Pergamon

PII: S0003–4878(98)00070–2

Ann. occup. Hyg., Vol. 43, No. 1, pp. 43–55, 1999
© 1999 British Occupational Hygiene Society
Published by Elsevier Science Ltd. All rights reserved
Printed in Great Britain.
0003–4878/99 $19.00+0.00

Hierarchical Cluster Analysis Applied to Workers' Exposures in Fiberglass Insulation Manufacturing

JYUN-DE WU,*‡§ DONALD K. MILTON,† S. KATHARINE HAMMOND* and ROBERT C. SPEAR*

**Center for Occupational and Environmental Health, School of Public Health, University of California, Berkeley, CA* 94720, *U.S.A.*; †*Department of Environmental Health, Harvard School of Public Health, Boston, MA* 02115, *U.S.A.*; ‡*UC Agricultural Health and Safety Center at Davis, University of California, Institute of Toxicology and Environmental Health, Old Davis Road, One Shield Avenue, Davis, CA* 95616, *U.S.A.*

The objectives of this study were to explore the application of cluster analysis to the characterization of multiple exposures in industrial hygiene practice and to compare exposure groupings based on the result from cluster analysis with that based on non-measurement-based approaches commonly used in epidemiology. Cluster analysis was performed for 37 workers simultaneously exposed to three agents (endotoxin, phenolic compounds and formaldehyde) in fiberglass insulation manufacturing. Different clustering algorithms, including complete-linkage (or farthest-neighbor), single-linkage (or nearest-neighbor), group-average and model-based clustering approaches, were used to construct the tree structures from which clusters can be formed. Differences were observed between the exposure clusters constructed by these different clustering algorithms. When contrasting the exposure classification based on tree structures with that based on non-measurement-based information, the results indicate that the exposure clusters identified from the tree structures had little in common with the classification results from either the traditional exposure zone or the work group classification approach. In terms of the defining homogeneous exposure groups or from the standpoint of health risk, some toxicological normalization in the components of the exposure vector appears to be required in order to form meaningful exposure groupings from cluster analysis. Finally, it remains important to see if the lack of correspondence between exposure groups based on epidemiological classification and measurement data is a peculiarity of the data or a more general problem in multivariate exposure analysis. © 1999 British Occupational Hygiene Society. Published by Elsevier Science Ltd.

Keywords: glass fibre; cluster analysis; exposure groups; occupational groups; epidemiology

INTRODUCTION

Hines *et al.* (1995) used hierarchical cluster analysis in exploring the concurrent exposure of female workers in the semiconductor fabrication industry to a number of chemical and physical agents in the context of an epidemiological investigation of spontaneous abortions. We have been further investigating the application of cluster analysis to the characterization of multiple exposures in other aspects of industrial hygiene practice. This work has been motivated by the fact that all workers are exposed to multiple hazards during a typical workday and the technology to measure these exposures is increasingly available. It is not clear, however, how best to summarize this type of multivariate data for use in exposure monitoring and surveillance. It is of some interest, for example, to speculate on the multivariate equivalent of the homogeneously exposed group or, from a different perspective, to consider the multivariate analog of the random effects model (Rappaport *et al.*, 1995) in exposure characterization.

In epidemiological studies, exposure assessments often are performed by classifying study subjects into discrete exposure categories. When dealing with exposures to multiple agents, the exposure classification problem becomes much more complex because of the increase in dimensionality. Although multiple regression or logistic regression models are often applied to explore exposure-relationships, a

Received 25 March 1998; in final form 10 August 1998.

§Author to whom correspondence should be addressed: Tel.: +1 530 754-8164; E-mail: jdwu@epm.ucdavis.edu

well-defined outcome (or dependent) variable is required before using these models. However, the construction of objective exposure classes does not require any information about health outcomes. Examples of multivariate statistical methods which can meet the purposes of exposure assessment include principal component analysis and hierarchical cluster analysis. For example, Simmons and Spear (1993) utilized principal components techniques to characterize workers' exposures to a variety of solvents in a printing plant. Sahl *et al.* (1994) also used the method to examine the intercorrelation between different summary measures of 60 Hz magnetic field exposure among utility workers. Recently, Bye (1996) emphasized the advantages of the technique for the management of large data sets and the application to the generation of new hypotheses for investigations of complex systems. While principal components analysis and its variants have been applied in various exposure-related applications, this has not been the case for cluster analysis. Hence, the purpose of this study was to explore, via cluster analysis, exposure patterns among workers exposed to three agents (airborne endotoxin, phenolic compounds and formaldehyde) in fiberglass insulation manufacturing, and to characterize exposures based on these patterns.

There are many variants of cluster analysis and one may generally expect differences in the final results of such an analysis depending on which particular procedure is used. Hence, one of the principal objectives in this study was to gain some sense of the magnitude of these differences when common clustering procedures were applied to typical exposure data. As will be seen, we will contrast exposure groupings for the fiberglass workers based on cluster analysis with non-measurement-based strategies commonly used in epidemiology. In making these comparisons, it is important to be confident that any differences observed between the two approaches are not artifacts of the statistical procedures used in identifying exposure clusters.

CLUSTER ANALYSIS

Everitt (1994) has given a clear definition of cluster analysis, '*Cluster analysis is a generic term used for a large number of techniques which attempt to determine whether or not a data set contains distinct groups, and, if so, to find the groups.*' That is, cluster analysis is a good statistical tool of searching for objects with similar attributes in a data set. The following paragraphs describe the general aspects of the approach with respect to the selections of distance metrics, clustering algorithms, or criteria for determining the number of clusters and for validating a tree structure.

In general, cluster analysis regards each object as a point in a multi-dimensional space defined by the values of each of its attributes. The distance between two objects is measured to determine the similarity of the objects in terms of each of its attributes. Therefore, the choice of a distance metric is the initial step of cluster analysis. There are a variety of distance metrics available, but Euclidean distance is the most common and intuitive and was used throughout this study.

Because differences in units and in the magnitude of the variance in each of the individual attributes may influence the computation of distance metrics, variable standardization is important for cluster analysis. Various standardization methods have been proposed. Milligan and Cooper (1988) conducted a Monte Carlo study to compare the performance of seven different variable (attribute) standardization methods in recovering known clusters of synthetic data. They found that the standardization approach which divides each variable by its range exhibited consistently superior recovery of the structures under different error conditions, separation distances, clustering algorithms, and coverage levels. In contrast to Milligan and Cooper's study, Schaffer and Green (1996) evaluated the variable standardization methods by using real data sets and external validation. They too discovered that no standardization did as well or better than the range standardization. Although the range standardization is considered to be a good method for variable standardization in cluster analysis, it is not possible to conclude that it is always the most effective approach.

