Proceedings of the Korean Nuclear Society Autumn Meeting
Seoul, Korea, October 1998

The Classification of CHF Data Using Principal Component Analysis (PCA) and Hierarchical Clustering Method

B.S. Jun, E.J. Park, S.G. Yang, H.J. Kim, K.H. Kim, J.R. Park
Korea Nuclear Fuel Co., Ltd.
P.O. Box 14, Yusong, Taejon, Korea, 305-600,

ABSTRACT

A new approach for classification of critical heat flux (CHF) data based on the principal component analysis (PCA) and the hierarchical clustering method is suggested. The PCA is used to describe the multivariate structure of CHF data and the characteristics of resulting CHF structures are identified. The agglomerative hierarchical clustering is performed to see the proximity of the CHF data with the obtained information. Clusters are represented by a dendrogram and grouped into three meaningful categories. Katto's CHF-regime map is applied to the resulted CHF group for a better understanding of the physical meaning of the clusters. The combination between principal component analysis and agglomerative hierarchical clustering method provides a meaningful grouping of CHF data which can be used for other applications.

I. Introduction

The critical heat flux (CHF) can be influenced by many independent variables such as the inlet flow rate, the inlet temperature, the system pressure, the tube internal diameter, the tube length, and so on. Because the influence of these variables on CHF mechanism is so complex and obscure, there are lots of attempts to understand the effect of system parameters on CHF, especially using experimental data or empirical correlation [1,2]. Recently, an advanced information processing technique such as artificial neural network or entropy minimax principle used to provide the possibility of valuable alternative for estimating CHF [3,4].

To understand the exact CHF mechanism and parametric trends, one must recognize the structure of CHF data and its relationship that may exist in each variables. Therefore, clustering or classification of CHF data based on its internal structure is necessary for properly estimating CHF phenomena.

In this paper, as a multivariate data analysis tool which provides information for organizing a large set of data, principal component analysis (PCA) is used to find out the multivariate structure of CHF data. PCA is a multivariate technique in which a number of related variables are transformed to set of uncorrelated variables [5,6]. The principal component method has been widely applied for many areas, especially for quality control, forestry, environmental science, and so on [7,8,9]. It can be applied to thermal-hydraulic area such as CHF data classification.

After obtaining information through the PCA of CHF data, cluster analysis using the agglomerative hierarchical clustering [10,11,12] which proceeds by a series of successive fusions of the individuals into groups is performed. To measure the proximity of CHF data, the Euclidean distance (ED) which is the distance between two points in CHF data set is computed and used. Clustering is represented by a two-dimensional diagram known as a dendrogram [10].

For application of the combination of the PCA and agglomerative hierarchical clustering method to CHF phenomena, vertical round tube CHF data for water has been adopted from the reference [13]. For convenience, CHF data at 1000 psia pressure are selected for evaluation of the clustering. After clustering is made, Katto's CHF-regime map [14] is applied to the obtained clusters for a better understanding of the physical meaning of the clusters.

II. Principal Component Analysis and Hierarchical Clustering Method

The essential feature of the Principal Component Analysis (PCA) is the transformation of the original variables x_1, x_2, ..., x_p into a new set of variables y_1, y_2, ..., y_k [5]. The new variables are linear transformations of the original ones with the characteristics that y_1, y_2, ..., y_k are uncorrelated with each other and they account for decreasing portions of the variance of the original variables [5]. The coefficients defining the linear transformations from x_1, x_2, ..., x_p to y_1, y_2, ..., y_k, are found from the eigenvectors of the correlation matrix of the original variables.

In the paper, data matrix (X) consists of the number of CHF data as p=1,2, ..., 134 variables (columns) and non-dimensional parameters as k=1,2,3,4 objects (rows). The objects are converted from the independent variables into 4 non-dimensional parameters ($\frac{\sigma\rho_f}{G^2 l}, \frac{l}{d}, \frac{\Delta H_i}{H_{fg}}, \frac{q_c}{GH_{fg}}$) [14] to make the highly correlated relationship within the correlation matrix. With the modified object data matrix, the correlation matrix is calculated and principal component scores(y_1, y_2, y_3, y_4) are determined for each data point considering the eigenvalues and associated eigenvectors. The principal component equation is as follows :

$$y = W' D^{-1}(x - \bar{x}) \tag{1}$$

where, y is principal component score,
W′ is transpose of eigenvector
D is diagonal matrix of standard deviation.

