Tutorials and worked examples for simulation, curve fitting, statistical analysis, and plotting.
http://www.simfit.org.uk

Given a swarm of n multivariate points, K-means clustering attempts to assign these data into K non-empty clusters where $1 < K < n$. The clusters are created by assigning cases to those groups that minimize the within-cluster sum of squared distances of the data from the means of the clusters. Starting values from which to commence the iterations to find such clusters must be provided.

Example 1

From the SIMFIT main menu choose [Statistics], [Multivariate], then [K-means clustering], and observe the format for the test data contained in `g03eff.tf1` which are observations of five variables on twenty soil types as follows.

```
77.3   13.0    9.7   1.5   6.4
82.5   10.0    7.5   1.5   6.5
66.9   20.6   12.5   2.3   7.0
47.2   33.8   19.0   2.8   5.8
65.3   20.5   14.2   1.9   6.9
83.3   10.0    6.7   2.2   7.0
81.6   12.7    5.7   2.9   6.7
47.8   36.5   15.7   2.3   7.2
48.6   37.1   14.3   2.1   7.2
61.6   25.5   12.9   1.9   7.3
58.6   26.5   14.9   2.4   6.7
69.3   22.3    8.4   4.0   7.0
61.8   30.8    7.4   2.7   6.4
67.7   25.3    7.0   4.8   7.3
57.2   31.2   11.6   2.4   6.5
67.2   22.7   10.1   3.3   6.2
59.2   31.2    9.6   2.4   6.0
80.2   13.2    6.6   2.0   5.8
82.2   11.1    6.7   2.2   7.2
69.7   20.7    9.6   3.1   5.9
begin{values}
82.5   10.0    7.5   1.5   6.5
47.8   36.5   15.7   2.3   7.2
67.2   22.7   10.1   3.3   6.2
end{values}
```

Note that starting cluster coordinates are appended to this data set in the section identified by

begin{values} . . . end{values}

as this is the most convenient way to perform K-means clustering. However, these can be supplied independently, e.g. in a file like `g03eff.tf2`, or generated randomly. Note that K-means clustering is an iterative technique and the outcome will depend on the starting clusters.

From the analysis the following results are displayed, where for each case (in the odd-numbered rows) the cluster number to which it is assigned is the corresponding figure below it (in the even-numbered rows).

Results for K-means clustering with `g03eff.tf1`

Variables included: 1 2 3 4 5

Number of clusters K = 3

Transformation: Untransformed

Weighting: Unweighted for replicates

Cases (odd rows) and Clusters (even rows)

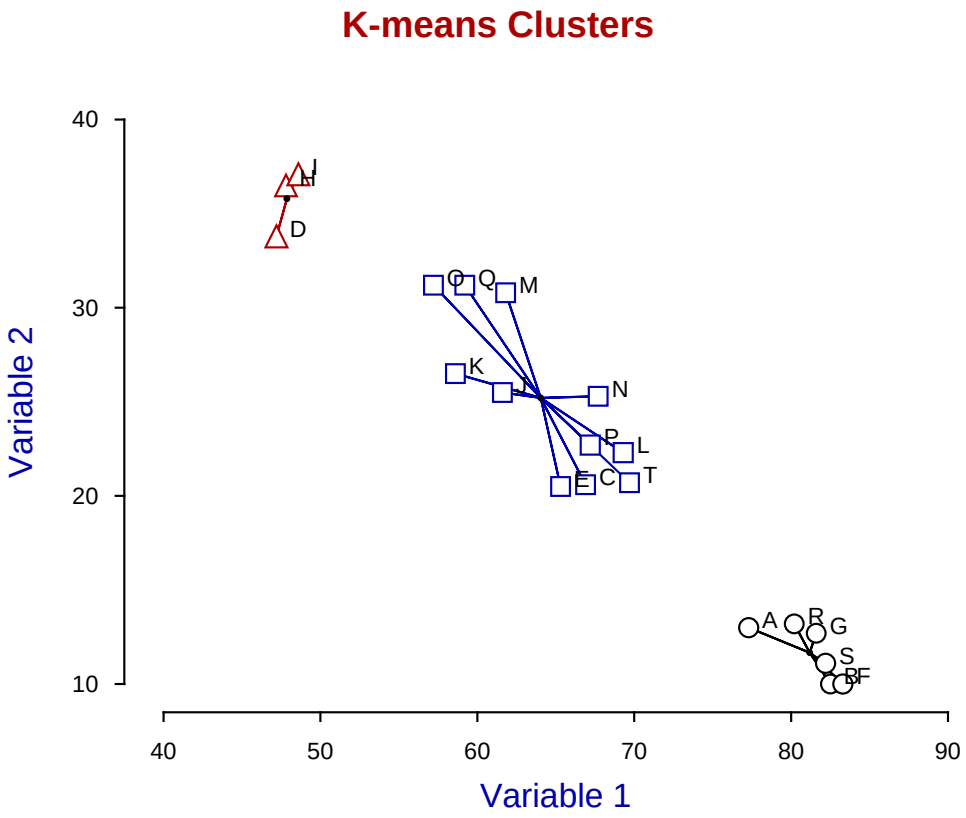
1	2	3	4	5	6	7	8	9	10	11	12
1	1	3	2	3	1	1	2	2	3	3	3
13	14	15	16	17	18	19	20				
3	3	3	3	3	1	1	3				

