

Clustering Files of Chemical Structures Using the Fuzzy *k*-Means Clustering Method

John D. Holliday, Sarah L. Rodgers, and Peter Willett*

Krebs Institute for Biomolecular Research and Department of Information Studies, University of Sheffield,
Western Bank, Sheffield S10 2TN, U.K.

Min-You Chen and Mahdi Mahfouf

Department of Automatic Control and Systems Engineering, University of Sheffield, Mappin Street,
Sheffield S1 3JD, U.K.

Kevin Lawson and Graham Mullier

Syngenta, Jealott's Hill International Research Centre, Bracknell RG42 6EY, U.K.

Received November 17, 2003

This paper evaluates the use of the fuzzy *k*-means clustering method for the clustering of files of 2D chemical structures. Simulated property prediction experiments with the Starlist file of logP values demonstrate that use of the fuzzy *k*-means method can, in some cases, yield results that are superior to those obtained with the conventional *k*-means method and with Ward's clustering method. Clustering of several small sets of agrochemical compounds demonstrate the ability of the fuzzy *k*-means method to highlight multicluster membership and to identify outlier compounds, although the former can be difficult to interpret in some cases.

INTRODUCTION

The clustering of chemical structures is a widely used technique that has found application in the selection of compounds for screening, the analysis of substructure search output, and the prediction of molecular properties, inter alia (see, e.g., refs 1–3). Cluster analysis involves three principal components: a similarity coefficient for quantifying the degree of similarity between pairs of compounds, between a compound and a cluster, or between a pair of clusters; a clustering method that processes the similarity data to identify groups of structurally related compounds; and an efficient algorithm for the implementation of the method so that it can be applied to data sets of nontrivial size. There is an extensive literature associated with all of these components (see, e.g., refs 4–6): here, we focus on the choice of clustering method.

There are many different types of clustering method^{7–9} and there have been several detailed comparisons of their effectiveness and efficiency when used for the processing of chemical structure databases, perhaps most notably the work of Willett and his collaborators (as summarized in ref 4) and of Brown and Martin.^{3,10} Studies such as these have resulted in the widespread operational use of the Jarvis-Patrick and Ward's clustering methods. There is, however, a further method whose computational complexity makes it particularly attractive when very large numbers of compounds need to be clustered, as is the case with many database applications. This is the *k*-means method,⁶ which is perhaps the archetypal relocation method¹¹ and which involves an initial partition of a data set that is then refined by iterative relocation of the compounds in the data set.

The *k*-means method, like most of those used previously for chemical clustering, is an example of a *crisp* clustering method, in which each compound is a member of just a single cluster. Alternatively, in a *fuzzy* clustering method, each compound can belong to one, some, or many of the clusters (albeit to a greater or lesser degree, rather than just belonging or not belonging in the case of a crisp classification).^{12–14} There has already been some interest in using these methods for the analysis of chemical data,^{15–23} with many of these studies using the fuzzy *k*-means clustering method that is probably the most common type of fuzzy clustering method and that forms the focus of this paper. Specifically, we report a detailed evaluation of the fuzzy *k*-means method for clustering files of chemical structures and compare its performance with that of established crisp methods using both quantitative and qualitative approaches. Full details of the work are presented by Rodgers.²⁴

THE FUZZY *K*-MEANS CLUSTERING METHOD

Fuzzy clustering is one of the principal applications of fuzzy set theory,^{25–27} an important part of which is the concept of a *membership function*. The membership function of an object describes to what degree that object is a member of a given set. In traditional set theory an object is either a member of a set, corresponding to a membership function of 1, or is not a member of that state, corresponding to a membership function of 0. Fuzzy logic extends this notion by allowing an object to belong to more than one set, with the membership function being allowed to take values between 0 and 1. Formally, given a set *A*, in a space of points, *X*, with a generic element of *X* denoted as *x*, i.e., *X* = {*x*}, then the membership function in a conventional crisp set assigns a value $\mu_A(x)$ to each *x*

* Corresponding author phone: +44-114-2222633; e-mail: p.willett@sheffield.ac.uk.

$$\mu_A(x) = \begin{cases} 1 & \text{if } x \in X \\ 0 & \text{if } x \notin X \end{cases}$$

i.e., $\mu_A(x): X \rightarrow \{0, 1\}$.

In fuzzy set theory, an object is assigned a value for its membership function of a given set anywhere between these two extremal values, i.e., $\mu_A(x): X \rightarrow [0, 1]$. The sum of the memberships for each object is unity, and the closer the value of $\mu_A(x)$ to this upperbound, the greater the degree to which x is a member of the fuzzy set A .

The idea of partial membership that underlies fuzzy set theory provides an obvious way of tackling the inherent problem of conventional clustering methods, where an object can only belong or not belong to a particular cluster. A fuzzy cluster is a fuzzy subset on the set of objects, with the membership function of each object representing the degree to which it belongs to that cluster. If a cluster is a group whose members share common properties, then the membership function of an object indicates the degree to which that object displays these properties, with similar objects having high membership of the same cluster(s).

All of the fuzzy clustering methods that have appeared in the literature are extensions of conventional methods that have incorporated fuzzy set theory, with the fuzzy k -means method being by far the most widely used. The method was first characterized by Dunn²⁸ and then generalized by Bezdek¹² and is simply an iterative procedure for finding memberships of all the objects in the feature space that optimize an objective function.^{13,28–30} The method is summarized below.

1. Select parameter values: k is the number of partitions, m is the fuzziness index, and ϵ is the stopping threshold.
2. Initialize the centroid matrix with the cluster seeds.
3. Calculate the membership, μ_{ij} , of each compound, j , in each cluster, i , using

$$\mu_{ij} = \frac{\left(\frac{1}{\sum_k (x_{jk} - v_{ik})^2} \right)^{1/(m-1)}}{\sum_i \left(\frac{1}{\sum_k (x_{jk} - v_{ik})^2} \right)^{1/(m-1)}} \quad (1)$$

where x_{jk} is the data point of the j th compound at the k th variable and v_{ik} is the centroid value in the i th cluster at the k th variable.

