Cluster Analysis and Workload Classification

Kimmo E. E. Raatikainen[†]

February 25, 1992

Abstract

Clustering techniques are widely recommended tools for workload classification. The k-means algorithm is widely accepted as the "standard" technique of detecting workload classes automatically from measurement data. This paper examines validity of the obtained workload classes, when the current system and workload is analyzed by a queueing network model and mean value analysis. Our results, based on one week's accounting data of a VAX 8600, indicate that the results of queueing network analysis are not stable when the classes of workload are constructed through the *k-means* algorithm. Therefore, we cannot recommended that the most widely used clustering technique should be used in any workload characterization study without careful validation.

1. Introduction

Cluster analysis has been the most popular statistical technique for automatically dividing the workload into workload classes. Since the late 1970's, a number of authors have published numerous papers on clustering in workload characterization: [AgMB76], [AgMo77], [AgMo78], [Arti78], [Arti86], [CaFe86], [Schr78], [Sera81], [Wigh81], among many others.

The ultimate objective of any clustering technique is to find "natural groupings" of observations such that the observations in the same group are "similar" and the observations in different groups are "dissimilar". In workload classification the clustering techniques are claimed to eliminate the dependency of classification results on the analyst's skills and subjective judgements. Unfortunately this is not true, in general. While the older techniques, such as introduced in [Ferr72], required expert knowledge of the examined system and common sense, the clustering techniques also require statistical knowledge and experience. The expert knowledge about clustering and its application is necessary to specify adequately the highly vague terms of *natural groupings, similar,* and *dissimilar*.

[†]Current address: Department of Computer Science, University of Helsinki, Teollisuuskatu 23, SF-00510 Helsinki, FINLAND; Kimmo.Raatikainen@Helsinki.FI

The extensive literature on clustering makes it hard to learn the essentials. During the last twenty years several textbooks – [Ande73], [Ever80], [Hart75], [Spät80], among many others – and hundreds of papers have been published. Review articles including [BlAl78], [DuJa80], and [Kana74] give hundreds of references and claim that their bibliographies are far from complete. Most papers, especially in engineering applications, introduce new algorithms and justify the results by demonstrating that the algorithms "work". On the contrary, only a few articles that help to gain understanding are available. Important aspects of clustering have been considered in [Corm71], [DuJa76], [DuJa79], and [Ever79]. These excellent papers, however, require some preliminary knowledge about clustering. Suitable introductions can be found, for example, in [ChCo80, Ch. 11] and [Ever80], while a comprehensive treatment is presented in [Ande73].

We examine the validity of obtained clusterings through analyzing the stability of mean response times in a multiclass queueing network model when the classes are provided by automatic clustering. We use one variant of the *k-means* algorithm, SAS PROC FASTCLUS [SAS85, Ch. 18] and a postprocessor described in [Arti86], to perform the classification. Two different scalings of variables are considered: robust standardization and logarithmic transform. The examined data is daily prime-time accounting files of one week from a VAX8600 running VAX/VMS. The data consists of 29,124 interactive tasks characterized by the cpu time, direct io-count, and buffered io-count.

2. Fundamental Issues in Clustering

The use of clustering techniques introduces several statistical issues. In workload classification most attention is paid to scaling of variables and to treatment of outliers. Although these two issues are important, they are not the fundamental issues. In fact, they are subtopics of the fundamental issue; the concept of similarity: What does similarity mean? How should similarity be measured? Other major issues include the choice of clustering technique and the assessment of stability and validity of detected clusters.

In addition to the concept of similarity, the object and objectives of the study determine the clustering techniques that are suitable for the situation. Each technique has its own concept of similarity. If the technique has a different concept than the study requires, the results of the study are at least uncertain, but usually misleading. Therefore, the analyst applying clustering must know which techniques have a compatible concept of similarity and how the measure of similarity can be tailored to the situation.

In this section we discuss the concept of similarity, principles of evaluating the results of clustering, and choice of clustering technique.