Final cluster centroids

81.183	11.667	7.1500	2.0500	6.6000
47.867	35.800	16.333	2.4000	6.7333
64.045	25.209	10.745	2.8364	6.6545

Note that the final cluster centroids minimizing the objective function, given the starting estimates supplied, are calculated, and the cases are assigned to these final clusters.

Plots of the clusters and final cluster centroids can be created as in the next figure for variables x_1 and x_2, with the optional labels as these were also supplied on the data file `g03eff.tf1` (as for dendrograms).



With two dimensional data representing actual distances, outline maps can be added and other special effects can be created, as shown later. Further, techniques are provided to perturb the default positions of labels if this is required in order to clarify the labeling.

Example 2

The next table is for analysis of the Fisher Iris data set in `iris.tf1`, using starting clusters in `iris.tf2`.

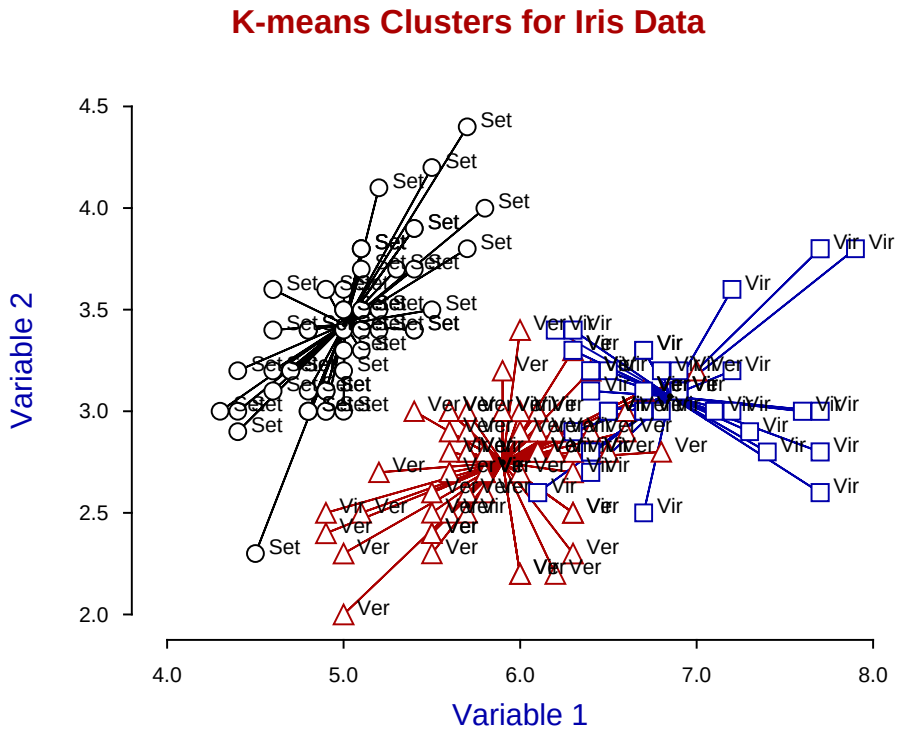
```
1   1   1   1   1   1   1   1   1   1   1   1
1   1   1   1   1   1   1   1   1   1   1   1
1   1   1   1   1   1   1   1   1   1   1   1
1   1   1   1   1   1   1   1   1   1   1   1
1   1   2   2   3*  2   2   2   2   2   2   2
2   2   2   2   2   2   2   2   2   2   2   2
2   2   2   2   2   3*  2   2   2   2   2   2
2   2   2   2   2   2   2   2   2   2   2   2
2   2   2   2   3   2*  3   3   3   3   2*  3
3   3   3   3   3   2*  2*  3   3   3   3   2*
3   2*  3   2*  3   3   2*  2*  3   3   3   3
3   2   3   3   3   3   2*  3   3   3   2*  3
3   3   2*  3   3   2*
```

