the most important and frequently used techniques in data analysis (Beringer & Hüllermeier, 2006). Data clustering is the task of organizing a dataset into different groups, such that the objects of the same cluster are more similar to each other compared to those in other clusters. K-Means (MacQueen, 1967; Hugo Steinhaus, 1957; Stuart Lloyd, proposed in 1957 published in 1982) and Fuzzy C-Means (FCM) (developed by Dunn in 1973; improved by Bezdek in 1981) clustering can be considered as the dominant algorithms in both theoretical and practical applications of data mining. K-Means partitions the data in a way that each point belongs only to one cluster. On the contrary, FCM possesses a fuzzy approach for reporting the memberships to different clusters. It allows each data point to belong to more than one cluster. Many of the clustering algorithms are based on knowing the number of clusters beforehand. K-Means and FCM algorithms are of that type and therefore require this initial value before clustering has been accomplished. At this point, the dilemma of determining the best number of clusters emerges.

Mountain clustering proposed by Yager & Filev (1994) was an improvement over earlier methods of clustering. This heuristic technique performs based upon the density of data points. It applies a mountain function (density function) to the customized griding data space in order to find the grid point with the highest density value as the first cluster center. This procedure continues by destructing the effect of each cluster mountain function to find the next greatest density value. While this approach is primarily considered as a stand-alone clustering technique, in another mode it can function as a tool to obtain initial number of clusters for more complex techniques (Yager & Filev, 1994). However, as the problem's dimension grows so does the computations for evaluating all grid points, a problem known as the "curse of dimensionality" (Bellman, 1961). Chiu (1994) presented *Subtractive Clustering* to mitigate this problem. Subtractive Clustering only deems the data points as candidates for cluster centers. In this way, computational complexity and effort grows proportionally to the size of the problem instead of its dimension (Hammouda & Karray, 2000).

The necessity of evaluating the "goodness" of clustering, or comparing between two sets of clusters, created the need for clustering validation. Clustering validation is the act of verifying how well an algorithm can recognize underlying patterns of data. Usually in 2D and 3D data spaces, visualization is used as an empirical way of validation. But in case of the large multidimensional data spaces, deficiency of an effective visualization leads to application of more formal approaches (Kovács, Legány, & Babos, 2005). Internal validation is an approach to evaluate the clustered dataset based on inherent features of the data itself (Halkidi, Batistakis, & Vazirgiannis, 2001). Different internal validity indices have been proposed as an assessment for *compactness* and *separation* among the data distribution (Kovács, Legány, & Babos, 2005). The former examines the members of each cluster to be as close as possible to each other and the latter evaluates clusters themselves to be widely spaced (Michael J. Berry, 1997). The main drawback of internal validation is that supreme values of an internal index do not necessarily conduct us to best information retrieval applications (Manning, Raghavan, & Schutze, 2008).

Clustering techniques extensively have enhanced hybrid intelligent systems. Takagi and Hayashi (1991) utilized clustering techniques to partition the inference rules for a Neural-Network-Driven Fuzzy Reasoning (NNDFR) model. In this model, the data is grouped by K-Means and the number of inference rules is equal to the number of clusters. Although the model is dynamic to the number of clusters, this research does not provide any optimization process to find this initial value. Takagi-Sugeno (TS) type fuzzy models define different regions in the data space each of which represents a linear input-output mapping (Takagi & Sugeno, 1985). It is prevalent in TS models that an automatic method, like fuzzy clustering, is exploited to attain candidates for linear regions (Jantzen, 1998). Elwakil and Zayed (2012) built a Fuzzy Knowledge Based Model in order to estimate the duration of construction operations. This model presents a fuzzification step in which crisp values were fuzzified using an integration of ANN and FCM forming hypersurface membership functions. Although this research presented design principles of fuzzy rule induction, it was not built on the determination of optimal number of clusters that makes a reasonable compromise between complexity and efficiency. The above-cited works show the great significance of clustering techniques in data analysis and modeling.