After a distance metric is selected and the variables are standardized, the next step is the determination of a clustering algorithm. Since the purpose of cluster analysis is to combine objects into groups or clusters, some rules or methods are required to determine how to form these groups. Clustering algorithms are the rules or procedures used for this purpose. In general, the issue is to decide when two objects are sufficiently similar to form a cluster and then to decide whether other objects should be added to this nucleus, to another, or to start a new cluster. Some of the popular algorithms are the centroid method, the single-linkage (or nearest-neighbor) method, the complete-linkage (or farthest-neighbor) method, the average-linkage method and the Ward's method. Complete-linkage, single-linkage, average-linkage and model-based clustering algorithms are the methods available in S-plus (Venables and Ripley, 1994) and were used in this study.

The detailed explanations of these algorithms are well described in Everitt's book (Everitt, 1993) and Banfield and Raftery's paper (Banfield and Raftery, 1993). Brief descriptions of these algorithms are given here to provide an understanding of how they work. The centroid method calculates the distance between two clusters based on the weighted centers of the clusters; two clusters with the smallest distance are grouped and a new centroid is computed. In the single-linkage method, the distance between two clusters is defined as the minimum of the distances between all possible pairs of objects in the two clusters. In the complete-linkage method, the distance between two clusters is represented by the maximum of the distances between all pairs of objects in the clusters. The average-linkage uses the mean distance from all objects in one cluster to all objects in another; two clusters with the smallest mean distance are then merged to form a new cluster. Ward's method does not compute distances between clusters, but instead forms clusters by maximizing within-clusters homogeneity where the within-cluster sum of squares is used as the measure of homogeneity. Unlike the algorithms previously described, the model-based clustering algorithm allows one to choose cluster features *a priori*, i.e., shape, size and orientation; this is achieved by reparameterization of the covariance matrix and utilizes information resulting from eigenvalue decomposition (Banfield and Raftery, 1993).

It has been pointed out that different clustering algorithms may produce different shaped and sized clusters. As described by Everitt (1993), single linkage clustering is good for finding elliptical clusters; both centroid and Ward's methods have a tendency to obtain spherical clusters. Therefore, it has been a concern that the orientation, size and shape, inherently existing in a data set determine whether or not a particular clustering algorithm gives useful results. However, these characteristics are not known *a priori*. Although this concern originated from a statistical perspective, it also has important implications for industrial hygiene and toxicology. For example, if the size of a cluster is 'large', within-cluster variability is 'large'. Thus, it may not be reasonable to claim that the members of the cluster share homogeneous exposures because the exposure can vary significantly in both magnitude and composition. In summary, however, a cluster is comprised of a set of sample points (in this study each point defined by an endotoxin-phenolic-formaldehyde level) which are 'close' together as defined by the application of clustering algorithms to the measured Euclidean distances between the points.

For hierarchical clustering algorithms, the number of clusters at each step is one less (or more) than the previous one. A dendrogram or hierarchical tree is the graphical presentation of various steps of the hierarchical clustering process. Normally, the vertical axis of a hierarchical tree indicates the Euclidean distance or level of dissimilarity where two objects or clusters merge to form a larger cluster. The tree shown in Fig. 1 is a typical hierarchical tree. Cutting a hierarchical tree horizontally creates a clustering. The horizontal axis of a hierarchical tree identifies the objects being classified. Objects connected by lines represent clusters which are nested together. The tree clearly shows the Euclidean distance between clusters and the numbers of clusters at each merging stage. In order to form clusters from a hierarchical tree, a threshold on the Euclidean distance or dissimilarity value needs to be specified. Hence, a method of determining the number of clusters in a data set is needed to form objective and representative exposure clusters or patterns.

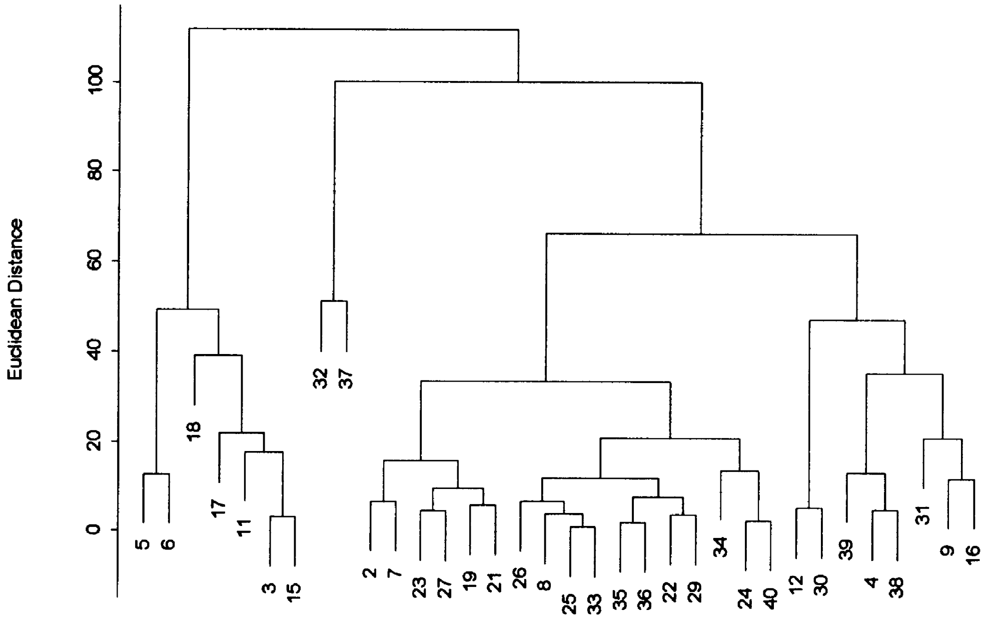
There are different approaches to determining the number of clusters. These include the variance ratio criterion (VRC) (Calinski and Harabasz, 1974), the point serial correlation coefficient (or called MH index) (Jain and Dubes, 1988), and the approximate weight of evidence (AWE) approach (Banfield and Raftery, 1993). Milligan and Cooper (1985) conducted a Monte Carlo study which evaluated 30 procedures for determining the number of clusters in data sets with different numbers of non-overlapping clusters. The detailed discussion of how these approaches perform is not within the scope of this paper. For the purpose of illustration, however, the VRC approach is presented to show the determination of the number of clusters in the fiberglass data set. This method determines the number of clusters by comparing the ratio of between cluster sum-of-squares to within cluster sum-of-squares.

The final important issue in cluster analysis is the determination of the validity of a tree structure. That is, how does one conclude that a tree structure obtained in a particular cluster analysis was not produced by chance? Methods have been proposed for accomplishing this task (Jain and Dubes, 1988) but their discussion is also beyond the scope of this paper although the Rand index (Hubert and Arabie, 1985) was applied to assess the validity of the tree structure of the fiberglass data set.