Cluster	Size	WSSQ
1	50	15.15
2	62	39.82
3	38	23.88

Final cluster centroids

5.0060	3.4280	1.4620	0.2460
5.9016	2.7484	4.3935	1.4339
6.8500	3.0737	5.7421	2.0711

The data were maintained in the known group order (as in `manova1.tf5`), and the clusters assigned are seen to be identical to the known classification for group 1 (setosa), while limited misclassification has occurred for groups 2 (versicolor, 2 assigned to group 3), and 3 (viginica, 14 assigned to group 2), as shown by the starred values. Clearly group 1 is distinct from groups 2 and 3 which show some similarities to each other, a conclusion also illustrated in the next figure.



Example 3

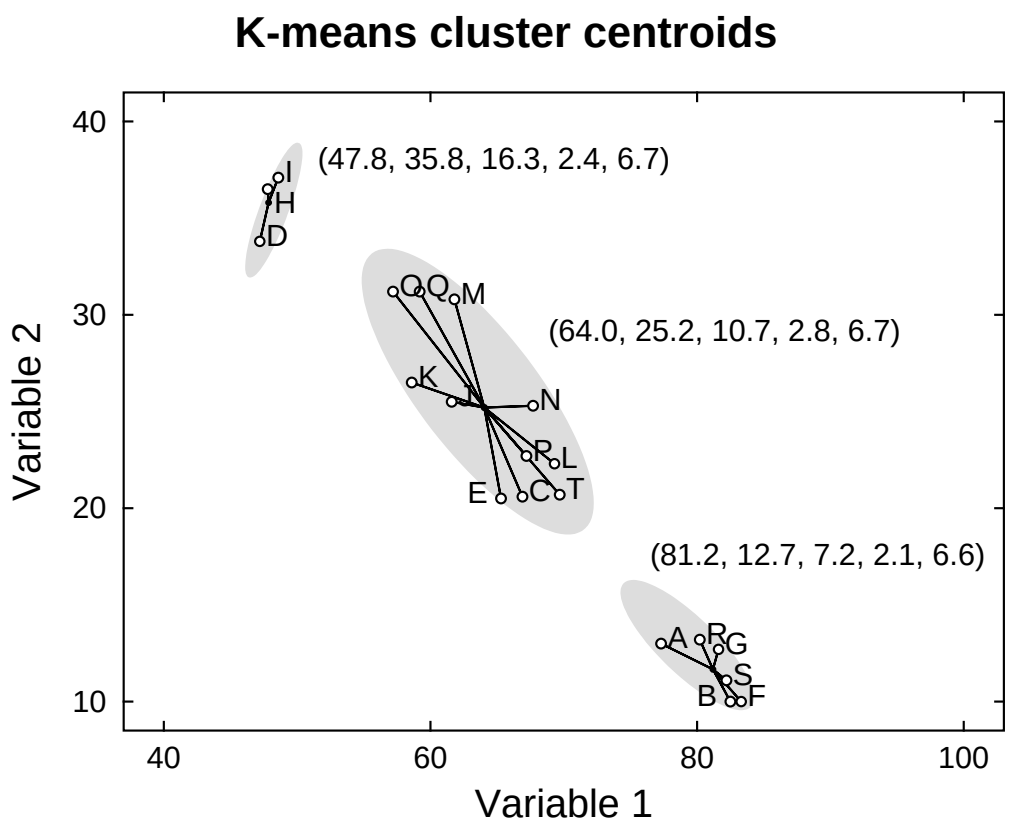
This example explains how to include additional features such as maps which are often added to plots to emphasize the meaning of clusters.

