ELSEVIER

European Journal of Operational Research 93 (1996) 402–417

EUROPEAN JOURNAL OF OPERATIONAL RESEARCH

A comparison of SOM neural network and hierarchical clustering methods

Paul Mangiameli [a,*], Shaw K. Chen [a], David West [b]

[a] *Department of Management Science, College of Business Administration, The University of Rhode Island, Kingston, RI 02881, USA*
[b] *Department of Decision Sciences, School of Business, East Carolina University, Greenville, NC 27858, USA*

Received 1 July 1994; revised 1 August 1995

Abstract

Cluster analysis, the determination of natural subgroups in a data set, is an important statistical methodology that is used in many contexts. A major problem with hierarchical clustering methods used today is the tendency for classification errors to occur when the empirical data departs from the ideal conditions of compact isolated clusters. Many empirical data sets have structural imperfections that confound the identification of clusters. We use a Self Organizing Map (SOM) neural network clustering methodology and demonstrate that it is superior to the hierarchical clustering methods. The performance of the neural network and seven hierarchical clustering methods is tested on 252 data sets with various levels of imperfections that include data dispersion, outliers, irrelevant variables, and nonuniform cluster densities. The superior accuracy and robustness of the neural network can improve the effectiveness of decisions and research based on clustering messy empirical data.

Keywords: Neural networks; Cluster analysis; Self organizing maps; Unsupervised learning

1. Introduction

Cluster analysis is an important statistical methodology used in a wide variety of fields including artificial intelligence, business, biology, psychology, and medicine. Cluster analysis involves grouping similar objects into distinct, mutually exclusive subsets referred to as clusters. Anderberg [1] (p. 11) states that: "the objective is to group either the data units or the variables into clusters such that elements within a cluster have a high degree of 'natural association' among themselves while the clusters are 'relatively distinct' from one another."

The cluster definition problem is NP-complete; thus no efficient and optimal algorithm exists to solve this problem. A number of heuristic methods that provide approximate solutions have been proposed. One group of heuristic methods, hierarchical agglomerative techniques, is commonly used today for cluster analysis. Although all the hierarchical methods work well for data sets with compact, isolated clusters, they fail to accurately define clusters for "messy data", that is data sets that depart from these ideal conditions [12,20]. It has also been shown

* Corresponding author. E-mail: mangia@uriacc.uri.edu

PII S0377-2217(96)00038-0

that the performance of any single hierarchical clustering heuristic is dependent on the empirical data conditions, and that no single method is robust over a wide range of data conditions [12,20]. Recognizing that most empirical data does not conform to the ideal conditions of compact, distinct natural groupings, Helsen and Green [5] (p. 1139) suggest: "the objective of cluster analysis is to define districts in the multidimensional variables space rather than compact and isolated clusters."

We need improved clustering methods that accurately identify cluster membership and that are robust for a wide range of "messy data" conditions. The purpose of this research is to investigate the Self Organizing Map (SOM) network as an alternative to conventional hierarchical clustering methods used today. We study the ability of SOM neural networks to correctly identify cluster membership in "messy data" sets that include different levels of data dispersion, the inclusion of irrelevant variables, the presence of outliers, and different cluster densities. For each data set, the cluster membership accuracy of the SOM network is compared to the accuracy of seven popular hierarchical clustering algorithms including: single linkage, complete linkage, average linkage, centroid method, Ward's method, two stage density, and Kth nearest neighbor. The performance of each technique is measured by the percentage of data points correctly assigned cluster membership. This research is a broad, benchmark study of the ability of SOM networks to determine cluster membership. To the best of our knowledge, no such investigation has been done.

In the next section we define the mathematical basis for the cluster definition problem and briefly discuss the hierarchical clustering heuristics commonly used today. We follow this with a discussion of problems that investigators have reported using hierarchical clustering methods with messy empirical data. We then describe the SOM network that we propose as an alternative to hierarchical clustering methods. Lastly, we show the superiority of the SOM network to accurately define cluster structure for 252 data sets. These "messy data" sets include various levels of data dispersion, irrelevant information, outliers, and nonuniform cluster densities.

2. The cluster definition problem

Following Mulvey and Crowder [14] and Rao [17], the clustering problem can be formulated as follows:

$$\min Z = \sum_I \sum_J d_{ij} x_{ij} \tag{1}$$

subject to

$$\sum_J x_{ij} = 1 \quad \forall i, \tag{2}$$

$$\sum_J x_{jj} = m, \tag{3}$$

$$x_{ij} \le x_{jj} \quad \forall i, j, \tag{4}$$

$$x_{ij} \in \{0,1\} \quad \forall i, j, \tag{5}$$

where m is the desired number of clusters; d_{ij} measures the distance or dissimilarity between object i and object j; x_{ij} is a binary variable indicating whether object i is assigned to cluster j; $x_{jj} = 1$ indicates a cluster median occurs at j; I is the set of n objects; and J is the set of eligible medians. Conceptually the cluster problem can be stated as follows: select m points from the set J, constraint (3), and assign all objects in set I to one and only one median, constraints (2) and (4), so that the sum of the distances from all points to their respective cluster median is minimized. The cluster definition problem is inherently combinatorial. The optimal solution is one arrangement from all the possible ways of forming m cluster medians from n objects. This type of problem is NP-complete and it is unlikely that an efficient optimal solution methodology will ever be found. Therefore, many heuristics and suboptimal methods are used for cluster definition.

Cluster analysis involves two distinct problems: 1) the determination of the number of clusters present in the data; and 2) the assignment of data observations to one and only one cluster. The focus of this study is the accuracy of the assignment of data observations to the appropriate clusters given that the number of clusters in the data is known. If the researchers do not have a priori information about the number of clusters in the data, they must first make this determination. There are no known statistical techniques to prove that their choice of the

number of clusters is correct. However, a number of methods such as dendograms, the pseudo F, and the pseudo t^2 can be used to aid their judgment. Often the physical aspects of the problem determine the number of clusters. For example Giuliano [4] uses a Census statistic to define 32 centers in his clustering analysis of subcenters in the Los Angeles region. It is easy to envision other problems involving school districts, post offices, or field services where the physical attributes of the problem provide an a priori definition of the number of clusters.

