

CLUSTER ANALYSIS

Analyses de classification

Cluster analysis

- Elles font partie des analyses de ressemblances : regrouper les observations en un nombre réduit de groupes (clusters) homogènes. Les observations d'un groupe sont le plus semblables possible et les groupes les plus différents possibles. → Découvrir les groupes naturels au sein d'une population donnée.
- Objectifs :
 - identifier une typologie ;
 - test d'hypothèses ;
 - analyse exploratoire des données ;
 - réduction des données ;
 - analyse prédictive sur base de groupes ;
 - ...
 -
 - Proximité des observations = similarité
 - Distance entre les groupes = dissemblance / stratégie d'agrégation

Les choix à faire :

- les individus à classer ;
- les variables à utiliser :
 - standardisation ??? (quand les unités de mesure sont très différentes ou quand les ordres de grandeurs sont très différents) ;
 - analyse en composantes principales / analyse factorielle des correspondances multiples ??? → travailler avec les composantes/ les facteurs ;
- la stratégie d'agrégation ;
- mesure de la distance entre les observations.

- Il est déconseillé d'effectuer une rotation dans le cadre d'une analyse de classification.
- Il est préférable de prendre en compte tous les facteurs jusqu'à obtenir une part de variance expliquée au moins égale à 90 %.

Les méthodes non hiérarchiques

- On les appelle aussi méthodes de partitionnement. Elles divisent les observations en un nombre fixé au préalable de classes.
- Le problème des méthodes de partitionnement est la définition a priori du nombre de classes. Ceci peut conduire à un partitionnement inadéquat. Il faut donc essayer plusieurs combinaisons avant de trouver celle qui correspond le mieux.
- Inertie totale du nuage de points : moyenne des carrés des distances entre chaque point et le centre de gravité du nuage de points. C'est aussi la somme de l'inertie intraclasse et de l'inertie interclasse. Une classe est d'autant plus homogène que son inertie est faible.
- Inertie intraclasse : moyenne des inerties de chaque classe calculées par rapport à leurs centres de gravité respectifs.
- Inertie interclasse : dispersion des k centres de gravité autour du centre de gravité du nuage de points.

- **Regroupement autour de centres mobiles**
- **Méthode des nuées dynamiques**
- ...

Les méthodes hiérarchiques

- On les appelle aussi agglomératives. A chaque étape, elles regroupent les deux classes les plus proches → fusion des individus dans des classes de plus en plus vastes.
- Les différentes partitions sont représentées sous la forme d'un arbre de classification (dendrogramme).
- Le niveau d'agrégation est le niveau auquel ont lieu les regroupements. C'est un indice numérique qui indique l'hétérogénéité des parties regroupées.
- Méthode du 'voisin le plus proche'
- Méthode du 'voisin le plus éloigné'
- Méthode de la 'moyenne du groupe'
- Méthode de la médiane
- ...

Méthode de Ward

- A chaque étape de l'analyse, il est possible de mesurer la perte d'information par la somme des carrés des écarts de chaque point-individu par rapport à la moyenne du groupe auquel il appartient. A chaque étape, c'est le regroupement qui créant la perte minimum d'information qui est effectué → Méthode basée sur la perte d'inertie la plus faible.
- Ce n'est pas la meilleure partition en k classes qui est obtenue mais la meilleure de celles obtenues par réunion de deux classes de la partition en $k+1$ classes.
- Cette méthode favorise surtout le regroupement des petites classes proches.
- Choix du nombre de groupes :
- à partir du dendrogramme : quand il y a des variations importantes de l'indice de dissimilarité. En ordonnée, on observe le niveau auquel a eu lieu chaque fusion successive, ce niveau correspond à la perte d'information (de variance / d'inertie) occasionnée par chaque fusion.
- Graphique représentant la perte de la variance (inertie) en fonction du nombre de fusions → basé sur l'analyse de la perte d'information à chaque niveau d'agrégation ou sur de la proportion de variance conservée à chaque étape.

ENAF 1986

agissant particulièrement sur des comportements comme l'allaitement et/ou la contraception.

3.5. Régimes de fécondité selon le secteur d'habitat.

Tableau 7 Répartition des mères (15-49) selon le secteur d'habitat et le régime de fécondité

	Métropoles	Urbain	Rural
Traditionnel	.19	.22	.58
Moderne	.18	.22	.60
Survivant	.16	.24	.59
	.18	.23	.59

Le tableau 7 montre une distribution presque équivalente des 3 régimes selon les secteurs d'habitat : environ 18 % en métropoles, 23 % en villes et 59 % en milieu rural. Les écarts sont vraiment limités.

Voilà un signe de l'inefficacité des critères a priori à saisir véritablement les courants profonds qui animent la dynamique de population. Autrement dit, les trois régimes se retrouvent dans chacun des secteurs et il n'est plus question d'associer aux métropoles, l'attribut "moderne" ni au milieu rural, l'attribut "traditionnel".

Comment alors expliquer les différences de niveau de fécondité entre secteurs d'habitat ? C'est que un régime de fécondité ne se définit pas par un indice unique, fut-il de fécondité, mais par une constellation d'indicateurs socio-biologiques (mariage, contraception, allaitement, attitudes et opinions etc...). De plus, nous avons établi au chapitre 1 la convergence qui se dessine entre la fécondité urbaine et rurale.

