TwoStep Cluster Analysis

Contents

TwoStep Cluster Analysis

The TwoStep Cluster Analysis procedure is an exploratory tool designed to reveal natural groupings (or clusters) within a data set that would otherwise not be apparent. The algorithm employed by this procedure has several desirable features that differentiate it from traditional clustering techniques:

- The ability to create clusters based on both categorical and continuous variables.
- Automatic selection of the number of clusters.
- The ability to analyze large data files efficiently.

Next

Clustering Principles

Using TwoStep Cluster Analysis to Classify Motor Vehicles

Related Procedures

1. Clustering Principles

In order to handle categorical and continuous variables, the TwoStep Cluster Analysis procedure uses a likelihood distance measure which assumes that variables in the cluster model are independent. Further, each continuous variable is assumed to have a normal (Gaussian) distribution and each categorical variable is assumed to have a multinomial distribution. Empirical internal testing indicates that the procedure is fairly robust to violations of both the assumption of independence and the distributional assumptions, but you should try to be aware of how well these assumptions are met.

The two steps of the TwoStep Cluster Analysis procedure's algorithm can be summarized as follows:

Step 1. The procedure begins with the construction of a Cluster Features (CF) Tree. The tree begins by placing the first case at the root of the tree in a leaf node that contains variable information about that case. Each successive case is then added to an existing node or forms a new node, based upon its similarity to existing nodes and using the distance measure as the similarity criterion. A node that

contains multiple cases contains a summary of variable information about those cases. Thus, the CF tree provides a capsule summary of the data file.

Step 2. The leaf nodes of the CF tree are then grouped using an agglomerative clustering algorithm. The agglomerative clustering can be used to produce a range of solutions. To determine which number of clusters is "best", each of these cluster solutions is compared using Schwarz's Bayesian Criterion (BIC) or the Akaike Information Criterion (AIC) as the clustering criterion.

Next

Parent topic: TwoStep Cluster Analysis

2. Using TwoStep Cluster Analysis to Classify Motor Vehicles

Car manufacturers need to be able to appraise the current market to determine the likely competition for their vehicles. If cars can be grouped according to available data, this task can be largely automatic using cluster analysis.

Information for various makes and models of motor vehicles is contained in *car_sales.sav*. See the topic Sample Files for more information. Use the TwoStep Cluster Analysis procedure to group automobiles according to their prices and physical properties.