THE DATA SET

The data set used in this study was collected in an epidemiological study of peak expiratory flow change among workers exposed to endotoxin, phenolic resin and formaldehyde in a fiberglass wool manufacturing plant (Milton *et al.*, 1996). Worker exposures were measured by taking personal air samples and recording time-activity worklogs. Sampling and analysis techniques for these samples were discussed in detail in Walters (1993). A total of 393 half-shift (4-hour) personal air samples were collected from 37 workers in four work groups for 5–6 days each; two measurements were taken per day for almost all workers. The four work groups include two production groups (A and D) and two maintenance groups (M and N). These groups were selected largely for sampling convenience and are reported here since they correspond roughly to traditional classifications based on job titles. The prior expectation would be that the two production groups would be similar to one another, as would the maintenance groups, but that production exposures would differ from maintenance exposures. Ten to twelve personal measurements were taken for most workers.

In a previous study Walters (1993) used job titles and observations of work tasks and areas to classify these workers into four exposure zones (B, F, O and X for basement, forehearth, curing ovens and maintenance, respectively). Detailed descriptions of the assignment of the exposure zones were given in a later study (Milton *et al.*, 1996). In summary, the workers



Note : The number at the end of each branch indicates a worker's identification number.

Fig. 1. Hierarchical tree based on workers' mean exposures using complete-linkage and non-standardized data.

in the exposure zones B, O and F mainly stayed in their work areas to conduct their work tasks. Therefore, they were considered to be 'fixed location workers.' The workers in the exposure zone X were considered to be 'mobile workers' because they moved around different work areas. Although the workers in this exposure group were not entirely maintenance workers, they shared a common characteristic of spending most of their time in low-exposure areas.

Statistical analyses were performed by using S-plus for Windows version 3.2. Summary statistics were calculated including the mean, variance, coefficient of variation and correlation coefficients. Quantile-quantile plots were generated to examine the distributions of workers' exposures. Among 380 valid measurements, no value was under the detection limit for endotoxin. There were a large (59%) and small (12%) proportions of the measurements under the detection limits of phenolic compounds and formaldehyde, respectively. The half-shift exposures of these workers were approximately lognormal when examining the quantile-quantile plots after removal of the values under the detection limits. Measurements below the detection limits were replaced by the values of half the detection limits for the cluster analysis. Maximum likelihood methods were used in estimating the means and variances in the summary statistics as described below.

RESULTS

Descriptive statistics

Exposure means and standard deviations of these three agents and four work groups are shown in Table 1. Because the high proportion of exposure measurements under the detection limit for phenolic compounds probably resulted in unreliable estimates of exposure means and standard deviations, a maximum likelihood estimation (MLE) algorithm was used under S-plus to estimate these means and standard deviations. The basic assumptions underlying the MLE algorithm is that the values under the detection limit follow the same distribution as those above the detection limit. Here, the distribution of the values above the detection limit is assumed to be lognormal. Although the MLE was used to estimate the means and standard deviations of the work groups, the group mean exposures were not subsequently used in cluster analysis. As expected, the median exposure of groups A and D were similar to each other and higher than groups M and N. Three univariate analyses of variance (ANOVA) were conducted on the logs of the individual concentration data to explore differences in exposure by zone. The model used for the ANOVA was a fixed-effects model. The results illustrate that there was a significant difference between the median exposures for different exposure zones for each agent (Tables 2 and 3).

However, analysis of consecutive four-hour measurements suggests that some degree of autocorrelation was present in the formaldehyde data in particular. Hence, the independence assumption underlying the ANOVA significance tests was compromised. The F values were sufficiently large and the strength of the autocorrelation sufficiently modest, however, to support the conclusion that a difference in median exposure exists between zones. The Pearson

Table 1. Mean, median and standard deviation of workers exposure measurements by work groups

Agents	Work groups	Number of samples	Mean (μg/m^3)	Median (μg/m^3)	Range (μg/m^3)	Standard deviation (μg/m^3)
Endotoxin						
	A	91	0.0323	0.0091	0.0003–0.8170	0.0901
	D	104	0.0467	0.0129	0.0003–1.9860	0.1983
	M	82	0.0202	0.0033	0.0006–0.7642	0.0859
	N	87	0.0164	0.0023	0.0002–0.2929	0.0394
	Total	364				
Phenolic Compounds	A	*91(35)	**50.10[66.85]	35.40	5.57–281.98	**58.48 [204.15]
	D	104(47)	38.23[46.32]	21.80	5.24–302.36	47.50 [127.92]
	M	82(71)	16.08[13.92]	7.13	5.78–135.65	24.97 [59.49]
	N	87(66)	***37.30[NA]	7.54	4.60–605.89	90.93 [NA]
	Total	364(219)				
Formaldehyde						
	A	91(13)	**64.60[115.21]	36.37	1.12–314.97	**68.21 [469.63]
	D	104(3)	75.65[95.25]	66.74	1.05–344.80	61.24 [164.61]
	M	82(15)	24.91[32.94]	12.44	1.29–148.25	33.57 [104.62]
	N	87(11)	29.08[36.06]	15.86	1.09–183.97	33.15 [84.01]
	Total	364(42)				

*The number in the parenthesis is the number of measurements under the detection limit.
**The number in the brackets is the maximum likelihood estimate.
***The maximum likelihood estimate is not available.

Table 2. Mean, median and standard deviation of workers exposure measurements by exposure zones

Agents	Exposure zones	Number of samples	Mean (μg/m^3)	Median (μg/m^3)	Range (μg/m^3)	Standard deviation (μg/m^3)
Endotoxin						
	B	77	0.08	0.02	0.0006–1.9860	0.24
	F	69	0.03	0.02	0.0006–0.3871	0.05
	O	34	0.01	0.01	0.0009–0.0857	0.02
	X	188	0.02	0.00	0.0002–0.7642	0.06
	Total	368				
Phenolic Compounds	B	77	59.91	31.46	5.68–302.36	70.90
	F	69	48.24	43.15	5.24–147.35	34.17
	O	34	20.85	8.21	5.53–125.54	27.45
	X	188	24.38	7.39	4.60–605.89	64.67
	Total	368				
Formaldehyde						
	B	77	74.33	49.96	1.13–314.97	67.24
	F	69	76.43	75.92	1.05–217.12	57.52
	O	34	74.54	48.70	1.85–344.80	73.07
	X	188	25.57	13.66	1.09–183.97	33.13
	Total	368				

product-moment correlation between the individual exposures was low when all exposure zones were combined. In general, a medium correlation (between 0.30 and 0.45) existed between exposures to endotoxin and phenolic compounds and between exposures to phenolic compounds and formaldehyde in exposure zones B, F and X (Table 4). However, a high correlation (0.709) between exposures to endotoxin and phenolic compounds was observed in exposure zone O (Table 4). It was not clear what caused the observed high exposure correlation. Because the correlation between the agent exposures was generally not high, the adjustment for the intercorrelations between the variables was not performed before calculating the Euclidean distance for cluster analysis.

