Advances in Methodology, Data Analysis, and Statistics
Anuška Ferligoj (Editor), Metodološki zvezki, 14
Ljubljana: FDV, 1998

The Evaluation of the Sensitivity of some Clustering Techniques to Irrelevant Variables

Istvan Hajnal[1] and Geert Loosveldt

Abstract

In sociology cluster analysis is often used in situations where the researcher does not have a clear theoretical view of the importance of variables in the analysis. An irrelevant variable might disguise a good clustering. Although some special purpose algorithms to deal with variable weighting and variable selection, have been proposed in the literature, we will focus on the classical (and widely available) algorithms such as the *k-means* algorithm and hierarchical algorithms like *single linkage*, *average linkage* and so on.

In this paper we summarize first some of the results that we found in the clustering literature on the effects of irrelevant variables on recovery in cluster analysis. Milligan (1980) found in a simulation study that all the algorithms involved were sensitive to irrelevant or noise variables. This was also found by Fowlkes (see Milligan and Cooper, 1987). In van Meter's (1984) study, on the other hand, the addition of arbitrary variables did not alter the results significantly.

We then present a simulation study in which a noise variable is added to a set of variables with an *a priori* cluster structure. The *a priori* cluster structure was generated with a mixture modelling framework. We included the following factors in our simulation: the dimensionality of the true clustering, the total sample size, the variance in the noise variable and the (Euclidean) distance between the clusters' centroids (see Hajnal and Loosveldt, 1996). These factors can affect the cluster solution. In this simulation we will evaluate the impact of these factors on some hierarchical and non-hierarchical cluster algorithms. We will also look at, whether the R^2 measure for predicting the variable from the cluster (SAS Institute Inc., 1989: 834) that is typically given in *k-means* output is informative for variable selection purposes.

1 Introduction

Cluster analysis can be a very useful tool in an exploratory analysis [2] when a researcher does not have a clear view of which variables are important in the cluster structure. In

[1] Department of Sociology, University of Leuven, Edward Van Evenstraat 2B, B-300 Leuven, Belgium, e-mail: istvan.hajnal@soc.kuleuven.ac.be

[2] For a general overview of the basic concepts and purposes of classification and typology construction in the social sciences, see amongst others Lorr(1983), van Meter et al. (1987), Bailey (1994).

these situations it is easy to imagine that an irrelevant variable could be added to an otherwise good set of variables. An often cited example is given by De Sarbo et al. (1984). In an automobile marketing study, attitudinal variables yielded clear and easy to interpret clusters of car owners. The clusters reflected the different brands of cars. After the addition of general attitudinal variables the original structure was obscured.

In this paper, we will study the sensitivity of a clustering solution to irrelevant variables by adding random noise to a data set with a known cluster structure. Furthermore, we will study one particular evaluation tool for deciding on the importance of the variables to the cluster solution: In the output of PROC FASTCLUS of the statistical package SAS, for each variable used in the cluster analysis, the R^2 for predicting the variable from the cluster is given (SAS Institute Inc., 1989: 834). We will look if this measure is informative for variable selection purposes.

2 Literature overview

Milligan (1980) studied the effects of six types of error perturbation on cluster solutions. In his study, the *k-means* algorithms performed well with respect to all types of error only when appropriate seeds were used. He also found that the *single linkage* method was only mildly affected by outliers, but that the method was very sensitive to perturbation in the distance matrix. This was also found by Baker in 1974. In Baker's study the *complete linkage* method performed better with respect to this type of error (see Milligan, 1981: 383-384). Another result of Milligan's simulation study was the finding that the selection of a proximity measure is less important than the selection of a clustering method (Milligan, 1980: 339-340). This was also argued by Punj and Stewart (see Milligan and Cooper, 1987: 344). Kaufman (1985) found that problems of measurement space distortion are less severe than the problems of including "too much" variation. Donoghue (1995) studied the effects of the within-group covariance structure on the cluster solution. He found that, in general, negative within-group correlation resulted in a lower recovery.

