4.2.4 Summary

The findings in this limited set of examples lack probative value in showing the conduct of **DetClus** for a specific application, but shed light on the tendency to cluster of the entities that are subject of the clustering in question. The results can be summarized as follows.

1) It is advisable to plot the value of the indices against the number of clusters, and assessing the plot by eye, looking for peaks and valleys and consensus among the various indices.

2) The values of k to be investigated should be kept small as some indices tend to assume a bizarre behavior as $k \rightarrow n$.

3) The user should consider both peaks and valleys as an indication of a good candidate values for k.

4) If the clustering is strong enough to withstand true confusion and plots used with discretion then there is a good chance of determining the true value of k by considering the point on which all indices are concordant.

5) In general, when partial agreement is found between stopping rules, then the user should opt for the smaller number of clusters.

For a fixed quality of the initial partition the recovery rate tends to be better for even sized cluster than for cluster of unequal size

DetClus has given satisfactory practical results although it is possible that for a particular application it will perform poorly. In fact, a good starting partition, an optimal reassigning scheme, faithfully reproduction of the algorithm in the programming language, and a high speed computer do not guarantee that the end partition will coincide with one of the best configuration compatible with the data set. On the other hand, if this is the case we have no way of knowing it, unless the structure of the data were perfectly known as in the case of artificial data sets used in simulation.

DetClus is not practical when it comes to solving large problems as all iterative schemes take a lot of time to get a solution and require a large amount of space for storage purposes. It take about two weeks to execute the largest problem *(n=12000, m=50, k=2,3,...,25)*.

6. Syntax

The present version, freely available on the internet network, can handle a maximum of $n=12000$ data points, $m=50$ variables, and $k=25$ clusters. Other versions can be requested if the program does not handle enough data for your applications (and you have enough memory storage and execution time). *Please send bug reports to agotar@unical.it*

The variables are not transformed so that any transformation of the raw data should be executed before using **DetClus**.

Missing values are not allowed. The user is expected eliminate all cases with one or more missing values.

DetClus satisfies the constrain 2b so that each entity is admitted to the cluster it has its smallest distance with, even if this distance is large. As a consequence, the software does not indicate which entities are obviously members of a cluster and which should be regarded as borderline entities, singletons or outliers. The user, however, is free to remove the suspected entities from the data set. *The default values are* $k_1=2$ *and* $k_2=Min\{[(n-m(m-1))/(3m)]+1,25\}$

The format of the data set

There are three file specification to be given

1. The input data file (according to the scheme *rows=entities, columns=variables*;
2. The output file (contains the results of the run);
3. The clusterings file (contains the cluster membership of all the entities for each value of *k*).

Additional specifications are required if the user provides a set of initial centroids or a starting configuration or a set of cardinalities. A separate file is also necessary for the true or known classification (for a fixed number of clusters) if comparison must be performed with the final partition of the algorithm.

DetClus requires a separate line (or set of lines) for each entity, containing the entity identification and variable values, utilizing a comma delimited format

$$(AAA\ldots AAA, Value_1,\ Value_2,\ \ldots,\ Value_m)_i, \qquad i=1,2,\ldots,n$$

The field *"AAA...AAA"* expects a maximum of 32 letters or numbers or special characters, but considers the symbols to form a name and not a value on which regular computations can be performed. A labels can have less than 32 characters (at least one is required). The comma cannot be used as part of a case name. The data Value_1, Value_2 etc. must be numeric. The data in the file should be formatted as text items delimited by commas and/or carriage-returns. Each item is assigned to a separate variable.

DetClus reads a line of text from the file, beginning at the current "file mark" position (which is usually at the beginning of the line), and ending when a carriage-return character is

encountered, or the end of the file is encountered, or 255 characters have been read, whichever occurs first. The file mark is then advanced to a position just past the last character read. **DetClus** then attempts to assign each of the comma-delimited items in the text line to one of the variables (Value_1, Value_2 etc.) in the variable list. If there are more items in the text line than variables in the list, the extra items are discarded. If there are fewer items in the text line than variables in the list, then zeros are assigned to the extra variables.