2.1 What Does Similarity Mean in Workload Classification?

The concept of similarity depends on several different factors related to the object and objectives of the study. How will the results of clustering be used? What is the phenomenon being studied? How much theory is known about the phenomenon? The answers to these and other similar questions related to the real-world problem determine what similarity means and how it should be measured.

Before we can specify the meaning of similarity we must specify our application domain of clustering. This study considers the performance of a computer system analyzed by a multiclass queueing network model. The clustering technique "automatically" constructs workload classes from the available resource-usage data of individual tasks.

Primarily we are interested in the means of the system response times for the classes detected by automatic clustering. Since the mean value of an index having a great variability is not informative, the similarity should mean that the service demands of tasks in the same workload class are close to each other. This implies that the Euclidean distance of properly scaled or transformed resource usages is an appropriate measure of similarity in clustering.

2.2 Stability and Validity of Clusters

Stability and validity of clusters are widely neglected issues in workload classification. We have found only a single and elementary technique, so-called *reassignment* [AgMo77, BCMS82], used to assess stability, reliability, and validity of obtained workload classifications. However, the reassignment only examines how completely the clusters detected in one set of data cover another set of data. Clearly, stability and validity are more complex issues than pure coverage.

Potential solutions to the problem of validating clusters are usually hampered by disagreement on the meaning of validity. Since the detected clusters are not an "end-product" but a tool in further analysis, the assessment of validity of clusters is insufficient, if only the results of clustering are considered. Therefore, validity of clusters should be assessed from the results of analysis applying the clusters. In this study we assess validity of clusters by analyzing how stable the results of analysis are. Actually, stability of results is only a necessary but not a sufficient condition for validity.

Although some attempts have been made to develop mathematical and statistical theories of cluster validity (see e.g. [DuJa79] and [DuJa80]), we have not found any measure of validity suitable to our situation. Our approach to examine stability of the results is based on resampling and graphical display.

2.3 Selecting the Clustering Technique

The real-world phenomenon being studied and the objectives of the study are the most important factors affecting the choice of clustering technique and measure of similarity. In workload classification these issues of clustering are widely ignored. The clustering is usually carried out by the *k-means* algorithm without any discussion of consequences. The algorithm uses Euclidean distance and tries to minimize the sum of the within-group variances. Therefore service demands that are close to each other are similar. When the analysis is based on mean values of service demands and the service demands are properly scaled before clustering, the *k-means* algorithm is a suitable technique.

However, the *k-means* algorithm is not a unique technique, but has several major versions and numerous variants. Program listings of three variants (Forgy and Jancey's method [Forg65, Janc66], the convergent k-means algorithm, and MacQueen's method [MacQ67]) are found in an appendix of [Ande73]. Other variants include Wishart's variant in the CLUSTAN package [Clus87], McRae's MIKCA [McRa71], the ISODATA point method [BaHa65], and the method given in [HaWo79]. Moreover, the major statistical packages like SAS, SPSS, and BMPD contain their own versions of the *k-means* algorithm.

Additional differences between clusterings by variants of the *k-means* algorithm are due to the actual clustering programs. The algorithm does not specify the strategy and tactics of reaching the goal of minimizing the sum of variances within the clusters. Therefore, programs implementing the algorithm do not behave equally. A strategy, heuristic tactics and *ad hoc* procedures are inherent in any clustering program. It is not evident how the differences in variants and implementations affect the results of clustering.

After careful reviewing of several alternative implementations, we decided to use the SAS PROC FASTCLUS [SAS85, Ch. 18] and the postprocessor outlined in [Arti86]. Advantages of this approach are presented in [Arti86]. Particularly, the number of clusters is determined automatically in an appropriate fashion.

The classification is controlled by two parameters; *max_rad* and *min_size*. With the option RADIUS = *max_rad*, FASTCLUS determines the minimum number of clusters so that the Euclidean distance from the cluster centroid to any member of that cluster is at most *max_rad*. In addition, the minimum Euclidean distance between centroids of clusters is at least *max_rad*. The parameter *min_size* is used by the postprocessor. The postprocessor removes all the clusters having fewer than *min_size* members. Observations belonging to the cluster being removed are allocated to the nearest remaining cluster. Finally the postprocessor calculates the cluster centroids in original scale (means of service demands) and the number of members in each cluster. These values are used in the queueing network analysis.