4. Update the positions of the centroids using

$$v_{ik} = \frac{\sum_j (\mu_{ij})^m x_{jk}}{\sum_j (\mu_{ij})^m} \quad (2)$$

5. Calculate the difference between the centroid matrix from the current and previous iterations:

$$\text{If } \sum_{ik} (v_{ik\text{CURRENT}} - v_{ik\text{PREVIOUS}}) < \epsilon$$

then stop, otherwise go to step 3.

In this fuzzy version of the k -means method the objective function (J_m) to be minimized is as follows:

$$J_m = \sum_i \sum_j (\mu_{ij})^m |x_j - v_i|^2 \quad (3)$$

By minimizing the function (3), the objects should have high membership in the resulting clusters. The fuzzy version of the conventional objective function includes two extra terms—the membership function of each object, μ , and an exponent weight, m (or “fuzziness” index). The objective function is the sum of the squared distances between each object and corresponding cluster centroid with the distances weighted by the fuzzy memberships. Thus objects that have a large distance will also have a small membership function for that fuzzy set and so will have less effect on the objective function, J_m . The membership function is first weighted by m , which is present to reduce the sensitivity to small differences in distance. The fuzzy objective function is then minimized, and the method compares the input vectors to the mean vectors for each class to allow adjustments to the partition matrix.

There are three parameters of importance. The first of these is the number of clusters, k , which (as with most relocation clustering methods) must be chosen by the user prior to the start of the clustering process. The weighting exponent, or fuzziness index, m ($1 < m < \infty$), weights the membership values so that the effect of noise on the centroid is reduced. A value of $m = 1$ corresponds to a normal crisp partition, but as m increases toward ∞ , the partition becomes fuzzier and the membership function of each compound in each cluster tends toward the limiting value of $1/k$. The value of m thus has a large impact on the cluster analysis: a value that is too low will not effectively handle noise in the data and a value that is too high will produce very poorly separated clusters. In fact, as will be seen below, all of our experiments have used only small values of m as previous work has shown that this gives the most appropriate results in a wide range of applications (see, e.g., refs 16, 20, 31–33). J_m is a squared error criterion and by minimizing it fuzzy clusters are being produced that are optimal in a generalized least-squares errors manner. The termination criterion is usually set to $\epsilon = 0.001$, but 0.01 can be sufficient. The membership calculation in step 3 of the algorithm has normally involved the Euclidean distance measure, but other similarity coefficients³⁴ could be used if required.

In crisp k -means analysis the Euclidean distance between each of the objects in the data set and the centroids is calculated, and then the object is assigned to the centroid to which it is closest. This can be thought of as “winner takes all” as only the closest cluster is given membership for the object. In fuzzy k -means, membership of each object to each cluster is assigned proportionately depending on this Euclidean distance value, with the closest cluster being assigned the highest membership and the furthest one the least. In this respect fuzzy k -means could be described as a “winner takes most” method.

SIMULATED PROPERTY PREDICTION

The prediction of chemical, biological, and physical properties has been extensively used for the quantitative evaluation and comparison of clustering methods^{3,4}—indeed,

cluster-based and nearest-neighbor searching methods have been suggested as an effective alternative to conventional prediction methods¹⁰—and we have used this approach in the work reported here.

Our experiments used a set of 1763 molecules randomly selected from the Starlist database³⁵ for which logP values are available. The Molconn-Z package³⁶ was used to calculate a wide range of topological indices and other physicochemical parameters for each of the molecules; principal components analysis was then carried out so that each molecule was characterized by the first 45 principal components, these accounting for 77.6% of the variance in the sets of calculated values.

A MATLAB implementation of the *k*-means fuzzy clustering method was used to classify the Starlist data set. This program takes as input the 1763×45 data matrix and produces two matrices as output: the fuzzy partition matrix, which contains the membership function of each compound for each cluster; and the centroid matrix, which contains the centroids (calculated using (2) above) for each of the current set of clusters, and thus the basis for the next iteration of the relocation procedure that drives this clustering method. These two matrices were initialized randomly so as create the centroids for the first iteration of the method. Iteration is continued until the difference, ϵ , between the current and previous centroid matrices is not greater than 0.01.

The results obtained using the fuzzy *k*-means clustering were compared with those obtained using two established crisp clustering methods, conventional *k*-means and Ward's method, with these results being obtained using the Barnard Chemical Information (BCI) implementations of these two methods.³⁷ The comparison involved simulated property prediction, in which the groupings resulting from a cluster analysis were used to predict the properties of the compounds within each cluster. These predicted values were then compared with the corresponding observed values to provide a measure of the effectiveness of the clustering method.^{3,4} The conventional leave-one out procedure used to do this is as follows:

1. Take each molecule, *j*, in turn, and note the cluster, *i*, that contains it.
2. Calculate the mean property value from all the other molecules in cluster *i*.
3. This is the predicted property value for molecule *j*.

Note that step 3 is only carried out if there is at least some minimal number—3 in our experiments—of molecules in the cluster *i*. The set of predicted values is compared with the observed values by calculating the product moment correlation coefficient.⁴ A perfect correlation, and thus an effective clustering, is denoted by a coefficient value of 1; a perfect inverse correlation will be shown by a coefficient value of -1 and a value of 0 shows that there is no correlation between the observed and predicted values. This is a simple way of measuring the effectiveness of clustering that has been used very extensively in the past;^{2–4,10,11} more complex performance measures are discussed by Johnson.³⁸

The correlation coefficients obtained with varying numbers of clusters for the (crisp) *k*-means and Ward's methods are shown in Table 1 with *k*, the number of clusters, in the range $10 \leq k \leq 90$. It will be seen, as would be expected, that the predictive power of the classification rises rapidly and is greatest with the largest number of, and hence the smallest

Table 1. Product Moment Correlation Coefficients for the Prediction of logP Using Crisp Clustering Methods

<i>k</i>	<i>k</i> -means	Ward's
10	0.23	0.17
30	0.59	0.40
50	0.64	0.58
70	0.67	0.64
90	0.71	0.69

Table 2. Product Moment Correlation Coefficients for the Prediction of logP Using the PM4 Prediction Method

<i>k</i>	<i>m</i>									
	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
10	0.40	0.34	0.31	0.23	0.20	0.17	0.17	0.17	0.17	0.16
30	0.62	0.64	0.40	0.26	0.25	0.21	0.19	0.17	0.17	0.16
50	0.68	0.66	0.55	0.26	0.24	0.23	0.18	0.17	0.17	0.17
70	0.73	0.72	0.57	0.28	0.24	0.22	0.18	0.17	0.17	0.17
90	0.74	0.75	0.65	0.32	0.26	0.21	0.17	0.17	0.17	0.17

and tightest, clusters. This approach to measuring the effectiveness of a cluster analysis is only suitable for crisp clusters, and hence needs modification if it is to be applied to the evaluation of fuzzy classifications. We have tested four such modified prediction methods (PM1–PM4), of which the best results were obtained with the last of these, PM4.²⁴ We hence describe the first three only briefly and provide a detailed account plus experimental results (see Table 2) for just PM4.