3 Proposed Framework

In this section, we propose an innovative framework so as to deal with the pre-mentioned gaps in optimization and validation of data clustering. The essence of this framework is finding an initial value for the number of clusters, which fits the natural patterns among the dataset. To this end, subtractive clustering is accomplished and cluster centroids will be identified. The number of centroids is counted and then fed to K-Means which is more expert in data clustering. The reasons that we chose K-Means are: 1) Its compatibility with validity indices checking separation and compactness of the dataset 2) Its accuracy in locating centroids compared to subtractive clustering. In this way, the limitations of each clustering technique are treated by the other one. The proposed procedure could be elaborated in the following steps.

3.1 Step 1: Subtractive Clustering

Subtractive clustering starts by calculating a density measure at each data point through the following function (Chiu, 1994):

$$[1] \quad D_i = \sum_{j=1}^{n} \exp\left(- \frac{\left\| x_i - x_j \right\|^2}{\left(r_a / 2 \right)^2} \right)$$

Where r_a is a positive constant, which represents the cluster radius. Thus, the more neighboring data points, the more density value. The data points beyond this radius will have less influence on density measure. The point with the highest density value D_{c_i} is selected as the first cluster center x_{c_i}. Then, the density measures of all data points are revised through (Chiu, 1994):

$$[2] \quad D_i = D_i - D_{c_i} \exp\left(- \frac{\left\| x_i - x_{c_i} \right\|^2}{\left(r_b / 2 \right)^2} \right)$$

Where r_b is a positive constant representing a neighborhood with measurable reductions in potential. Hence, the nearer data points to x_{c_i} will have more reduction of density measure. r_b is usually described by a coefficient of r_a. This coefficient is named "*Squash Factor"* in Matlab software. After amending the

density functions, the point having the greatest value will represent the next center. This procedure continues until the predefined termination condition is met or an appointed number of centroids is acquired (Hammouda & Karray, 2000).

3.2 Step 2: Combining Subtractive Clustering with K-means

K-Means commence running only when the initial value for the number of clusters K is enacted beforehand. However, nobody can neglect the effect of an improper determination of the initial value, which may result in wrong decisions. This matter can be resolved when the subtractive method acts as an initial value generator for K-Means. Hence, the K-Means obtains the output of the Subtractive method and then starts clustering (Liu, Xiao, Wang, Shi, & Fang, 2003).

3.3 Step 3: K-means

K-means, as one of the simplest unsupervised clustering algorithms, partitions the data space in hard clusters. This iterative algorithm locates centroids via minimizing the following objective function (Matteucci, 2006):

$$[3]\ J = \sum_{j=1}^{k}\sum_{i=1}^{n}\left\|x_i^{(j)} - c_j\right\|^2$$

Where $x_i^{(j)}$ is the i^{th} measured data, c_j is the center of the j^{th} cluster, and ||*|| is a kind of distance between the d-dimensional vector $x_i^{(j)}$ and the d-dimensional vector c_j. The objective function can be minimized through the following steps (Matteucci, 2006):

1. Randomly place K points representing initial centroids in the data space
2. Assign each data point to the cluster that has closest centroid
3. Calculate the revised position of each centroid
4. If the positions of centroids didn't change go to the next step, else to the step 2
5. End.

3.4 Step 4: Validation

In this step, the optimized number of clusters derived from subtractive clustering is validated. For this purpose, the dataset is partitioned to different numbers of clusters and the desired validity indices are calculated for all of the sets. The highest rank values represent better separation and compactness within the clusters. We predict that the result of subtractive clustering would be supported by internal validity indices. The following indices are selected to assess resulted clusters based on internal criteria:

3.4.1 Dunn Index:

The Dunn index (Dunn J. C., 1974) is proposed to identify compact and well-separated clusters. For each partitioning, the index can be defined by the following formula (Dunn J. C., 1973):

$$[4]\ D_{nc} = \min_{1\le i\le nc}\left\{\min_{1\le j\le nc, i\ne j}\left\{\frac{d(i,j)}{\max_{1\le k\le nc} d'(k)}\right\}\right\}$$

Where $\mathrm{d(i,j)}$ represents any distance measure considering the distance between two clusters, such as the distance between the centroids of the clusters; and $\mathrm{d'(k)}$ represents any distance measure considering within-cluster distance, such as the distance between any pair of elements in cluster k. Considering the Dunn's index definition, it may be concluded that higher values of the index are more favorable.