3.6 Régimes de fécondité et niveau d'instruction des femmes.

Tableau 8 Répartition des mères (15-49) selon le niveau d'instruction de la femme et le régime de fécondité.

	Illettrées	Coran- Primaire	Moyen secondaire	Supérieur	
<u>régime 1</u>					
moins de 30 ans	-	-	-	-	
30 ans et plus	.87	.11	.02	-	(1019)
<u>régime 2</u>					
moins de 30 ans	.72	.23	.05	-	(192)
30 ans et plus	.65	.27	.07	.01	(785)
<u>régime 3</u>					
moins de 30 ans	.61	.22	.15	-	(1304)
30 ans et plus	.57	.30	.13	.01	(362)

L'effet de la politique de scolarisation massive est visible : indépendamment des régimes démographiques, il y a un effet de générations qui l'emporte. Les plus jeunes sont systématiquement plus instruits.

Autrement dit, l'instruction à elle seule, comme la résidence ainsi qu'il est apparu plus haut, n'est pas suffisante pour garantir l'accès à un régime "moderne". La qualité de cette instruction est certainement à mettre en cause. Le prix payé pour l'éducation de masse a été la qualité de l'enseignement. C'est ainsi qu'on retrouvera des femmes en moyenne plus instruites, parce que plus jeunes et ayant bénéficié de la politique en matière d'éducation nationale mais appartenant à un régime traditionnel.

Les mêmes remarques restent valables lorsqu'on s'intéresse à l'instruction du mari.

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.

TwoStep Cluster Analysis >

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.

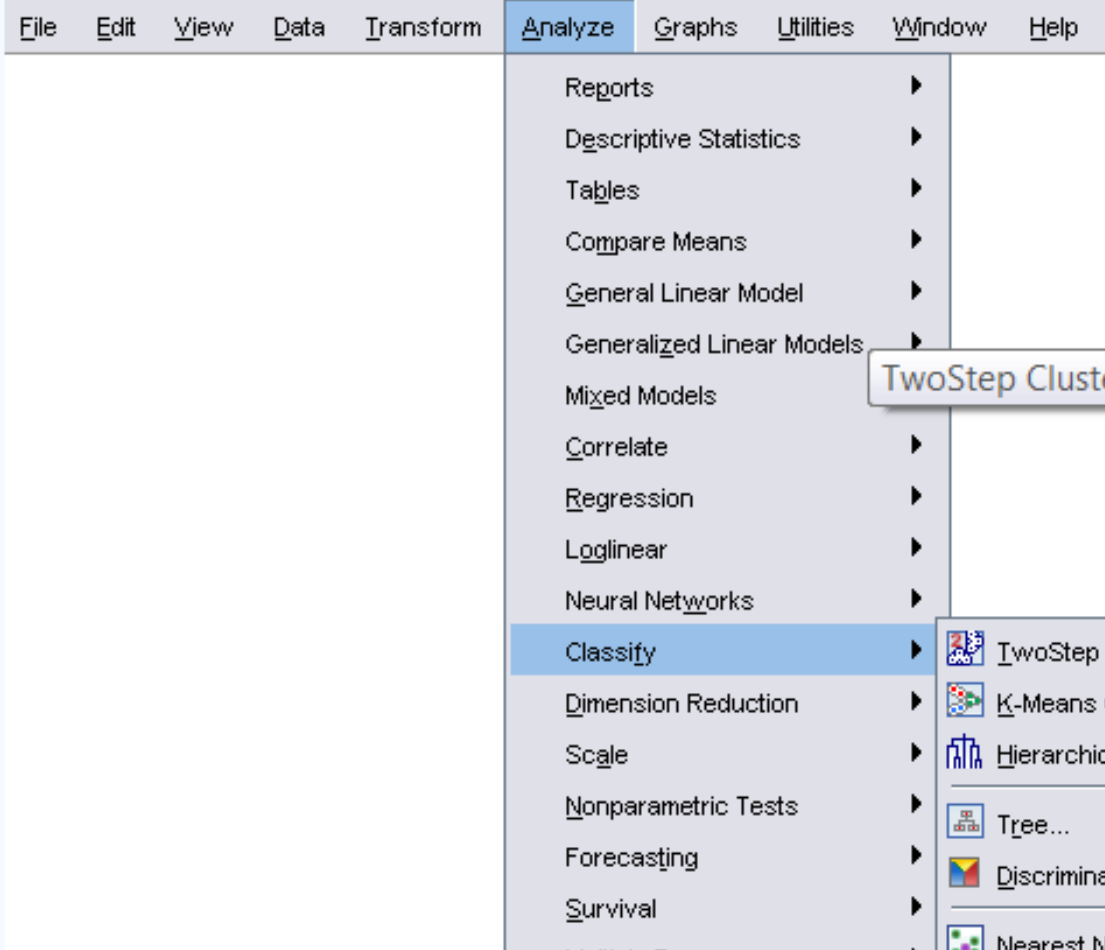
TwoStep Cluster Analysis >

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 [Sample Files](#) for more information. Use the TwoStep Cluster Analysis procedure to group automobiles according to their prices and physical properties.



TwoStep Cluster Analysis > Using TwoStep Cluster Analysis to Classify Motor Vehicles >



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

Analyze
Classify

TwoStep Cluster Analysis menu selection

Cluster...