PM1. The first approach that can be used is to “defuzzify” the fuzzy partition after the clustering has taken place. Each compound is assigned to the cluster in which it has the highest membership function (the “home” cluster) and assigned $\mu = 1$ for that cluster and $\mu = 0$ for all other clusters. The data would then be equivalent to that from a crisp clustering (and can hence be evaluated in the normal way) but will have benefited from being clustered using the fuzzy algorithm.

PM2. The second prediction approach involves assigning each molecule to its home cluster as above but retaining its membership functions in all of the other clusters for the predictions of other compounds. The procedure is outlined below:

1. Apply a threshold to the data, μ_t , below which membership to clusters is ignored.
2. Cluster *i* is the home cluster of compound *j*.
3. Calculate the mean property value for cluster *i* by multiplying each qualifying membership function ($\mu > \mu_t$) by the property value of that compound (excluding compound *j*). These values are then summed and divided by the sum of the qualifying memberships in the cluster.

4. Predict property for *j* from the mean value in cluster *i*. All molecules are thus included in the calculations as long as their membership is high enough, but only the cluster in which compound *j* has the greatest membership is considered for the prediction. The use of μ_t eliminates the many low cluster memberships, while retaining the more important ones; the approach is thus less crude than PM1 but still uses only the home cluster in the final prediction stage.

PM3. The third approach includes the membership functions in the prediction stage:

1. Take each cluster in turn.

2. Calculate the average property value for each cluster by summing the property value multiplied by membership function for each of the molecules and divide this by the sum of the memberships.

3. Multiply the average property value in each cluster by the membership function of compound *j* for that cluster and sum the values from each cluster to obtain an overall property value for *j*.

Here the membership function of each molecule is being considered and thus each molecule contributes more or less to the prediction depending upon the degree to which it is a member of the specific cluster. This approach is more complex than the two previous ones but uses all of the information provided in the fuzzy clustering matrix in the prediction stage.

PM4. The final approach is an extension of PM3 and involves the following steps:

1. Select a minimal membership μ_{min} , the minimum membership function from which to make predictions.

2. For compound *j* put membership functions in descending order. Then sum these memberships until μ_{min} is reached: these are the clusters that will be used for prediction.

3. For the clusters to be used, scale the memberships to 1.

4. Calculate the average property value of each cluster by summing the membership function multiplied by the property value and then dividing by the sum of the memberships for all of the compounds present.

5. Multiply the average property value by the membership function of compound *j* for each of the clusters to be used. These values can then be summed to obtain an overall value.

Many of the membership values for many of the molecules are very small, implying that it may be inappropriate to use the corresponding clusters for predictive purposes. PM4 hence uses just the most important significant memberships for each molecule, together with the properties of all of the other molecules in the data set.

The results obtained using the PM4 approach are shown in Table 2. The fuzzy clusters created with a high value of *m* are not well separated, with the membership functions of the individual molecules here having similar values (and similar behavior is also observed with PM1–PM3). At lower levels of *m*, the predictive ability of the clusters improves as the clusters become better separated, i.e., exhibit a much lower degree of overlap. Indeed, there is some slight improvement over the crisp-cluster results (see Table 1) at all partition levels for *m* = 1.1 and *m* = 1.2. It should be noted that the precise correlations depend on the value that is chosen for the threshold membership function, and experiments were hence carried out with $0.5 \leq \mu_{min} \leq 0.95$ across the range of *m* values. The overall effect was small, with the largest variation in the correlation coefficient (a total range of 0.015 units) occurring at *m* = 1.3 and *m* = 1.8; the correlations peaked at $\mu_{min} = 0.80$ for these *m* values, and we have hence used this value across the whole range of *m* values to obtain the results shown in Table 2.

As noted in the Introduction, one of the principal advantages of fuzzy clustering is that it can describe a wider range of similarity relationships than is possible with a crisp clustering method, where each molecule belongs to only a single cluster. As an example, consider the Starlist molecule labeled 1 in Figure 1. This was assigned to three clusters,

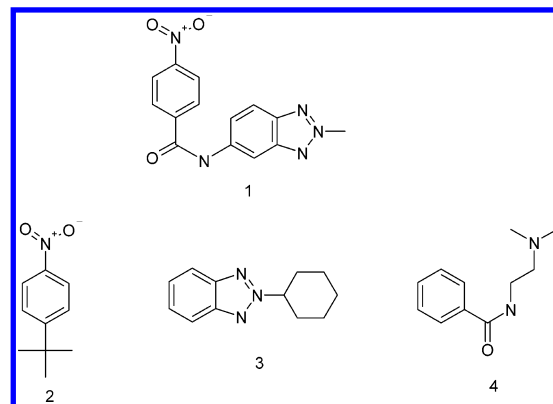


Figure 1. Allocation of a typical Starlist molecule (1) to three different fuzzy clusters with membership functions of 0.289, 0.166, and 0.394. The representative molecules for these three clusters are shown at 2–4.

with membership values of 0.289, 0.166, and 0.394 for the groups represented by the molecules labeled 2–4 in Figure 1. An inspection of this figure shows that molecule 1's structural characteristics are clearly reflected in this classification. However, when the data set was clustered using crisp *k*-means or Ward's methods, the molecule was assigned to a single cluster that consisted of nitrobenzenes of various sorts, i.e., that reflected only part of molecule 1's characteristics.

Additional Experiments with Other Structure Representations. The experiments thus far have used just a single structure representation, viz. the principal components resulting from the Molconn-Z parameters, and we have hence carried out additional studies using a broader range of structural descriptors and also using different procedures for the initialization of the relocation that lies at the heart of *k*-means clustering.