In this study we have used *max_rad* = 1 (the value of *max_rad* must be compatible with the scaling used) and *min_size* = 0.5% of observations. These values are recommended in [Arti86]. For further details, see also [SAS85, Ch. 18].

2.4 Scaling of Variables

The importance of scaling the variables is discussed in many papers. An improper scaling and treatment of outliers are the most usual reasons for bizarre results of clustering.

The most popular scaling is the classical standardization. Each original variable x is scaled independently of other original variables to a new variable z:

$$z = \frac{x - \bar{x}}{s_x} , \tag{1}$$

where $\bar{x}$ is the mean value and s_x the standard deviation.

As discussed in [Arti86], the classical standardization is sensitive to outliers. It is not uncommon that measurement data of computer resource usage contains a few observations that are an order of magnitude greater than all other observations. However, the standard deviation is dominated by these few outliers.

To overcome the problem of outliers, the mean and standard deviation in Eqn (1) are replaced by their robust equivalents, usually by their trimmed equivalents. It is generally recommended that 1–5% of observations from the right-hand tail of the distribution should be removed before the mean and standard deviation are calculated. In this study we have trimmed 2.5% of observations as proposed in [Arti86].

The standardization, either classical or robust, makes the clustering completely automatic, but does not take into account any piece of real-world common-sense knowledge about the phenomenon under study. Therefore the full automation is not a real advantage, but a serious disadvantage. The standardization has two serious shortcomings: the magnitudes of original variables are not taken into account and the standard deviation determines the relative importance of the variable.

Since the magnitudes have no effect, the difference of one unit is of the same importance everywhere. For example, the difference between 1 and 2 seconds is as important as the difference between 3600 and 3601 seconds. In statistical terms the standardization implies that both small and large values have the same inherent variability.

Since standardization scales the variables into units of (robust) standard deviation, the deviation determines the relative importance of each variable. Therefore the classification may be dominated by variables that have only a small impact in the analysis applying the detected clusters.

Because of the shortcomings that the standardization has, we examine as an alternative scaling a logarithmic transform. Although logarithmic transform is mentioned in many studies of workload modeling recommending cluster analysis, its use is not widely reported. Intuitively the logarithmic transform is more plausible than standardization. The differences are measured in orders of magnitudes and the inherent variability is assumed to be proportional to the order of magnitude.

We use the following weighted and shifted logarithmic transform:

$$z = w_x \log_{10}(1 + c_x x) , \tag{2}$$

where w_x and c_x are subjectively determined coefficients. Later this transform will be referred to as *logarithmization* and the transformed variables as *logarithmized* variables.

In Eqn (2) the relative importances of variables are controlled by the weights w_x. The constant 1 is necessary to overcome the problem of x being zero. The coefficient c_x affects the resolution of the transform for small values of x. In this study, where the cpu time (seconds) and two io-counts (number of operations) are measured, each w_x is 1. The two counts have $c_x = 1$ and the cpu time has $c_x = 100$.

3. Stability of Results Based on Detected Clusters

As discussed, the validity of clusters should be assessed by examining the results of analysis using the detected clusters. Therefore, stability of results is a necessary condition for validity. In this section we examine how stable the mean system response times of the detected workload classes are in an open queueing network model.

The examined data consists of resource usages of interactive tasks executed during one week, daily from 10 am to 3 pm. The system examined is a VAX8600 running VAX/VMS and used in academic research; primarily in physics, chemistry and engineering. The data contains the cpu time, buffered io-count, and direct io-count of 29,124 tasks.

As an initial phase, we classified all the available data by the *k-means* algorithm and the two scalings of variables (standardization and logarithmization) described in Section 2. The number of detected clusters was 12 in the case of standardization and 14 in the case of logarithmization.