Hierarchical agglomerative heuristics are commonly used today to assign data observations to clusters. They all assign data to clusters using the same iterative process whereby they start with N clusters where N is the number of data points. They then merge the two most similar clusters to form $N-1$ clusters. On the next iteration, $N-2$ clusters are formed, again by merging the two most similar clusters. This process continues until there is only 1 cluster remaining (containing all N observations). In no way does this agglomerative process determine how many clusters actually exist in the data. None of the hierarchical clustering methods have the ability to determine the appropriate number of clusters in the data. At best, dendograms and pseudo statistics can be used to guide estimates of the actual number of clusters existing in the data.

The agglomerative methods differ only in the decision rules used to merge clusters. For example, in single linkage the merged clusters are the pair with a minimum distance between an observation in one cluster and an observation in the second cluster. Average linkage, another hierarchical method, merges the pair with a minimum average distance between all pairs of observations in the respective clusters. A number of agglomerative methods exist; each is based on a different distance criterion. They all have unique strengths and weaknesses. None are robust across a wide range of data conditions [20]. Ward's minimum variance method has been shown to be one of the better techniques [12]. For this method, the distance is the ANOVA sum of squares between the two clusters, added up over all the variables. Ward's method is known to be biased towards producing clusters with roughly the same number of observations, and is particularly sensitive to outliers [20]. Unfortunately, to obtain the best cluster results, the investigator must have considerable knowledge about the empirical data including the number of natural clusters, the statistical distribution of observations within the natural clusters, the presence of outliers, and the density of observations among the natural clusters. The information required for an intelligent choice of cluster heuristic is usually not available. This is illustrated in the next section, which describes some of the problems field researchers have encountered using clustering techniques.

3. Issues with hierarchical clustering methods

In recent publications hierarchical cluster methods are used to analyze market entry strategies [19], to design group technology manufacturing cells [7], to provide decision support for large scale R&D projects [11], and to increase the effectiveness of credit decisions [6].

Berry et al. [2] use cluster analysis to identify the focus of a factory on the operational requirements imposed by different market segments. They acknowledge that a compact set of clusters does not exist, and that evaluating cluster solutions remains a subjective task. They also report that the hierarchical and K-means clustering algorithms produce different cluster assignments. This misassignment creates ambiguity in decisions necessary for product management. Giuliano [4] uses cluster analysis to define employment subcenters in the Los Angeles region. He refers to problems with the wide dispersion of sizes and locations within each type of subcenter. The number of employment subcenters (clusters) is not obvious, and the assignment of particular centers to subcenters may be in error. The significance of a misclassification is that inappropriate policy measures are used at those centers. Spisak [21] reports the use of cluster analysis to reduce the overpayment of unemployment insurance benefits by the U.S. Department of Labor. He states that diversity in payment performance among the 50 states blurs the data. Again, it is not obvious how many clusters exist. States may be assigned to the wrong clusters. A state that is misclassified by the cluster analysis will use the wrong strategies to reduce overpayment. Chandra et al. [3] identify a generic measure for the

compactness of a block diagonal form matrix used to cluster a binary machinepart matrix. The compactness of the matrix relates to the effectiveness of the formation of manufacturing cells. They report significant dispersion and problems resulting from misassignment of machines to clusters outside their natural groups. In some cases, dispersion caused the complete failure of the clustering heuristic. Helsen and Green [5] investigate the merits of replicated starting configurations for K-means clustering. They apply this technique to data that defines market segments for a new computer system. This market survey uses a constant sum scale to rate the importance of product attributes. Dispersion results naturally from differences in the response of the 319 users surveyed. Other survey problems also contribute to the data dispersion. Respondents may fail to assign all the points to the product attributes, or may simply have cognitive limitations evaluating all product attributes simultaneously. This dispersion creates problems in evaluating the cluster results. It was unclear whether 2 or 3 clusters existed. The 3 cluster solutions produced highly different cluster assignments between each replication. The issue is to decide which of the various alternative solutions is most valid.

In all of these problems, the researcher is concerned that the empirical data does not conform to the "ideal conditions" of distinct cluster structures required for accurate cluster definition. Errors during cluster analysis contribute to ineffective decisions. In the next section we describe the characteristics of the SOM network, which we propose to improve the accuracy of cluster definition.

4. Self organizing map neural networks

Cluster definition is an unsupervised learning problem, which means that the training set H_u has the following properties:

1. Observations from all clusters m are represented in H_u.
2. Subsets of H_u form natural groupings or clusters.
3. The training set H_u is unlabeled, meaning that no a priori information about cluster membership is available.

The last property precludes the use of backpropogation neural architectures, which require labeled information to train the network. However, a class of neural networks employing training algorithms that use self organization, a form of unsupervised learning, can be used for cluster analysis. One such neural network is the Self Organizing Map neural network (SOM network) defined by Kohonen [9].

The SOM network consists of an input layer and the Kohonen layer. The Kohonen layer is usually designed as a two-dimensional arrangement of neurons that maps N-dimensional input to two dimensions, preserving topological order. For the purpose of identifying cluster membership, we use a one-dimensional Kohonen layer. The SOM input layer of neurons is fully connected to the Kohonen layer. The Kohonen layer computes the Euclidean distance between the weight vector for each of the Kohonen neurons and the input pattern. The Kohonen neuron that is closest, (i.e., minimum distance) is the winner with an activation value of one while all other neurons have activations of zero.

The SOM network is trained by an unsupervised competitive learning algorithm, a process of self organization. Consider the Kohonen layer a one-dimensional array of neurons, each of which receives the same input vector $\boldsymbol{X}$. The index i measures the dimensionality of the input vector $\boldsymbol{X}$ such that $i = 1, 2, \ldots, m$. Let the N Kohonen layer neurons be indexed by the numbers $j = 1, 2, \ldots, n$. Any particular Kohonen neuron j, n_j, has an input weight vector $\boldsymbol{W}_j$. The neuron c, n_c, is the neuron with weight vector $\boldsymbol{W}_c$ that is closest to the input signal vector $\boldsymbol{X}$. This distance is calculated as follows:

$$|\boldsymbol{X} - \boldsymbol{W}_c| = \min_j |\boldsymbol{X} - \boldsymbol{W}_j| \quad \forall j. \tag{6}$$

Define N_c as the subset of neurons that includes n_c and its adjacent neighbors. The process of self organization is accomplished as follows:

$$\frac{\mathrm{d}\boldsymbol{W}_j}{\mathrm{d}t} = \alpha(t)(\boldsymbol{X} - \boldsymbol{W}_j) \quad \text{for } j \in N_c, \tag{7}$$

$$\frac{\mathrm{d}\boldsymbol{W}_j}{\mathrm{d}t} = \boldsymbol{0} \quad \text{for } j \notin N_c, \tag{8}$$

where $0 < \alpha < 1$. The magnitude of the learning coefficient $\alpha(t)$ determines how rapidly the system adjusts over time. Typically alpha is decreased as learning proceeds. The neighborhood function that defines N_c starts with a large area and decreases over

time. Kohonen [10] demonstrates that the dynamic adjustments defined by Eqs. (6)–(8) have the following properties:

1. With sufficient time, the set of weights W_1, W_2, ..., W_n become ordered in an ascending or descending sequence.
2. Once the set of weights is ordered it remains so for all subsequent iterations.
3. The point density function of the ordered weights will approximate some function of the input distribution.