Two additional representations were tested: fragment bit-strings and molecular holograms. These were both generated using Tripos software,³⁹ with the bit-strings containing a total of 992 bits and with the holograms containing 43, 199, and 991 bins. The hologram lengths of 43 and 991 were selected as the closest prime numbers to the lengths of the principal components and bit-strings, respectively, while 199 bins is one of the standard Tripos defaults. The runs in this second set of experiments used a C implementation of the fuzzy *k*-means method, this being about 10 times slower than the corresponding crisp implementation.

The *k*-means method, whether crisp or fuzzy, describes a general clustering approach, and many different versions of the method are possible. First, the algorithm may commence with a randomly initialized partition or with a centroid matrix based on selected seed molecules. The experiments in the previous section used the former approach, but this obviously means that the final clusters are nondeterministic in character: here, the Tripos Selector system was used³⁹ to identify a fixed, structurally diverse set of initial seeds. Second, the centroid and membership matrices can be updated either as each individual molecule is allocated to a cluster or at the end of an iteration, when all of the molecules have been assigned. The former approach is strongly order-dependent, and we have hence carried out the cluster updating step only at the end of each iteration.

The prediction experiments used a range of values for the fuzziness index, starting from *m* = 1.05 and increasing in

Table 3. Product Moment Correlation Coefficients for the Prediction of logP Using the Optimal Number of Clusters (*k*) and Using the Optimal *m* Values for the Fuzzy Clustering Runs^a

method	bit strings	holograms			Molconn-Z
		43 bins	199 bins	991 bins	
crisp <i>k</i> -means	0.76 (91.0)	0.69 (98.8)	0.72 (99.0)	0.70 (99.1)	0.73 (98.4)
fuzzy <i>k</i> -means	0.81 (90.5)	0.75 (99.0)	0.74 (98.1)	0.80 (98.9)	0.80 (96.7)
Ward's	0.75 (93.1)	0.69 (97.0)	0.72 (97.9)	0.74 (98.8)	0.71 (97.8)

^a The numbers in brackets indicate the percentage of molecules for which predictions were made.

increments of 0.05. This continued until the maximally fuzzy position was reached, i.e., the point at which all of the compounds were equal members of each of the clusters: for 30 clusters, this implies an average membership function that had converged to 1/30, i.e., ca. 0.033. In the case of the bit-strings, the maximally fuzzy clusters were achieved with *m* = 1.35; for the hologram data the analysis stopped at *m* = 2.50 and for the Molconn-Z data at *m* = 2.30. However, the best predictive results were (as before) obtained with much lower values of the fuzziness index: *m* = 1.10 for the bit-string data set; *m* = 1.15, 1.05, and 1.15 for the hologram-43, hologram-199, and hologram-991 data sets, respectively; and *m* = 1.20 for the Molconn-Z data set. It is the results with these parameter values that are discussed below.

Our experiments thus far have used a fixed value for the number of clusters in the partition, i.e., the value of *k*, but methods are available to suggest the number of clusters present in a data set. In particular, Wild and Blankley⁴⁰ have investigated stopping criteria for use with Ward's method and demonstrated the general effectiveness of the Kelley statistic.⁴¹ This statistic is implemented in the BCI OPTCLUS software, which was used to identify the optimal number of clusters for each of the representations of the 1763 molecules: these numbers (305 for the bit-strings; 134, 120, and 108 for the 43-, 199-, and 991-bin holograms; and 157 for the Molconn-Z data) were then used for the cluster-based prediction experiments. As before, crisp classifications were generated for comparison with the fuzzy results.

The results obtained with the optimum values for *m* and *k* are detailed in Table 3. It will be seen that the best fuzzy property prediction is better than the best crisp predictions in all cases and that the fuzzy and crisp approaches allowed predictions to be made for similar numbers of molecules. The significance of the differences was tested using a Wilcoxon Signed Ranks Test⁴² based on the mean absolute deviations (MADs) of the predictions. The MAD from each membership matrix was calculated using

$$\text{MAD} = \frac{\sum |\log P_{\text{pred}} - \log P_{\text{obs}}|}{n} \quad (4)$$

where *n* is the number of molecules in each partition for which logP values were predicted. The absolute deviation between the predicted and observed logP values was calculated for each molecule and then the mean was calculated over the entire data set. The Wilcoxon test was used to check the significance of the differences between the deviations calculated from the crisp Ward's and the various fuzzy cluster methods. These results are shown in Table 4, which confirms the statistical significance of the differences observed in Table 3 (significant differences are also observed for many of the other fuzzy clusterings when

Table 4. Wilcoxon Signed Ranks Test of Difference between Predicted and Observed logP Values of the Fuzzy and Best Crisp Clusters at the Ward Optimal Number of Partitions^a

data set	Z
bit-strings	4.36
holograms-43	6.06
holograms-199	2.98
holograms-991	6.77
Molconn-Z	7.34

^a All of the calculated Z values are significant at the 0.001 level of two-tailed statistical significance.

m < 1.25). Although significant, these differences are not large; however, any improvement in performance is worthwhile if this increased effectiveness can be achieved at an acceptable computational cost. Fuzzy cluster generation is more time-consuming, as noted previously, but the costs are not unreasonable for data sets of the sizes considered here.

Further, analogous experiments were carried out in which the SYBYL package was used to calculate molar refractivity values for each of the Starlist molecules. The results of these experiments are shown in Table 5, where it will be seen that the fuzzy correlations are often higher than the crisp correlations. However, the differences here are noticeably less than in Table 3, with none of these being significant at the 0.05 level of statistical significance in the Wilcoxon test.

CLUSTERING OF SYNGENTA DATA SETS

Thus far, we have considered the use of crisp and fuzzy methods for the prediction of physicochemical properties; in addition, we have carried out a more qualitative evaluation of the clusters produced when the fuzzy *k*-means method was applied to several Syngenta in-house data sets. The data sets are detailed in Table 6, where the right-hand column is the mean intradata set similarity calculated using Daylight fingerprints (the representation for all of the experiments in this section) and the Tanimoto coefficient. The core structure for each data set is included in the table, with the exception of the last, which contains a heterogeneous set of standard pesticide compounds from the corporate file; two cores are included for the imidazoline/oxazolidinedione data set as this contains representatives of two, closely related types of compound.