3.4.2 Davies-Bouldin Index:

The Davies-Bouldin index (Davies & Bouldin, 1979) represents the average of similarity between each cluster and its most similar one. It may be calculated via the following simplified formula (Davies & Bouldin, 1979):

$$[5]\quad DB_{nc} = \frac{1}{nc}\sum_{i=1}^{nc} \max_{1\le j\le nc, i\ne j}\left(\frac{s_i + s_j}{d(c_i, c_j)}\right)$$

Where nc is the number of clusters and $\mathrm{c_x}$ is the centroid for cluster x. $\mathrm{s_x}$ represents the average distance between all elements in cluster and centroid $\mathrm{c_x}$, and $\mathrm{d}(*,*)$ indicates the distance between different centroids. Hence, lower values of Davies-Bouldin index are more desirable.

Internal validation of clustered data spaces solely cannot ensure the best information retrieval application. This fact highlights the need for a performance assessment of the model under study, before and after the optimization process. The cluster-based model selected for this research is a NNDFR system, which is briefly described in the following section.

3.5 Neural-Network-Driven Fuzzy Reasoning

NNDFR was the first application of neural networks in self-regulating design of membership functions. This approach forms a nonlinear multi-dimensional membership function, which internally combines all the fuzzy variables. The design procedure of NNDFR could be summarized in three steps: clustering the training dataset, training the membership neural networks (NN_{mem}), training the consequent neural network ($NN_{1\text{-}k}$) of each cluster. In the first step, input data space is partitioned to hard clusters. In this fuzzy system, the number of rules equals the number of clusters. In the second step, NN_{mem} is trained between each input vector and its corresponding cluster assignment vector, as illustrated in figure 1. For example, the supervised part of the learning process for a vector which belongs to cluster 3 is (0,0,1). In the third step, the consequent neural networks are trained between the members of each cluster, previously partitioned in the first step, and their corresponding outputs.

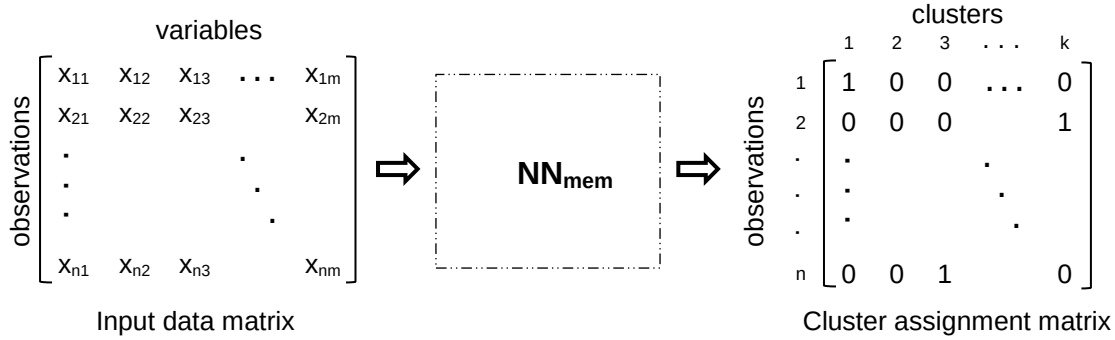

Figure 1: Second step, training the membership neural network.
(m, number of input variables; n, number of observations; k, number of clusters)

Figure 2 shows a holistic diagram of the NNDFR. The NN_{mem} generates the membership functions of the premise (IF) parts of the rules and $NN_{1\text{-}k}$ prepare consequent input-output relationships (THEN parts). This system calculates the final estimated output based on a weighted average of the output of THEN parts, such that the weights are the membership values produced by NN_{mem}.

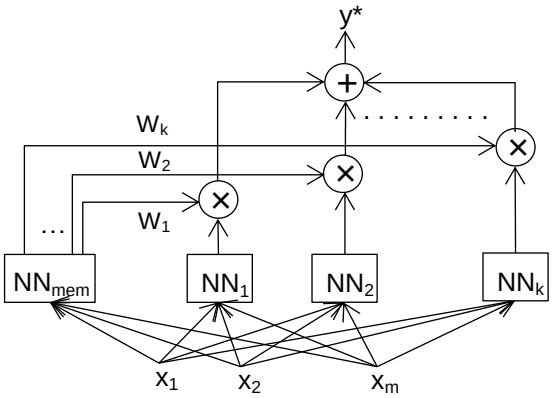

Figure 2: The structure of NNDFR system
(W is the membership value and y* is final estimated output)