Table 1 shows the result of PCA operation with the data matrix. For visual inspection of the structure of CHF data, principal component scores can be plotted as in Figure 1.

As a result of PCA operation, 4 principal component score matrices are produced. To classify the uncorrelated variables of CHF data within the matrices, an agglomerative hierarchical clustering method which produces a series of partitions of the data is introduced. The starting point is to find the nearest pair of distinct clusters as a function of distance or similarity. For constructing distance measure, Euclidean distance is used. The Euclidean

distance (ED) is

$$d_{ij} = \sqrt{\sum_{i=1}^{p}\sum_{i=1}^{n}(x_i - x_j)^2}\ . \quad (2)$$

In this case, data consist of p(=134) x n(=4) matrix and d_{ij} means the distance between ith group and jth group. After computing all the distances in data matrix, the distance matrix can be constructed. The next step is to select the entry with the smallest distance and form a two-member cluster. And then, another closest distance is searched. This procedure continues until the proper stage is produced. The corresponding dendrogram is shown in Figure 2. The dendrogram indicating those three clusters is shown in Figure 2. The agglomerative hierarchical clustering algorithm is programmed for evaluation of the analysis.

III. Results and Discussion

Table 1 shows the result of PCA operation. In Table 1, the eigenvalues and eigenvectors corresponding to 4 non-dimensional parameter's correlation matrix are presented. Table 1 shows that $\frac{\sigma\rho_f}{G^2 l}$ and $\frac{q_c}{GH_{fg}}$ are highly correlated each other among the variables. The first two components account for 90% of the variance in the non-dimensional parameters. From the visual inspection of the 4 principal component scores in Figure 1, some interesting aspects are inferred from involving the component number. For principal component score y1, there is some strong relationship with the mass velocity(G). As the y1 is increased, mass velocity is decreased as shown in y1-y3 plot of Figure 1. It is judged that the non-dimensional parameter $\frac{q_c}{GH_{fg}}$ has an dominant effect on component score y1. For y2, some tendency can be extracted from y1-y2 plot of Figure 1. The inlet subcooling(△Tin) is increased as y2 increases, which means that y2 is closely related with non-dimensional parameter $\frac{\Delta H_i}{H_{fg}}$. Non-dimensional parameters $\frac{\sigma\rho_f}{G^2 l}$ and $\frac{l}{d}$ have an influence on y3 as shown in y2-y3 plot of Figure 1. And the last component y4 seems to have a major dependency on $\frac{q_c}{GH_{fg}}$ as shown in y3-y4 plot of Figure 1. Based on the above inspection of the plot, the structure of CHF data can be recognized in some degree. Table 2 summarizes the characteristics of CHF data which is used in this analysis

With 4 component scores obtained, the simplest agglomerative hierarchical single linkage clustering is performed. As a consequence of analysis, the smallest distance is 0.00079 and it turns out that two CHF points have a similar characteristics. The dendrogram of Figure 2 shows the data into three main clusters according to the distance. The first cluster (C1) consists of 13 CHF points and characterize by very low mass velocity comparing with other data. The second cluster (C2) contains 8 CHF data and its $\frac{l}{d}$ is lower than that of the other data. The third cluster (C3) is made of 4 different test sections and can be divided into

sub-clusters.

For a better understanding of the physical meaning of the clusters, Katto's CHF-regime map is applied to the obtained clusters. Katto [14] made the 4 characteristic CHF regimes which can be classified as L, N, H and HP-regimes and also defined CHF mechanism followed by CHF regimes. Figure 3 shows that the application of Katto's CHF regime map to the obtained three clusters. Cluster 2 and cluster 3 belong to N-regime and cluster 1 is in H-regime. According to CHF mechanism by Katto, cluster 2 and cluster 3 correspond to the DNB type CHF data and cluster 1 is LFD type. Even though cluster 2 contains two different test geometries, the combination of PCA and agglomerative hierarchical clustering gives one meaningful cluster.