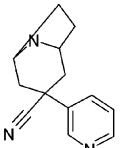
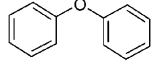
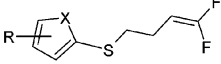
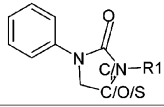
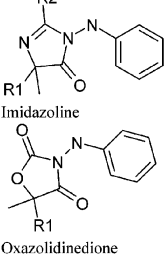
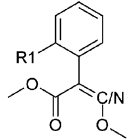
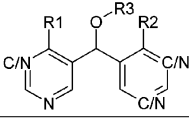
Each of the data sets was clustered on its own and also when merged with some or all of the other data sets. Here, we describe the results obtained in just a few of the cases: specifically, the most self-similar and dissimilar data sets (the cyanotropans and the pesticide standards, respectively), these two data sets merged together, and the first seven data sets (those that are each based on a common scaffold) merged together. The full sets of runs are discussed by Rodgers.²⁴

Table 5. Product Moment Correlation Coefficients for the Prediction of MR Using the Optimal Number of Clusters (k) and Using the Optimal m Values for the Fuzzy Clustering Runs^a

method	bit-strings	holograms			Molconn-Z
		43 bins	199 bins	991 bins	
Ward	0.81 (93.1)	0.85 (97.0)	0.82 (97.9)	0.82 (98.8)	0.92 (97.8)
k -means	0.79 (91.0)	0.84 (98.8)	0.83 (99.0)	0.82 (99.1)	0.93 (98.4)
Fuzzy k -means	0.80 (90.5)	0.87 (99.0)	0.84 (98.1)	0.84 (99.8)	0.94 (96.7)

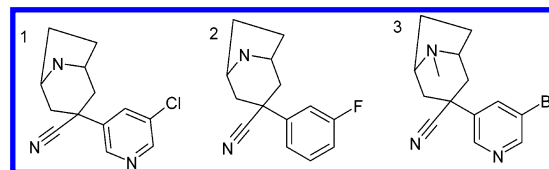
^a The numbers in brackets indicate the percentage of molecules for which predictions were made.

Table 6. Characteristics of the Syngenta Data Sets that Were Clustered^a

Name	Number of Compounds	Core structure	Mean Tanimoto coefficient
Cyanotropanes	84		0.65
Diphenylethers	311		0.34
Heterocyclic fluorovinyl nematocides	162		0.35
N-aryl lactam herbicides	272		0.46
Imidazoles and Oxazolidinediones	383		0.44
Strobilurin fungicides	226		0.47
Heteroaryl carbinol herbicides	303		0.56
Pesticide Standards	710		0.19

^a The right-hand column is the mean intradata set similarity calculated using Daylight fingerprints and the Tanimoto coefficient.

The fuzzy and crisp k -means methods both require the user to specify the number of clusters. The (mostly) small data sets used here are typical of those that might be generated in a project or retrieved by a substructure search of a corporate file and were hence processed to give a small number of clusters (3 or 4) that could easily be understood by a chemist using a clustering tool in a typical research project. The large, diverse pesticide standards data set was tested with several different numbers of clusters: the results here are based on specifying a total of 98 cluster seeds. The Jarvis-Patrick method was run with the default clustering parameters.

**Figure 2.** Typical compounds from each of the three cyanotropane clusters.

We have seen previously that the choice of fuzziness index, m , can have a large effect on the clusters produced. Based on the property-prediction results, all of the analyses here were carried out with $m = 1.2$; experiments with other values in the range 1.10–1.25 for the small data sets gave results little different from those reported below. The choice of m did, however, affect the memberships of the clusters produced for the standard pesticides data set, particularly for the small clusters. This is, perhaps, hardly surprising given the very diverse nature of these data set, especially as comparable variations in membership were occasioned by the choice of different random starting points for the clustering.

Clustering of Similar Compounds. The cyanotropanes are the most homogeneous set of compounds (as determined by the mean intra-data set Tanimoto coefficients): the clusters produced from the analysis of this data set are examined here to exemplify the performance of the fuzzy k -means method with a set of highly similar compounds.

All of the compounds have the cyanotropane core structure (or small variations thereof), but there are small structural variations that were analyzed with $k = 3$, the clustering producing one large cluster and two smaller ones of similar size. Typical structures from each of the clusters are shown in Figure 2. The first compound is from the large cluster and thus represents the majority of the cyanotropanes. The second cluster contains very similar compounds to the first but has a benzene, rather than a pyridine ring, while the tropane ring in the third cluster contains a double bond that is absent from the other two clusters.

An analysis of the membership functions for each of the compounds shows that the clusters are well-defined with each compound typically having one membership function close to unity and the other two functions close to zero: this behavior is shown in Figure 3, which demonstrates clearly the crisp nature of this classification with only a few compounds showing any evidence of shared cluster-membership. It is thus hardly surprising that the clusters here are very little different from those resulting from application of the crisp k -means and Jarvis-Patrick methods to this data set.

Clustering of Dissimilar Compounds. The pesticide standards is by far the most diverse of the data sets considered, without any common core structure. A whole

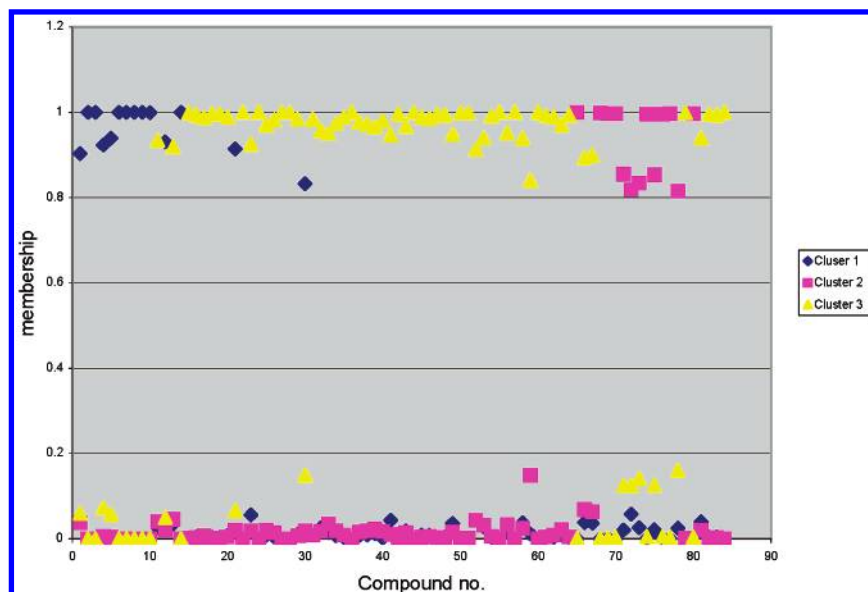


Figure 3. Membership functions of the cyanotropane data set. The *X* axis denotes the compounds, arranged in increasing ID order, and each such compound has three values plotted on the *Y* axis, these values being the membership function for each of the three clusters.