Preliminary analyses

In order to apply cluster analysis to the multivariate data with repeated measurements, mean values of worker exposures to each agent were used in most analyses. A data matrix consisted of 37 rows (workers) and 3 columns (agents) was thereby created. By viewing each worker's mean exposure as a point in a three-dimensional (three-agent) space, the distance between

Table 3. Analysis of Variance of Exposure Zones for Log(Agent Exposure)

	DF	Sum of square	Mean square	F value	Pr(F)
Log (endotoxin)					
Exposure zones	3	233.17	77.72	39.37	0.000
Residuals	364	718.60	1.97	—	—
Log (phenolic compounds)					
Exposure zones	3	101.49	33.83	33.67	0.000
Residuals	364	365.72	1.00	—	—
Log (formaldehyde)					
Exposure zones	3	144.71	48.24	26.34	0.000
Residuals	364	666.53	1.83	—	—

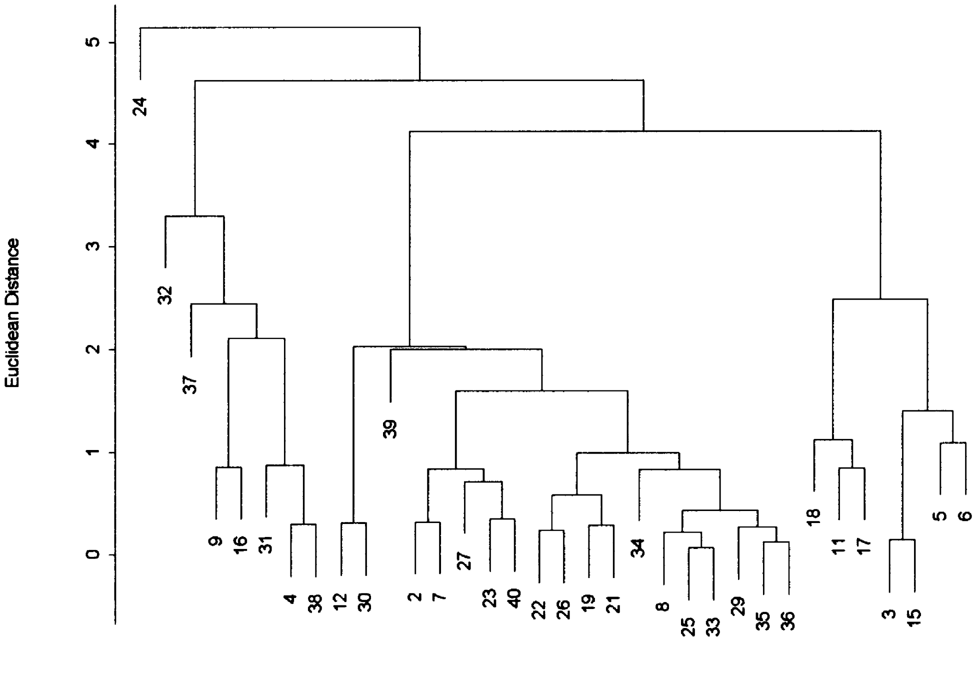
Table 4. Correlation matrix of agent exposures

All exposure zones combined	Endotoxin	Phenolic compounds	Formaldehyde
Endotoxin	1.000	0.286	0.156
Phenolic compounds		1.000	0.330
Formaldehyde			1.000
Exposure Zone B			
Endotoxin	1.000	0.329	0.149
Phenolic compounds		1.000	0.424
Formaldehyde			1.000
Exposure zone F			
Endotoxin	1.000	0.263	−0.059
Phenolic Compound		1.000	0.351
Formaldehyde			1.000
Exposure zone O			
Endotoxin	1.000	0.709	0.020
Phenolic compounds		1.000	0.155
Formaldehyde			1.000
Exposure zone X			
Endotoxin	1.000	0.253	0.238
Phenolic compounds		1.000	0.240
Formaldehyde			1.000

each pair of workers was calculated. Thus, a distance or similarity matrix was obtained. By applying clustering algorithms to the distance matrix, a hierarchical tree was constructed from which the similarity of workers' exposures can be assessed. Examining the hierarchical trees of the unstandardized and standardized (z-standardization) arithmetic mean exposures based on Euclidean distance and three different clustering algorithms: complete-linkage, group-average and single-linkage, showed three workers (#1, 13 and 14) had very different exposures from all others. These three points formed long branches which did not join the main tree until very late stages. Hence, the algorithms tended to identify clusters comprised of only one worker. Therefore, we chose to treat these workers as having unique exposures and they were deleted from the following analyses. Figure 1 is the hierarchical tree of the unstandardized mean exposures based on Euclidean distance and the complete-linkage (or farthest-neighbor) clustering algorithm with the three outliers deleted. The numbers at the end of each node are the identification numbers of the workers. Examining the hierarchical tree, we can see the 34 workers can be classified into three, four or five exposure clusters (or groups) by cutting the tree at Euclidean distances of about 65, 55 or 50.

Effects of standardization

Because the difference in the magnitude of exposure and in variability between the three agents are likely to influence clustering, the data were standardized to mean zero and variance one (z-standardization) and analyzed by the complete-linkage algorithm. The hierarchical tree obtained from the standardized data is showed in Fig. 2. Comparing this tree with that obtained from the unstandardized data with the same clustering algorithm (Fig. 1), it can be seen that some



Note : The number at the end of each branch indicates a worker's identification number.

Fig. 2. Hierarchical tree based on workers' mean exposures using complete-linkage and z-standardization.

workers changed clusters. This difference is largely due to the fact that endotoxin exposures in the unstandardized data were small compared with those of other two variables, hence that exposure component did not play an important role in the unstandardized clustering. After standardization, the differences of endotoxin exposures among these workers made a contribution to the exposure clustering.

Effects of clustering methods

The influence of clustering algorithms on the formation of a tree structure has been discussed extensively (Everitt, 1993; Jain and Dubes, 1988). Here, this issue was explored by the production of tree structures using complete-linkage, group-average and single-linkage clustering algorithms applied to standardized data. Because this data set is three dimensional the use of scatter plots supplements the tree diagrams in displaying the effects of different clustering procedures. Figure 3a is one such plot which illustrates the clusters found by the complete-linkage algorithm when four exposure clusters were chosen (i.e., the tree was cut at a distance of 3.5 in Fig. 2). Figure 3b and 3c are the plots from group-average and single-linkage clustering algorithms, respectively, when four clusters were specified. When comparing the latter two plots with that produced by complete-linkage (Fig. 3a), there are minor differences between the group-average and complete-linkage clustering algorithms, but, very significant differences between the single-linkage and complete-linkage clustering algorithms. The exposure clusters produced by the single-linkage algorithm contain three single-member clusters which illustrate the danger of specifying the number of clusters *a priori*.

Finally, the model-based clustering algorithm was applied to the data set and asked to identify ellipsoidal clusters. Figure 3d shows the three-dimension exposure cluster plot produced by this algorithm. Comparing these exposure clusters with that from complete-linkage clustering algorithm, we can see significant changes on the members of clusters where there are high exposures to three agents. Therefore, the specification of the cluster shape had an impact on the outcome.