The self organization process begins with all network weights (w_{ij}) initialized to a small random value. Training proceeds by repeatedly exposing the network to the entire set of input vectors. For each input X, the neurons compete for the right to respond. The neuron with the weight vector W that is a minimum distance from the input vector X, Eq. (6), is the winner. The weights of this winning neuron are adjusted in the direction of the input vector. The weights of neurons included in the set N_c defined by the neighborhood function are also adjusted. The weight adjustments for all J neurons are calculated according to Eqs. (7) and (8). The result of a single competitive learning step is that the neighborhood of neurons surrounding the winner moves towards the input vector. The learning process continues with the presentation of input vectors in random order until the Kohonen weight vectors stabilize.

4.1. Convergence of weight vectors

For the simplest case of univariate input, Kohonen [10] proves that the value of the weights will converge to a unique limit. The generalization of this result to nonconstant and higher order input is that at equilibrium, the asymptotic value of every weight vector will coincide with the center of gravity of the influence region. The influence region is the weight surface in the vicinity of neuron j defined as follows; if input X is in this influence region, the weight vector W_j of neuron j is modified. The influence region is controlled by the user's definition of the neighborhood function N_c. The asymptotic values of the weight vectors constitute a vector quantization of the pattern space referred to as a Voronoi tessellation [10]. The Voronoi tessellation partitions the weight space $\mathbb{R}^m$ into regions such that

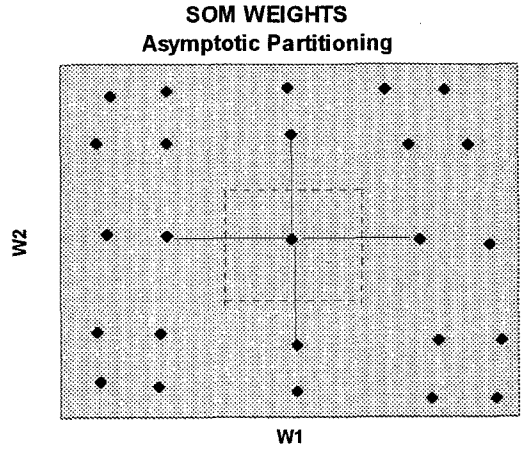

Fig. 1. Partitioning of weight vectors.

all points X within the same region have a single weight vector W_i as their nearest neighbor. The borders of these regions are hyperplanes that are orthogonal to connecting neighboring weight vectors. For a general input form, the asymptotic value of any neuron's weight vector W_i coincides with the centroid of the input signal $\rho(X)$ over the partition P_i of the tessellation. An example of a two-dimensional partition for neuron j is shown in Fig. 1 for uniformly distributed input data. This ability to partition data space is utilized for cluster analysis described in the next section.

4.2. Self organizing maps for cluster analysis

Cluster analysis typically consists of two distinct problems: 1) the determination of the number of clusters present in the data; and 2) the assignment of data observations to one and only one cluster. Our preliminary SOM designs attempted to do both by using a large two-dimensional Kohonen layer that would map the clusters into distinct regions [8]. Unfortunately, the SOM results from our messy and overlapping data consistently mapped all over the Kohonen surface and we were unable to determine the number of clusters in the data. We therefore chose to represent each data cluster with a single unique Kohonen neuron rather than the larger feature map. The resulting Kohonen layer is a one-dimensional map of N neurons, where N is the number of data clusters [18]. This SOM architecture is consis-

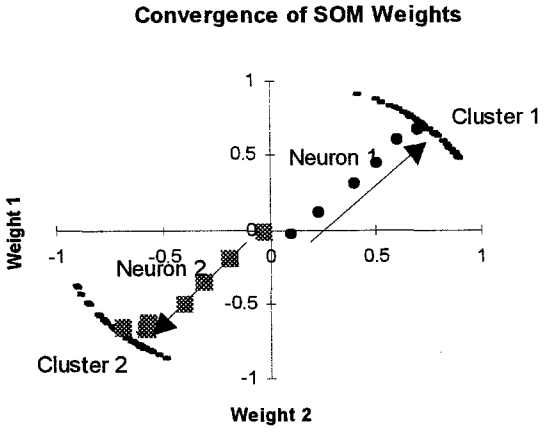

Fig. 2. SOM weight convergence two cluster problem.

tent with the network models described by Pal et al. [16], who also represent each data cluster with a single competitive layer neuron.

During the self organization process, the Kohonen weight vectors obtain asymptotic values that approximate the cluster centroids. These weight vectors partition the data space into regions representing the data clusters. Within any region, all data points are a minimum distance from the same weight vector and are therefore assigned to the same cluster. Since self organization maintains topological order, data clusters that are similar in data space are mapped to Kohonen neurons that are adjacent in the one-dimensional map. This can be important for problems involving three or more clusters. In many problems the cost of misclassification is not constant but depends on the distance between the true cluster and the cluster where the observation is misclassified. In these situations, self organization increases the likelihood that a misclassification will end up in the "next most similar" cluster. A demonstration of the self organization process for a two cluster problem is shown in Fig. 2. The weight vectors progress over time from their initial values near the origin to the final asymptotic values at the cluster centroid. The centroid locations of the two example clusters are $(1, 1)$ and $(-1, -1)$. The data points are normally distributed about the cluster centroids, and the clusters are clearly distinct. The curvature results from normalization of the data, which positions all data observations on the surface of a hypersphere.