IV. Conclusions and Recommendations

The combination of the PCA and agglomerative hierarchical clustering method has been applied to CHF data and gives meaningful classification of CHF data. The PCA operation for handling CHF data seems to be a fruitful approach to understand the structure of CHF data, and the agglomerative hierarchical clustering method provides reasonable clusters. For further work, the followings are recommended: (a) Expand the raw CHF data and find out more correlated non-dimensional parameters which show the structure of CHF data clearly. (b) The parametric trend and correlation scheme should be developed to explain the obtained classification properly. (c) These methods can be used in other thermal-hydraulic areas for example, flow regime classification or flooding data classification.

References

1. J.G. Collier and J.R. Thome, Convective Boiling and Condensation (3rd Ed.), Clarendon Press, Oxford (1994)
2. G.F. Hewitt, Burnout in Handbook of Multiphase Systems, Vol. 1 (Ed. by G. Hetsroni), pp 6.66-6.141, Hemisphere Publishing Corporation, Washington, D.C., USA (1982)
3. S.K. Moon, W.P. Baek, and S.H. Chang, Parametric Trends Analysis of the Critical Heat Flux based on Artificial Neural Network, Nucl. Eng. Des., 163, 29-49, (1996)
4. S.H. Chang et al., Classification of Critical Heat Flux Patterns using Entropy Minimax Principle, Int. comm. Heat Mass Transfer, 18, 185-193, (1991)
5. J.E. Jackson, A User's Guide to Principal Components, A Wiley-Interscience Publication, (1991)
6. P.E. Green and J.D. Carroll, Analyzing Multivariate Data, The Dryden Press, (1978)
7. T.L. Coleman et al., Evaluation of Remote Sensing Methods used to Differentiate Forested Wetlands, SPIE Vol. 1819, (1992)
8. H. Lamparczyk et al, Classification of Marine Environment Samples Based on Chromatographic Analysis of Hydrocarbons and Principal Component Analysis, Oil & Chemical Pollution 6, 177-193, (1990)
9. S.S. Ismail, Pattern Recognition Analysis for Characterization of Coal and Ash Samples, J. Radioanalytical & Chemistry, Articles, Vol. 169, No. 2, 381-395, (1993)
10. B.S. Everitt, Cluster Analysis, (3rd Ed.), Halsted Press, (1993)
11. H. Spath, Cluster Analysis Algorithms for Data Reduction and Classification of Objects, Ellis Horwood Limited, (1980)
12. A.D. Gordon, Classification Methods for the Exploratory Analysis of Multivariate Data, Chapman & Hall, (1981)
13. R.A. DeBortoli et al., Forced-Convection Heat Transfer Burnout Studies for Water in Rectangular Channels and Round Tubes at Pressures above 500 psia, WAPD-188, (1958)
14, Y. Katto, Critical Heat Flux of Forced Convection Boiling in Uniformly Heated Vertical Tubes (Correlation of CHF in HP-regime and Determination of CHF Regime map), Int. J. Heat Mass Transfer 23, 1573-1580 (1980)

Table 1. The result of the PCA operation with the data matrix

Correlation Matrix

	$\frac{\sigma\rho_f}{G^2 l}$	$\frac{l}{d}$	$\frac{\Delta H_i}{H_{fg}}$	$\frac{q_c}{GH_{fg}}$
$\frac{\sigma\rho_f}{G^2 l}$	1.0000			
$\frac{l}{d}$	-0.6482	1.0000		
$\frac{\Delta H_i}{H_{fg}}$	0.4096	-0.1111	1.0000	
$\frac{q_c}{GH_{fg}}$	0.8381	-0.6952	0.5559	1.0000

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
PRIN 1	2.69891	1.79708	0.674728	0.67473
PRIN 2	0.90183	0.62138	0.225458	0.90019
PRIN 3	0.28046	0.16165	0.070114	0.97030
PRIN 4	0.11880	.	0.029700	1.00000

Eigenvectors

	PRIN 1	PRIN 2	PRIN 3	PRIN 4
$\frac{\sigma\rho_f}{G^2 l}$	0.554722	-0.094556	0.704596	-0.432304
$\frac{l}{d}$	-0.473907	0.552465	0.623003	0.286464
$\frac{\Delta H_i}{H_{fg}}$	0.355836	0.827459	-0.337153	-0.273899
$\frac{q_c}{GH_{fg}}$	0.584017	0.033954	0.041714	0.809957