Stretching and clipping are also valuable when graphs have to be re-sized to achieve geometrically correct aspect ratios, as in the map shown in this next figure, which can be generated by the K-means clustering procedure using program **simstat** as follows.

- Input `ukmap.tf1` with coordinates for UK airports.
- Input `ukmap.tf2` with coordinates for starting centroids.
- Calculate centroids then transfer the plot to advanced graphics.
- Read in the UK coastal outline coordinates as an extra file from `ukmap.tf3`.
- Suppress axes, labels, and legends, then clip away extraneous white space.
- Stretch the PS output using the [Shape] then [Portrait +] options, and save the stretched eps file.

Example 4

It is frequently useful to be able highlight groups of data points in a two dimensional swarm, as in this figure.



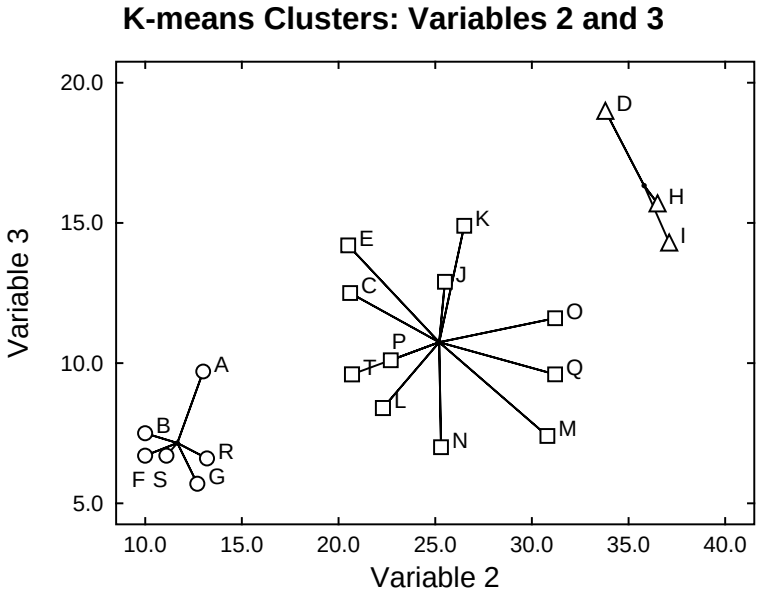
In this case a partition into three groups has been done by K-means clustering, and to appreciate how to use this technique, note that this figure can be generated by the K-means clustering procedure using program **simstat** as follows.

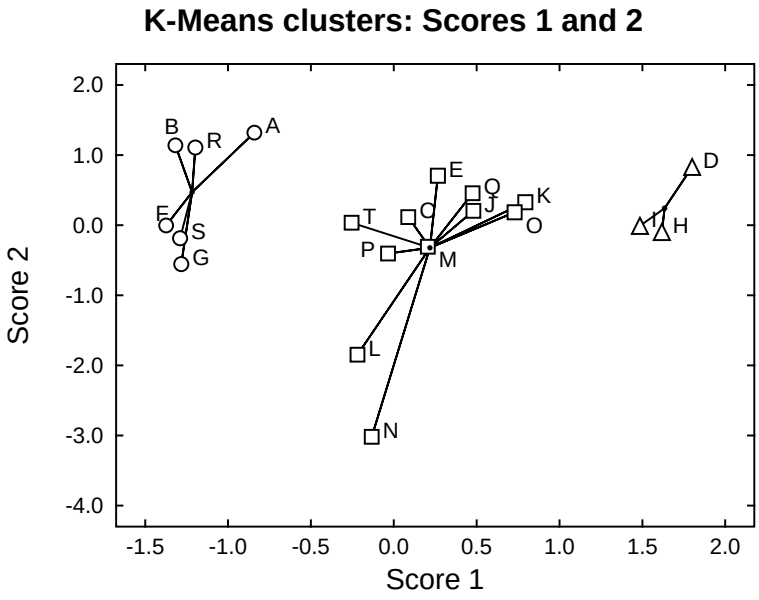
- Input the K-means clustering test file `kmeans.tf1`.
- Calculate the centroids, using the starting estimates appended to the test file. View them, which then adds them to the results file, then record the centroid coordinates from the results file.
- Select to plot the groups with associated labels, but then it will prove necessary to move several of the labels by substituting new labels, or shifting the *x* or *y* coordinates to clarify the graph.
- Add the solid background ellipses using the lines/arrows/boxes option because both head and tail coordinate must be specified using the red arrow, as well as an eccentricity value for the ellipses. Of course, any filled shapes such as circles, squares, or triangles can be chosen, and any size or color can be used.
- Add the centroid coordinates as extra text strings.