The SOM neural network architecture is completely defined by the cluster problem; the number of Kohonen neurons is set to equal the number of clusters in the data set, while the number of dimensions of the input determines the number of input neurons. In this study, a SOM network learning coefficient, α, of 0.06 is used for all data sets. The number of training iterations during the self organization process is thirty times the cluster population, including outliers when present. This is an adequate number of training steps for the weight vectors to converge to their limiting values [15].

5. Experimental design

The purpose of this research is to investigate the accuracy of SOM neural networks in defining cluster membership for test data sets that include various levels of data dispersion combined with outliers, irrelevant information, and unequal cluster densities including both very small and very large clusters. The cluster accuracy of the neural networks is compared to the performance of seven conventional hierarchical clustering algorithms commonly used today. These are single linkage, complete linkage, average linkage, centroid method, Ward's minimum variance, two stage density linkage, and the Kth nearest neighbor density linkage.

The definition of what constitutes a cluster structure is an important operational definition in the development of the data sets. We define cluster structure around the properties of external isolation and internal cohesion. This definition requires that members of one cluster be separated from members of all other clusters by fairly empty areas of space. Internal cohesion requires that entities within the same cluster should be similar to each other. To satisfy this requirement, cluster overlap is not permitted on the first dimension of the space, but is permitted on the remaining dimensions.

We will use the term "messy data" to refer to empirical data that has one or more of the following characteristics: 1) data dispersion that causes some degree of cluster overlap, 2) the inclusion of irrelevant variables that provide no information about the cluster structure, 3) the presence of outliers that

Table 1
Data set design

Data set	Design factors	# of data sets
Base data	2, 3, 4, or 5 clusters; 4,6, or 8 variables; low, med, or high dispersion	36
Irrelevant variables	Basic data plus 1 or 2 irrelevant variables	72
Cluster density	Basic data plus 10% or 60% density	72
Outliers	Basic data plus 10% or 20% outliers	72

belong to no natural grouping of data, and 4) clusters with different sized populations. The definition of these properties and the design of the data sets is consistent with previous research on clustering methodology [5,12].

5.1. Generating cluster data sets

The data sets are constructed to simulate the "messy data" conditions found in empirical data sets. We follow a well established methodology for creating test data sets for cluster analysis [5,12,13]. A total of 252 empirical data sets are used to test the clustering techniques. Thirtysix basic data sets were constructed from design factors that include the number of clusters (2, 3, 4, or 5 clusters), the number of variables in the Euclidean space (4, 6, or 8-dimensional space), and the level of intracluster dispersion (low, medium, and high dispersion). To each of the 36 base data sets we add the following effects: two levels of outliers (20%, and 40% outliers), two levels of irrelevant information (1 and 2 irrelevant variables), and two different cluster densities (10% and 60% cluster density). The data set construction is summarized in Table 1.

To compare the cluster accuracy of the SOM network and the hierarchical clustering algorithms, we must control the level of cluster dispersion in the experimental data. This is accomplished by fixing the cluster centroids and varying the dispersion of observation about the centroids. Following is a summary of this procedure.

1. Randomly fix the cluster centroid. This centroid will remain fixed for all levels of cluster dispersion.
2. Use a multivariate normal random number generator to disperse observations about the fixed cluster centroids. Higher levels of relative dispersion are obtained by using larger variances to generate the cluster members. This expands the distance of the members from the cluster centroid and creates cluster overlap.

The 252 data sets are constructed as follows. For each cluster, a range value is randomly generated for each dimension of space. This range value is a random variable uniformly distributed between 10 and 40 units. The centroid of each cluster is the midpoint of this range. The distances between clusters are constant, independent of the level of dispersion. The average distance between clusters is summarized in Table 2.

Prior research [5,12] as well as our own preliminary investigation shows that clustering algorithms are not sensitive to the number of data observations. We therefore follow previous clustering methodologies [5,12,13] by setting the total number of observations in the data set to 50 times the number of clusters. The population is distributed to individual clusters according to the cluster density assumptions discussed later. Outliers are not considered members of the cluster population. The points within a cluster are normally distributed about the cluster centroid. Each observation is tagged with a cluster ID to facilitate the identification of cluster membership. Differences among the 3 levels of dispersion are caused by the value used for the standard deviation in the normal distribution to generate the cluster points. At the medium level of dispersion, the stan-

Table 2
Average distance between clusters

Number of clusters	Average distance (units)
2	5.03
3	5.31
4	5.33
5	5.78

Table 3
Basis for data set construction

Level of dispersion	Avg. cluster std. dev. (in units)	Range in std. dev.
High	7.72	3
Medium	3.72	6
Low	1.91	12

dard deviation is equal to 1/6 of the cluster range for that dimension. Note that the cluster ranges are a random variable, and therefore the standard deviations within the cluster will be different for each dimension and will vary in proportion to the range. In order to ensure separation between clusters on the first dimension, the standard deviation in this dimension is maintained at 0.5 for all clusters. Data sets for the high and low dispersion cases were generated with the same technique but using standard deviations that are 1/3 and 1/12 of the cluster length respectively. This provides levels of dispersion relative to the medium case that are factors of 2 and 1/2. The basis for data set construction is summarized in Table 3.

Two different levels of outliers are added to each of the 36 basic data sets creating an additional 72 data sets. At the 20% outlier level each cluster has 10 outliers, normally distributed about the cluster centroid with a standard deviation that is nine times the standard deviation for that dimension of space. This ensures that all outliers fall outside the cluster boundaries. With 40% outliers, each cluster has 20 outliers. The outliers are too distant from the cluster boundaries to be members of the cluster. These outliers are not directly considered in the calculation of cluster membership accuracy. Rather, we measure the effect that the presence of outliers has on the accuracy of each clustering algorithm in assigning cluster membership to the "true cluster members".

Two levels of irrelevant variables are added to the 36 basic data sets producing 72 more data sets. Irrelevant variables involve the inclusion of additional variables to the basic data sets of 4, 6, or 8 variables. These irrelevant variables are uniformly distributed across the data space and therefore provide no cluster membership information.

An additional 72 data sets result from altering the uniform cluster density (50 points per cluster) of the 36 basic data sets. For the 10% cluster density, the observations in cluster one are reduced to 10% of the total population where the total population is 50 times the number of clusters. The number of observations in the remaining clusters is increased uniformly to maintain the population totals. With a 60% cluster density, the observations in cluster one are increased to a number that represents 60% of the population total. The remaining clusters have their observations reduced uniformly to maintain the population total.