The resulting fuzzy model is expressed by the following (Takagi & Hayashi, 1991):

$\text{Rule 1: IF } x = (x_1, x_2, \dots, x_m) \text{ is } C_1, \text{THEN } y_1 = NN_1(x_1, x_2, \dots, x_m)$
$\text{Rule 2: IF } x = (x_1, x_2, \dots, x_m) \text{ is } C_2, \text{THEN } y_2 = NN_2(x_1, x_2, \dots, x_m)$
.
.
.
$\text{Rule k: IF } x = (x_1, x_2, \dots, x_m) \text{ is } C_k, \text{THEN } y_k = NN_k(x_1, x_2, \dots, x_m)$

Where C_{1-k} denote existing clusters. The final estimated value can be delivered through equation 6 (Takagi & Hayashi, 1991).

$$[6] \quad y_i^* = \frac{\sum_{s=1}^{k} W_s(x_i) \,.\, y_s(x_i)}{\sum_{s=1}^{k} W_s(x_i)}$$

Mean Squared Error (MSE) and Average Validity Percent (AVP), as two statistical criteria, can quantify the deference between estimated values and actual ones. Equation 7 and 8 represent AVP, which shows the prediction error. If the AVP value is closer to 100, the model is sound and a value closer to 0 shows that the model is not fitting (Zayed & Halpin, 2005).

$$[7] \quad AIP = \frac{\sum_{i=1}^{n} \left| 1 - \frac{E_i}{C_i} \right|}{n} \times 100$$

$$[8] \quad AVP = 100 - AIP$$

Where; AIP=Average Invalidity Percent; AVP=Average Validity Percent; E_i=Estimated Values; C_i=Estimated Values.

4 Case Study and Framework Implementation

This case study analyses the data gathered from construction processes of Engineering, Computer Science and Visual Arts complex of Concordia University. The dataset consists of several quantitative and qualitative variables affecting concrete pouring operations and their corresponding daily productivity. Nine factors are considered: Temperature, Humidity, Precipitation, Wind Speed, Gang Size, Labor Percentage, Work Type, Floor Level and Work Method. Table 1 presents a small sample of these 131 record points.

In the first step, subtractive clustering generates the initial value for K-Means algorithm. To perform the computations, the subclust() function in MATLAB is applied. Chiu recommended the values 0.5 and 1.25 for the cluster radius and the squash factor, respectively (Ren, Baron, & Balazinski, 2012). These values are set by default in MATLAB. The other two parameters are Accept Ratio and Reject Ratio. The former sets a fraction of the potential of the first cluster center as a minimum for acceptance of the next center. The latter sets a fraction of the potential of the first cluster, below which data points are rejected for being a center. Regarding the fact that we are searching for proportionally dense clusters, we adopt the number 0.7 for both these parameters. The results shows that 3 numbers of clusters with the centers reported in Table 2 are appropriate.

During the second and third steps, the data is repartitioned via K-Means into 2 to 10 clusters. Of course, our target number is three and other numbers only provide a framework for comparison. This procedure lets us validate the partitioning with Davies-Bouldin and Dunn indices, which are mainly created to examine hard clusters.

Table 1: A sample of conceret pouring data

Temperature (ºC)	Humidity (%)	Precipitation	Wind Speed (km/h)	Gang Size (workers)	Labor Percentage (%)	Work Type	Floor Level	Work Method	Daily Productivity (m^3/man/hr.)
-8	87	2	14.2	22	36	1	3	1	1.27
-6	37	0	19.9	19	33	1	8	2	1.23
25	77	0	24	20	30	1	14	1	1.65

- Precipitation: *No precipitation = 0, Light rain = 1, Rain = 2, and Snow = 3*
- Labor Percentage: *The percentage of the labor (non-skilled workers) in the gang*
- Work Type: *Reported in terms of activity type: Slabs= 1 and Walls = 2*
- Work Method: *Crane and bucket arrangement=1 and Pumping=2*

Table 2: Cluster centers generated by Subtractive Clustering

Centers	Temperature (ºC)	Humidity (%)	Precipitation	Wind Speed (km/h)	Gang Size (workers)	Labor Percentage (%)	Work Type	Floor Level	Work Method
C1	3	79	0	13	11	37	1	12	2
C2	21	71	0	10	21	33	1	13	2
C3	5.5	46	0	12	19	33	2	10	1

In the fourth step, the validity indices are calculated and then plotted as is shown in figure 3. As we expected, three numbers of clusters produce the highest value for Davies-Bouldin and the lowest for Dunn index. In other words, the desired separation and compactness among the dataset is attained through 3 clusters.