Next

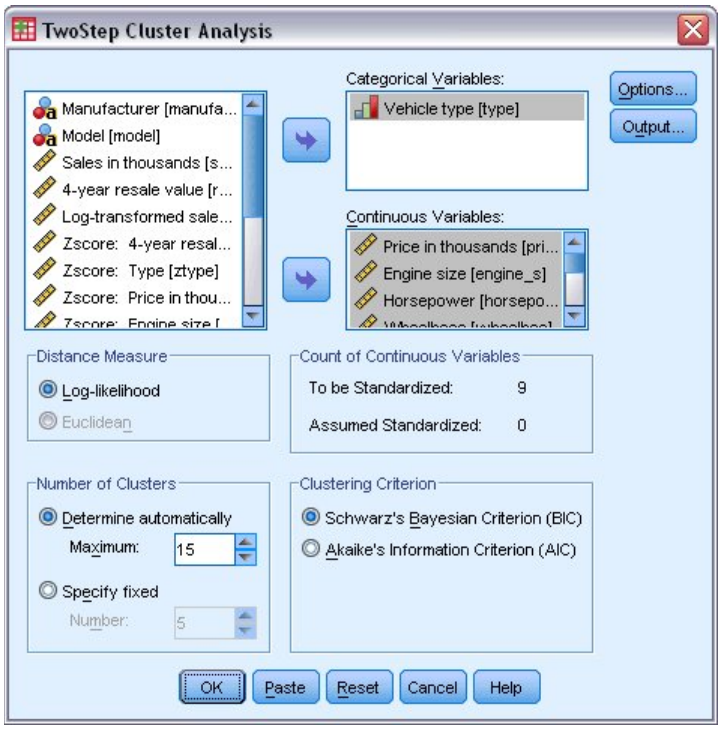
Parent topic: TwoStep Cluster Analysis

2.1. Running the Analysis

1. To run a TwoStep Cluster Analysis, from the menus choose:

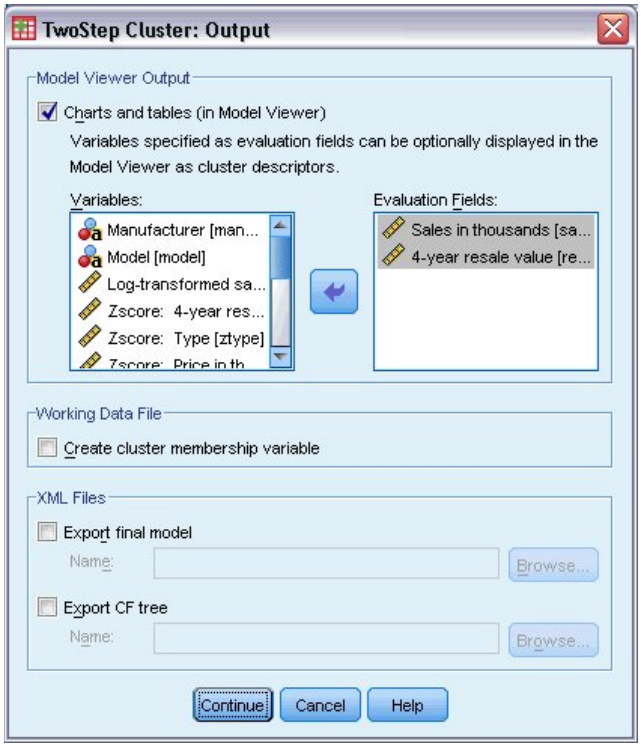
   Analyze > **Classify** > **TwoStep Cluster...**

   Figure 1. TwoStep Cluster Analysis main 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 *Vehicle type* as a categorical variable.
4. Select *Price in thousands* through *Fuel efficiency* as continuous variables.
5. Click **Output**.

Figure 2. TwoStep Cluster Analysis main dialog box



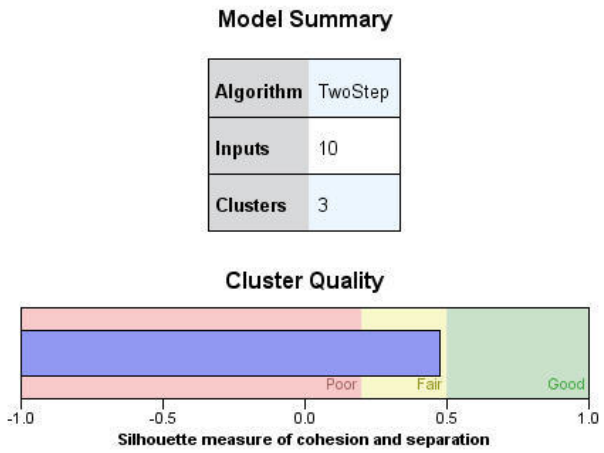
6. Select *Sales in thousands [sales]* and *4-year resale value [resale]* as evaluation fields. These fields will not be used to create the cluster model, but can give you further insight to the clusters created by the procedure.
7. Click **Continue** and then click **OK**.

Next

Parent topic: Using TwoStep Cluster Analysis to Classify Motor Vehicles

2.2. Model Summary and Cluster Quality

Figure 1. Model Summary view



The Viewer contains a Model Viewer object. By activating (double-clicking) this object, you gain an interactive view of the model. The default main view is the Model Summary view.