The variables should be in free-field format with valued separated by a comma. A comma is also required between the case name and the value of the first variable. The same position in successive data rows need not contain value for the same variable. Only the order of variable specification must be identical.

```
Setosa,51.000,35.000,14.000,2.000
Setosa,49.000,30.000,14.000,2.000
----------------------------------------
```

DetClus cannot read Excel files. In Excel save the data in a text file in tab separated format (or space separated or separated any other delimiter) and then, if necessary, open the file with a text editor and change the delimiter to a comma.

All numeric data are considered in double precision. It is important that the data row does not end with a comma and that a comma is not included in the entity label because the pairing value of the variable/datum on the row will be shifted. The consequence of this error may be dramatic. Sometimes an input impairment occurs allowing the detection of the input errors; otherwise, particularly if the identification labels are numbers, the run will be executed and blindly accepted if the findings look like what expected. The file which contains the initial partition and that of the true classification file (the two specifications may coincide) have only one integer column (the decimal point, if present, will be ignored).

Unfortunately, the user interface of **DetClus** is the part of the program that has received scant attention thus far. The Lions's share of effort were received by the initialization and relocation methods.

While reasonable effort has been made to ensure this software operates substantially as described the Author cannot guarantee proper operation in every possible configuration. For this reason, the **DetClus** is provided "As is" without warranty of any kind, either expressed or impled, including, but not limited to, the implied warranties of merchantability, fitness for a particular purpose, or non-infringement. In no event will Agostino Tarsitano be liable for any special, incidental, consequential, indirect or similar damages due to loss of data or any other reason.

If a user encounters problems with this program, the author is willing to help solve them.

References

Al-Daoud M.B. Roberts S.A. (1996). New methods for the initialization of clusters. *Pattern Recognition Letters*, 17, 451-455.

Art D. , Gnanadesikan R., Kettenring J.R. (1982). Data-based metrics for cluster analysis. *Utilitas Mathematica*, 21A, 75-99.

Anderson E.J. (1996). Mechanisms for local search. E*uropean Journal of Operational Research*, 88, 139-151.

Andrews D.F. Herzberg A.M. (1985). Data. A collection of problems from many fields for the student and researh worker. Springer-Verlag, New York. (http://lib.stat.cmu.edu/DASL/)

Arnold S.J. (1979). A test for clusters. *Journal of Marketing Research*, 16, 545-551.

Banfield J.D. Raftery A.E.(1993). Model-based Gaussian and non Gaussian clustering. *Biometrika*, 49, 803-828

Bayley T.A., Dubes R. (1982. Cluster validity profiles. *Pattern Recognition*, 15, 61-83.

Bayne C.K., Beauchamp J.J., Begovich C.L., Kane V.E. (1980). Monte Carlo comparisons of selected clustering procedures. *Pattern Recognition* 12, 51-62.

Beale E.M.L. (1972). Euclidean Cluster Analysis. *Bulletin of the International Statistical Institute*, 43, 92-94.

Bezdek J.C. Pal N.R. (1998). Some new indexes of cluster validity. *IEEE Transactions on Systems, Man, and Cybernetics, Part B,* 28,301-315.

Bissell A.F.(1976). Ordered random selction without replacement. *Applied Statistics*, 35,73-75.

Blake, C.L. & Merz, C.J. (1998). UCI Repository of machine learning databases. Irvine, CA: University of California, Department of Information and Computer Science. http://www.ics.uci.edu/~mlearn/MLRepository.html

Blashfield R.K., Aldenderfer M.S., Morey L.C. (1982). Cluster Analysis Software. P.R. Krishnaiah and L.N. Kanal eds. Handbook of Statistics. Vol.2, 245-266. North-Holland Publishing Company.

Bock H.H. (1998). Probability models and hypothesuis testing in partitioning cluster analysis. In P.Arabie, J. Hubert & G. De Soete (Eds.), Clustering and Classification (pp. 377-453). World Scientific Publishing, Singapore.

Calinski T. (1969). On the application of cluster analysis to experimental results. *Bulletin of the International Statistical Insitute,* 43, 101-103.

Casgrain P. (2001). Le Progiciel R v4.0d6. http://ProgicielR.webhop.org/

Castagnoli E. (1978). Un'osservazione sull'analisi classificatoria. Seminario su due temi di analisi statistica multivariata. CLEUP, Bologna (I), 135-139.