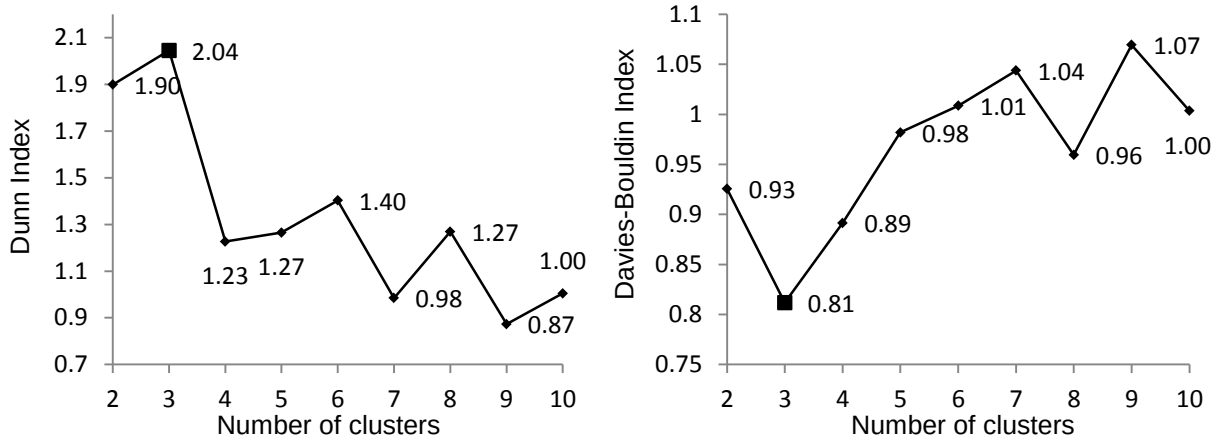

Figure 3: Calculated values for internal validity indices

At the end, the operation is modeled by means of a NNDFR system to clarify how this optimization technique can affect the model performance. We randomly divide our collection of 131 data points to two fractions of 118 and 13. The model utilizes the bigger fraction to be trained. The model performance would be then reflected through MSE of the targets and the simulated outputs of the testing sample. Eventually, we develop different models with different number of clusters from 2 to 10 which enables us to see the effect of any deviation from optimum cluster numbers. Table 3 tabulates MSE Train, MSE Test, AVP and AIP of all developed models versus their number of clusters. As seen in figure 4, the model with 3 clusters has the best performance in terms of the testing MSE. The MSE and AVP of the three-cluster model are improved by 60 and 2.5 percent respectively, over the four-cluster model, which owns the second rank. Thus, the empirical result supports the proposed procedure of detecting optimum number of clusters when it is embedded in a cluster-based model. Table 4 reports the estimated outputs of testing sample modeled by three-cluster NNDFR against the actual values of daily productivity.

Table 3: NNDFR result vs. number of clusters

Number of Clusters	MSE Train	MSE Test	AVP	AIP
2	0.0200	0.0636	86.76	13.24
3	0.0091	0.0218	92.59	7.41
4	0.0169	0.0544	90.07	9.93
5	0.0297	0.0771	86.34	13.66
6	0.0335	0.1021	83.77	16.23
7	0.0212	0.1193	83.60	16.40
8	0.0373	0.1598	80.20	19.80
9	0.0318	0.0686	86.07	13.93
10	0.0935	0.1905	78.45	21.55

Table 4: Result of the three-cluster NNDFR model

Daily Productivity (m3/man/hr.)			Daily Productivity (m3/man/hr.)		
Index	Actual	Estimated	Index	Actual	Estimated
1	1.550	1.550	8	1.80	2.044094
2	1.370	1.370	9	1.88	2.105987
3	1.250	1.330	10	2.02	2.272661
4	1.490	1.570	11	1.97	2.016129
5	1.210	1.275	12	1.73	1.902600
6	1.340	1.525	13	1.23	1.295237
7	1.510	1.656			

5 Conclusion

The current research presents a framework to investigate the impacts and benefits of optimizing the number of clusters in cluster-based models. Regarding the sensitivity of prevalent clustering techniques to initial number of clusters, subtractive clustering is applied to generate this initial value in advance. In this case, subtractive clustering is used to compute the number of clusters rather than partition the data.