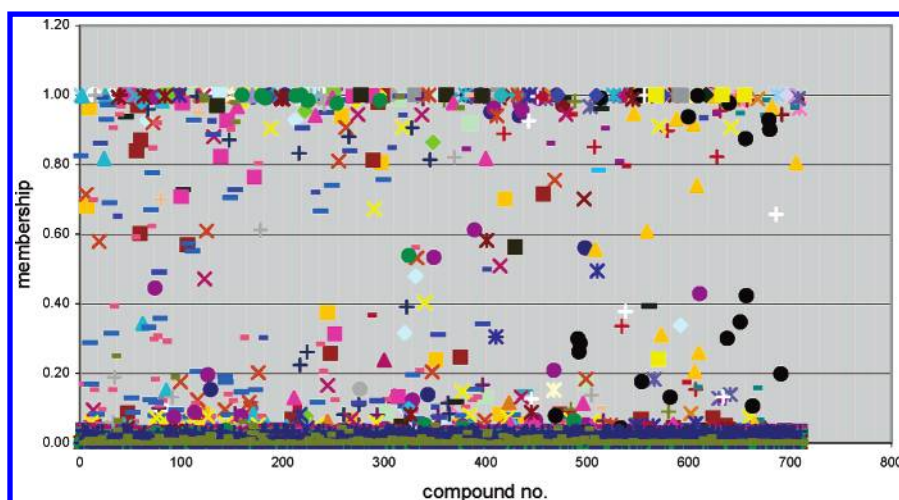


Figure 4. Membership functions of the pesticide standards data set.

series of clustering experiments were carried out, with $10 \leq k \leq 400$. The larger values resulted in many clusters that were effectively empty (i.e., there were no compounds that had significant memberships for that cluster), while small values resulted in the grouping of markedly dissimilar compounds; the results discussed here were obtained with $k = 98$. This partition contained a fair number of *singletons*. A singleton in crisp clustering is simply a cluster that contains just a single member, but in fuzzy clustering every cluster contains every compound to a greater or lesser extent. Thus a singleton here can be a cluster with only one significant membership or a compound with membership close to $1/k$ for all clusters present, indicating that that compound has no strong relationship to any of the clusters present and that it should thus not be allocated to any of them. This provides a simple, direct way of identifying outliers in a data set and avoids a common problem with some crisp clustering methods: compounds being allocated to clusters with which they have only a limited degree of similarity, this in turn resulting in “ragbag” clusters of compounds that are grouped together only because they cannot be easily grouped anywhere else.

There is, of course, a need to define at what level of membership a compound is considered to be a singleton. We have not found any discussion of this point in the literature; inspection of the many clusters produced during this study suggests the following, purely heuristic definition: a compound is a singleton if

$$\mu_{\max} - \mu_{\min} < \frac{1}{k} \text{ for } k < 10 \text{ or if } \mu_{\max} < \frac{3}{k} \text{ for } k \geq 10$$

Using this definition, there were 281 singleton pesticide standards; this is a large fraction of the total data set of 710 compounds but noticeably less than the 477 singletons identified by the Jarvis-Patrick method. There were 8 clusters with only one significant member and a further 24 clusters containing only molecules with membership < 0.1 .

The partition matrix resulting from the cluster analysis of this data set has a much greater spread of membership values than in the case of the cyanotropenes, where the memberships were concentrated around zero and unity. This behavior is shown in Figure 4.

The results obtained with this large set of unrelated structures suggest that the fuzzy *k*-means approach has some

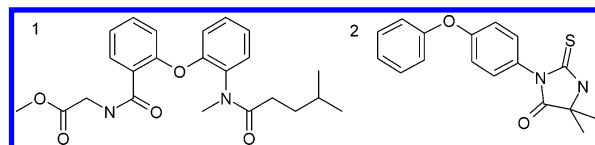
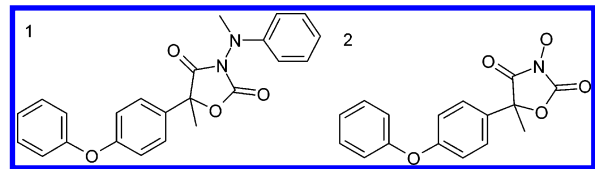
Table 7. Clusters Produced from the Seven Homogeneous Data Sets

cluster	principal data set(s) represented in the cluster	total membership	maximum membership
1	strobilurin fungicides	61.3	0.99
2	diphenyl ethers	84.9	0.99
3	lactam herbicides	50.2	1.00
4	imidazolines	81.9	0.73
5	imidazolines	101.4	1.00
6	diphenyl ethers	90.6	0.36
7	imidazolines	94.5	1.00
8	diphenyl ethers	44.5	0.99
9	oxazolidinediones	74.1	1.00
10	fluorovinyl nematocides	97.3	0.95
11	cyanotropanes	79.5	1.00
12	carbinol herbicides	130.5	0.98
13	diphenyl ethers and strobilurin fungicides	105.3	1.00
14	lactam herbicides	43.7	1.00
15	fluorovinyl nematocides	60.8	0.98
16	lactam herbicides	61.4	1.00
17	lactam herbicides	103.6	1.00
18	diphenyl ethers and strobilurin fungicides	87.9	0.97
19	carbinol herbicides	152.7	1.00
20	diphenyl ethers	105.0	0.93

useful advantages, particularly when it comes to qualitative interpretation of clustering results. It handles singletons better than the conventional *k*-means, it generally produces better clusters than the Jarvis-Patrick method, and it is faster than Ward's or other hierarchical methods. These points together with the improved qualitative interpretability make the approach a useful addition to the armory of chemoinformatics tools.

