K-Means Cluster Analysis

Contents

K-Means Cluster Analysis

K-means cluster analysis is a tool designed to assign cases to a fixed number of groups (clusters) whose characteristics are not yet known but are based on a set of specified variables. It is most useful when you want to classify a large number (thousands) of cases.

A good cluster analysis is:

- **Efficient.** Uses as few clusters as possible.
- **Effective.** Captures all statistically and commercially important clusters. For example, a cluster with five customers may be statistically different but not very profitable.

Next

Clustering Principles

Using K-Means to Classify Customers

Related Procedures

Recommended Readings

1. Clustering Principles

The K-Means Cluster Analysis procedure begins with the construction of initial cluster centers. You can assign these yourself or have the procedure select *k* well-spaced observations for the cluster centers.

After obtaining initial cluster centers, the procedure:

- Assigns cases to clusters based on distance from the cluster centers.
- Updates the locations of cluster centers based on the mean values of cases in each cluster.

These steps are repeated until any reassignment of cases would make the clusters more internally variable or externally similar.

Next

Parent topic: K-Means Cluster Analysis

2. Using K-Means to Classify Customers

A telecommunications provider wants to segment its customer base by service usage patterns. If customers can be classified by usage, the company can offer more attractive packages to its customers.

Standardized variables indicating service usage are contained in *telco_extra.sav*. See the topic Sample Files for more information. Use the K-Means Cluster Analysis procedure to find subsets of "similar" customers.

Next

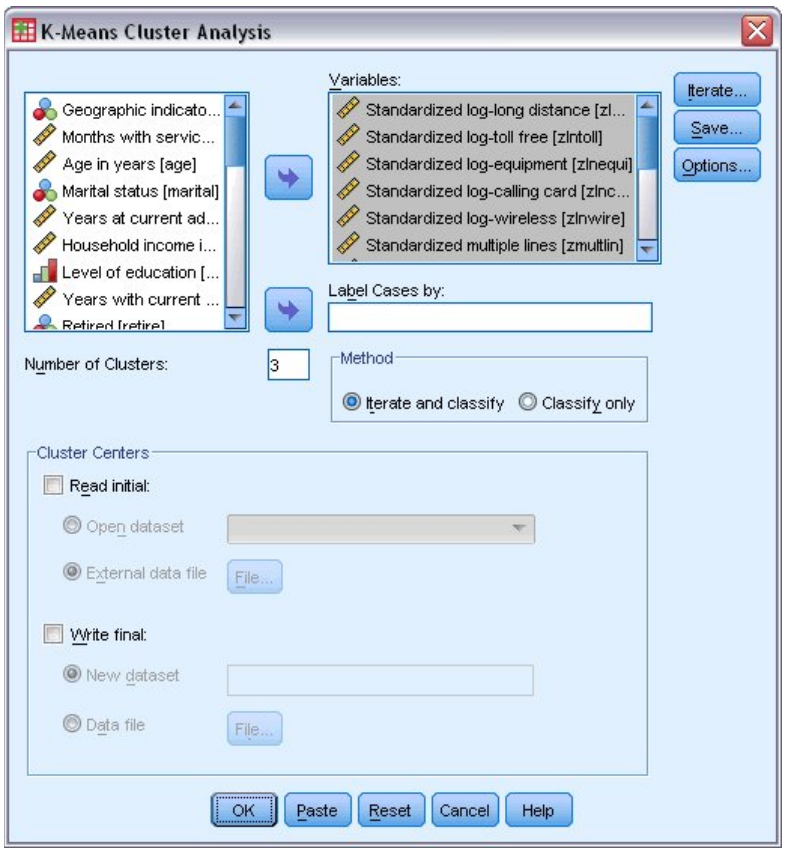
Parent topic: K-Means Cluster Analysis

2.1. Running the Analysis

1. To run the cluster analysis, from the menus choose:

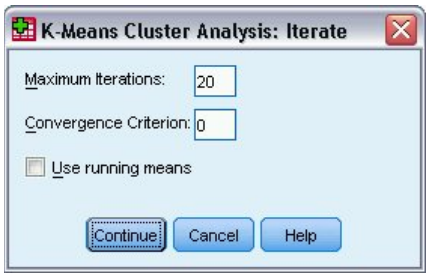
   Analyze > **Classify** > **K-Means Cluster...**