Cormack R.M. (1971) *A Review of Classification*. Journal of the Royal Statistical Society. Series A, Vol. 134, 321-353.

Chen H., Gnanadesikan R., Kettenring J.R. (1974). Statistical methods for grouping corporations. *Sankhyã*, 36, series B, 1-28.

Chernoff H. (1970). Metric considerations in cluster analysis. Pp. 621-629 in: *Proceedings of the 6h Berkley symposium on mathematical statistics and probability*, UCLA Press, Berkley, CA.

Chernoff H. (1973) The use of faces to represent points in k-dimenional space graphically. *Journal of the American Statistical Association*, 73, 361-368.

Coleman D., Dong X. ,Hardin J. ,Rocke D.M. ,Woodruff D.L. (1999). Some computational issue in cluster analysis with no a priori metric. *Computational Statistics & Data Analysis*, 31, 1-11.

Cox D.R. (1957). Note on grouping. *Journal of the American Statistical Association*, 52, 543-547.

Dagnelie P. Merckx A. (1991). Using generalized distances in classification of groups. *Biometrical Journal*, 33, 683-695.

Davenport M. Studdert-Kennedy G. (1972). The Statistical Analysis of Aesthetic Judgements: An Exploration. *Applied Statistics*, 21, 324-333.

Davies D.L. Bouldin D.W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1, 224-227.

Digby P.G.N. Kempton R.A. (1987. Multivariate Analysis of Ecological Communities. Chapman & Hall, London.

Dubes R.C. (1987): How many cluters are best? An experiment. *Pattern Recognitio*n, 20, 645-663.

Duda R.O. Hart P.E. (1973). Pattern classification and scene analysis. John Wiley & Sons, New York.

Duda R.O. Hart P.E. Stork D.G. (2001). Pattern classification. 2nd editions. John Wiley & Sons, New York.

Engelman L. Hartigan J.A. (1959). Percentage points of a test for clusters.*journal of the American Statistical Association*, 64, 1647-1648.

Everitt B.S. (1979) Unresolved probles in cluster analysis. *Biometrics,*35, 169-181.

Everitt B.S., Landau S., Lase M. (2001). Cluster analysis, 4th ed. Arnold, London.

Fisher L. Van Ness J.W. (1971) Amissible Clustering Procedures. *Biometrika,* 58, 91-104.

Fisher D., Xu L., Zard N. (1992). Ordering effects in clustering. In: *Proceedings of the9th International Conference on MachineLearning*, pp. 163-168. Morgan Kauffman, San Matero, Ca.

Forgy E. (1965). Cluster analysis of multivariate data. Efficiency vs interpretability of classification. WNAR Meetings, UCLS Riverside, Ca. June 22-23, 1965 (Abstract in *Biometrics*, 21, 768).

Friedman H.P. Rubin J. (1967). On some invariant criteria for grouping data. *Journal of the American Statistical Association*, 62, 1159-1178.

Gentleman J.F. (1975). Algorithm AS 88. Generation of all $_{n}C_{r}$ combinations by symulating nested Fortran Do loops. *Applied Statistics*, 24, 374-376 (http://www.Stat.unipg.it/pub/stat/statlib/apstat)

Gnanadesikan R., Kettenring J.R., Landwehr J.M. (1977). Interpreting and assessing the results of cluster analysis. *Bulletin of the International Statistical Institute*, 47, 451-463.

Gordon A.D. (1996). How many clusters? An investigation of five procedures for detecting nested cluster structure. In C. Hayashi et al. (Eds.). *Data science, classification, and related methods*. Proceeedings of the fifth conference of the international federation of classification societies, pp 109-116. Kobe, Japan, March, 27-30, 1996.

Gordon A.D. (1999). Classification, 2nd Edition. Chapman and Hall, London.

Hall D.J. Khanna D. (1977). The ISODATA method computation for the relative perception of similarities and differences in complex and real data. In K. Enslein, A. Ralston, and H.W. Wilf (eds.) *Statistical Method for Digital Computers*. John Wiley & Sons. New York.