5.2. Measuring cluster assignment accuracy

Notice there is always overlap on some variables regardless of the level of dispersion. As the average cluster standard deviation increases, the dispersion increases as does the amount of cluster overlap. Cluster structure, therefore, becomes increasingly difficult to identify. The addition of the other effects (outliers, irrelevant variables, and cluster density) compounds the problem of defining the original cluster structure. The performance of each cluster algorithm and the SOM neural network is measured by its ability to correctly identify the cluster membership for each observation in the data set. Cluster assignments for each of the hierarchical methods is determined using Proc Cluster from the SAS Institute. The cluster assignments of the SOM neural network is determined using Neural Ware's "NeuralWorks Professional II/PLUS" software. The cluster assignment results are scanned for misclassifications. For example, in the two cluster uniform density case there are 50 observations belonging to cluster 1, and 50 that belong to cluster 2. Any cluster 1 observation assigned to cluster 2 is a misclassification as is any cluster 2 observation assigned to cluster 1. If there is a total of two misclassifications in the results, the cluster accuracy is 98%

6. Results and discussion

The results of the experimental cluster analysis are expressed as the percentage of cluster observations that are correctly classified, that is assigned to

Table 4
SOM neural network cluster recovery rank

Dispersion	Low	Medium	High	Total	Percent
First	29	59	73	161	63.9%
Tied for first	24	5	1	30	11.9%
Second	19	9	1	29	11.5%
Not in top 2	12	11	9	32	12.7%

the correct cluster. Each of the 252 data sets was tested with the SOM neural network and with the following seven popular clustering methods: single linkage, complete linkage, average linkage, centroid method, Ward's method, two stage density, and Kth nearest neighbor. The dominance of the SOM neural network in identifying the correct cluster membership is evident in Table 4. This table summarizes the rank of the SOM neural network cluster definition for each of the 252 data sets. A rank of first means no other technique was as accurate as the SOM network in identifying cluster membership. A rank of tied for first means that one or more of the other methods did as well as the SOM neural network, but

Table 5
Cluster definition results for base data percent correctly assigned

Dispersion	Low				
# Clusters	2	3	4	5	Average
Single linkage	67.7	84.4	75.1	80.3	76.9
Complete linkage	95.7	97.1	93.7	94.8	95.3
Average linkage	99.3	99.6	96.8	97.2	98.2
Centroid method	99.0	99.3	99.2	86.5	96.0
Ward's method	99.3	99.6	96.8	97.2	98.2
Two stage density	100.0	99.8	97.0	97.5	98.6
Kth neighbor	51.0	67.3	67.3	86.7	68.1
SOM network	100.0	99.3	95.0	97.7	98.0
Dispersion	Medium				
# Clusters	2	3	4	5	Average
Single linkage	51.0	34.9	26.3	21.2	33.4
Complete linkage	85.7	72.4	61.5	59.9	69.9
Average linkage	51.7	44.9	35.3	41.6	43.4
Centroid linkage	51.3	34.4	28.5	32.7	36.7
Ward's method	95.3	96.0	79.2	74.4	86.2
Two stage density	64.3	73.8	38.5	43.6	55.1
Kth neighbor	51.0	34.4	25.8	20.8	33.0
SOM network	100.0	100.0	93.0	90.9	96.0
Dispersion	High				
# Clusters	2	3	4	5	Average
Single linkage	51.0	34.2	26.0	21.3	33.1
Complete linkage	58.7	43.6	40.5	38.1	45.2
Average linkage	51.7	35.8	29.8	28.9	36.6
Centroid method	51.3	35.3	27.3	22.1	34.0
Ward's method	61.0	51.6	42.0	46.9	50.4
Two stage density	51.0	34.2	26.0	21.3	33.1
Kth neighbor	51.0	34.4	26.2	21.2	33.2
SOM network	100.0	80.9	77.3	71.9	82.5

none did better. A rank of second indicates that one other technique was more accurate. For the last rank, Not in Top 2, at least two better performances existed.

The SOM neural network determines cluster membership more accurately than any of the other seven techniques in 63.9% of the data sets. For 75.8% of the data sets, none of the other methods could beat the SOM neural network performance. The neural network does not achieve a rank of first or second in only 12.7% of the data sets. It is also evident that the dominance of the neural network increases as the level of dispersion in the data increases. For the high dispersion level, the SOM neural network was first or tied for first in 88% of the data sets. High dispersion levels are characteristic of field research data encountered in many research applications. While the ranking of the SOM network may not seem impressive for the low dispersion level, many techniques achieve near perfect cluster definition at low dispersion. The opportunity to dominate is therefore reduced.

6.1. Cluster accuracy for base data sets

The base data sets investigate the effect of data dispersion on cluster accuracy. The results are summarized for low, medium, and high levels of dispersion in Table 5. The results tables define the percentage of data observations that are assigned the correct cluster membership for each of the eight clustering methods investigated. The results are reported for data sets with 2, 3, 4, and 5 clusters. Each cluster result is an average of 4, 6, and 8 variables. The average classification reported is an arithmetic average across the cluster levels.

Even at low levels of dispersion, two hierarchical techniques, single linkage and Kth nearest neighbor, are inaccurate with recoveries of 76.9% and 68.1%. This performance is substantially below the 95–98% recovery of all other methods. The average recovery for the SOM network is 98.0%. The average recovery for the five conventional methods excluding single linkage and Kth nearest neighbor is 97.2%. For all cluster levels, the performance of the SOM network is comparable to the hierarchical methods. Any clustering technique except single linkage and Kth nearest neighbor performs well for data with low dispersion.

The performance of the hierarchical methods changes dramatically at medium levels of dispersion. The average accuracy of the same five conventional methods decreases to 58.3% while the SOM network shows only slight degradation to 96.0%. Only Ward's method and the SOM network produce respectable accuracy for all cluster levels. The SOM network outperforms Ward's method at every level from 2–5 clusters. The SOM performance is also robust, decreasing from 100% at two clusters to 90.9% at five clusters. Ward's method achieves a 95.3% recovery at two clusters but falls to 74.4% recovery at five clusters.

At high levels of data dispersion the accuracy of the five conventional methods (excluding single linkage and Kth nearest neighbor) is only 39.8%. The SOM network achieves an average accuracy of 82.5% for these same conditions. At high levels of data dispersion the SOM network is significantly more accurate than any hierarchical method at every cluster level.