The main topic of this paper is one particular type of error perturbation; namely, irrelevant variables. In Milligan's (1980) simulation, two situations were considered. In the error-free condition, the data sets were generated on the basis of a true clustering; in the noise condition, random noise dimensions were added to the data sets. The best clustering method in the error-free condition was the *group average* method with an adjusted Rand statistic of 0.998. In the noise condition, this method again had the best recovery. The adjusted Rand statistic in that condition was 0.930. All the other methods showed a similar decrease in cluster recovery. Milligan concluded that all the methods were sensitive to this type of error, but it was possible to classify the clustering methods according to their sensitivity to noise. The *group average* method turned out to be the best performing method, followed by the *weighted average* method, the *beta-flexible* method, and the *minimum average sum of squares*[3] method. Milligan's finding that the use of all available data can possibly obscure the clustering present in a subset of variables was also found by Fowlkes (see Milligan and Cooper, 1987: 344). The only study that we found in the literature where the addition of arbitrary variables hardly modified the results was by van

[3] See Anderberg (1973:148).

Meter (1984). In his study, he used two types of classification methods; namely, *dynamic clusters* and *fixed centre typology analysis*.

In the classification literature, a few clustering models have been proposed that could be used for detecting noisy variables. The *synclus* model (De Sarbo, et al., 1984), for example, uses a weighting scheme to enhance the cluster structure. In this case, the irrelevant variables would have low weights. Some methods try to cluster objects in a lower dimensional space (see Bock,1987; De Soete and Caroll, 1994). The data reduction used in these methods would, unlike principal components, preserve the cluster structure. Although these models are very interesting, they have serious drawbacks for sociologists. A major problem is that implementations of these algorithms are not widely available yet. Another, even more serious, drawback is that the current implementations of these models can handle only a few hundred objects. In sociological research, especially in settings where large scale surveys are used (i.e. where the total sample size tends to be high), this is problematical. It is, therefore, interesting to see to what extent algorithms such as the *k-means* algorithm and hierarchical algorithms like *single linkage*, *average linkage* and the like (in the remainder of this paper refered to as 'the classical clustering techniques'), are vulnerable to irrelevant variables. Furthermore, it is also interesting to see in what conditions the solutions become very bad. Another research question is: if a particular clustering technique is vulnerable to irrelevant variables, can the R^2 measure for each variable be used to detect important variables?

3 Design

The aim of this paper is to study the effect of an irrelevant variable on an otherwise good cluster solution and to evaluate the usefulness of the R^2 measure for detecting the irrelevant variable. To this end we generated artificial data and clustered them with 5 hierarchical clustering methods from PROC CLUSTER[4] in SAS mentioned in Table 1. We also used a *k-means* algorithm from PROC FASTCLUS[5]. Below we discuss the design factors, the data generation, and the outcome measure.

Of the several factors that affect cluster solutions we only varied those given in Table 1. Other factors remained equal in the simulation, such as the number of clusters (3), the variances of the relevant variables (1), the dissimilarity measure for the hierarchical cluster analyses (Euclidean distance), the number of irrelevant variables (1), the mean of the irrelevant variable (0), and, finally, the relative cluster sizes in the data set (1/3).

The true mean vectors were specified in such a way that the Euclidean distances $\| \mu_i - \mu_j \|$ between any two cluster centroids are the same in the three situations of the number of relevant variables (p = 2, 4 and 6). A little geometry and algebra will show

[4]For more information on PROC CLUSTER see SAS Institute Inc., (1989: 519-614).

[5]Because of the implementation of the algorithm and the particular method of generating seeds used in PROC FASTCLUS (see SAS Institute Inc., 1989: 823-850), we decided to refer to this method as *fastclus* instead of *k-means*.