TwoStep Cluster Analysis > Using TwoStep Cluster Analysis to Classify Motor Vehicles >

TwoStep Cluster Analysis

Sales in thousands [s...]
 4-year resale value [r...]
 Log-transformed sale...
 Zscore: 4-year resale...
 Zscore: Type [ztype]
 Zscore: Price in thou...
 Zscore: Engine size [...]
 Zscore: Horsepower ...
 Zscore: Wheelbase [...]
 Zscore: Width [zwidth]

Categorical Variables:
 Vehicle type [type]

Continuous Variables:
 Price in thousands [pr...]
 Engine size [engine_s]
 Horsepower [horsepo...]
 Wheelbase [wheelbas]

Distance Measure
 Log-likelihood
 Euclidean

Count of Continuous Variables
 To be Standardized: 9
 Assumed Standardized: 0

Number of Clusters
 Determine automatically
 Maximum: 15
 Specify fixed
 Number: 5

Clustering Criterion
 Schwarz's Bayesian Criterion (BIC)
 Akaike's Information Criterion (AIC)

- ▶ 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**.
- ▶ Select *Vehicle type* as a categorical variable.
- ▶ Select *Price in thousands* through *Fuel efficiency* as continuous variables.



TwoStep Cluster: Plots

Within cluster percentage chart
 Cluster pie chart

Variable Importance Plot

Rank of variable importance


Rank Variables

By cluster
 By variable

Importance Measure

Chi-square or t-test of significance
 Significance

Confidence level
Percentage:
 Omit insignificant variables

 Selection of a chart will result in the creation of a cluster membership variable.

- ▶ Select **Rank of variable importance**.
- ▶ Select **By variable** in the Rank Variables group.
- ▶ Select **Confidence level**.
- ▶ Click **Continue**, then click **Output** in the TwoStep Cluster Analysis dialog box.

TwoStep Cluster Analysis > Using TwoStep Cluster Analysis to Classify Motor Vehicles >

TwoStep Cluster: Output

Statistics

- Descriptives by cluster
- Cluster frequencies
- Information Criterion (AIC or BIC)

Working Data File

- Create cluster membership variable

XML Files

- Export final model
Name:
- Export CF tree
Name:

- ▶ Select **Information criterion (AIC or BIC)** in the Statistics group.
- ▶ Click **Continue**.
- ▶ Click **OK** in the TwoStep Cluster Analysis dialog box.



TwoStep Cluster Analysis > Using TwoStep Cluster Analysis to Classify Motor Vehicles >

Number of Clusters	Schwarz's Bayesian Criterion (BIC)	BIC Change ^a	Ratio of BIC Changes ^b	Ratio of Distance Measures ^c
1	1214.377			
2	974.051	-240.326	1.000	1.829
3	885.924	-88.128	.367	2.190
4	897.559	11.635	-.048	1.368
5	931.760	34.201	-.142	1.036
6	968.073	36.313	-.151	1.576
7	1026.000	57.927	-.241	1.083
8	1086.815	60.815	-.253	1.687
9	1161.740	74.926	-.312	1.020
10	1237.063	75.323	-.313	1.239
11	1316.271	79.207	-.330	1.046
12	1396.192	79.921	-.333	1.075
13	1477.199	81.008	-.337	1.076
14	1559.230	82.030	-.341	1.301
15	1644.366	85.136	-.354	1.044

- The changes are from the previous number of clusters in the table.
- The ratios of changes are with respect to the change at the two clusters.
- The ratios of distance measures are based on the current number of clusters against the previous number of clusters.

The Auto-clustering table summarizes the process by which the number of clusters is chosen.

TwoStep Cluster Analysis > Using TwoStep Cluster Analysis to Classify Motor Vehicles >

Number of Clusters	Schwarz's Bayesian Criterion (BIC)	BIC Change ^a	Ratio of BIC Changes ^b	Ratio of Distance Measures ^c
1	1214.377			
2	974.051	-240.326	1.000	1.829
3	885.924	-88.128	.367	2.190
4	897.559	11.635	-.048	1.368
5	931.760	34.201	-.142	1.036
6	968.073	36.313	-.151	1.576
7	1026.000	57.927	-.241	1.083
8	1086.815	60.815	-.253	1.687
9	1161.740	74.926	-.312	1.020
10	1237.063	75.323	-.313	1.239
11	1316.271	79.207	-.330	1.046
12	1396.192	79.921	-.333	1.075
13	1477.199	81.008	-.337	1.076
14	1559.230	82.030	-.341	1.301
15	1644.366	85.136	-.354	1.044

- The changes are from the previous number of clusters in the table.
- The ratios of changes are with respect to the change at the two clusters.
- The ratios of distance measures are based on the current number of clusters against the previous number of clusters.

The clustering criterion (in this case the BIC) is computed for each potential number of clusters. Smaller values of the BIC indicate better models, and in this situation, the "best" cluster solution has the smallest BIC.