At low levels of dispersion the cluster definition accuracy of the SOM network is comparable to the better hierarchical clustering methods. As data dispersion increases to moderate and high levels, the SOM network is the only technique that accurately defines cluster structure. The SOM network provides robust performance for a wide range of data dispersion.

The cluster classification results for the medium level of dispersion with outliers, irrelevant variables, and nonuniform cluster density are reported in Tables 6–8 respectively. Space limitations preclude printing comparable tables for the low and high dispersion levels. These results are available from the authors and are briefly described in the following subsections.

6.2. Cluster accuracy for data sets with outliers

Table 6 reports the cluster performance at the medium level of data dispersion for two levels of data outliers: 20% of observations are outliers and 40% of observations are outliers. As the proportion of outliers in the data sets increases from 20% to

Table 6
Cluster definition results with outliers and medium dispersion level percent correctly assigned

Outliers	20%				
# Clusters	2	3	4	5	Average
Single linkage	50.0	44.4	25.0	20.0	34.9
Complete linkage	50.0	44.4	25.0	20.0	34.9
Average linkage	50.0	44.4	25.0	20.0	34.9
Centroid linkage	50.0	44.4	25.0	20.0	34.9
Ward's method	50.0	44.4	25.0	20.0	34.9
Two stage density	63.7	72.4	45.2	54.8	59.0
Kth neighbor	50.0	44.4	25.0	20.0	34.9
SOM network	83.3	88.9	91.7	81.2	86.3
Outliers	40%				
# Clusters	2	3	4	5	Average
Single linkage	50.0	33.3	25.0	20.0	32.1
Complete linkage	50.0	33.3	25.0	20.0	32.1
Average linkage	50.0	33.3	25.0	20.0	32.1
Centroid method	50.0	33.3	25.0	20.0	32.1
Ward's method	50.0	33.3	25.0	20.0	32.1
Two stage density	63.7	73.6	45.2	43.1	56.4
Kth neighbor	50.0	33.3	25.0	20.0	32.1
SOM network	83.7	88.7	91.5	70.0	83.5

40%, the average accuracy of the SOM network decreases slightly from 86.3% to 83.5%. Except for the five cluster levels, there is no difference in the accuracy of the SOM network at 20% and 40% outlier levels. With outliers in the data, the best hierarchical clustering method is Two Stage Density

Table 7
Cluster definition results with irrelevant variables and medium dispersion percent correctly assigned

Irrelevant	1 variable				
# Clusters	2	3	4	5	Average
Single linkage	51.0	34.7	26.0	21.1	33.2
Complete linkage	60.0	78.4	62.5	65.2	66.5
Average linkage	50.3	47.1	41.2	62.1	50.2
Centroid method	50.3	34.4	27.0	22.8	33.6
Ward's method	91.0	88.7	77.3	78.0	83.8
Two stage density	50.3	58.9	31.5	43.5	46.1
Kth neighbor	50.3	34.2	26.2	21.1	33.0
SOM network	100.0	97.6	74.5	63.9	84.0
Irrelevant	2 variables				
# Clusters	2	3	4	5	Average
Single linkage	50.3	34.7	26.3	21.2	33.1
Complete linkage	65.0	60.2	53.7	59.6	59.6
Average linkage	50.0	54.9	49.3	54.7	52.2
Centroid method	50.3	34.7	28.0	21.5	33.6
Ward's method	80.7	90.0	70.2	76.3	79.3
Two stage density	50.3	52.9	25.8	39.9	42.3
Kth neighbor	50.3	34.4	25.8	21.3	33.0
SOM network	100.0	90.7	69.0	57.2	79.2

with averages of 59.0% and 56.4%. The presence of data outliers confounds the hierarchical cluster algorithms and causes chaining. Chaining results in the formation of one large cluster that contains most of the data with the remaining clusters having only a few observations. Therefore, the results of the hierarchical methods are almost identical, and represent a lower limit of performance achieved by assigning most of the data observances to a single cluster.

At low dispersion and 20% outliers the SOM network average is 99.4% versus 98.5% for Two Stage Density. Ward's is 56.1%, and no other method is better than 32.1%. At low dispersion and 40% outliers the SOM network average is 91.3%, and the Two Stage Density average is 98.5%. No other technique is above 41% accuracy. At high levels of dispersion, the SOM network clearly dominates all other methods; for cluster levels of 2, 3, and 4 it achieves perfect classification. The hierarchical methods, including Two Stage Density, have poor capability to determine cluster membership at high dispersion. The average accuracy of the hierarchical methods is 32.1% at both outlier levels. The SOM average accuracy is 94.3% and 94.8% at 20% and 40% outlier levels.

The SOM network is clearly superior at defining cluster structure when outliers are present in the data. The average recovery of the SOM varies between 83.5% and 99.4%, and its accuracy is robust as data dispersion increases. The accuracy of the hierarchical methods is limited by the tendency for chaining. The only acceptable performance for a hierarchical method is achieved by Two Stage Density at low levels of dispersion.

6.3. Cluster accuracy for data sets with irrelevant variables

Detail results for medium data dispersion with one and two irrelevant variables are reported in Table 7. The average performance of the SOM network is 84% and 79.2% with one and two irrelevant variables. The performance of Ward's method is comparable to the SOM network with average recoveries of 83.8% and 79.3%. The SOM network is better at cluster levels of 2 and 3; Ward's method is better at cluster levels of 4 and 5. The performance of all other methods ranges from 33% to 66%.

With low levels of dispersion and one irrelevant variable, the SOM network has an average accuracy of 90.9% and Ward's method 91.1%. However, with two irrelevant variables, the SOM network maintains an accuracy of 88.3% while Ward's method falls to 56.3%. The best performance of the hierarchical methods is by Two Stage Density at 75%. At high levels of dispersion, the performance of the SOM network is dominant at every cluster level. Its average accuracy for one irrelevant variable is 74.0% with Ward's method second at 48.3%. For two irrelevant variables the SOM network averages 64.7% Vs 48.2% for Ward's method.

Cluster definition accuracy decreases when some of the variables in the data set are irrelevant, meaning they provide no significant information about cluster membership. Under these conditions the SOM network is again superior in identifying cluster accuracy. The SOM average ranges from 90.9% at low dispersion and one irrelevant variable to 64.7% at high dispersion and two irrelevant variables. Comparable performances are achieved by hierarchical methods only under the following extremely limited conditions: Two Stage Density at low dispersion and 1 irrelevant variable (91.1%), and Ward's at medium levels of dispersion (83.8%, 79.2%). Not only is the SOM network most accurate, it maintains robust performance for a wide range of data that includes irrelevant variables. No other method demonstrates robust capabilities.