Table 1: The experimental design factors

Factor	Name	Number of levels	Description of level
p	Number of relevant variables	3	2
			4
			6
n	Sample size for each cluster	3	50
			100
			150
v	Variance in the noise	3	1
			3
			5
m	Cluster procudure	6	-Average linkage
			-Centoid methode
			-Complete linkage
			-Single linkage
			-Ward's methode
			-Fastclus
d	Euclidean distance between the true clusters' centroids	3	4
			6
			8

that, for any positive real number d, the set of μ's:

$$\mu_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}; \mu_2 = \begin{bmatrix} d\cos 60° \\ d\sin 60° \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}; \mu_3 = \begin{bmatrix} d \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{1}$$

in R^6 will satisfy this condition:

$$d_{ij} = \| \mu_i - \mu_j \| = d \tag{2}$$

4 Data generation

We generated 3 spherical clusters around these centroids using a multivariate normal mixture approach (with fixed cluster sizes n) with the identity matrix as the covariance matrix. The five-factor design resulted in 3x3x3x3x6=486 different setups. Each setup was replicated 20 times. This resulted in 9720 data sets. To each data set one random noise variable was added. Two cluster analyses were performed: one with and one without the noise variable[6].

[6]In total almost 20,000 analyses were carried out for this simulation. The R^2 measure for predicting the variable from the cluster (SAS Institute Inc., 1989: 834) is only computed in PROC FASTCLUS and

In Figure 1, an example is given of the effect of adding a noise variable for $p = 2$, $n = 100$, and $d = 8$. In Figure 1(a), three clusters are shown in two dimensions. In Figure 1(b), the same data are shown in three dimensions. In Figure 1(c), the noise variable is added.

5 Outcome measures

Many outcome measures have been suggested to evaluate cluster solutions (see, for example, Rand, 1971; Milligan, 1980; Milligan, 1981; Kaufman, 1985; Milligan, Cooper, 1987; Fisher, Hoffman, 1988). We measured the recovery with the value of Cramer's V (see SAS Institute Inc., 1989: 866):

$$V = \sqrt{\frac{\chi^2}{N \times min[(r-1), k-1)]}} \tag{3}$$

where N is the total sample size, r is the number of rows, and c is the number of columns in a crosstabulation. Since in this case we use the crosstabulation of the true clustering and the cluster solution and we have 3 clusters, $N = 3 \times n$ and $min[(r-1), k-1)] = 2$. Cramer's V amounts here to $V = \sqrt{\frac{\chi^2}{6n}}$. The measure has an lower bound of 0 and an upper bound of 1. The former will occur when the cluster recovery is minimal, the latter when the recovery is maximum. Cramer's V was also used by Mezzich to assess cluster validation (see Milligan, 1981: 392). The sensitivity was measured as the difference between the value of Cramer's V of the cluster solution with the noise variable and the value of Cramer's V of the cluster solution without the noise variable. We call this difference Δ. In this way, the deterioration of the clustering solution caused by adding an irrelevant variable can be measured. If Δ is close to zero, the addition of a noise variable changed the cluster solution very little. If the value is close to one, the solution was altered in a very strong way.

6 Results

In a preliminary analysis, the value of Cramer's V **without** noise was analyzed. Because we wanted to work with clear *a priori* clusters, we chose the values for d in such a way that in all cases the mean of the value of Cramer's V was relatively high. The overall mean of Cramer's V without noise was 0.86. It is worth noting, however, that a significant effect for the clustering procedure was found. The *fastclus* procedure showed the best performance, second came *Ward's method*, followed by *average* and *complete linkage*. The *single linkage* procedure performed less well than the other methods[7].

not in PROC CLUSTER. For the hierarchical clustering methods we had to to extract that measure from an analysis of variance (with the cluster solution as the independent variable and the p variables as the dependent variables). A SAS macro was written to generate the data sets and to analyse them subsequently. This took more than 11 CPU-hours on a IBM9672-mainframe computer.

[7]We should stress here that these results are highly influenced by the cluster configuration. If we would have generated 'chained' clusters instead of spherical clusters, the *single linkage* method would probably have performed much better.