Figure 1. K-Means Cluster Analysis dialog box



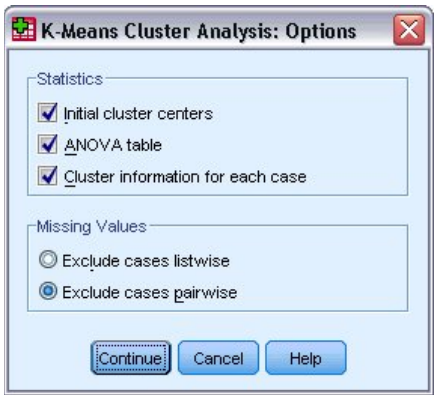
2. If the variable list does not display variable labels in file order, right-click anywhere in the variable list and from the context menu choose **Display Variable Labels** and **Sort by File Order**.
3. Select *Standardized log-long distance* through *Standardized log-wireless* and *Standardized multiple lines* through *Standardized electronic billing* as analysis variables.
4. Type `3` as the number of clusters.
5. Click **Iterate**.

Figure 2. Iterate dialog box



6. Type `20` as the maximum iterations.
7. Click **Continue**.
8. Click **Options** in the K-Means Cluster Analysis dialog box.

Figure 3. Options dialog box



9. Select **ANOVA table** and **Cluster information for each group** in the Statistics group.
10. Select **Exclude cases pairwise** in the Missing Values group. There are many missing values due to the fact that most customers do not subscribe to all services, so excluding cases *pairwise* maximizes the information you can obtain from the data... at the cost of possibly biasing the results.
11. Click **Continue**, then click **OK** in the K-Means Cluster Analysis dialog box.

Next

Parent topic: Using K-Means to Classify Customers

2.2. Initial Cluster Centers

Figure 1. Initial cluster centers for three-cluster solution

	Cluster		
	1	2	3
Standardized log-long distance	2.48	-1.70	.12
Standardized log-toll free	2.34	-.20	-.39
Standardized log-equipment	1.34	-.65	.59
Standardized log-calling card	2.49	-.86	-1.28
Standardized log-wireless	1.14	-1.75	1.42
Standardized multiple lines	1.05	-.95	1.05
Standardized voice mail	1.51	1.51	1.51
Standardized paging	1.68	1.68	1.68
Standardized internet	1.31	-.76	1.31
Standardized caller id	1.04	1.04	-.96
Standardized call waiting	1.03	-.97	1.03
Standardized call forwarding	1.01	1.01	-.99
Standardized 3-way calling	1.00	1.00	-1.00
Standardized electronic billing	-.77	-.77	1.30

The initial cluster centers are the variable values of the *k* well-spaced observations.

Next

Parent topic: Using K-Means to Classify Customers

2.3. Iteration History

Figure 1. Iteration history for three-cluster solution

Iteration	Change in Cluster Centers		
	1	2	3
1	3.298	3.590	3.491
2	1.016	.427	.931
3	.577	.320	.420
4	.240	.180	.195
5	.119	.125	.108
6	.093	.083	.027
7	.069	.094	.032
8	.059	.051	.018
9	.035	.085	.063
10	.025	.359	.333
11	.068	.439	.287
12	.079	.368	.177
13	.125	.139	.078
14	.077	.096	.020
15	.041	.047	.015
16	.014	.027	.000
17	.019	.038	.000
18	.000	.000	.000

The iteration history shows the progress of the clustering process at each step. In early iterations, the cluster centers shift quite a lot. By the 14th iteration, they have settled down to the general area of their final location, and the last four iterations are minor adjustments.

If the algorithm stops because the maximum number of iterations is reached, you may want to increase the maximum because the solution may otherwise be unstable. For example, if you had left the maximum number of iterations at 10, the reported solution would still be in a state of flux.