Validity of detected clusters is analyzed using 512 random samples of size 5,800 drawn from the available data. This sample size represents a typical number of interactive tasks executed daily during the five hours of prime-time.

The open queueing network model consists of three service centers: CPU, DIO, and BIO. BIO is an infinite server executing the buffered io-operations. CPU and DIO are queueing devices servicing the cpu time demands and direct io-operations, respectively. The mean response time of each class can be solved from the following response time formulas given in [LZGS84, p. 137]:

$$\begin{aligned}
R_c &= R_{c,cpu} + R_{c,dio} + R_{c,bio} \\
R_{c,cpu} &= D_{c,cpu}/(1 - U_{cpu}) \\
R_{c,dio} &= D_{c,dio}/(1 - U_{dio}) \\
R_{c,bio} &= D_{c,bio} = V_{c,bio} S_{bio} \\
D_{c,dio} &= V_{c,dio} S_{dio} \\
U_{cpu} &= \lambda_1 D_{1,cpu} + \cdots + \lambda_C D_{C,cpu} \\
U_{dio} &= \lambda_1 D_{1,dio} + \cdots + \lambda_C D_{C,dio} \\
\lambda_c &= (N_c/N)\lambda
\end{aligned} \tag{3}$$

The unknowns of Eqn (3) are either design parameters of the experiment (λ, S_{dio}, S_{bio}, N) or obtained as results of sampling and clustering. N is the size of random samples generated; $N = 5\,800$. It represents the typical number of interactive tasks executed daily during the five hours of prime-time in the system examined. The three other design parameters calibrate the queueing network model to the measurement data.

The task arrival rate λ is the measured rate: 29,124 tasks in 25 hours implying $\lambda = 0.3236$. The mean service time of direct io-operations (S_{dio}) is calculated using the properties of disk io-subsystem of the installation examined (a HSC50-controller, two disk channels and 7 RA80 disk-units) and measured request rates to different disk units and hit ratios to caches. Our calculations give $S_{dio} = 0.0290642$. The mean service time of buffered io-operations (S_{bio}) is determined so that the model has the same mean system response time as the measurement data. We obtained $S_{bio} = 0.455955$.

The sampling and clustering provides:

- C: number of workload classes,
- N_c: number of tasks in workload class c, $c = 1, \ldots, C$,
- $D_{c,cpu}, V_{c,dio}, V_{c,bio}$: means of service demands in workload class c, $c = 1, \ldots, C$.

The distributions of mean response times are summarized in Figure 1. The figure displays the average numbers of workload classes detected in different classes of response time. In addition, the figure shows the mean response times of classes detected, when the whole set of available data was classified. The shadowed bars give the reference shape of absolute stability. The coverage reported, which is the percentage of shadowed area covered by the bars, is a heuristic indicator of how stable the results are.

Figure 1 clearly shows that the mean response times of the two automatic classifications are not stable. In both these cases, there are 3-5 relatively short ranges of the response time, in which we obtain one or two classes. However, the response times of the other classes (typically 5-9) are spread almost randomly over a few relatively wide ranges.

4. Conclusions

We have carried out an empirical study of validity of clusters detected by *k-means* algorithm. The two most usual scalings of variables were analyzed. Our focus was on mean value analysis of a multiclass queueing network model, when the workload classes are determined by the clustering technique. Our results are in contrast to most of earlier recommendations. We cannot confirm that clustering techniques are useful tools in workload characterization. However, our conclusion is based on data in one installation, but we have no reason to believe that the examined installation has an exceptional workload.