Table 2. Summary of the characteristics of CHF data used in analysis

	△Tin (°F)		G (Mlb/hr-ft^2)		q_c(MBtu/hr-ft^2)		L/D	No. Data
	Range	Avg.	Range	Avg.	Range	Avg.		
T01	132-222	192.5	0.97-4.54	2.13	2.53-4.21	2.77	21	4
T04	52-446	233.8	0.03-0.07	0.04	0.08-0.35	0.16	52	17
T05	0-140	40.6	1.55-4.96	2.55	0.74-1.93	1.30	64.5	44
T08	43.6-133.5	85.6	1.23-4.71	3.35	1.36-2.19	1.78	67	15
T10	0-361	107.3	0.415-1.64	0.72	0.3-1.74	0.93	76	41
T12	85-434	247.7	0.925-7.79	5.19	0.68-2.74	1.83	109	13

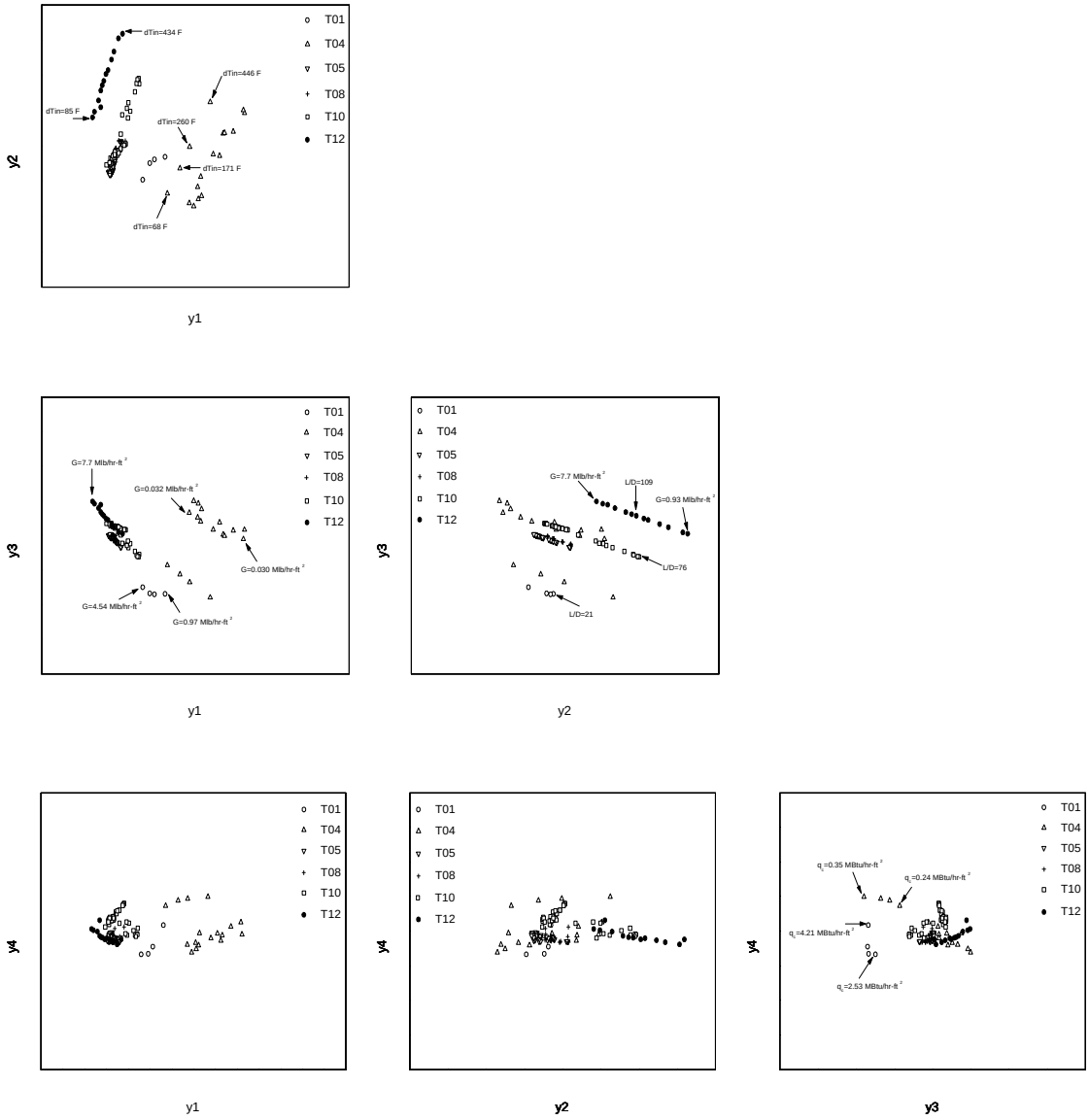

Figure 1. Lower triangular plot of first four principal component scores

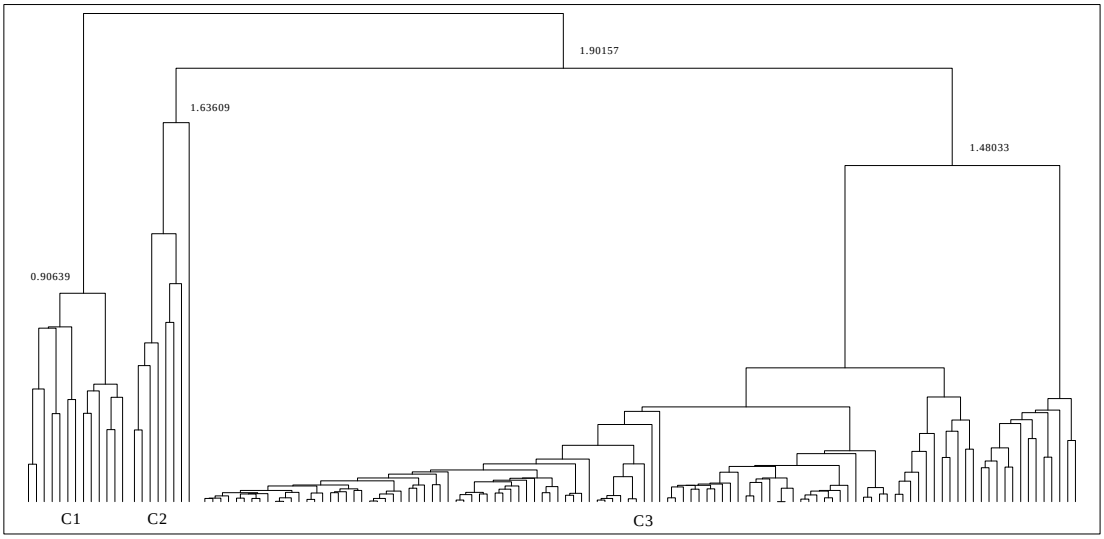

Figure 2. The dendrogram of resulted three clusters

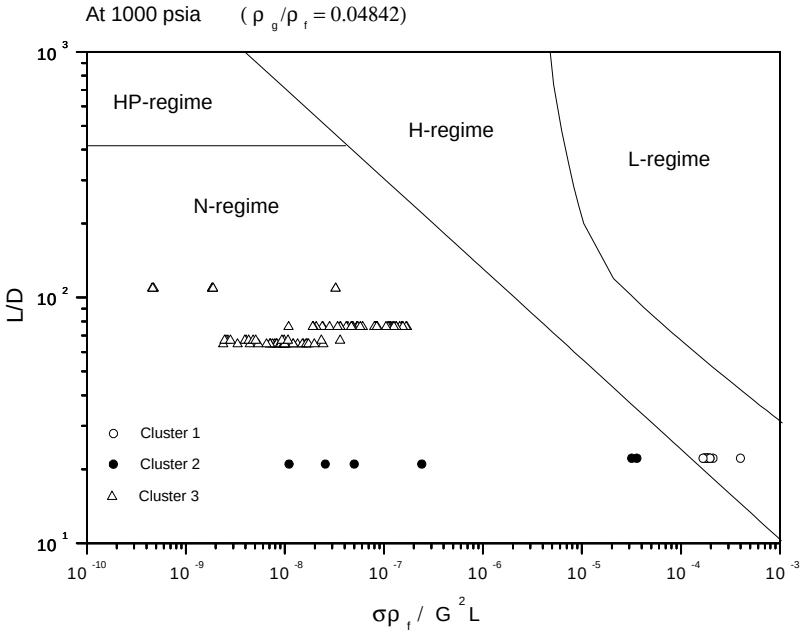

Figure 3. Application of Katto's CHF regime map to obtained three clusters