TwoStep Cluster Analysis > Using TwoStep Cluster Analysis to Classify Motor Vehicles >

Number of Clusters	Schwarz's Bayesian Criterion (BIC)	BIC Change ^a	Ratio of BIC Changes ^b	Ratio of Distance Measures ^c
1	1214.377			
2	974.051	-240.326	1.000	1.829
3	885.924	-88.128	.367	2.190
4	897.559	11.635	-.048	1.368
5	931.760	34.201	-.142	1.036
6	968.073	36.313	-.151	1.576
7	1026.000	57.927	-.241	1.083
8	1086.815	60.815	-.253	1.687
9	1161.740	74.926	-.312	1.020
10	1237.063	75.323	-.313	1.239
11	1316.271	79.207	-.330	1.046
12	1396.192	79.921	-.333	1.075
13	1477.199	81.008	-.337	1.076
14	1559.230	82.030	-.341	1.301
15	1644.366	85.136	-.354	1.044

- The changes are from the previous number of clusters in the table.
- The ratios of changes are with respect to the change at the two clusters.
- The ratios of distance measures are based on the current number of clusters against the previous number of clusters.

However, there are clustering problems in which the BIC will continue to decrease as the number of clusters increases, but the improvement in the cluster solution, as measured by the BIC Change, is not worth the increased complexity of the cluster model, as measured by the number of clusters.



TwoStep Cluster Analysis > Using TwoStep Cluster Analysis to Classify Motor Vehicles >

Number of Clusters	Schwarz's Bayesian Criterion (BIC)	BIC Change ^a	Ratio of BIC Changes ^b	Ratio of Distance Measures ^c
1	1214.377			
2	974.051	-240.326	1.000	1.829
3	885.924	-88.128	.367	2.190
4	897.559	11.635	-.048	1.368
5	931.760	34.201	-.142	1.036
6	968.073	36.313	-.151	1.576
7	1026.000	57.927	-.241	1.083
8	1086.815	60.815	-.253	1.687
9	1161.740	74.926	-.312	1.020
10	1237.063	75.323	-.313	1.239
11	1316.271	79.207	-.330	1.046
12	1396.192	79.921	-.333	1.075
13	1477.199	81.008	-.337	1.076
14	1559.230	82.030	-.341	1.301
15	1644.366	85.136	-.354	1.044

- The changes are from the previous number of clusters in the table.
- The ratios of changes are with respect to the change at the two clusters.
- The ratios of distance measures are based on the current number of clusters against the previous number of clusters.

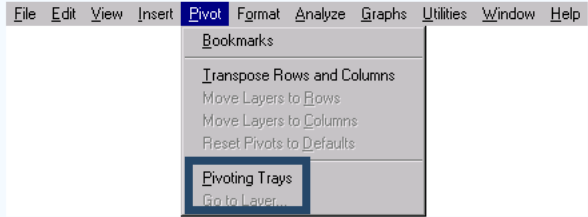
In such situations, the changes in BIC and changes in the distance measure are evaluated to determine the "best" cluster solution. A good solution will have a reasonably large Ratio of BIC Changes and a large Ratio of Distance Measures.

TwoStep Cluster Analysis > Using TwoStep Cluster Analysis to Classify Motor Vehicles >

	N	% of Combined	% of Total
Cluster 1	62	40.8%	39.5%
Cluster 2	39	25.7%	24.8%
Cluster 3	51	33.6%	32.5%
Combined	152	100.0%	96.8%
Excluded Cases	5		3.2%
Total	157		100.0%

The cluster distribution table shows the frequency of each cluster. Of the 157 total cases, 5 were excluded from the analysis due to missing values on one or more of the variables. Of the 152 cases assigned to clusters, 62 were assigned to the first cluster, 39 to the second, and 51 to the third.

TwoStep Cluster Analysis > Using TwoStep Cluster Analysis to Classify Motor Vehicles >



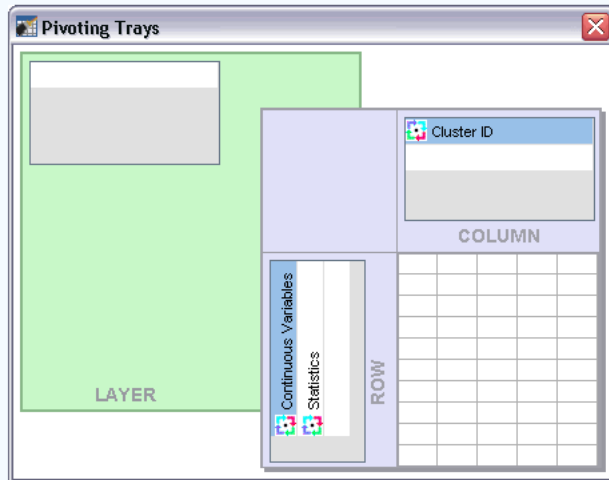
The Centroids table displays the mean and standard deviation for the cases in each cluster. It is currently too wide for easy viewing, so you are going to pivot it.

- ▶ In the output window, double-click the Centroids table to activate it.
- ▶ From the Viewer menus choose:

Pivot
Pivoting Trays...



TwoStep Cluster Analysis > Using TwoStep Cluster Analysis to Classify Motor Vehicles >



- ▶ Move *Cluster ID* to the Columns area.
- ▶ Move *Continuous Variables* and *Statistics* to the Rows area.
- ▶ Close the Pivoting Trays window.
- ▶ Deactivate the table by clicking outside of its boundaries.