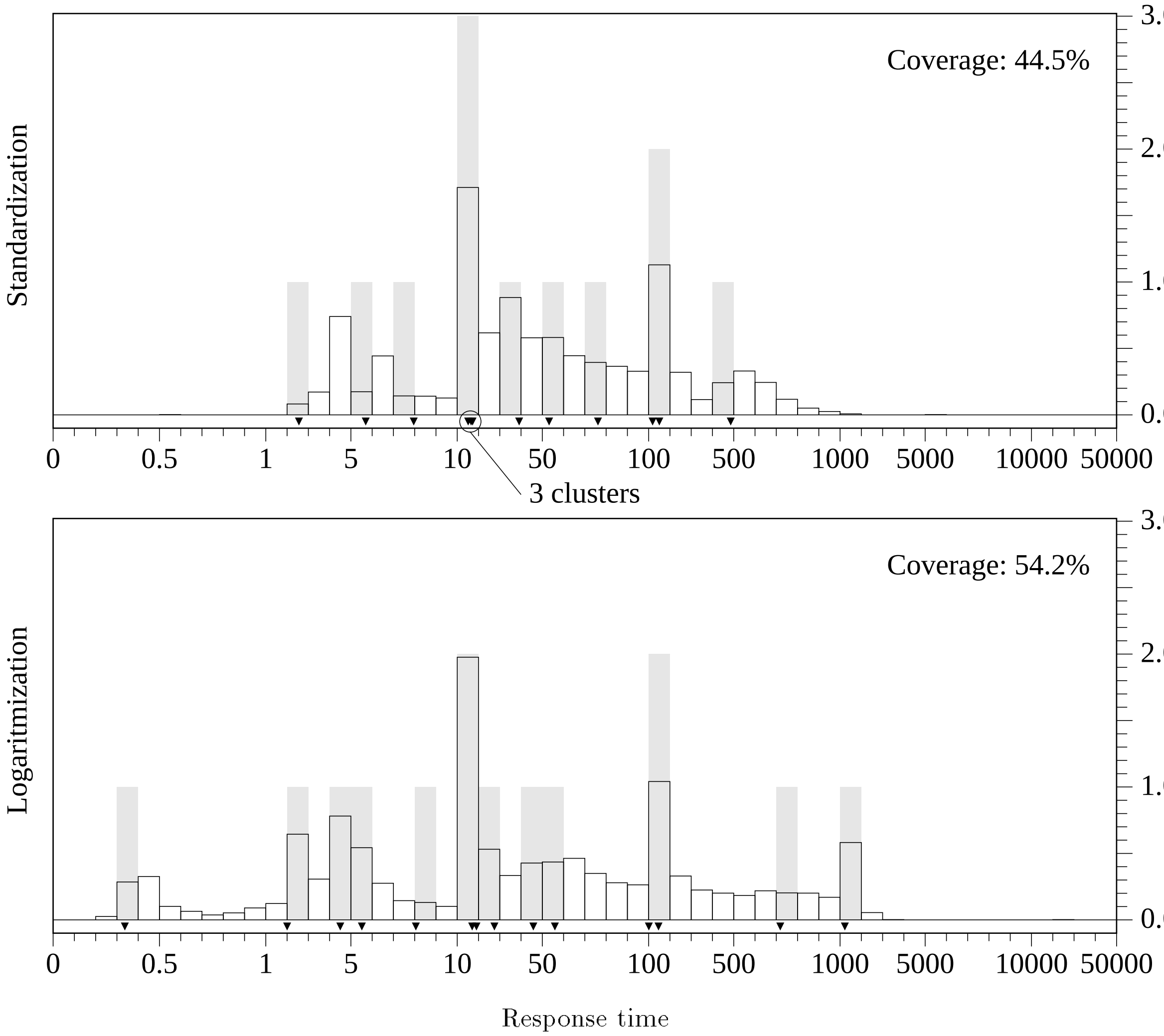

Figure 1: Average number of workload classes detected in different response time classes
Shadowed bars represent the reference shape of absolute stability
▼ : response time of workload class detected by clustering the whole measurement data

We hope that our discussion and examples have shown that the "automatic" clustering techniques do not provide an easy solution to the workload classification problem. The fundamental danger of automatic clustering is that clusters are always detected but without any indicator of their stability and validity. If clustering techniques are used, hard thinking, expert knowledge about the phenomenon being studied and clustering techniques are required. Moreover, clustering programs consume a lot of cpu time, particularly if validity of the detected clusters is examined. On the contrary, subjective methods require "only" hard thinking and expert knowledge about the phenomenon being studied. **The most important advantage of subjective methods for workload classification is that the analyst is forced to think.**