Hand D.J.,Daly F., Lunn A.D., McConway K.J., OstrowskiE. (Editors)(1994). Handbook of small data sets. Chapman & Hall, New York.

http://www2.ncsu.edu/eos/service/pams/stat/sicl/small_data/index.html

Hansen P. Mladenovic (2001). J-means: a non local search heuristic for minimum sum of squares clustering. *Pattern recognition,* 34, 405-413.

Hardy A.(1996). On the number of clusters. *Computational Statistics and Data Analysis*, 23, 83-96.

Hartigan J.A.Wong M.A. (1979). A k-Means clustering algorithm. *Applied Statistics*,28, 100-108.

Heeler R.M. Day G.S. (1975). A supplementary note on the use of cluster analysis for stratification. *Applied Statistics*, 3, 342-343.

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

Hubert L.J Levin j.R. (1976). A general statistical framework for assessing categorical clustering in free recall. *Psychological Bulletin*, 83, 233-253

Ismail M.A., Kamel M.S. (1989). Multidimensional data classification utilizing hybrid search strategies. *Pattern Recognition*, 22, 75-89.

Kaufman L. Rouseeauw P.L. (1990). Finding Group in Data. An Introduction to Cluster Analysis. John Wiley & Sons, New York.

Kennard R.W., Stone L.A. (1969). Computer aided design of experiments. *Technometrics*. 11, 137-148.

Klastorin T.D. (1983). Assessing cluster analysis results. *Journal of Marketing Research*, 20, 92-98.

Knuth D.E. (1981). The Art of Computer Programming. Vol.2, 2nd Edition. Addison-Wesley Publishing Company, Reading.

Kotari R; Pitts D. (1999). On finding the number of clusters. *Pattern Recognition Letters*, 20, 405-416.

Kruskal J. (1977). Multidimensional scalind and clustering. In Classification and Clustering. J. Van Ryzin (Ed.). Academic Press, New York. 17-44.

Lance G. Williams W.T. (1967*).* A general theory of classificatory aorting strategies. II. Clustering systems. *Computer Journal,* 10, 271-277.

Lubishew A.A. (1962): On the Use of Discriminant Function in Taxonomy. *Biometrics,* 18, 455-477..

MacQueen J. (1967). Some methods for classification and analysis of multivariate observations. In:L.M. LeCam, J. Neyman (Eds.) *Proceedings of the 5th Berkeley sysmposium on mathematical statistics and probability, Vol. 1-Statistics* pp. 281-297, UCLA Press, Berkeley, Ca.

Mahalanobis P.C. (1936). On the generalizedistance in statistics. *Proceedings of the National Institute of Science* (India), 12, 49-55.

Manly B.F.J. (1985). The statistics of natural selection on animal populations. Chapman and 5Hall, London.

Maronna R. Jacovkis (1974). Multivariate Clustering Procedure with Variable Metrics. *Biometrics,* 30, 499-505.

Marriott F.H.C. (1971). Practical Problems in a Method of Cluster Analysis. *Biometrics*, 27, 501-514.

Martin M.J. Pablos F. Gonzalesz A.G. (1998). Discrimination between arabica and robusta green coffee varieties according to their chemical composition. *Talanta*, 46, 1259-1264.

McRae D.J. (1971). MIKCA: A Fortran IV Iterative k-means Cluster Analysis Program. Behavioral Science. Vol. 16, 423-424.

Meza J.C. Olkin I. (1993). Numerical procedures for estimanting the parameters in a multivariate homogeneous correlation model with unequal variances. *Sankhya, A,* 506-515.

Milligan G.W. (1980). An examination of the effect of six types of error perturbation on fifteen 5clustering algorithms. *Psychometrika,* 45, 325-342.

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

Milligan G.W.(1996). Clustering Validation: Results and Implications for Applied Analysis. In P.Arabie, J. Hubert & G. De Soete (Eds.), Clustering and Classification (pp. 341-373). World Scientific Publishing, Singapore.

Mineo A. (1985) A New Criterion for the Choice of Seed Points for a Nearest Centroid Cluster Analysis. *Rivista di Statistica Applicata*, 18,191-198.

O'Flaherty and MacKenzie (1982). Algorithm As172: direct simulation of nested fortran Do-Loops. Applied Statistics, 31, 71-74.