Next

Parent topic: Using K-Means to Classify Customers

2.4. ANOVA

Figure 1. ANOVA table for three-cluster solution

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Standardized log-long distance	13.063	2	.976	997	13.387	.000
Standardized log-toll free	43.418	2	.820	472	52.932	.000
Standardized log-equipment	99.056	2	.488	383	202.999	.000
Standardized log-calling card	6.301	2	.984	675	6.402	.002
Standardized log-wireless	52.879	2	.646	293	81.873	.000
Standardized multiple lines	38.032	2	.926	997	41.084	.000
Standardized voice mail	236.301	2	.528	997	447.554	.000
Standardized paging	298.992	2	.402	997	743.348	.000
Standardized internet	123.447	2	.754	997	163.642	.000
Standardized caller id	308.104	2	.384	997	802.474	.000
Standardized call waiting	294.674	2	.411	997	717.172	.000
Standardized call forwarding	288.343	2	.424	997	680.718	.000
Standardized 3-way calling	262.397	2	.476	997	551.678	.000
Standardized electronic billing	112.782	2	.776	997	145.381	.000

The ANOVA table indicates which variables contribute the most to your cluster solution. Variables with large *F* values provide the greatest separation between clusters.

Next

Parent topic: Using K-Means to Classify Customers

2.5. Final Cluster Centers

Figure 1. Final cluster centers for three-cluster solution

	Cluster		
	1	2	3
Standardized log-long distance	.05	.22	-.16
Standardized log-toll free	.24	.12	-1.05
Standardized log-equipment	.81	-.19	-.69
Standardized log-calling card	.17	.02	-.17
Standardized log-wireless	.42	-.75	-1.00
Standardized multiple lines	.48	-.29	-.05
Standardized voice mail	1.26	-.24	-.44
Standardized paging	1.43	-.38	-.44
Standardized internet	.81	-.59	-.02
Standardized caller id	.82	.71	-.81
Standardized call waiting	.76	.72	-.80
Standardized call forwarding	.78	.69	-.79
Standardized 3-way calling	.74	.67	-.75
Standardized electronic billing	.70	-.63	.05

The final cluster centers are computed as the mean for each variable within each final cluster. The final cluster centers reflect the characteristics of the typical case for each cluster.

- Customers in cluster 1 tend to be big spenders who purchase a lot of services.
- Customers in cluster 2 tend to be moderate spenders who purchase the "calling" services.
- Customers in cluster 3 tend to spend very little and do not purchase many services.

Next

Parent topic: Using K-Means to Classify Customers

2.6. Distances between Final Cluster Centers

Figure 1. Distances between final cluster centers for three-cluster solution

Cluster	1	2	3
1		3.500	4.863
2	3.500		3.396
3	4.863	3.396	

This table shows the *Euclidean distances* between the final cluster centers. Greater distances between clusters correspond to greater dissimilarities.

- Clusters 1 and 3 are most different.
- Cluster 2 is approximately equally similar to clusters 1 and 3.

These relationships between the clusters can also be intuited from the final cluster centers, but this becomes more difficult as the number of clusters and variables increases.