From these results it is clear that different clustering algorithms produce different results when applied to typical exposure data as they do in other statistical applications. The next issue was to investigate whether any of these statistically-defined clusters correlated with the epidemiological classifications of exposure based on work area or work group.

Clusters versus epidemiological classification

When Fig. 2 was labeled according to the exposure zones, it is easily seen that workers in the same exposure zones were classified into different exposure clusters [Fig. 4(A)]. Again, when the exposure clusters were labeled by the work groups A, D, M and N [Fig. 4(B)], one cluster was comprised of only workers from the production work groups (A and D) and another cluster had about two-thirds of the members from the maintenance work groups (M and N). However, 6 out of 16 workers in the production work groups were classified into the exposure clusters of the maintenance

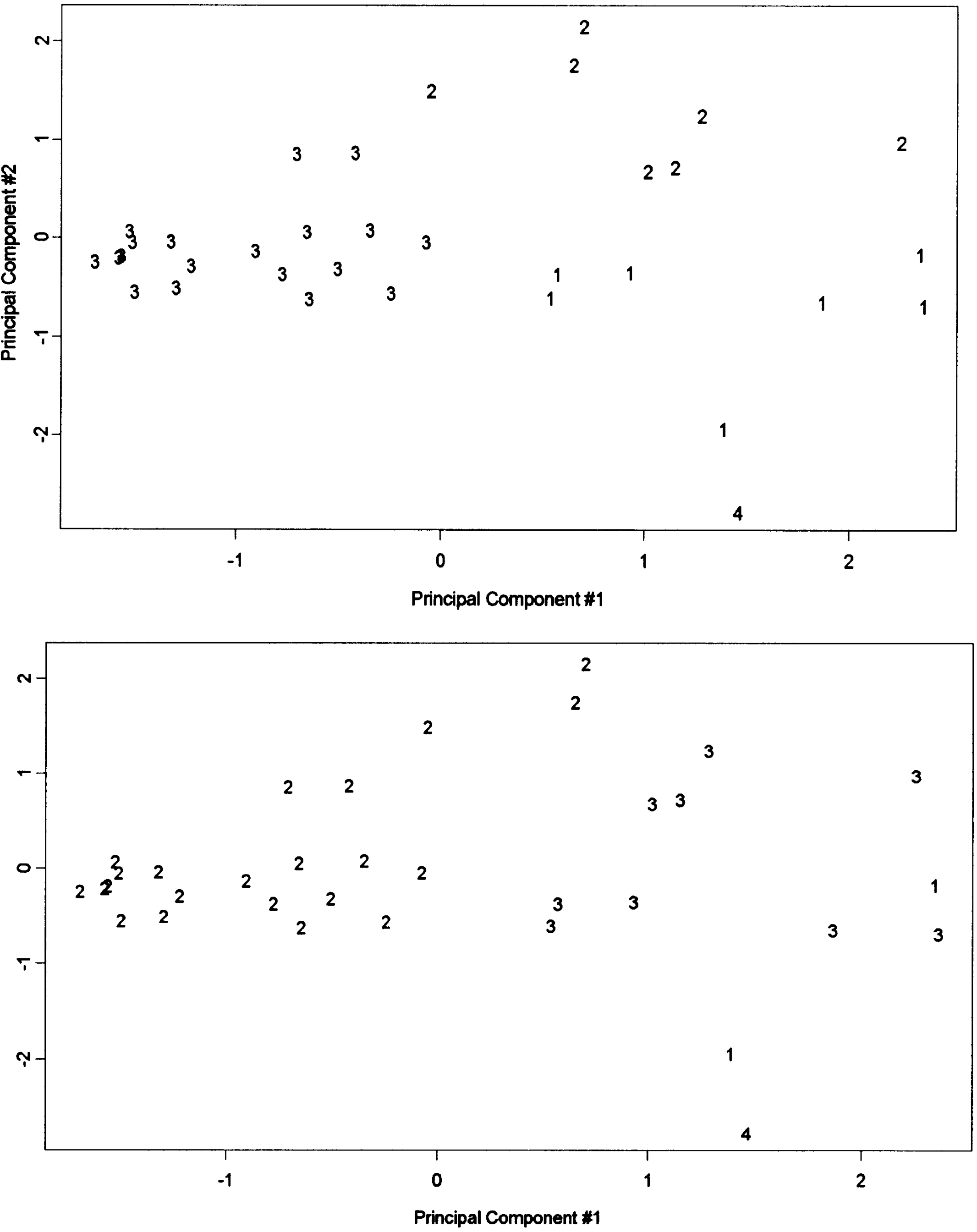
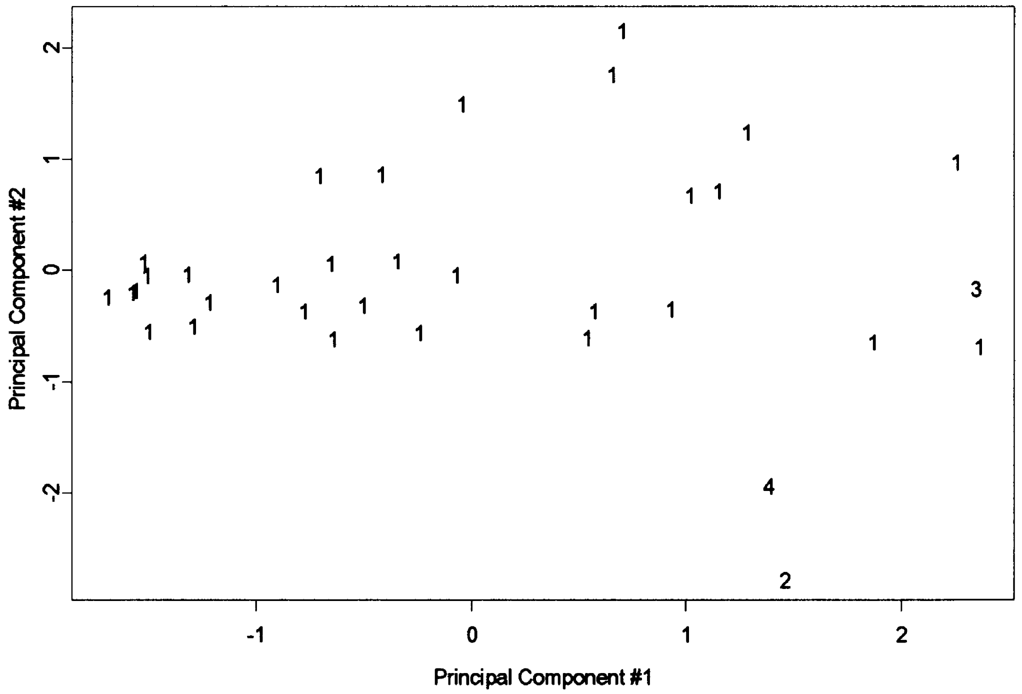


Fig. 3. (a) Principal component plot of workers' standardized mean exposures based on cutting the tree into four clusters (complete-linkage). (b) Principal component plot of workers' standardized mean exposures based on cutting the tree into four clusters (group-average).

work groups and another cluster, formed at high Euclidean distance, consisted of about equal number of workers from both the production and maintenance work groups. All these results illustrate that the exposure classification based on actual measurements differed from that based on subjective criteria. Although the small numbers of workers in some of the exposure zones and clusters made the determination of a representative exposure cluster of an exposure zone difficult, it was evident that the exposure clusters identified from the standardized data set had little relationship with the classification results from either the exposure zone or the work group classification approach. While we cannot allege that one exposure classification approach is superior to another in all settings, most experts in exposure assessment prefer methods based on field measurements versus those based on expert opinion (Kromhout *et al.*, 1993; Post *et al.*, 1991).