We want to conclude our paper with the following quotations from statistical literature:

- The availability of computer packages of classification techniques has led to the waste of more valuable scientific time than any other "statistical" innovation (with the possible exception of multiple-regression techniques). [Corm71, p. 321]
- But how in practice does one tailor statistical methods to the real needs of the user, when the real need of the user is to be forced to sit and think? [Corm71, p. 346]

– SOME people have to be taught that a computer solution to the wrong problem isn't worth much. [Hook80, p. 40]

– UNLIKE most problems in mathematics, problems in data analysis generally do not have a unique answer. [Hahn85, p. 15]

References

[AgMB76] Agrawala, A. K.; J. M. Mohr; and R. M. Bryant. 1976. "An Approach to the Workload Characterization Problem". *Computer Magazine* **9**, *no. 6* (June): 18–32.

[AgMo77] Agrawala, A. K. and J. M. Mohr. 1977. "Some Results on the Clustering Approach to Workload Modeling". In *Proceedings of the CPEUG 13th Meeting* (October 11–14, 1977, New Orleans, La.). NBS Special Publication 500-18, U. S. Department of Commerce, National Bureau of Standards, Washington, D. C.: 23–38.

[AgMo78] Agrawala, A. K. and J. M. Mohr. 1978. "A Comparision of the Workload on Several Computer Systems". In *Proceedings of CMG IX Conference* (December, 5–8, 1978, San Francisco, Calif.). Computer Measurement Group, Inc., Bethesda, Md.: 177–183.

[Ande73] Anderberg, M. R. 1973. *Cluster Analysis for Applications*. Academic Press, New York, N. Y.

[Arti78] Artis, H. P. 1978. "Capacity Planning for MVS Systems". In *Performance of Computer Installations*, D. Ferrari, Ed. North-Holland, Amsterdam, The Netherlands: 25–53.

[Arti86] Artis, H. P. 1986. "Workload Characterization Using SAS PROC FASTCLUS". In *Workload Characterization of Computer Systems and Computer Networks*, G. Serazzi, Ed. North-Holland, Amsterdam, Netherlands: 21–32.

[BaHa65] Ball, G. H. and D. J. Hall. 1965. "ISODATA, A Novel Method of Data Analysis and Pattern Classification". AD 699616, Stanford Research Institute, Menlo Park, Calif.

[BlAl78] Blashfiels, R. K. and M. S. Aldenderfer. 1978. "The Literature on Cluster Analysis". *Multivariate Behavioral Research* **13**: 271–295.

[BCMS82] Bolzoni, M. L.; M. C. Calzarossa; P. Mapelli; and G. Serazzi. 1982. "A Package for the Implementation of Static Workload Models". In *Proceedings of the 1982 ACM SIGMETRICS Conference on Measurement and Modeling of Computer Systems* (August 30 – September 1, 1982, Seattle, Wash.). Performance Evaluation Review **11**, no. 4: 58–67.

[CaFe86] Calzarossa, M. and D. Ferrari. 1986. "Sensitivity Study of the Clustering Approach to Workload Modeling". *Performance Evaluation* **6**, *no. 1* (Mar.): 25–33.

[ChCo80] Chatfield, C. and A. J. Collins. 1980. *Introduction to Multivariate Analysis*. Chapman and Hall, London, U. K.

[Clus87] *Clustan User Manual*, 4th edition. 1987. Computing Laboratory, University of St. Andrews, St. Andrews, Scotland, U.K.

[Corm71] Cormack, R. M. 1971. "A Review of Classification" (with discussion). *Journal of the Royal Statistical Society, Series A*, **134**, *part 3*: 321–367.

[DuJa76] Dubes, R. and A. K. Jain. 1976. "Clustering Techniques: The User's Dilemma". *Pattern Recognition* **8**, *no. 4*: 247–260.