The optimized number of clusters is then validated through internal validation indices which statistically examine the results based on the inherent features of the data. The developed framework is further validated and verified using a case study implemented to a cluster-based model. Our data, which is a set of variables affecting daily productivity of concrete pouring process, is optimally clustered via proposed framework. It is then fed to an NNDFR model as our choice of cluster-based models. Developing different models against different number of clusters reveals that model's prediction improves by 60 percent in terms of MSE using this innovative technique. The case study is provided as a response to the need of real-world validation besides statistical and model-free approaches. The developed research helps researchers and practitioners by providing them with an effective way of data partitioning.

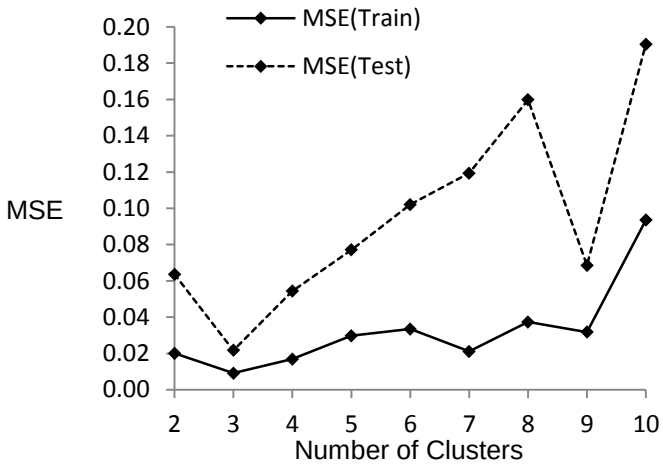

Figure 4: MSE vs. number of clusters

References

Bellman, R. E. 1961. *Adaptive control processes: a guided tour*, Princeton University Press, Princeton, NJ, USA.

Beringer, J. and Hüllermeier, E. 2006. Online clustering of parallel data streams. *Data & Knowledge Engineering, 58*(2): 180-204.

Berry, M. J. and Linoff, G. 1997. *Data Mining Techniques: For Marketing, Sales, and Customer Support.* John Wiley & Sons, New York, NY, USA.

Bezdek, J. C. 1981. *Pattern recognition with fuzzy objective function algorithms,* Plenum Press, New York, NY, USA.

Boussabaine, A. H. 1996. The use of artificial neural networks in construction management: a review. *Construction Management and Economics, 14*(5): 427-436.

Bowena, P. A. and Edwardsa, P. J. 1985. Cost modelling and price forecasting: practice and theory in perspective. *Construction Management and Economics, 3*(3): 199-215.

Chan, A. P., Chan, D. W. and Yeung, J. F. 2009. Overview of the application of "fuzzy techniques" in construction management research. *Journal of Construction Engineering and Management, 135*(11): 1241-1252.

Cheng, M. Y. and Ko., C. H. 2003. Object-oriented evolutionary fuzzy neural inference system for construction management. *Journal of Construction Engineering and Management, 129*(4): 461-469.

Chiu, S. L. 1994. Fuzzy Model Identification Based on Cluster Estimation. *Journal of intelligent and Fuzzy systems, 2*(3): 267-278.

Davies, D. L. and Bouldin, D. W. 1979. A Cluster Separation Measure. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*(2): 224-227.

Dunn, J. C. 1973. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *Journal of Cybernetics, 3*(3): 32–57.

Dunn, J. C. 1974. Well-separated clusters and optimal fuzzy partitions. *Journal of cybernetics, 4*(1): 95-104.

ELWAKIL, E. and Zayed, T. 2012. Data Management for Construction Processes Using Fuzzy Approach. *Construction Research Congress 2012@ sConstruction Challenges in a Flat World,* ASCE, Purdue University, West Lafayette, IN, USA, 1: 1222-1231.