TwoStep Cluster Analysis > Using TwoStep Cluster Analysis to Classify Motor Vehicles >

		Cluster			
		1	2	3	Combined
Price in thousands	Mean	19.61671	26.56182	37.29980	27.33182
	Std. Deviation	7.644070	10.185175	17.381187	14.418669
Engine size	Mean	2.194	3.559	3.700	3.049
	Std. Deviation	.4238	.9358	.9493	1.0498
Horsepower	Mean	143.24	187.92	232.96	184.81
	Std. Deviation	30.259	39.049	54.408	56.823
Wheelbase	Mean	102.595	112.972	109.022	107.414
	Std. Deviation	4.0799	9.6537	5.7644	7.7178
Width	Mean	68.539	72.744	72.924	71.089
	Std. Deviation	1.9366	4.1781	2.1855	3.4647
Length	Mean	178.235	191.110	194.688	187.059
	Std. Deviation	9.6534	14.4415	10.3512	13.4712
Curb weight	Mean	2.83742	3.96759	3.57890	3.37618
	Std. Deviation	.310867	.671766	.297204	.636593
Fuel capacity	Mean	14.979	22.064	18.443	17.959
	Std. Deviation	1.8699	4.2894	2.0445	3.9376
Fuel efficiency	Mean	27.24	19.51	23.02	23.84
	Std. Deviation	3.578	2.910	2.060	4.305

The centroids show that the clusters are well separated by the continuous variables.



TwoStep Cluster Analysis > Using TwoStep Cluster Analysis to Classify Motor Vehicles >

		Cluster			
		1	2	3	Combined
Price in thousands	Mean	19.61671	26.56182	37.29980	27.33182
	Std. Deviation	7.644070	0.185175	17.381187	14.418669
Engine size	Mean	2.194	3.559	3.700	3.049
	Std. Deviation	.4238	.9358	.9493	1.0498
Horsepower	Mean	143.24	187.92	232.96	184.81
	Std. Deviation	30.259	39.049	54.408	56.823
Wheelbase	Mean	102.595	112.972	109.022	107.414
	Std. Deviation	4.0799	9.6537	5.7644	7.7178
Width	Mean	68.539	72.744	72.924	71.089
	Std. Deviation	1.9366	4.1781	2.1855	3.4647
Length	Mean	178.235	191.110	194.688	187.059
	Std. Deviation	9.6534	14.4415	10.3512	13.4712
Curb weight	Mean	2.83742	3.96759	3.57890	3.37618
	Std. Deviation	.310867	.671766	.297204	.636593
Fuel capacity	Mean	14.979	22.064	18.443	17.959
	Std. Deviation	1.8699	4.2894	2.0445	3.9376
Fuel efficiency	Mean	27.24	19.51	23.02	23.84
	Std. Deviation	3.578	2.910	2.060	4.305

Motor vehicles in cluster 1 are cheap, small, and fuel efficient.



TwoStep Cluster Analysis > Using TwoStep Cluster Analysis to Classify Motor Vehicles >

		Cluster			
		1	2	3	Combined
Price in thousands	Mean	19.61671	26.56182	37.29980	27.33182
	Std. Deviation	7.644070	10.185175	7.381187	14.418669
Engine size	Mean	2.194	3.559	3.700	3.049
	Std. Deviation	.4238	.9358	.9493	1.0498
Horsepower	Mean	143.24	187.92	232.96	184.81
	Std. Deviation	30.259	39.049	54.408	56.823
Wheelbase	Mean	102.595	112.972	109.022	107.414
	Std. Deviation	4.0799	9.6537	5.7644	7.7178
Width	Mean	68.539	72.744	72.924	71.089
	Std. Deviation	1.9366	4.1781	2.1855	3.4647
Length	Mean	178.235	191.110	194.688	187.059
	Std. Deviation	9.6534	14.4415	10.3512	13.4712
Curb weight	Mean	2.83742	3.96759	3.57890	3.37618
	Std. Deviation	.310867	.671766	.297204	.636593
Fuel capacity	Mean	14.979	22.064	18.443	17.959
	Std. Deviation	1.8699	4.2894	2.0445	3.9376
Fuel efficiency	Mean	27.24	19.51	23.02	23.84
	Std. Deviation	3.578	2.910	2.060	4.305

Motor vehicles in cluster 2 are moderately priced, heavy, and have a large gas tank, presumably to compensate for their poor fuel efficiency.



Next

TwoStep Cluster Analysis > Using TwoStep Cluster Analysis to Classify Motor Vehicles >

		Cluster			
		1	2	3	Combined
Price in thousands	Mean	19.61671	26.56182	37.29980	27.33182
	Std. Deviation	7.644070	10.185175	17.381187	14.418669
Engine size	Mean	2.194	3.559	3.700	3.049
	Std. Deviation	.4238	.9358	.9493	1.0498
Horsepower	Mean	143.24	187.92	232.96	184.81
	Std. Deviation	30.259	39.049	54.408	56.823
Wheelbase	Mean	102.595	112.972	109.022	107.414
	Std. Deviation	4.0799	9.6537	5.7644	7.7178
Width	Mean	68.539	72.744	72.924	71.089
	Std. Deviation	1.9366	4.1781	2.1855	3.4647
Length	Mean	178.235	191.110	194.688	187.059
	Std. Deviation	9.6534	14.4415	10.3512	13.4712
Curb weight	Mean	2.83742	3.96759	3.57890	3.37618
	Std. Deviation	.310867	.671766	.297204	.636593
Fuel capacity	Mean	14.979	22.064	18.443	17.959
	Std. Deviation	1.8699	4.2894	2.0445	3.9376
Fuel efficiency	Mean	27.24	19.51	23.02	23.84
	Std. Deviation	3.578	2.910	2.060	4.305

Motor vehicles in cluster 3 are expensive, large, and are moderately fuel efficient.