Peña J.M., Lozano J.A., Larrañaga P. (1999). An empirical comparison of four initialization

methods for the k-means algorithm. *Pattern Recognition Letters*, 20, 1027-1040.

Pop F. Dumitrescu D.Sarbu C. (1995). A study of roman pottery (terra sigillata) using hieararchical fuzzy clustering. *Analytica Chimica Acta*, 310, 269-279.

Rencher A.C. (1992). Interpretation of canonical discriminant functions, canonical variates, and principal components. The American Statisticiam, 217-225.

Rao M.M. (1971). Cluster Analysis and Mathematical Programming. *Journal of the American Statistical Association*, 66, 622-626.

Rasson J.P. Kubushishi T. (1996). The gap test: an optimal method for determining the number of natural classes in cluster analysis.

Richards L.E. (1972). Refinement and Extensions of Distribution-free Discrimant Analysis. *Applied Statistcs*, 21, 174-176.

Rouncefield M. (1995). The statistics on poverty and inequality. *Journal of Statistics Education*, 3. http://lib.stat.cmu.edu/DASL/DataArchive.html

Rousseeuw P.J. Kaufman L. Trauwaert E. (1996). Fuzzy clustering using scatter matrices. Computational statistics and data analysis. 23, 135-151.

Rubin J. (1967). Optimal classification into groups: an approach for solving the taxonomy problem. *Journal of Theoretical Biology*, 15, 103-144.

Sadocchi S. (1977). Un metodo di cluster analysis non stratificato derivato dal metodo del legame singolo. *Rivista di Statistica Applicata,*10, 228-241.

Sarle W.S. (1983). Cubic clustering criterion. SAS technical Report A+108, Cary, Nc: SAS institute Inc.
http://www.csc.fi/cschelp/sovellukset/stat/sas/sasdoc/sashtml/hrddoc/indfiles/5903.htm

Scott A.J., Symons M.J. (1971). Clustering method based on likelihood ratio criteria. *Biometrics*, 27, 387-397.

Symons M.J. (1981). Clustering criteria and multivariate normal mixtures. *Biometrics*. Vol. 37, 35-43.

Seber G.A.F. (1984). Multivariate Observations. John Wiley & Sons. New York.

Selim S.Z., Ismail M.A. (1984). K-means type algorithm: generalized convergence theorem and characterization of local optimality. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6, 81-87.

Späth H. (1985). Cluster Dissection and Analysis. Theory, FORTRAN Programming, Examples. Ellis Horward Limited, Chichester (U.K.).

Struyf A. Hubert M., Rousseeuw P.J. (1997).Integrating robust clustering techniques in S-Plus. *Computational Statistics & Data Analysis*, 26, 17-37

Tarsitano A. Anania G. (1995) Tecniche di analisi statistica multivariata per l'individuazione dei sistemi agricoli territoriali in Italia. G. Cannata (a cura di). I sistemi territorali agricoli italiani degli anni '90. Contributi metodologici. Rubettino Editore, Soveria Mannelli (Cz), Italy.

Tarsitano A. (2002). A computational study of several relocation methods for k-means algorithms. Submitted for publication.

Theodoris S. Koutroumbas K. (1998). Pattern recognition. Academic Press

Van Rijsbergen (1970). A clustering algorithm. *Computer Journal,* 13, 113-115.

Vicari D. (1990). Indici per la scelta del numero dei gruppi (indices for determining the numebr of clusters). *Metron,*47, 473-492.

Vichi M. (1985). Cluster analysis and the graphical approach for identification of two types of multivariate analysis. *Metron*, 43, 165-188.

Warkenar C.S. Krishna G. (1979): A heuristic clustering algorithm using union of overlapping pattern-cells. *Pattern Recognition* 11, 85-93.

Wong C. Che C. Su M. (2001). A novel algorithm for data clustering. *Pattern recognition*, 34, 425-442.

Zhang Q., Boyle R.D. (1991). A new clustering algorithm with multiple runs of iterative procedures. *Pattern Recognition*, 24, 835-848.

Zupan J. (1981). Clustering large data sets. John Wiley & Sons, New York.