Of course, this technique can be used to highlight or draw attention to any subsets of data points, for instance groups in principal component analysis.

Example 5

This example considers the plotting of principal component scores instead of original variables, illustrated in the next figure for `kmeans.tf1`.





Note that, in the upper figure, symbols F, S, and P have been translated for clarity, and it should be compared to an earlier figure, for the same data with variables 1 and 2. This highlights an important point when plotting clusters for more than 2 variables: the plot shape depends on the variables chosen. So, for a more representative plot when there are more than 2 variables it is better to plot principal component scores instead of variables. The SIMFIT default is to plot the scores obtained using the correlation matrix technique, as this can prevent the analysis being dominated by columns with unduly large values.

In the lower figure, symbols B, O, M, and P have been translated for clarity, but now the principal component scores 1 and 2 have been plotted, which will usually be a better representation of the clustering, as the shape of the plot is not so strongly influenced by the variables chosen.

Theory

Once a n by m matrix of values a_{ij} for n cases and m variables has been provided, the cases can be sub-divided into K non-empty clusters where $K < n$, provided that a K by m matrix of starting estimates b_{ij} has been specified. The procedure is iterative, and proceeds by moving objects between clusters to minimize the objective function

$$\sum_{k=1}^{K} \sum_{i \in S_k} \sum_{j=1}^{m} w_i (a_{ij} - \bar{a}_{kj})^2$$

where S_k is the set of objects in cluster k and $\bar{a}_{kj}$ is the weighted sample mean for variable j in cluster k. The weighting factors w_i can allow for situations where the objects may not be of equal value, e.g., if replicates have been used to determine the a_{ij}.

Certain other aspects of the SIMFIT implementation of K-means clustering should be made clear.

1. If variables differ greatly in magnitude, data should be transformed before cluster analysis but note that, if this is done interactively, the same transformation will be applied to the starting clusters. If a transformation cannot be applied to data, clustering will not be allowed at all, but if a starting estimate cannot be transformed (e.g., square root of a negative number), then that particular value will remain untransformed.

2. If, after initial assignment of data to the starting clusters some are empty, clustering will not start, and a warning will be issued to decrease the number of clusters requested, or edit the starting clusters.

3. Clustering is an iterative procedure, and different starting clusters may lead to different final cluster assignments. So, to explore the stability of a cluster assignment, you can perturb the starting clusters by adding or multiplying by a random factor, or you can even generate a completely random starting set. For instance, if the data have been normalized to zero mean and unit variance, then choosing uniform random starting clusters from $U(-1, 1)$, or normally distributed values from $N(0, 1)$ might be considered.

4. After clusters have been assigned you may wish to pursue further analysis, say using the groups for canonical variate analysis, or as training sets for allocation of new observations to groups. To do this, you can create a SIMFIT MANOVA type file with group indicator in column 1. Such files also have the centroids appended, and these can be overwritten by new observations (not forgetting to edit the extra line counter following the last line of data) for allocating to the groups as training sets.

5. If weighting, variable suppression, or interactive transformation is used when assigning K-means clusters, all results tables, plots and MANOVA type files will be expressed in coordinates of the transformed space.

6. When viewing two dimensional plots of clusters where there are more than two variables, users can choose which coordinates to display, and this can give a misleading impression where it can seem that some cases have been wrongly assigned. This is to forget that the assignment is based on a selection process that uses all of the variables, and is a good reason to view using several pairs of coordinates to get a better overall picture, or to plot using principal components.

7. When displaying clusters as principal components, the loadings used to plot scores for data and centroids are calculated interactively from the data correlation matrix and standardized for unit variance. The scores are not used for further iterations to refine the clustering procedure.