TwoStep Cluster Analysis > Using TwoStep Cluster Analysis to Classify Meter Vehicles >

	Automobile		Truck	
	Frequency	Percent	Frequency	Percent
Cluster 1	61	54.5%	1	2.5%
2	0	.0%	39	97.5%
3	51	45.5%	0	.0%
Combined	112	100.0%	40	100.0%

The cluster frequency table by *Vehicle type* further clarifies the properties of the clusters.



TwoStep Cluster Analysis > Using TwoStep Cluster Analysis to Classify Motor Vehicles >

	Automobile		Truck	
	Frequency	Percent	Frequency	Percent
Cluster 1	61	54.5%	4	2.5%
Cluster 2	0	.0%	39	97.5%
Cluster 3	51	45.5%	0	.0%
Combined	112	100.0%	40	100.0%

Cluster 2 is comprised entirely of trucks.



TwoStep Cluster Analysis > Using TwoStep Cluster Analysis to Classify Motor Vehicles >

	Automobile		Truck	
	Frequency	Percent	Frequency	Percent
Cluster 1	61	54.5%	1	2.5%
2	0	.0%	39	97.5%
3	51	45.5%	0	.0%
Combined	112	100.0%	40	100.0%

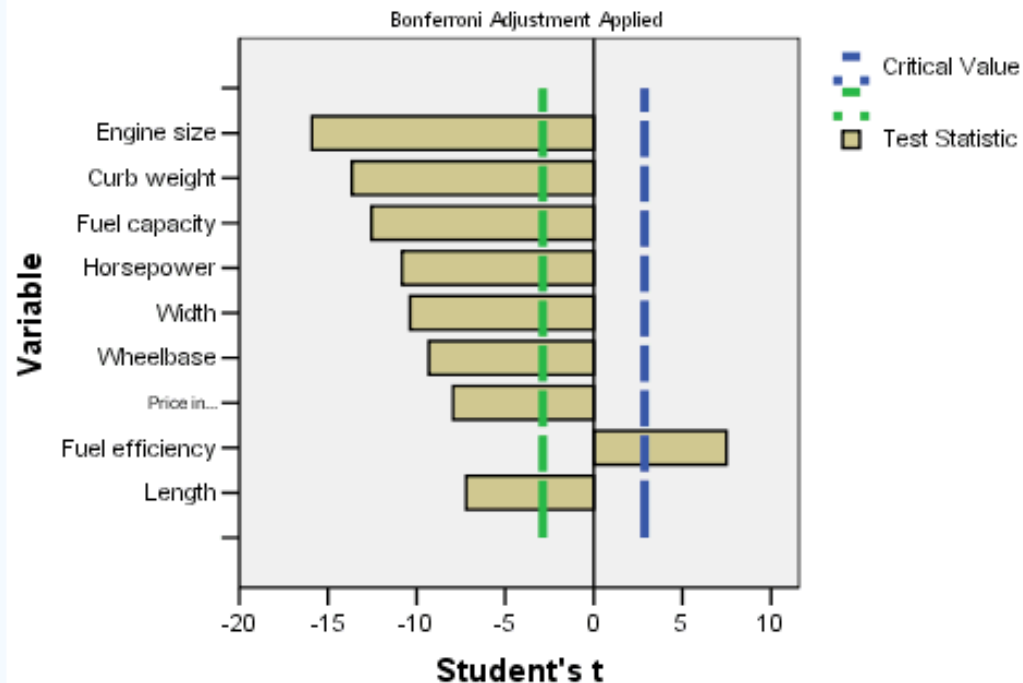
Clusters 1 and 3 contain automobiles, save for a single truck in Cluster 1. Examination of the data file reveals this to be the Toyota RAV4.



Next

TwoStep Cluster Analysis > Using TwoStep Cluster Analysis to Classify Motor Vehicles >

TwoStep Cluster Number = 1

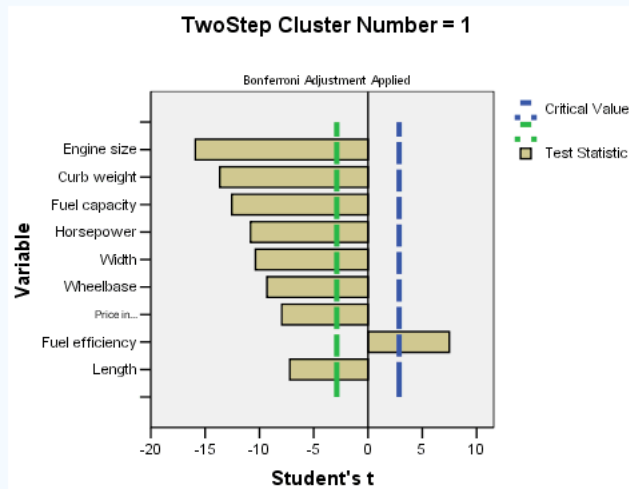


The "by variable" importance charts are produced with a separate chart for each cluster. The variables are lined up on the Y axis, in descending order of importance.