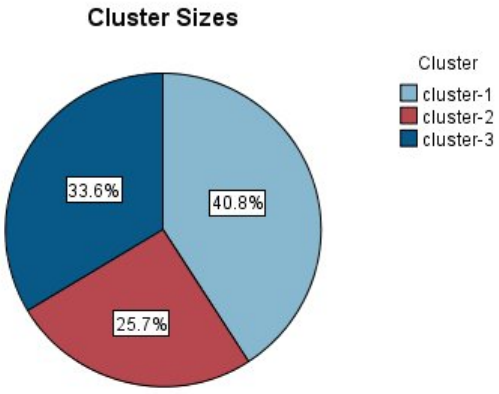
- The model summary table indicates that three clusters were found based on the ten input features (fields) you selected.
- The cluster quality chart indicates that the overall model quality is "Fair".

Next

Parent topic: Using TwoStep Cluster Analysis to Classify Motor Vehicles

2.3. Cluster Distribution

Figure 1. Cluster distribution table



Size of Smallest Cluster	39 (25.7%)
Size of Largest Cluster	62 (40.8%)
Ratio of Sizes: Largest Cluster to Smallest Cluster	1.59

The Cluster Sizes view shows the frequency of each cluster. Hovering over a slice in the pie chart reveals the number of records assigned to the cluster. 40.8% (62) of the records were assigned to the first cluster, 25.7% (39) to the second, and 33.6% (51) to the third.

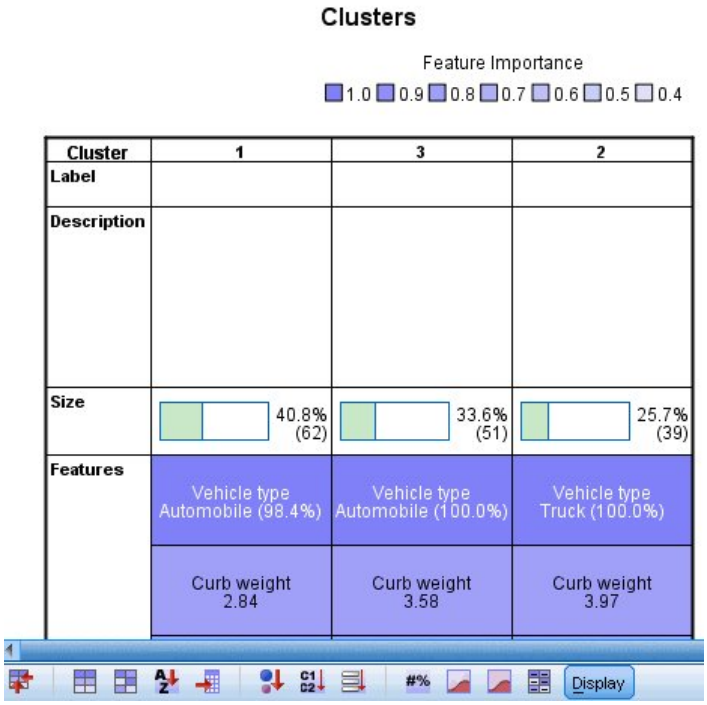
Next

Parent topic: Using TwoStep Cluster Analysis to Classify Motor Vehicles

2.4. Cluster Profiles

1. In the main view, select Clusters from the dropdown to display the Clusters view.

   Figure 1. Clusters table



By default, clusters are sorted from left to right by cluster size, so they are currently ordered 1, 3, 2.