Next

Parent topic: Using K-Means to Classify Customers

2.7. Number of Cases in Each Cluster

Figure 1. Number of cases in each cluster for three-cluster solution

Cluster	1	226.000
	2	292.000
	3	482.000
Valid		1000.000
Missing		.000

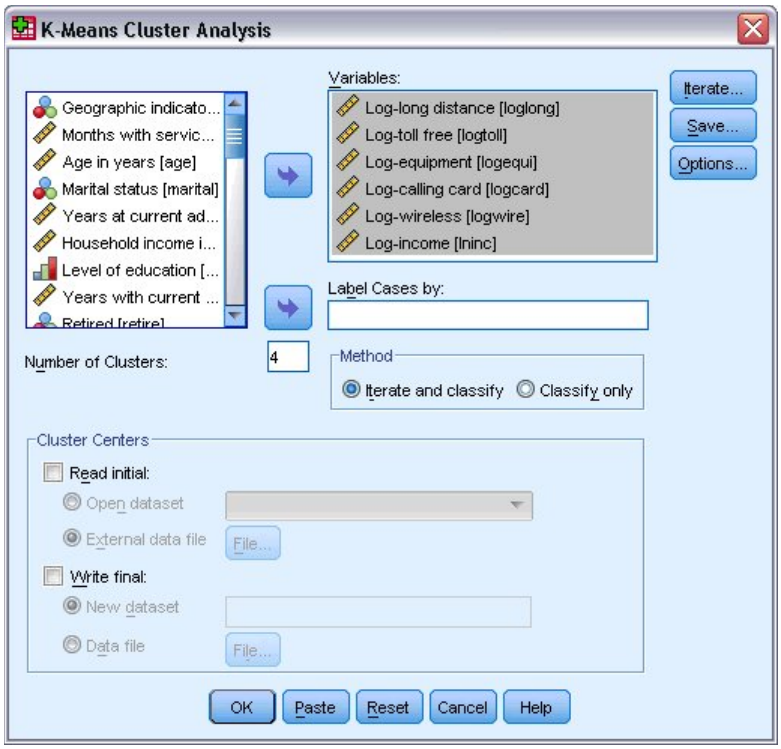
A large number of cases were assigned to the third cluster, which unfortunately is the least profitable group. Perhaps a fourth, more profitable, cluster could be extracted from this "basic service" group.

Next

Parent topic: Using K-Means to Classify Customers

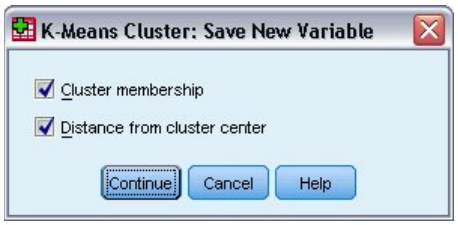
2.8. Building a Four-Cluster Solution

Figure 1. K-Means Cluster Analysis dialog box



1. To run a cluster analysis with four clusters, reopen the K-Means Cluster Analysis dialog box.
2. Type `4` as the number of clusters.
3. Click **Save**.