Next

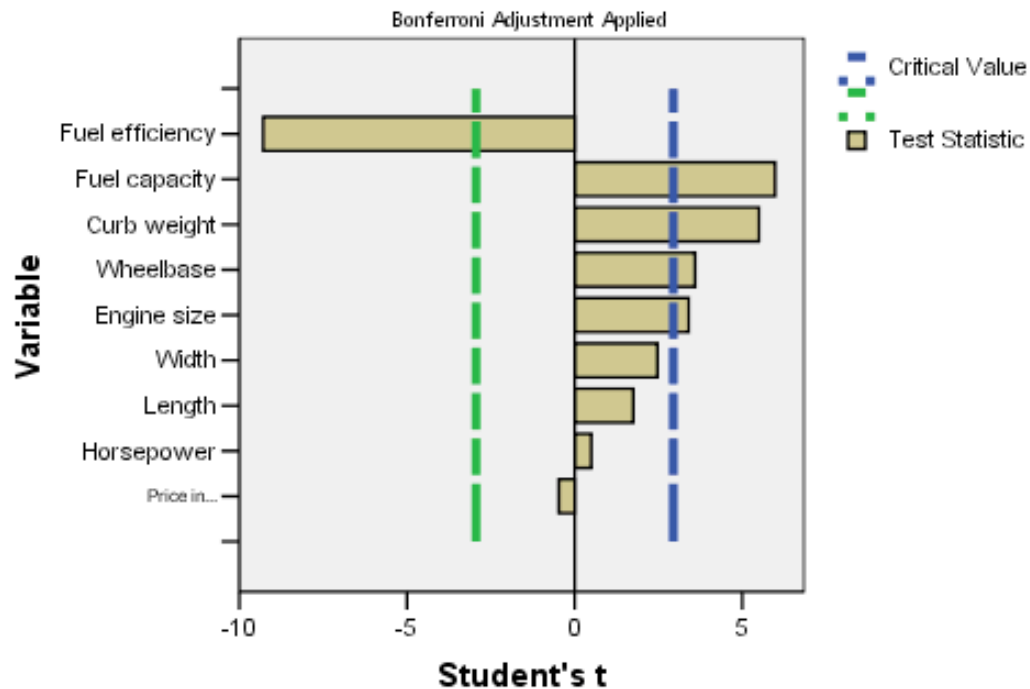
TwoStep Cluster Analysis > Using TwoStep Cluster Analysis to Classify Motor Vehicles >



A negative t statistic indicates that the variable generally takes smaller than average values within this cluster, while a positive t statistic indicates the variable takes larger than average values. Thus, for Cluster 1, *Fuel efficiency* takes larger than average values while all of the other variables take smaller than average values. These results confirm the trends observed in the Centroids table.

TwoStep Cluster Analysis > Using TwoStep Cluster Analysis to Classify Motor Vehicles >

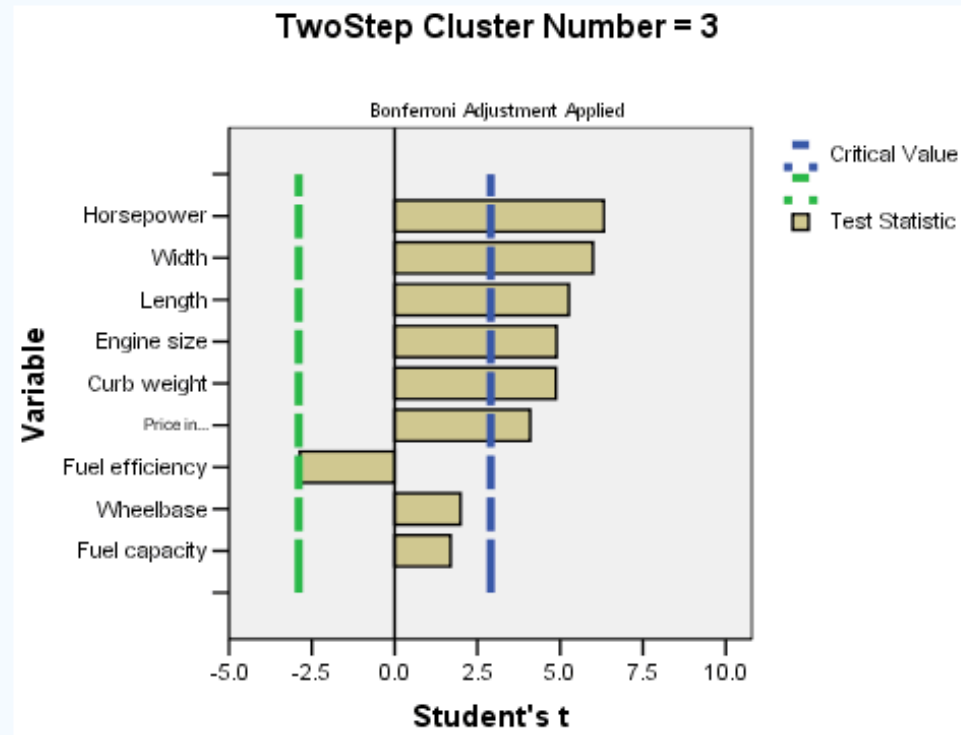
TwoStep Cluster Number = 2



The chart for Cluster 2 shows that *Width*, *Length*, *Horsepower*, and *Price in thousands* are not important to the formation of this cluster.



TwoStep Cluster Analysis > Using TwoStep Cluster Analysis to Classify Motor Vehicles >



The chart for Cluster 3 shows that *Wheelbase* and *Fuel capacity* are not important to the formation of this cluster, while *Fuel efficiency* is just barely significant.