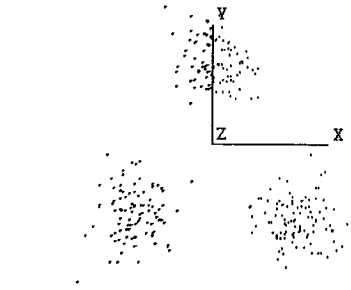
Table 2: Main effects, first and second order interactions for ANOVA of Δ

Source	DF	SS	F	Pr > F	Eff. Size
d	2	45.37449	1084.07	0.0001	0.183784
m	5	24.57980	234.9	0.0001	0.099558
v	2	53.53757	1279.1	0.0001	0.216848
p	2	0.385489	9.21	0.0001	0.001561
n	2	0.355982	8.5	0.0002	0.001442
$m \times d$	10	44.31847	211.77	0.0001	0.179507
$v \times d$	4	19.50892	233.05	0.0001	0.079019
$d \times p$	4	1.418802	16.95	0.0001	0.005747
$d \times n$	4	0.180177	2.15	0.0717	0.000730
$v \times m$	10	12.44142	59.45	0.0001	0.050393
$m \times p$	10	2.240457	10.71	0.0001	0.009075
$m \times n$	10	0.847368	4.05	0.0001	0.003432
$v \times p$	4	0.179146	2.14	0.0732	0.000726
$v \times n$	4	0.23578	2.82	0.0238	0.000955
$p \times n$	4	0.111443	1.33	0.2556	0.000451
$v \times m \times d$	20	19.65571	46.96	0.0001	0.079613
$m \times d \times p$	20	14.62272	34.94	0.0001	0.059228
$m \times d \times n$	20	2.317153	5.54	0.0001	0.009385
$v \times d \times p$	8	1.292221	7.72	0.0001	0.005234
$v \times d \times n$	8	0.146099	0.87	0.5387	0.000592
$d \times p \times n$	8	0.483008	2.88	0.0033	0.001956
$v \times m \times p$	20	1.192081	2.85	0.0001	0.004828
$v \times m \times n$	20	0.575066	1.37	0.1227	0.002329
$m \times p \times n$	20	0.700113	1.67	0.0303	0.002836
$v \times p \times n$	8	0.187199	1.12	0.3471	0.000758

Below we will present the results of our analysis of the effect of the noise variable on the recovery by giving the mean value of Δ for each level of a factor (or combination of factors). Instead of presenting all 31 tables (+ one overall mean) we will make a selection based on an analysis of variance. The degree of deterioration due to the noise variable, measured by Δ, was used as the dependent variable in the ANOVA. The independent variables were the design factors described in a previous section. In Table 2 we summarize the results of the ANOVA for the five design factors that we considered in the simulation. Because of the large number of cluster analyses that were used in this simulation it comes as no surprise that most of the main effects of Table 2 are significant. We only want to summarize the results of our simulation. That's why, instead of using traditional significance tests, we tried to capture the important effects with the following measure of effect size (see Donoghue, 1995):

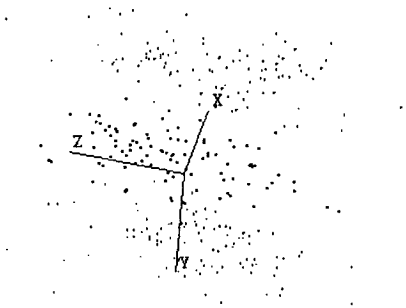
$$\eta^2 = \frac{SS_{effect}}{SS_{tot}} \tag{4}$$

where SS_{effect} is the sum of squares of an effect and SS_{tot} is the total sum of squares.



(a)

(b)



(c)

Figure 1: An example of three clusters in two dimensions without noise (a), in three dimensions without noise (b), and in three dimensions with noise (c)

Table 3: Mean values of Δ for three levels of distances d between the cluster centroids

d	Mean Δ
4	0.189338
6	0.091305
8	0.022853

Table 4: Mean values of Δ for the six cluster procedures m

m	Mean Δ
Centroid method	0.182810
Average Linkage	0.130260
Complete Linkage	0.111535
Single Linkage	0.082501
Fastclus	0.081224
Ward's method	0.018661

Applying, the somewhat arbitrary (but see Donoghue, 1995: 233), criterion of $\eta^2 > .05$, the number of relevant variables p and cluster size n are "less important" effects. It is clear that the variance in the noise variable v, the cluster procedure m and distance d are very important. In the remainder of this paragraph we describe only the factors (and factor combinations) with an associated effect size $\eta^2 > .05$ and the overall mean.