6.4. Cluster accuracy for data sets with nonuniform density

The cluster density levels test the accuracy of the cluster method when one cluster is extremely small (10% of observations) and when one cluster is extremely large (60% of observations). The cluster density results are reported in Table 8. At the 10% cluster density and medium data dispersion, the SOM network achieves an average cluster classification accuracy of 88.4%. Ward's method and Two Stage Density are next at 78.0% and 77.5%. There is significant variability for both of these methods at different cluster levels. Ward's method starts at 66.7% for two clusters and increases continuously to 82.4% at five clusters. Two Stage Density does just the opposite with an 89.7% accuracy at two clusters, decreasing continuously to 62.8% at five clusters. The accuracy of the SOM network is stable ranging

Table 8
Cluster definition results with cluster density and medium dispersion percent correctly assigned

Density	10%				
# Clusters	2	3	4	5	Average
Single linkage	90.3	48.0	31.3	28.4	49.5
Complete linkage	60.7	73.0	68.5	67.5	67.4
Average linkage	90.3	75.1	52.8	75.2	73.4
Centroid method	90.3	46.2	43.0	32.8	53.1
Ward's method	66.7	80.7	82.2	82.4	78.0
Two stage density	89.7	93.1	64.5	62.8	77.5
K th neighbor	91.0	47.8	31.2	28.0	49.5
SOM network	89.0	83.1	90.8	90.5	88.4
Density	60%				
# Clusters	2	3	4	5	Average
Single linkage	60.3	61.1	60.8	61.1	60.8
Complete linkage	74.3	68.2	44.8	50.5	59.5
Average linkage	60.3	75.6	81.3	74.3	72.9
Centroid method	60.3	61.3	63.3	66.5	62.9
Ward's method	93.0	87.6	71.3	64.5	79.1
Two stage density	72.0	75.6	60.2	59.5	66.8
K th neighbor	60.3	61.1	57.8	60.9	60.0
SOM network	100.0	93.6	72.7	54.7	80.3

from 83.1% to 90.5%. At the 60% cluster density level the SOM network averages 80.3% accuracy. This is slightly better than Ward's method at 79.1%. The accuracy of both methods decreases significantly from two clusters to five clusters. The SOM network accuracy is better at cluster levels of 2, 3, and 4; Ward's method is better at the five cluster level.

At low data dispersion levels and 10% cluster density, all cluster methods achieve high levels of accuracy. The lowest accuracy is complete linkage at 91.6%, and the highest is the SOM network at

Table 9
Network weights and cluster centroids, high dispersion, 2 clusters, 6 variables

Cluster	#1					
Variable	1	2	3	4	5	6
$\alpha = 0.05$	0.7188	−0.0044	0.1630	−0.5761	0.1316	0.3215
$\alpha = 0.1$	0.7180	−0.0039	0.1644	−0.5754	0.1314	0.3210
$\alpha = 0.2$	0.7175	−0.0042	0.1655	−0.5753	0.1313	0.3209
$\alpha = 0.5$	0.7276	−0.0048	0.1643	−0.5759	0.1330	0.3202
Centroid	0.7184	−0.0043	0.1654	−0.5761	0.1328	0.3157
Cluster	#2					
variable	1	2	3	4	5	6
$\alpha = 0.05$	−0.7091	−0.3185	−0.0762	0.0286	−0.0733	−0.3389
$\alpha = 0.1$	−0.7093	−0.3197	−0.0766	0.0292	−0.0733	−0.3387
$\alpha = 0.2$	−0.7094	−0.3212	−0.0761	0.0295	−0.0735	−0.3389
$\alpha = 0.5$	−0.7096	−0.3220	−0.0740	0.0306	−0.0739	−0.3397
Centroid	−0.7006	−0.3166	−0.0736	0.0315	−0.0740	−0.3405

97.6%. At low dispersion and 60% cluster density, the SOM network achieves its poorest relative performance. The average accuracy of 85.9% ranks seventh of the eight methods. This weakness is traced to low accuracy for cluster levels of four and five (78.8% and 65.3%). The SOM accuracy for cluster levels of two and three is outstanding (100% and 99.6%). The overall accuracy of 85.9% is still good in the context of the overall experimental results. With high data dispersion and 10% cluster density the SOM accuracy dominates all other methods. Its average accuracy is 72.4% while second best performance is Two Stage Density at 52.4%. Except for the two cluster level, the SOM accuracy is significantly better than all other methods. For high dispersion and a 60% cluster density the SOM average is 67.5%. Average linkage is second with 62.4%. The SOM network is perfect at the two cluster level, but again has trouble with four and five clusters achieving an accuracy of 52% and 49.5%. The accuracy of Average Linkage under these same conditions is 65% and 61.3%.

Overall, the SOM network is the most accurate technique for data conditions in which the cluster densities are nonuniform. The SOM average ranges from 97.6% at low dispersion and 10% density to 67.5% at high dispersion and 60% density. The accuracy of the hierarchical methods is not affected by cluster density at low dispersion, but their accuracy deteriorates significantly at medium and high dispersion levels.

6.5. Comparison of SOM and centroid results

We might question the performance differences between the SOM network and the centroid method. The centroid method uses distance between cluster centroids to determine similarity and therefore the clusters to be merged during the next iterative process. The SOM network self organizes the weight vectors to approximate the cluster centroids. In fact, there is a very important distinction between the two which accounts for the different results. When the centroid method (or any hierarchical clustering method) makes an assignment of a data observation to a cluster, it is permanent and irreversible. At each iteration, the centroid method merges the two clusters whose centroid is a minimum distance. If this merger is incorrect, there is no opportunity to undo it at future iterations. Consider a situation where two data points from different clusters are close together because they are both extreme distances from their respective cluster centroids. The centroid method would combine these two observations into a single cluster and create a misclassification that remains in all future iterations. With the SOM network, observations are continuously reassigned as the network weight vectors gradually converge to the cluster centroids. Cluster assignments are not final until the learning process terminates.