Clustering the Seven Homogeneous Data Sets. All of the data sets from Table 6, excluding the pesticide standards, were merged together and then clustered to determine how successfully the fuzzy *k*-means method could group these seven sets into appropriate clusters. Sensible groupings were obtained with $k = 20$, as detailed in Table 7, which lists the majority of the compounds present in each cluster, and the total and the maximum memberships for each cluster, i.e., the sum of the memberships of the molecules in that cluster and the largest such membership, respectively. It may be thought that 20 is a small number of clusters for the number of compounds (1711 in all), but this does mean that the output is readily comprehensible to a chemist (as well as being intuitively sensible). As will be seen in the table, each cluster consisted mainly of compounds from a specific data set apart from cluster-13 and cluster-18, both of which contained both diphenyl ethers and strobilurin fungicides: this is quite reasonable as these two data sets contain a lot of overlap, with many compounds containing the core structures of both groups.

A molecule with a membership value close to unity for a particular cluster means that that molecule is very closely associated with, and hence an appropriate representation of, that cluster. All of the clusters here contained molecules with high membership functions apart from cluster-6, where the maximum membership function was only 0.36. However, visual examination of these compounds confirms that there is a fair degree of structural similarity between the compounds here (at least for those with membership greater than 0.20). For example, two of the compounds are shown in Figure 5 where it will be seen that they both have the

**Figure 5.** A diphenyl ether ($\mu = 0.36$) and an imidazoline ($\mu = 0.26$) grouped together in cluster-6 when all of the homogeneous data sets were merged.**Figure 6.** Two structurally similar diphenyl ethers having very different membership functions ($\mu = 0.99$ and $\mu = 0.19$) for the predominantly oxazolidinedione cluster-9 when all of the homogeneous data sets were merged.

diphenyl ether moiety (although the second compound is, in fact, from the imidazoline/oxazolidinedione data set). The cyanotropanes contributed to the fewest clusters, as would be expected given the small size of this data set. The diphenyl ethers produced the most clusters, with four all-diphenyl ether and two joint with the strobilurin fungicides; this is again as expected given that this is the most heterogeneous of the seven data sets.

The membership functions normally assist in delineating the structural relationships between compounds, but there are occasional apparent exceptions to this general rule. One example of this behavior is illustrated in Figure 6, which shows two diphenyl ethers in cluster-9: these molecules are clearly very similar but have very different memberships for this particular cluster. An inspection of the other molecules strongly associated with this predominantly oxazolidinedione cluster reveals that while both of these compounds have this ring (as well as the diphenyl ether moiety), the second compound has an O substitution, rather than the N-benzyl substitution common to most members of the oxazolidinedione data set.

Rodgers²⁴ also discusses the analysis of these merged Syngenta data sets using a conventional, crisp *k*-means method and the crisp Jarvis-Patrick method (which is the standard Syngenta in-house clustering tool). There was not very much difference in the performance of the crisp and fuzzy *k*-means methods (except that the fuzzy method is rather more time-consuming); however, the Jarvis-Patrick clusters were often more difficult to interpret, yielding one very large cluster, many smaller ones and very large numbers of singletons (a well-known characteristic of the method).

The comparative studies summarized here serve to highlight the strengths and weaknesses of the fuzzy *k*-means approach. The availability of the membership function allows compounds that display structural features of more than one cluster to belong to several different groups, makes it easy to identify singletons that do not contribute noticeably to any cluster (i.e., do not have any significant memberships), and also makes it easy to identify representative compounds (i.e., those with a membership ≈ 1 in a cluster). Against this, however, the multiple memberships can be difficult to interpret, and there is the need to specify a value for the fuzziness index and a criterion to determine at what level a membership becomes significant; the fuzzy method is also

noticeably slower in operation, albeit still entirely feasible for data sets of the sort considered here.

CONCLUSIONS

In this paper, we have reported a detailed evaluation of the fuzzy *k*-means method when used for clustering sets of chemical compounds. Simulated property prediction experiments suggest that, when appropriately parametrized, the method is at least as effective as traditional, crisp clustering approaches based on 2D fingerprints. Clustering of several typical in-house project data sets shows that the method is again competitive with existing approaches, with the membership function information providing a useful explanatory tool for rationalizing the clusters that are identified. Our experiments certainly do not suggest that the fuzzy *k*-means method should supplant existing tools for chemical clustering; however, we do believe that it provides a useful complement to them.

ACKNOWLEDGMENT

We thank the following: the Engineering and Physical Sciences Research Council and Syngenta for funding; Barnard Chemical Information Ltd. and Tripos Inc. for software; and the Royal Society and the Wolfson Foundation for hardware and laboratory support.