The overall mean of Δ was 0.10, which means that the addition of a noisy variable decreased the value of Cramer's V by 0.10 on average in our simulation. Table 3 gives the mean value of Δ for the diffferent levels of distances that were used in the simulation. When the initial clusters were very clear ($d = 8$), the addition of noise hardly affected the cluster recovery (Δ=0.02); when the distances between the clusters were the smallest ($d = 4$), the value of Δ increased to 0.19. An interesting finding, shown in Table 4, is that the method of *Ward* yielded an average Δ of only 0.02, which suggests that the method is not attenuated by irrelevant variables within the framework of this simulation. With a value of 0.08 for Δ, the *fastclus* procedure loses the first position it held in the no noise situation to *Ward's method.*

In Table 5, the means for Δ are shown for the three levels of variance v in the noise variable. It clearly shows the obvious result that the higher the variance, the more the cluster solutions are blurred.

A few interactions were also substantial (see Table 2). The variance v by cluster procedure m interaction displayed in Figure 2 shows that the increase in variance v in the noise had differential effects on the cluster methods. The *single linkage* and, especially, the method of *Ward* seem to be only slightly affected by the increase in variance in the noise variable. The *centroid* and *complete linkage* methods perform poorly once the variance is high.

In Figure 3, which shows the interaction between the noise variance v and the distance d, the combined effect of the two variables becomes very clear. In fact, this shows that

Table 5: The mean values of Δ for the three levels of variance v in the noise

v	Mean Δ
5	0.196315
3	0.091954
1	0.015226

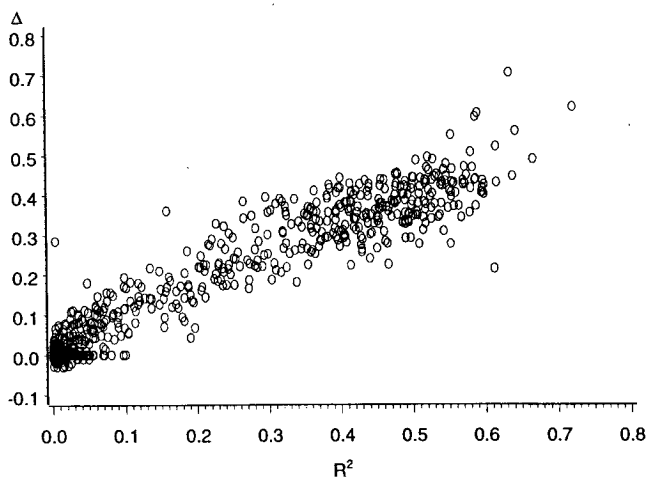


Figure 2: Mean values of Δ for the interaction of cluster procedure m and variance v in the noise

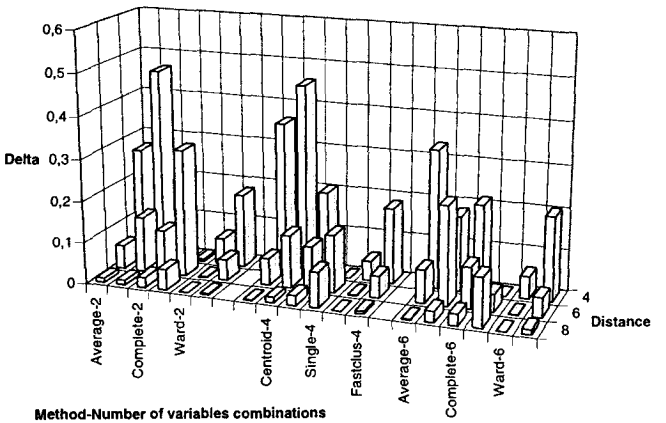


Figure 3: Mean values of Δ for the interaction of distances d between the clusters and variance v in the noise