Halkidi, M., Batistakis, Y. and Vazirgiannis, M. 2001. On Clustering Validation Techniques. *Journal of Intelligent Information Systems, 17*(2): 107-145.

Hammouda, K. and Karray, F. 2000. A Comparative Study of Data Clustering Techniques. *Tools of intelligent systems design, In Course Project, SYDE 625*, Department of Systems Design Engineering, University of Waterloo, Waterloo, Ontario, Canada

Jantzen, J. 1998. Neurofuzzy modelling. *Tech report no 98-H-874*, Technical University of Denmark, Department of Automation, Lyngby, Denmark.

Ketchen, D. J. 1996. The application of cluster analysis in Strategic Management Research: An analysis and critique. *Strategic management journal, 17*(6): 441-458.

Kim, G. H., An, S. H. and Kang, K. I. 2004. Comparison of construction cost estimating models based on regression analysis, neural networks, and case-based reasoning. *Building and Environment, 39*(10): 1235–1242.

Kovács, F., Legány, C. and Babos, A. 2005. Cluster Validity Measurement Techniques. *6th International Symposium of Hungarian Researchers on Computational Intelligence,* Budapest, Hungary.

Leu, S. S., Chen, A. T. and Yang, C. H. 2001. A GA-based fuzzy optimal model for construction time–cost trade-off. *International Journal of Project Management, 19*(1): 47-58.

Li, H. 1995. Neural networks for construction cost estimation. *Building Research & Information , 23*(5): 279-284.

Liu, W. Y., Xiao, C. J., Wang, B. W., Shi, Y. and Fang, S. F. 2003. Study on combining subtractive clustering with fuzzy c-means clustering. *2003 International Conference on Machine Learning and Cybernetics,* IEEE, Xi'an, Shaanxi, China*, 5*:2659-2662.

Lloyd, S. 1982. Least squares quantization in PCM. *IEEE Transactions on Information Theory, 28*(2): 129–137.

MacQueen, J. 1967. Some Methods for classification and Analysis of Multivariate Observations. *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability,* University of California Press, Berkeley, CA, USA, *1*: 281–297.

Manning, C. D., Raghavan, P. and Schutze, H. 2008. *Introduction to Information Retrieval.* Cambridge University Press Cambridge, Cambridge, United Kingdom.

Martin Skitmore, R. and Thomas Ng, S. 2003. Forecast models for actual construction time and cost. *Building and Environment, 38*(8): 1075–1083.

Mathworks. Subtractive Clustering. Retrieved Dec 12th, 2012, from http://www.mathworks.com/help/fuzzy/subclust.html

Matteucci, M. 2006. *A tutorial on clustering algorithms.* Politecnico di Milano, Dipartimento di Elettronica e Informazione, Como, Como, Italy. Retrieved from http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/

Moselhi, O., Hegazy, T. and Fazio, P. 1992. Potential applications of neural networks in construction. *Canadian Journal of Civil Engineering, 19*(3): 521-529.

Ren, Q., Baron, L. and Balazinski, M. 2012. Fuzzy identification of cutting acoustic emission with extended subtractive cluster analysis. *Nonlinear Dynamics, 64*(7): 2599-2608.

Steinhaus, H. 1957. Sur la division des corps matériels en parties. *Bulletin L'Académie Polonaise des Science, 4*(12): 801–804.

Tah, J. H. and Carr, V. 2000. A proposal for construction project risk assessment using fuzzy logic. *Construction Management & Economics, 18*(4): 491-500.

Takagi, H. and Hayashi, I. 1991. NN-driven fuzzy reasoning. *International Journal of Approximate Reasoning, 5*(3): 191–212.

Takagi, T. and Sugeno, M. 1985. Fuzzy identification of systems and its applications to modeling and control. *IEEE Transactions on Systems, Man and Cybernetics, 15*(1): 116-132.

Yager, R. R. and Filev, D. P. 1994. Approximate Clustering Via the Mountain Method. *IEEE Transactions on Systems, Man, and Cybernetics, 24*(8): 1279-1284.

Zayed, T. M. and Halpin, D. W. 2005. Pile construction productivity assessment. *Journal of construction engineering and management, 131*(6): 705-714.