Figure 2. Save dialog box

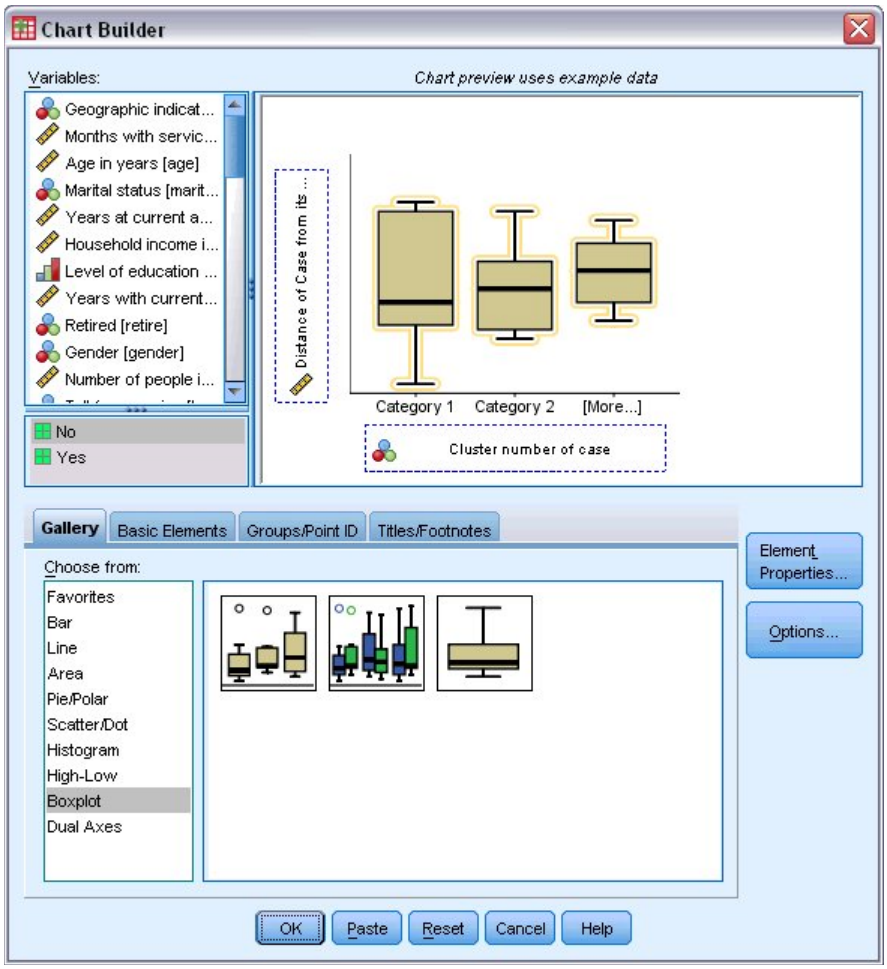


4. Select **Cluster membership** and **Distance from cluster center**.
5. Click **Continue**.
6. Click **OK** in the K-Means Cluster Analysis dialog box.
7. The saved variables can be used to create a useful boxplot. From the menus, choose:

   Graphs > **Chart Builder...**

8. Click the Gallery tab, select Boxplot from the list of chart types, and drag and drop the Simple Boxplot icon onto the canvas.
9. Drag and drop *Distance of Case from its Classification Cluster Center* onto the y axis.
10. Drag and drop *Cluster Number of Case* onto the x axis.
11. Click **OK** to create the boxplot.

Figure 3. Chart Builder



Next

Plot of Distances from Cluster Center by Cluster Membership

Final Cluster Centers

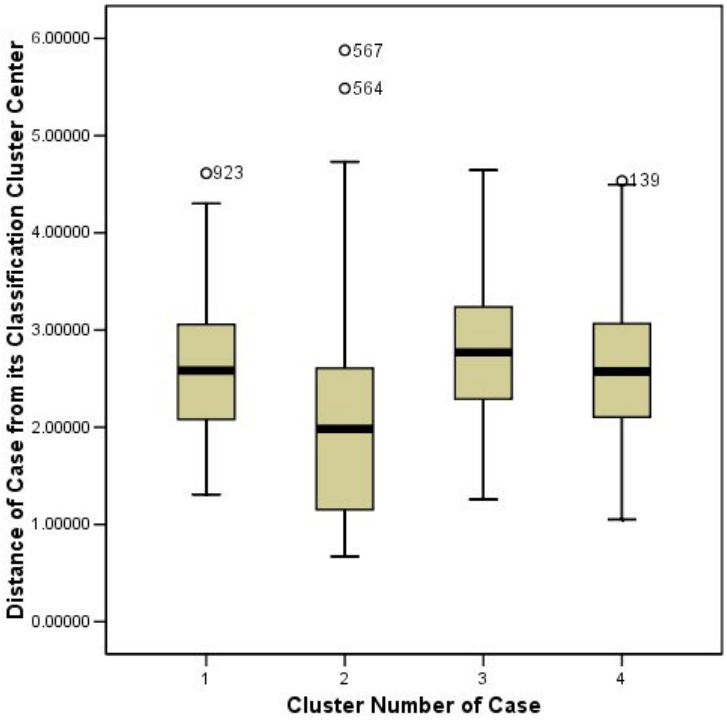
Distances between Final Cluster Centers

Number of Cases in Each Cluster

Parent topic: Using K-Means to Classify Customers

2.8.1. Plot of Distances from Cluster Center by Cluster Membership

Figure 1. Plot of distances from cluster center by cluster membership for four-cluster solution



This is a diagnostic plot that helps you to find outliers within clusters. There is a lot of variability in cluster 2, but all the distances are within reason.