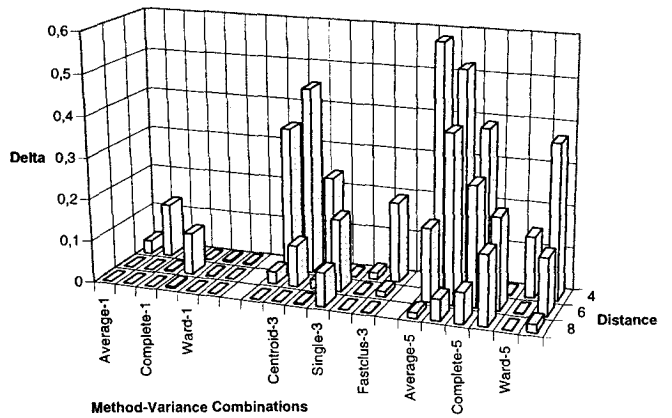


Figure 4: Mean values of Δ for the interaction of cluster procedure m and distances d between the clusters

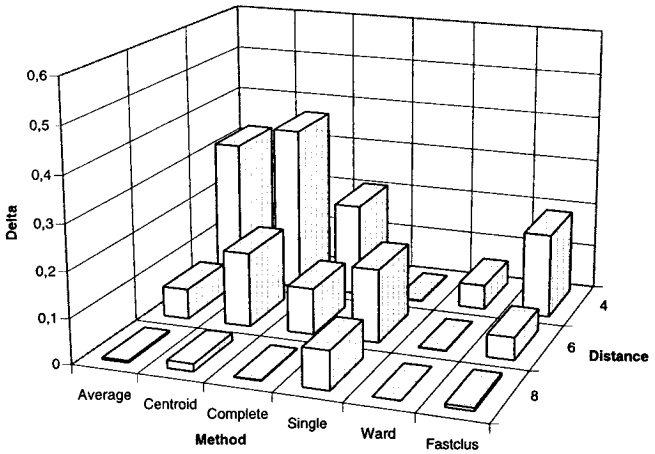


Figure 5: Mean values of Δ for the interaction of variance v in the noise, cluster procedure m and distances d between the clusters

what is really important is the relationship between the variance of the noise variable and the distances between the clusters. Figure 4 shows the interaction between distance d and cluster procedure m. It is clear that the *average linkage* and the *centroid* method were very negatively influenced by the noise variable when d was small. In the cases of smallest distance ($d = 4$), these methods decreased the value of Cramer's V by 0.40. *Ward's method* was only mildly affected. A very interesting observation is that *single linkage* performs best when the true clustering is weak. A possible explanation is that the *single linkage* method performed so poorly even without the noise variable that a further decline was almost impossible. Finally, the last two figures show second order interactions. The interpretation of these effects seems to be less straightforward than the previous ones. In Figure 5, one can see that the combined effect of a high variance v in the noise variable and a weak (true) clustering yields very high values for Δ, especially for the average linkage and the centroid method. On the other hand, when the variance of the noise variable is small ($s = 1$), there is almost no effect of distance (except for the *centroid* and *average linkage* methods).

Although the effect of the number of variables p as such is not significant, it's interaction with clustering method m and distance d, as shown in Figure 6, is significant. The effect of additional relevant variables seems to be different for the different clustering methods and the true clustering distances.