REFERENCES AND NOTES

- (1) Shemetulskis, N. E.; Dunbar, J. B.; Dunbar, B. W.; Moreland, D. W.; Humblet, C. Enhancing the Diversity of a Corporate Database Using Chemical Database Clustering and Analysis. *J. Comput.-Aided. Mol. Des.* **1995**, *9*, 407–416.
- (2) Willett, P.; Winterman, V.; Bawden, D. Implementation of Non-Hierarchical Cluster Analysis Methods in Chemical Information Systems: Selection of Compounds for Biological Testing and Clustering of Substructure Search Output. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 109–118.
- (3) Brown, R. D.; Martin, Y. C. Use of Structure–Activity Data to Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572–584.
- (4) Willett, P. *Similarity and Clustering in Chemical Information Systems*; Research Studies Press: Letchworth, 1987.
- (5) *Molecular Similarity in Drug Design*; Dean, P. M. Ed.; Chapman and Hall: Glasgow, 1994.
- (6) Downs, G. M.; Barnard, J. M. Clustering Methods and their Uses in Computational Chemistry. *Rev. Comput. Chem.* **2002**, *18*, 1–40.
- (7) Sneath, P. H. A.; Sokal, R. R. *Numerical Taxonomy*; W. H. Freeman: San Francisco, 1973.
- (8) Everitt, B. S. *Cluster Analysis*, 3rd ed.; Edward Arnold: London, 1993.
- (9) Kaufman, L.; Rousseeuw, L. *Finding Groups in Data. An Introduction to Cluster Analysis*; John Wiley: New York, 1990.
- (10) Brown, R. D.; Martin, Y. C. The Information Content of 2D and 3D Structural Descriptors Relevant to Ligand–Receptor Binding. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1–9.
- (11) Willett, P. An Evaluation of Relocation Clustering Algorithms for the Automatic Classification of Chemical Structures. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 29–33.
- (12) Bezdek, J. C. *Pattern Recognition with Fuzzy Objective Function Algorithms*; Plenum Press: New York, 1981.
- (13) Bezdek, J. C.; Pal, S. K. *Fuzzy Models For Pattern Recognition*; Institute of Electrical and Electronics Engineers: New York, 1992.
- (14) Duda, R. O.; Hart, P. E.; Stork, D. G. *Pattern Classification*; John Wiley: New York, 2001.
- (15) Zhang, C.; Chou, K.; Maggiora, G. M. Predicting Protein Structural Classes from Amino Acid Composition: Applications of Fuzzy Clustering. *Prot. Eng.* **1995**, *8*, 425–435.
- (16) Friederichs, M.; Franzle, O.; Salski, A. Fuzzy Clustering of Existing Chemicals According to their Ecotoxicological properties. *Ecol. Model.* **1996**, *85*, 27–40.
- (17) Sarbu, C.; Horowitz, O.; Pop, H. E. A Fuzzy Cross-Classification of the Chemical Elements, Based on Their Physical, Chemical, and Structural Features. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1098–1108.
- (18) Doman, T. N.; Cibulskis, J. M.; Cibulskis, M. J.; McCray, P. D.; Spangler, D. P. Algorithm5: A Technique for Fuzzy Similarity Clustering of Chemical Inventories. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1195–1204.
- (19) Pop, H. F.; Sarbu, C. The Fuzzy Hierarchical Cross-Clustering Algorithm. Improvements and Comparative Study. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 510–516.
- (20) Linusson, A.; Wold, S.; Norden, B. Fuzzy Clustering of 627 Alcohols, Guided by a Strategy for Cluster Analysis of Chemical Compounds for Combinatorial Chemistry. *Chemomet. Intell. Lab. Syst.* **1998**, *44*, 213–227.
- (21) Daszykowski, M.; Walczak, B.; Massart, D. L. On the Optimal Partitioning of Data with K-Means, Growing K-Means, Neural Gas and Growing Neural Gas. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1378–1389.
- (22) Lin, T.-H.; Wang, G.-M.; Hsu, Y.-H. Classification of some Active HIV-1 Protease Inhibitors and Their Inactive Analogues Using some Uncorrelated Three-Dimensional Molecular Descriptors and a Fuzzy C-Means Algorithm. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1490–1504.
- (23) Feher, M.; Schmidt, J. M. Fuzzy Clustering as a Means of Selecting Representative Conformers and Molecular Alignments. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 810–818.
- (24) Rodgers, S. L. *Application of the Fuzzy K-Means Clustering Algorithm to the Analysis of Chemical Structures*, Ph.D. Thesis, University of Sheffield, in preparation.
- (25) Zadeh, L. A. Fuzzy Sets. *Inf. Control* **1965**, *8*, 338–353.
- (26) Dubois, D.; Prade, H. *Fuzzy Sets and Systems: Theory and Applications*; Academic Press: New York, 1980.
- (27) Zimmerman, H.-J. *Fuzzy Set Theory and Its Applications*; Kluwer Academic Publishers: Boston, 1996.
- (28) Dunn, J. C. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *J. Cybern.* **1973**, *3*, 32–57.
- (29) Windham, M. P. Cluster Validity for the Fuzzy c-Means Clustering Algorithm. *IEEE Trans. Patt. Anal. Mach. Intell.* **1982**, *PAMI-4*, 357–363.
- (30) Cannon, R. L.; Dave, J. V.; Bezdek, J. C. Efficient Implementation of the Fuzzy c-Means Clustering Algorithms. *IEEE Trans. Patt. Anal. Mach. Intell.* **1986**, *PAMI-8*, 248–255.
- (31) Babuska, R.; Alic, L.; Lourens, M. S.; Verbraak, A. F. M.; Bogaard, J. Estimation of Respiratory Parameters via Fuzzy Clustering. *Artif. Intell. Med.* **2001**, *21*, 91–105.
- (32) Bezdek, J. C.; Castela, P. F. Prototype Classification and Feature Selection with Fuzzy Sets. *IEEE Trans. Syst., Man, Cybernet.* **1977**, *SMC-7*, 87–92.
- (33) Dalezios, I.; Siebert, K. J. Comparison of Pattern Recognition Techniques for the Identification of Lactic Acid Bacteria. *J. Appl. Microbiol.* **2001**, *91*, 225–236.
- (34) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.
- (35) The Starlist database is available from BioByte Corp. at <http://www.biobyte.com>.
- (36) Molconn-Z is available from eduSoft at <http://www.eslc.vabiotech.com/>.
- (37) Barnard Chemical Information Ltd. is at <http://www.bci.gb.com/>.
- (38) Johnson, M. A. Similarity-Based Methods for Predicting Chemical and Biological Properties: a Brief Overview from a Statistical Perspective. In *Chemical Information Systems: Beyond the Structure Diagram*; Bawden, D., Mitchell, E. M., Eds.; Ellis Horwood: New York, 1990; pp 149–159.
- (39) Tripos Inc. is at <http://www.tripos.com>.
- (40) Wild, D.; Blankley, C. J. Comparison of 2D Fingerprint Types and Hierarchy Level Selection Methods for Structural Grouping Using Ward's Clustering. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 155–162.
- (41) Kelley, L. A.; Gardner, S. P.; Sutcliffe, M. J. An Automated Approach for Clustering an Ensemble of NMR-Derived Protein Structures into Conformationally-Related Subfamilies. *Prot. Eng.* **1996**, *9*, 1063–1065.
- (42) Siegal, S.; Castellan, N. J. *Nonparametric Statistics for the Behavioral Sciences*; McGraw-Hill: New York, 1988.

CI0342674