Next

Parent topic: Building a Four-Cluster Solution

2.8.2. Final Cluster Centers

Figure 1. Final cluster centers for four-cluster solution

	Cluster			
	1	2	3	4
Standardized log-long distance	.23	-.48	.05	.24
Standardized log-toll free	-.75	-1.10	.26	.11
Standardized log-equipment	-.36	-1.28	.86	-.20
Standardized log-calling card	-.06	-.37	.18	.06
Standardized log-wireless	-.84	-1.54	.45	-.71
Standardized multiple lines	.75	-.74	.45	-.26
Standardized voice mail	-.21	-.59	1.30	-.23
Standardized paging	-.32	-.51	1.50	-.37
Standardized internet	.57	-.52	.77	-.56
Standardized caller id	-.78	-.73	.86	.72
Standardized call waiting	-.76	-.73	.80	.74
Standardized call forwarding	-.71	-.82	.83	.76
Standardized 3-way calling	-.69	-.79	.84	.72
Standardized electronic billing	.58	-.38	.67	-.63

This table shows that an important grouping is missed in the three-cluster solution. Members of clusters 1 and 2 are largely drawn from cluster 3 in the three-cluster solution, and they are unlikely to be big spenders. However, members of cluster 1 are highly likely to purchase Internet-related services, which establishes them as a distinct and possibly profitable group.

Clusters 3 and 4 seem to correspond to clusters 1 and 2 from the three-cluster solution.

Next

Parent topic: Building a Four-Cluster Solution

2.8.3. Distances between Final Cluster Centers

Figure 1. Distances between final cluster centers for four-cluster solution

Cluster	1	2	3	4
1		2.568	4.454	3.631
2	2.568		5.746	3.675
3	4.454	5.746		3.515
4	3.631	3.675	3.515	

The distances between the clusters have not changed greatly.

- Clusters 1 and 2 are the most similar, which makes sense because they were combined into one cluster in the three-cluster solution.
- Clusters 2 and 3 are the most dissimilar, since they represent opposite spending behaviors.
- Cluster 4 is still equally similar to the other clusters.

Next

Parent topic: Building a Four-Cluster Solution

2.8.4. Number of Cases in Each Cluster

Figure 1. Number of cases in each cluster for four-cluster solution

Cluster	1	236.000
	2	272.000
	3	212.000
	4	280.000
Valid		1000.000
Missing		.000

Nearly 25% of cases belong to the newly created group of "E-service" customers, which is very significant to your profits.

Next

Parent topic: Building a Four-Cluster Solution

2.9. Summary

Using *k*-means cluster analysis, you initially grouped the customers into three clusters. However, this solution was not very satisfactory, so you reran the analysis with four clusters. These results were

better, and from the final cluster centers, you saw that a potentially profitable "Internet" grouping was missed in the three-cluster solution.

This example underscores the exploratory nature of cluster analysis, since it is impossible to determine the "best" number of clusters until you have run the analyses and examined the solutions.

The next step for the company is to try to construct a model that classifies the customers according to their demographic information. With such a model, the company can customize offers for individual prospective customers. For information on how the company builds such a model, see Using Discriminant Analysis to Classify Telecommunications Customers.

Next

Parent topic: Using K-Means to Classify Customers

3. Related Procedures

The K-Means Cluster Analysis procedure is a tool for finding natural groupings of cases, given their values on a set of variables. It is most useful when you want to classify a large number (thousands) of cases.

- The TwoStep Cluster Analysis procedure allows you to use both categorical and continuous variables, and can automatically select the "best" number of clusters.
- If you want to cluster variables instead of cases, or have a small number of cases, try the Hierarchical Cluster Analysis procedure.
- If your *k*-means analysis is part of a segmentation solution, these newly created clusters can be analyzed in the Discriminant Analysis procedure.

Next

Parent topic: K-Means Cluster Analysis

4. Recommended Readings

See the following texts for more information on *k*-means cluster analysis:

Aldenderfer, M. S., and R. K. Blashfield. 1984. *Cluster Analysis*. Newbury Park: Sage Publications.

Parent topic: K-Means Cluster Analysis