If cluster procedures are sensitive to noise variables, one can wonder whether the use of the R^2 measure (for predicting the variable from the cluster) is informative for variable selection purposes. In general, the correlation coefficient between Δ and R^2 for the noise

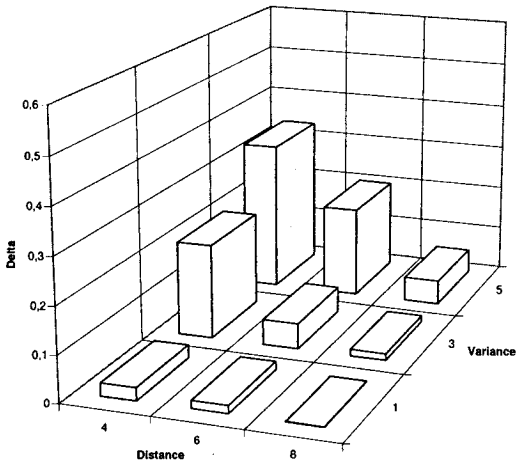


Figure 6: Mean values of Δ for the interaction of the number of variables p, cluster procedure m and distances d between the clusters

Table 6: Correlation coefficients between Δ and R^2 for the noise variable for the six cluster procedures m

m	Corr(Δ, R^2)
Average Linkage	0.70
Centroid method	0.60
Complete Linkage	0.91
Single Linkage	0.25
Ward's method	0.94
Fastclus	0.97

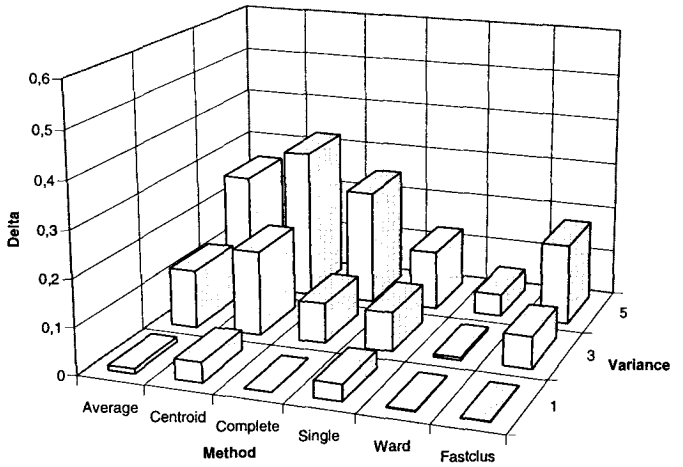


Figure 7: Scatter plot of Δ and R^2 for the noise variable for k-means procedure

variable was 0.55 [8]. However, there were considerable differences between the different clustering methods. In Table 6, the correlation coefficients for the different clustering methods are shown. For the *fastclus* procedure, the correlation coefficient for Δ and the R^2 of the noise variable is .97 (see Figure 7). In this case, selecting the variable(s) with a high R^2 will not always lead to a correct decision: when the deterioration caused by the noise variable is high, the R^2 value for the noise variable will be (relatively) high, and the R^2's of the relevant variables will be somewhat lower. When the effect of noise is low, the R^2 value of the noise variable will be low, and the R^2's of the relevant variables will be high.

7 Conclusion

This simulation clearly shows that irrelevant variables can be a problem in the sense that they can obscure an otherwise good cluster solution. Researchers should, therefore, carefully select variables that are of theoretical interest to the research problem, before attempting a cluster analysis. In our simulation, the *single linkage* method performed rather poorly in most cases, even without the addition of a noise variable. We also found that *Ward's method* was only mildly affected in our simulation. *Ward's method* performed better than any of the others. Even when the variance v of the noise variable was high, the recovery was rather good. We should stress again that the results of this simulation

[8] Although there is no reason to assume a linear relationship, we decided to use the correlation coefficient based on visual inspection of the scatterplots.

study are highly influenced by the cluster configuration. This is probably also why the results are not always in concordance with some of the results mentioned in the literature overview.

From a social scientist's perspective, these results are rather unpleasant. Hypothesis testing with classical cluster analysis is not possible because it lacks a thorough statistical theory. We have to use a model-based approach (see, for example, Bock, 1996) in these situations. Moreover, we have to be very careful if we want to use classical cluster analysis for exploratory data analysis because the addition of a noisy variable can blur a cluster solution.This conclusion is important. If classical cluster analysis methods are sensitive to this type of error, it is not very useful in an exploratory setting, since researchers should already know which variables are important for the cluster structure and which variables are not.