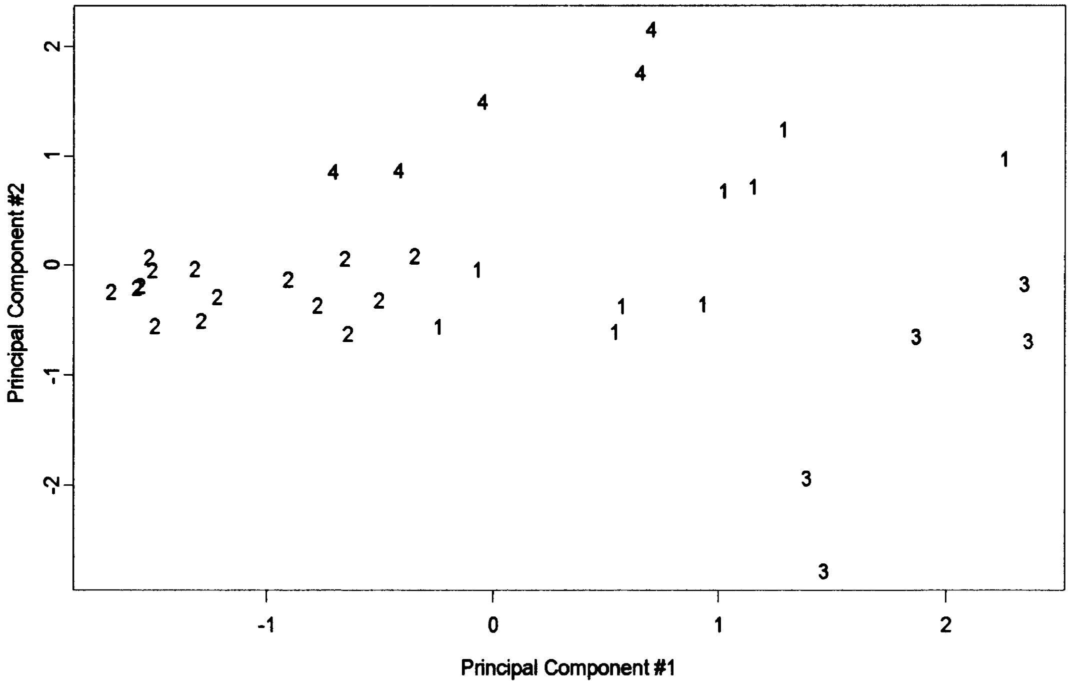


Fig. 3. (c) Principal component plot of workers' standardized mean exposures based on cutting the tree into four clusters (single-linkage). (d) Principal component plot of workers' standardized mean exposures based on cutting the tree into four clusters (model-based).

It is interesting to compare the difference of between-worker geometric standard deviations (GSD_b) of the marginal distributions of each of the three agents between the groupings based on the exposure zone and cluster analysis approaches. Table 5 shows the GSD_b of these two classification approaches. As can be seen, the cluster analysis based on the complete linkage approach tends to give a smaller GSD_b than does the exposure zone approach. While these differences in GSD_b are not large in this case, it must be recalled that the membership in the cluster is quite different than that based on exposure

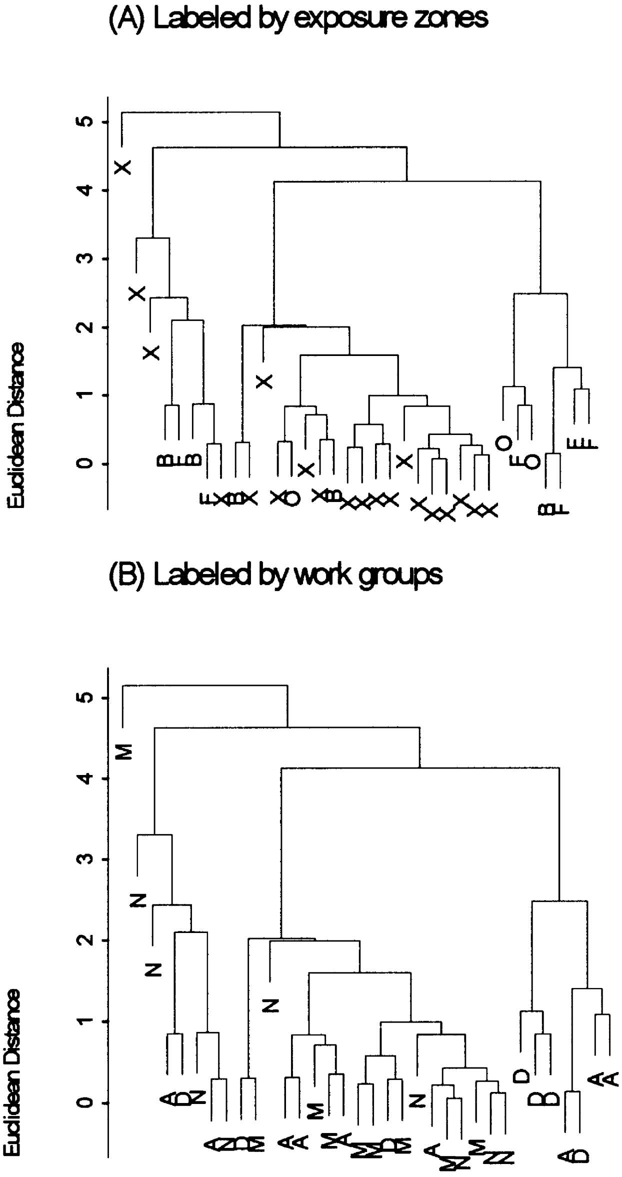


Fig. 4. Hierarchical tree based on workers' mean exposures using complete-linkage and z-standardization.