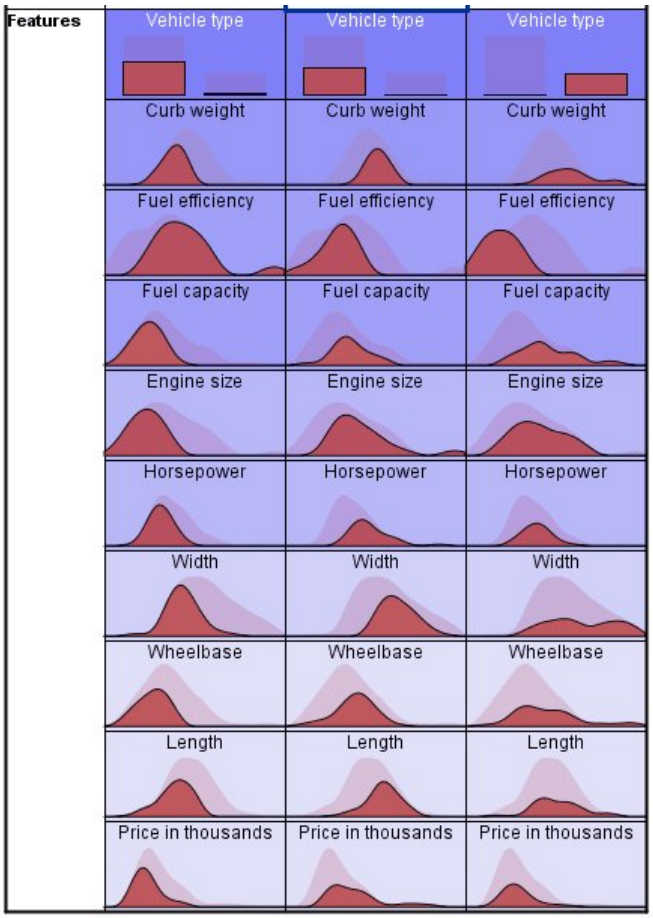
Figure 2. Cluster profiles: cells show cluster centers

Features	Vehicle type Automobile (98.4%)	Vehicle type Automobile (100.0%)	Vehicle type Truck (100.0%)
	Curb weight 2.84	Curb weight 3.58	Curb weight 3.97
	Fuel efficiency 27.24	Fuel efficiency 23.02	Fuel efficiency 19.51
	Fuel capacity 15.00	Fuel capacity 18.40	Fuel capacity 22.10
	Engine size 2.20	Engine size 3.70	Engine size 3.60
	Horsepower 143.24	Horsepower 232.96	Horsepower 187.92
	Width 68.50	Width 72.90	Width 72.70
	Wheelbase 102.60	Wheelbase 109.00	Wheelbase 113.00
	Length 178.20	Length 194.70	Length 191.10
	Price in thousands 19.62	Price in thousands 37.30	Price in thousands 26.56

The cluster means suggest that the clusters are well separated.

- Motor vehicles in cluster 1 are cheap, small, and fuel efficient automobiles, except for a single truck (the 1.6% of the cluster not comprised of automobiles).
- Motor vehicles in cluster 2 (column 3) are moderately priced, heavy, and have a large gas tank, presumably to compensate for their poor fuel efficiency. Cluster 2 is also entirely comprised of trucks.
- Motor vehicles in cluster 3 (column 2) are expensive, large, and are moderately fuel efficient automobiles.

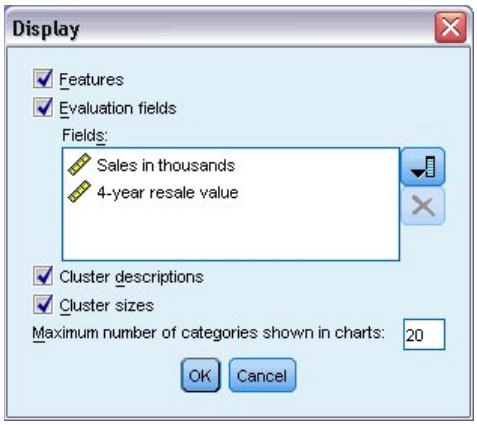
Figure 3. Cluster profiles: cells show absolute distributions



2. The cluster means (for continuous fields) and modes (for categorical fields) are useful, but only give information about the cluster centers. In order to get a visualization of the distribution of values for each field by cluster, click on the **Cells show absolute distributions** button in the toolbar.

   Now you can see, for example, that there is some overlap between clusters 1 and 3 on curb weight, engine size, and fuel capacity. There is considerably more overlap between clusters 2 and 3 on these fields, with the difference that the vehicles with the very highest curb weight and fuel capacity are in cluster 2 (column 3) and the vehicles with the very highest engine size appear to be in cluster 3 (column 2).