References

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

[2] Bailey K.D. (1994): *Typologies and Taxonomies: An Introduction to Classification Techniques.* Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-102, Thousand Oaks, CA.

[3] Bock H.H. (1987): On the interface between cluster analysis, principal components analysis and multidemnsional scaling. In: H. Bozdogan and A.K. Gupta (Eds.). *Multivariate Statistical Modeling and Data Analysis.* D.Reidel, Dordrecht, 17-34.

[4] Bock H.H. (1996): Probabilistic models in partitional cluster analysis. In: A. Ferligoj and A. Kramberger (Eds.). *Developments in Data Analysis.* FDV, Ljubljana, 3-25.

[5] De Sarbo W.S., Caroll J.D., Clark L.A. and Green P.E. (1984): Synthesised clustering: A method for amalgating alternative clustering bases with differential weighting of variables. *Psychometrika*, **49**, 57-78.

[6] De Soete G. and J.D. Caroll (1994): K-means clustering in a low-dimensional Euclidean space. In: E. Diday, Y. Lechevallier, M. Schader, P. Bertrand and B. Burtschy (Eds.). *New Approaches in Classification and Data Analysis.* Springer-Verlag, Berlin, 212-219.

[7] Donoghue J.R. (1995): The effects of within-group covariance structure on recovery in cluster analysis. I. The bivariate case. *Multivariate Behavioral Research*, **30**, 227-254.

[8] Everitt B.S. (1993): *Cluster Analysis.* Edward Arnold, London.

[9] Fisher D.G. and Hoffman P. (1988): The adjusted Rand statistic: A SAS macro. *Psychometrika*, **53**, 417-423.

[10] Hajnal I. and G. Loosveldt (1996): The sensitivity of hierarchical clustering solutions to irrelevant variables. *Bulletin de Méthodologie Sociologique*, **50**, 56-70.

[11] Jain A.K. and Dubes, R.C. (1988): *Algorithms for Clustering Data.* Prentice-Hall, Englewood Cliffs.

[12] Kaufman R.L. (1985): Issues in multivariate cluster analysis. *Sociological Methods and Research*, **14**, 467-486.

[13] Kaufman L. and P.J. Rousseeuw (1989): *Finding Groups in Data. An Introduction to Cluster Analysis.* John Wiley and Sons, New York.

[14] Lorr M. (1983): *Cluster Analysis for the Social Scientists.* Jossey-Bass, San Francisco.

[15] McLachlan G.J. and K.E. Basford (1987): *Mixture Models. Inference and Applications to Clustering.* Marcel Dekker, New York.

[16] Milligan G.W. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika*, **45**, 325-342.

[17] Milligan G.W. (1981): A review of Monte Carlo tests of cluster analysis. *Multivariate Behavorial Research*, **16**, 379-407.

[18] Milligan G.W. and Cooper M.C. (1987): Methodology review: clustering methods. *Applied Psychological Measurement*, **11**, 329-354.

[19] Rand W.M. (1971): Objective criteria for the evaluation in clustering methods. *Journal of the American Statistical Association*, **66**, 846-850.

[20] Romesburg H.C. (1990): *Cluster Analysis for Researchers.* Krieger Publishing Company, Malabar.

[21] SAS Institute Inc. (1989): *SAS/STAT User's Guide, Version 6, Fourth Edition, Volume 1.* SAS Institute, Cary, NC.

[22] Sneath P.H.A. and R.S. Sokal (1973): *Numerical Taxonomy. The Principles and Practice of Numerical Classification.* W.H. Freeman and company, San Francisco.

[23] van Meter K.M. (1984): Structure des données et stabilité des résultats - La typologie de base en science sociales. *Bulletin de méthodologie sociologique*, **4**, 41-57.

[24] van Meter K.M., M.W. de Vries, C.D. Kaplan, and C.I.M. Dijkman (1987). States, syndromes, and polythetic classes: The operationalization of cross-classification analysis in behavioral science research. *Bulletin de méthodologie sociologique*, **15**, 22-38.