[DuJa79] Dubes, R. and A. K. Jain. 1979. "Validity Studies in Clustering Methodologies". *Pattern Recognition* **11**, *no. 4*: 235–254.

[DuJa80] Dubes, R. and A. K. Jain. 1980. "Clustering Methodologies in Exploratory Data Analysis". In *Advances in Computers, Volume* **19**, M. C. Yovits, Ed. Academic Press, New York, N. Y.: 113–228.

[Ever79] Everitt, B. S. 1979. "Unresolved Problems in Cluster Analysis". *Biometrics* **35**, *no. 1* (Mar.): 169–181.

[Ever80] Everitt, B. S. 1980. *Cluster Analysis*, 2nd Edition. Heinemann Educational Books Ltd., London, U. K.

[Ferr72] Ferrari, D. 1972. "Workload Characterization and Selection in Computer Performance Measurement". *Computer Magazine* **5**, *no. 4* (July/Aug.): 18–24.

[Forg65] Forgy, E. W. 1965. "Effiency Versus Interpretability of Classification". Paper presented at the Meeting of Biometric Society (Riverside, Calif.). Abstract in *Biometrics* **21**: 768.

[Hahn85] Hahn, G. J. 1985. "More Intelligent Statistical Software and Statistical Expert Systems: Future Directions" (with Comments and Reply). *The American Statistican* **39**, *no. 1* (Feb.) 1–16.

[Hart75] Hartigan, J. A. 1975. *Clustering Algorithms*. Wiley, New York, N. Y.

[HaWo79] Hartigan, J. A. and M. A. Wong. 1979. "A K-Means Clustering Algorithms". *Applied Statistics* **28**: 100–108.

[Hook80] Hooke, R. 1980. "Getting People to Use Statistics Properly". *The American Statistican* **34**, *no. 1* (Feb.): 39–42.

[Janc66] Jancey, R. C. 1966. "Multidimensional Group Analysis". *Australian Journal of Botany* **14**, *no. 1*: 127–130.

[Kana74] Kanal, L. N. 1974. "Patterns in Pattern Recognition: 1968–1974". *IEEE Transactions on Information Theory* **IT-20**, *no. 6* (Nov.): 697–722.

[LZGS84] Lazowska, E. D.; J. Zahorjan; G. S. Graham; and K. C. Sevcik. 1984. *Quantitative System Performance. Computer System Analysis Using Queueing Network Models.* Prentice-Hall, Englewood Cliffs, N. J.

[MacQ67] MacQueen, J. B. 1967. "Some Methods for Classification and Analysis of Multivariate Observations". In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, Vol 1*. University of Clafifornia Press, Berkeley, Calif.: 281–297.

[McRa71] McRae, D. J. 1971. "MIKCA: A FORTRAN IV Iterative K-Means Cluster Analysis Program". *Behavioral Science* **16**, *no. 4*: 423–424.

[SAS85] *SAS User's Guide: Statistics*, Version 5 Edition. 1985. SAS Institute Inc., Cary, N. C.

[Schr78] Schroeder, A. 1978. "How Multidimensional Data Analysis Techniques Can Be of Help in the Study of Computer Systems". In *Proceedings of CPEUG 14th Meeting* (October 24–27, 1978, Boston, Mass.) NBS Special Publication 500–41. U. S. Department of Commerce, National Bureau of Standards, Washington, D. C.: 149–165.

[Sera81] Serazzi, G. 1981. "A Functional and Resource-Oriented Procedure for Workload Modeling". In *Performance '81*, F. J. Kylstra, Ed. North-Holland, Amsterdam, The Netherlands: 345–361.

[Spät80] Spät, H. 1980. *Cluster Analysis Algorithms for Data Reduction and Classification of Objects.* Ellis Horwood Publ., Chichester, U.K.

[Wigh81] Wight, A. S. 1981. "Cluster Analysis for Characterising Computer System Workloads – Panacea or Pandora ?". In *Proceedings of the CMG XII International Conference* (December 1–4, 1981, New Orleans, Louisiana). Computer Measurement Group, Inc., Phoenix, Ariz.: 183–189.