   Figure 4. Cluster profiles: cells show absolute distributions



3. To see this information for the evaluation fields, click on the **Display** button in the toolbar.
4. Select **Evaluation fields**.
5. Click **OK**.

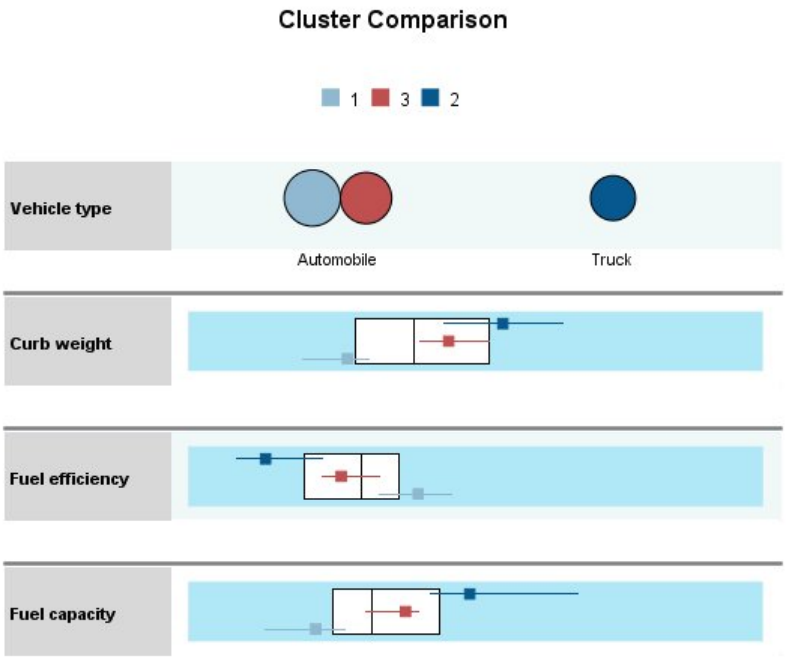
The evaluation fields should now appear in the cluster table.

Figure 5. Cluster profiles for evaluation fields: cells show absolute distributions

The distribution of sales is similar across clusters, except that clusters 1 and 2 have longer tails than cluster 3 (column 2). There is a fair amount of overlap in the distributions of 4-year resale value, but clusters 2 and 3 are centered on a higher value than cluster 1, and cluster 3 has a longer tail than either cluster 1 or 2.

6. For another way to compare clusters, select (control-click) on the cluster numbers (column headings) in the clusters table.
7. In the auxiliary view, select **Cluster Comparison** from the dropdown.

Figure 6. Cluster comparison view : first four fields shown



For each categorical field, this shows a dot plot for the modal category of each cluster, with dot size corresponding to the percentage of records. For continuous fields, this shows a boxplot for the distribution of values within each cluster overlaid on a boxplot for the distribution of values overall. These plots generally confirm what you've seen in the Clusters view. The Cluster Comparison view can be especially helpful when there are many clusters, and you want to compare only a few of them.

Next

Parent topic: Using TwoStep Cluster Analysis to Classify Motor Vehicles

2.5. Summary

Using the TwoStep Cluster Analysis procedure, you have separated the motor vehicles into three fairly broad categories. In order to obtain finer separations within these groups, you should collect information on other attributes of the vehicles. For example, you could note the crash test performance or the options available.

Next

Parent topic: Using TwoStep Cluster Analysis to Classify Motor Vehicles

3. Related Procedures

The TwoStep Cluster Analysis procedure is useful for finding natural groupings of cases or variables. It works well with categorical and continuous variables, and can analyze very large data files.

- If you have a small number of cases, and want to choose between several methods for cluster formation, variable transformation, and measuring the dissimilarity between clusters, try the Hierarchical Cluster Analysis procedure. The Hierarchical Cluster Analysis procedure also allows you to cluster variables instead of cases.
- The K-Means Cluster Analysis procedure is limited to scale variables, but can be used to analyze large data and allows you to save the distances from cluster centers for each object.

Parent topic: TwoStep Cluster Analysis