Table 5. Comparison of gEometric Mean and between-worker geometric standard deviation (GSD_b) between exposure groupings based on cluster analysis and exposure zone approach

		Cluster analysis					
	Number of	Endotoxin		Phenolic compounds		Formaldehyde	
Clusters	workers	GM	GSD_b	GM	GSD_b	tGM	GSD_b
#1	7	0.035	1.415	52.272	1.496	40.597	1.481
#2	7	0.016	1.583	35.233	1.543	86.125	1.147
#3	19	0.007	1.886	14.121	1.754	26.428	1.544
#4	1	NA	NA	NA	NA	NA	NA

			Exposure zones				
Number of	Endotoxin		Phenolic compounds		Formalehyde		
Zones	workers	GM	GSD_b	GM	GSD_b	GM	GSD_b
F	6	0.023	1.801	46.539	1.281	72.042	1.289
B	5	0.024	1.657	30.666	1.550	49.889	1.452
X	20	0.009	2.63	17.147	2.31	24.692	1.522
O	3	0.012	1.302	20.475	1.188	70.808	1.528

zones. Also, the difference of GMs between groups was larger for the cluster analysis than the traditionally constituted exposure zones.

Determination of number of exposure clusters

Although our results to this point suggest that statistically-based definitions of clusters may not be particularly helpful in analyzing exposure data, there is one additional aspect of cluster methodology that might provide further insight. This relates to the number of clusters that exist in a data set. In all of the foregoing analyses the number of clusters was chosen to match the epidemiological classification and based on the original tree structure of Fig. 2. In exploring a more organized approach to defining the number of clusters appropriate to the data set, the tree structure produced by using the complete-linkage clustering algorithm was used. Figure 5 is the result of applying the VRC (variance ratio criterion) approach mentioned earlier. According to the criterion described by Calinski and Harabasz (1974), we interpret this figure to infer that the most plausible number of clusters is four because the curve flattens at that point. Because the VRC approach was originally based on the single-linkage clustering algorithm, the tree structure produced by using the single-linkage clustering algorithm was also tested by using the VRC approach. The result (not shown here) also indicated that clusters were not well separated and that it was difficult to determine the number of clusters of the tree structure.

The Rand index (Hubert and Arabie, 1985) was used to assess the validity of the tree structure. Briefly, the basic question is, given the fiberglass exposure data, how likely is it that a given number of clusters would be determined by a particular clustering algorithm if there was no underlying structure to the data? In this case, the complete-linkage algorithm was used in a bootstrap scheme to estimate the probability that 3, 4 or 5 clusters would be determined by chance. The bootstrap approach estimated the number of clusters in the data using calculations based on the hypothesis that there was no underlying structure in the data. That is, we calculated the probability of observing certain structures in the data by chance. The result of this bootstrap approach was that, for this data set, the probability of observing 3 clusters by chance was 0.55, of 4 clusters 0.04, and of 5 clusters 0.02. The 3-cluster structure identified in the data was also frequently observed in the random samples but the 4- and 5-cluster structures were not. Hence, a result yielding either 4 or 5 clusters is likely to represent the data structure well, since such a result is unlikely to be observed by chance.

The lack of consistency in the number of clusters of the different determination approaches suggests that there is little separation between the clusters. However, the result is based on a relative small data set for multivariate statistical analysis. The unstable result may limit our ability to make conclusions but it also motivates more consideration of multivariate statistical approachs to classifying multiple occupational exposures and forming homogeneous exposure groups.

DISCUSSION AND CONCLUSIONS

To the extent that a cluster is regarded as identifying an exposure group, it is clear that different clustering

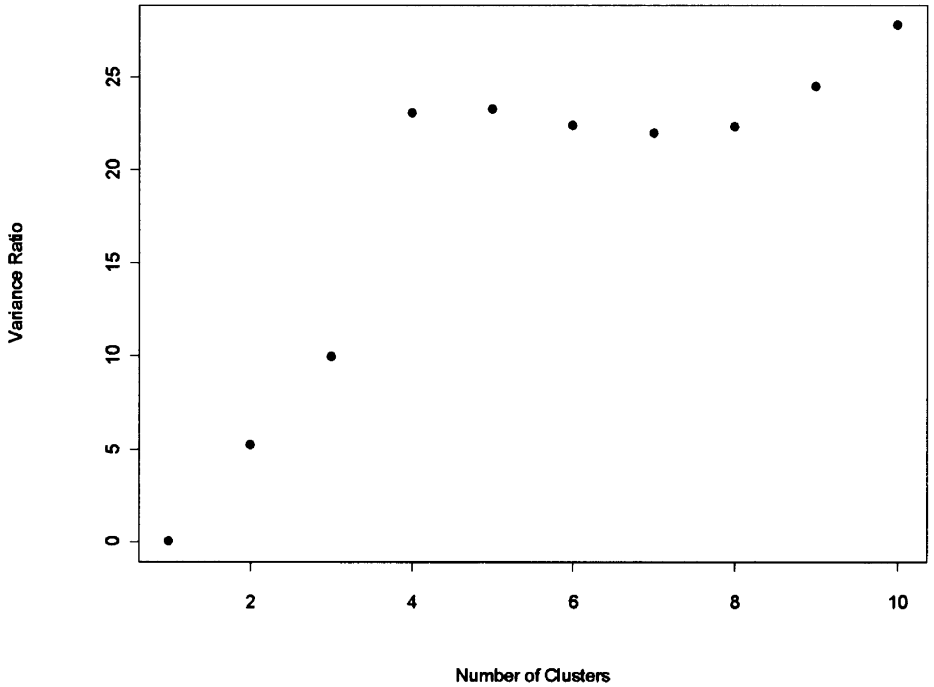


Fig. 5. Relationship between variance ratio (BGSS/WGSS) and number of clusters.

algorithms assign individuals to different exposure groups. That is, the assignment of individuals to exposure groups is algorithm-dependent. Hence, in this application cluster analysis must be considered an exploratory method since the interpretation of any particular clustering result is dependent not on statistical issues, but on toxicological or other external justification.

Despite the algorithm-specific differences in assigning individuals to different exposure clusters, it was clear that none of the measurement-based groupings corresponded to those based on an individual's work group or on the exposure zone in which an individual worked. That is, perhaps, not surprising based on the work of Rappaport *et al.* (1993) who found similar disparities in comparisons of measurement-based exposure classifications versus zoning or job-class approaches for single agent exposures using much more extensive data sets. In the present case, Walters (1993), who collected the data used herein, confirmed that the exposure zones were relatively homogeneous viewed from the perspective of the physical environment and the nature of the contaminants. However, there were variations in activities day-to-day which might well have led to differences in exposure intensity. On the other hand, recalling that five to twelve half-shift samples were collected per person, it seems reasonable to expect some stability in the mean exposures used as a basis for the analyses reported herein.

The foregoing results raise the general question of how exposure groups should be defined based on measurement data in the multivariate context. In the single agent case, Rappaport (1991) applied the random effects model and suggested that groups of workers be considered to be uniformly exposed if the between-person geometric standard deviation (GSD_b) of their exposures was 1.2 or less. Compared with a homogeneous exposure group, an uniform exposure group signifies truly homogeneous exposure among group members. While this is an attractive approach, the multivariate case is qualitatively different since it seems likely that there will be a need to deal with differences in the composition of exposure as well as differences in intensity, at least if these differences are 'large.' The distance metrics used in cluster analysis make no distinction between composition and intensity differences. From the perspective of exposure assessment, however, we argue that the notion of 'nearness' of two exposures, whether nearness in composition or in intensity, is a reasonable point of departure in the absence of detailed data relating to toxicological or other biological mechanisms of interaction. If one accepts this premise, then Rappaport's approach can be extended since it echoes the underlying concept of those variations of cluster analysis which separate clusters based on differences in within- and between-cluster distances.