6.6. Sensitivity of SOM network

A common problem with neural networks is that results are not repeatable. This is usually attributed to the stochastic nature of the network and the propensity for local optima. We investigated the sensitivity of the cluster results to the SOM network learning coefficient, α, for the high dispersion data level. The high dispersion level was chosen because it represents the most severe data conditions to test the repeatability of results. Two significant conclusions of the sensitivity analysis are:

1. The cluster results are insensitive to learning rates that vary from 0.05 to 0.5 in the self organization process.
2. The cluster results are repeatable. In all cases the weight vectors self organize to closely approximate the cluster centroids. The variability between different cluster runs is minimal.

The sensitivity analysis was conducted for all levels of clusters and variables. To simplify the presentation, Table 9 shows the results for the two cluster, six variable data case only. Other data conditions exhibit similar results. The asymptotic weight vectors are reported for SOM networks with initial learning coefficients of 0.05, 0.1, 0.2, and 0.5. The known cluster centroid is also reported. Note how closely the SOM network weights approximate the known cluster centroid and that there is very little difference in the asymptotic weights for the various levels of the learning coefficient. This ability results in accurate definition of cluster structure, even with "messy data".

7. Conclusion

The SOM network can improve the quality of decisions that require the cluster analysis of "messy data" such as market segmentation, credit analysis, quality problems, and operations problems. Identifying cluster membership in messy empirical data is a difficult problem that is confounded by data imperfections such as dispersion, outliers, irrelevant information, and nonuniform cluster densities. The experimental results are unambiguous; the SOM network is superior to all seven hierarchical clustering algorithms commonly used today. Furthermore, the performance of the SOM network is shown to be robust across all of these data imperfections. The SOM superiority is maintained across a wide range of "messy data" conditions that are typical of empirical data sets. Additionally, as the level of dispersion in the data increases, the performance advantage of the SOM network relative to the hierarchical clustering methods increases to a dominant level.

The SOM network is the most accurate method for 191 of the 252 data sets tested, which represents 75.8% of the data. The SOM network ranks first or second in accuracy in 87.3% of the data sets. For the high dispersion data sets, the SOM network is most accurate 90.2% of the time, and is ranked first or second 91.5% of the time. The SOM network frequently has average accuracy levels of 85% or greater, while other techniques average between 35% and 70%. Of the 252 data sets investigated, only six data sets resulted in poor SOM results. These six data sets occurred at low levels of data dispersion, with a dominant cluster containing 60% or more of the observations and four or more total clusters. Despite the relatively poor performance at these data conditions, the SOM network did average 72% of observations correctly classified.

Not only is the SOM network effective, it is easy to use. In fact, the SOM network does not require the analyst to have extensive experience with neural networks. There are no ambiguities concerning network architecture, numbers of neurons, or patterns of connectivities. We show that the cluster definition problem uniquely defines the SOM network architecture. The SOM network is also not sensitive to its starting conditions, and cluster results are therefore repeatable. We also show that cluster results are not sensitive to the initial network learning coefficient. Consistent cluster results are produced with learning coefficients that vary from 0.05 to 0.5.

In summary, our results show the SOM network is more accurate at assigning data observations to clusters for the "messy data" conditions that are typical of empirical field studies. Furthermore, the accuracy of the SOM network is more robust than the hierarchical methods to all the possible data imperfections. Investigators should be cautious using hierarchical agglomerative methods for cluster analysis when the SOM network is more accurate, more robust, and yet is easy to use.

References

[1] Anderberg, M.R., *Cluster Analysis for Applications*, Academic Press Inc., New York, 1973.

[2] Berry, W., Bozarth, C., Hill, T. and Klompmaker, J., "Factory focus: segmenting markets from an operations perspective", *Journal of Operations Management* 10/3 (1991) 363–387.

[3] Chandra, C., Shahrukh, A. and Arora, S., "Clustering effectiveness of permutation generation heuristics for machinepart matrix clustering", *Journal of Manufacturing Systems* 12/5 (1993) 388–407.

[4] Giuliano, G., "Subcenters in the Los Angeles region", *Regional Science and Urban Economics* 21 (1991) 163–182.

[5] Helsen, K. and Green, P.A., "Computational study of replicated clustering with an application to market segmentation", *Decision Science* 22/5 (1991) 1124–1141.

[6] Ibis, D., "Cluster analysis targets prospects", *Credit World* 80/2 (1991) 38–41.

[7] Kamrani, A.K., Parsaei, H.R. and Chaudry, M., "A survey of design methods for manufacturing cells", *Computers and Industrial Engineering* 25 (1993) 487–490.

[8] Kangas, J.A., Kohonen, T. and Laaksonen, J.T. "Variants of self organizing feature maps", *IEEE Transactions on Neural Networks* 1/1 (1990) 93–99.

[9] Kohonen, T., "Adaptive, associative, and self organizing functions in neural computing", *Applied Optics* 26/23 (1987) 4910–4918.

[10] Kohonen, T., *Self Organization and Associative Memory*, Springer-Verlag, New York, 1988.

[11] Mathieu, R., "A methodology for large scale r&d planning based on cluster analysis", *IEEE Transactions on Engineering Management* 40/3 (1993) 283–291.

[12] Milligan, G.W., "An examination of the effect of six types of error perturbation on fifteen clustering algorithms", *Psychometrika* 43/5 (1980) 325–342.

[13] Milligan, G.W., "An algorithm for generating artificial test clusters", *Psychometrika* 50/1 (1985) 123–127.

[14] Mulvey, J.M. and Crowder, H.P., "Cluster analysis: an

application of Lagrangian relaxation", *Management Science* 25/4 (1979) 329–340.

[15] *Neural Computing*, NeuralWare, Pittsburgh, PA., 1993.

[16] Pal, N.R., Bezdek, J.C. and Tsao, E.C.-K., "Generalized clustering networks and Kohonen's self organizing scheme", *IEEE Transactions on Neural Networks* 4/4 (1993) 549–557.

[17] Rao, M.R., "Cluster analysis and mathematical programming", *Journal of the American Statistical Association* 66 (1971) 622–626.

[18] Ritter, H., Martinetz, T. and Schulten, K., "*Neural computation and self organizing maps*", Addison Wesley, Reading, MA, 1992.

[19] Robles, F., "International market entry strategies and performance of united states catalog firms", *Journal of Direct Marketing* 8/1 (1994) 59–70.

[20] *SAS/STAT Users Guide*, Release 6.03 Edition, SAS Institute, Inc., Cary, N.C., 283–357, 1988.

[21] Spisak, A., "Cluster analysis as a quality management tool", *Quality Progress* 25/12 (1992) 33–38.