The notion of the *a priori* definition of membership, the GSD_b of 1.2 in the univariate case, for example, has considerable attraction in a multivariate extension where there is inadequate toxicological data to judge the relative potency of the components of the mixture in causing the effect in question. If such data is available, however, some sort of toxicological normalization would appear to be required as a first step both from the statistical perspective discussed above and from the standpoint of relative risk. Suppose, for example, that the set of agents in question each had an Occupational Exposure Limit (OEL), e.g., PEL, TLV, etc. Then one might divide each measurement by its OEL prior to exposure analysis in order to place each component of the exposure on a comparable basis from the perspective of known health risks. However, one should be aware that this toxicological normalization does not avoid the problem of exposure misclassification when high exposure variability is encountered (Peretz *et al.*, 1997). The final step is to define the generic exposure group based on the normalized exposure levels. For example, an exposure group might consist of all individuals whose mean exposures lie within a hypercube of 0.10 OEL units on a side. Because a published OEL or equivalent value may not always be available for each contaminant, the application of the potency normalization approach to multivariate exposure analysis has its limitation. How to cope with this limitation is an important issue for future study.

While we cannot argue that the fiberglass data set is typical of all multiple exposure data nor that cluster analysis is typical of all multivariate statistical procedures, we do conclude that several generic issues in multivariate exposure analysis have been raised by this example. The normalization of the components of the exposure measurement vector to some common risk-based units seems a necessity as does some *a priori* definition of what constitutes an exposure group, at least in applications where the assignment of individuals to groups is required. In addition, it remains to be seen if the lack of correspondence between exposure groupings based on epidemiological classification and measurement data is a peculiarity of this data set or a more general problem.

Finally, in epidemiological applications where a single agent cannot be linked with health effects, it may be informative to examine whether there is a different rate of health effects in different clusters. If so, it indicates that biological interactions of the exposure agents should be the focus of research on exploring the etiology of the health effects.

Acknowledgements—The authors thank Dr. Nong Shang for his helpful advice pertaining to statistical analyses. This research was supported in part by a grant from the National Institute for Occupational Safety and Health (R01/CCR912034-03).

REFERENCES

Banfield, J. D. and Raftery, A. E. (1993) Model-based gaussian and non-gaussian clustering. *Biometrics* **49(3),** 803–821.

Bye, E. (1996) Chemometrics in occupational hygiene-how and why! A picture can tell more than a thousand words and figures!. *Annals of Occupational Hygiene* **40(2),** 145–169.

Calinski, T. and Harabasz, J. (1974) A dendrite method for cluster analysis. *Communications in Statistics* **3(1),** 1–27.

Everitt, B. S. (1993) Hierarchical clustering techniques. In: *Cluster Analysis*, Third edition, chapter 4, pp. 55–89. John Wiley & Son, New York. (ISBN 0 470 22043 0).

Everitt, B. S. (1994) *A Handbook of Statistical Analyses Using S-Plus*. First edition, Chapman & Hall, London. (ISBN 0 412 56310 X).

Hines, C. J., Selvin, S., Samuels, S. J., Hammond, S. K., Woskie, S. R., Hallock, M. F. and Schenker, M.B. (1995) Hierarchical cluster analysis for exposure assessment of workers in the semiconductor health study. *American Journal of Industrial Medicine* **28,** 713–722.

Hubert, L. and Arabie, P. (1985) Comparing partitions. *Journal of Classification* **2,** 193–218.

Jain, A. K. and Dubes, R. C. (1988) Cluster validity. In: *Algorithms for Clustering Data*, chapter 4. pp. 143–222. Prentice Hall, Englewood Cliffs, New Jersey. (ISBN 0 13 022278 X)

Kromhout, H., Symanski, E. and Rappaport, S. M. (1993) A comprehensive evaluation of within- and between-worker components of occupational exposure to chemical agents. *Annals of Occupational Hygiene* **37(3),** 253–270.

Milligan, G. W. and Cooper, M. C. (1985) An examination of procedures for determining the number of clusters in a data set. *Psychometrika* **50(2),** 159–179.

Milligan, G. W. and Cooper, M. C. (1988) A study of standardization of variables in cluster analysis. *Journal of Classification* **5,** 181–204.

Milton, D. K., Wypij, D., Kriebel, D., Walters, M. D., Hammond, S. K. and Evans, J. S. (1996) Endotoxin exposure-response in a fiberglass manufacturing facility. *American Journal of Industrial Medicine* **29,** 3–13.

Peretz, C., Goldberg, P., Kahan, E., Grady, S. and Goren, A. (1997) The variability of exposure over time: a prospective longitudinal study. *Annals of Occupational Hygiene* **41(4),** 485–500.

Post, W., Kromhout, H., Heederik, D., Noy, D., Duijzentkunst, R. S. (1991) Semiquantitative estimates of exposure to methylene chloride and styrene: the influence of quantitative exposure data. *Applied Occupational and Environmental Hygiene* **6(3),** 197–204.

Rappaport, S. M. (1991) Assessment of long-term exposures to toxic substances in air. *The Annals of Occupational Hygiene* **35(1),** 61–121.

Rappaport, S. M., Kromhout, H. and Symanski, E. (1993) Variation of exposure between workers in homogeneous exposure groups. *American Industrial Hygiene Association Journal* **54(11),** 654–662.

Rappaport, S. M., Lyles, R. H. and Kupper, L. L. (1995) An exposure-assessment strategy accounting for within- and between-worker sources of variability. *Annals of Occupational Hygiene* **39(4),** 469–495.

Sahl, J. D., Kelsh, M. A., Smith, R. W. and Aseltine, D. A. (1994) Exposure to 60 Hz magnetic fields in the electric utility work environment. *Bioelectromagnetics* **15,** 21–32.

Schaffer, C. M. and Green, P. E. (1996) An empirical comparison of variable standardization methods in cluster analysis. *Multivariate Behavior Research* **31(2),** 149–167.

Simmons, B. P. and Spear, R. C. (1993) Source identification for multiple chemical exposure using pattern recognition and classification techniques. *Environmental Science & Technology* **27(12),** 2430–2434.

Venables, W. N. and Ripley, B. D. (1994) *Modern Applied Statistics with S-Plus*. Springer Verlag, New York, pp. 313–315. (ISBN 0 387 94350 1).

Walters, M. D. (1993) *Worker Exposure to Endotoxin and Other Contaminants in Fiberglass Insulation Manufacturing*. Dissertation, The Harvard School of Public Health, Boston, Massachusetts.