Next

TwoStep Cluster Analysis > Using TwoStep Cluster Analysis to Classify Motor Vehicles >

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](#)

TwoStep Cluster Analysis >

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.



Next

Hierarchical Cluster Analysis >

Hierarchical cluster analysis (HCA) is an exploratory tool designed to reveal natural groupings (or clusters) within a data set that would otherwise not be apparent. It is most useful when you want to cluster a small number (less than a few hundred) of objects.

The **objects** in hierarchical cluster analysis can be cases or variables, depending on whether you want to classify cases or examine relationships between the variables.



Hierarchical Cluster Analysis >

Hierarchical cluster analysis (HCA) is an exploratory tool designed to reveal natural groupings (or clusters) within a data set that would otherwise not be apparent. It is most useful when you want to cluster a small number (less than a few hundred) of objects.

The **objects** in hierarchical cluster analysis can be cases or variables, depending on whether you want to classify cases or examine relationships between the variables.

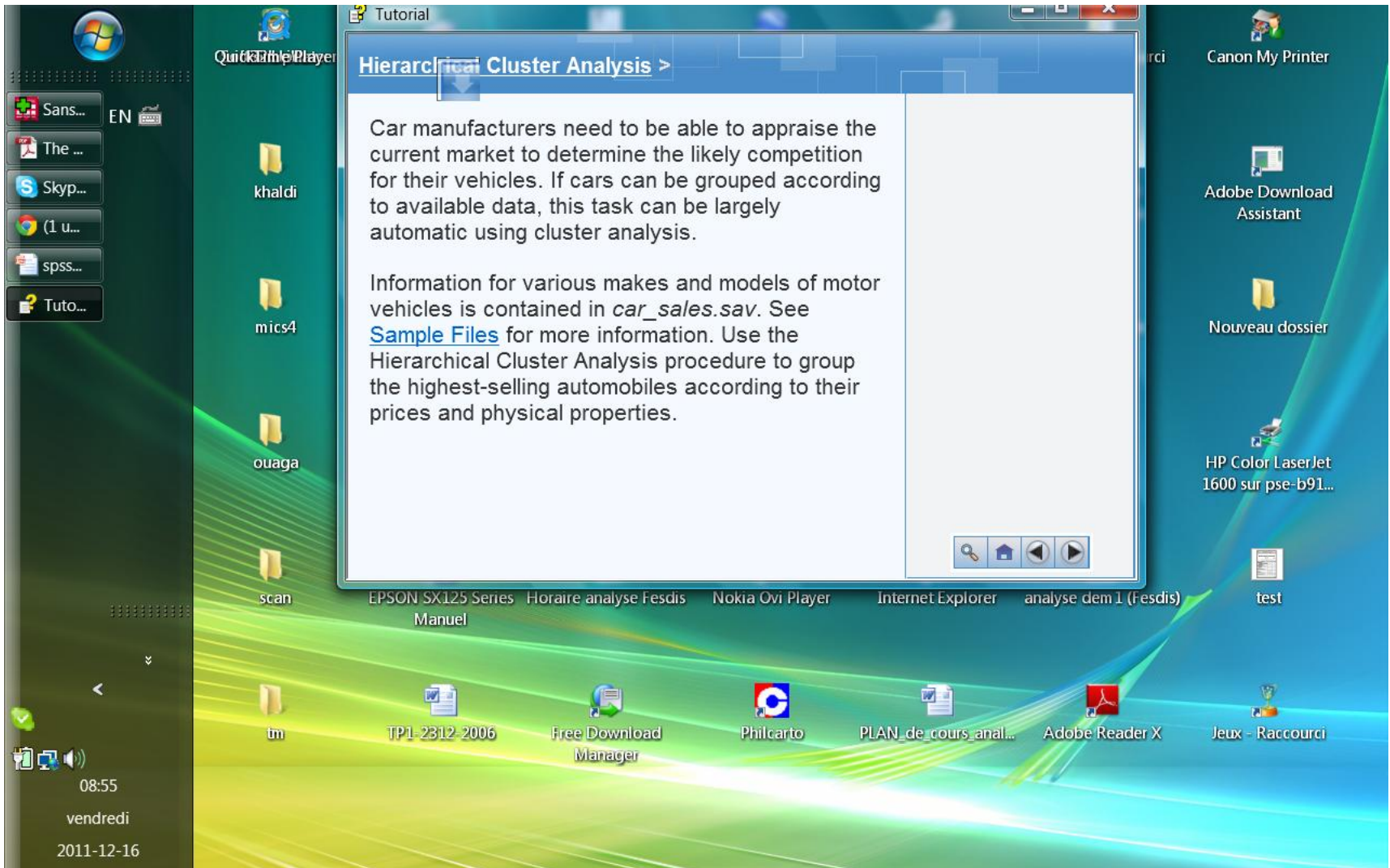


Hierarchical Cluster Analysis >

Hierarchical cluster analysis begins by separating each object into a cluster by itself. At each stage of the analysis, the criterion by which objects are separated is relaxed in order to link the two most similar clusters until all of the objects are joined in a complete classification tree.

The basic criterion for any clustering is distance. Objects that are near each other should belong to the same cluster, and objects that are far from each other should belong to different clusters. For a given set of data, the clusters that are constructed depend on your specification of the following parameters:

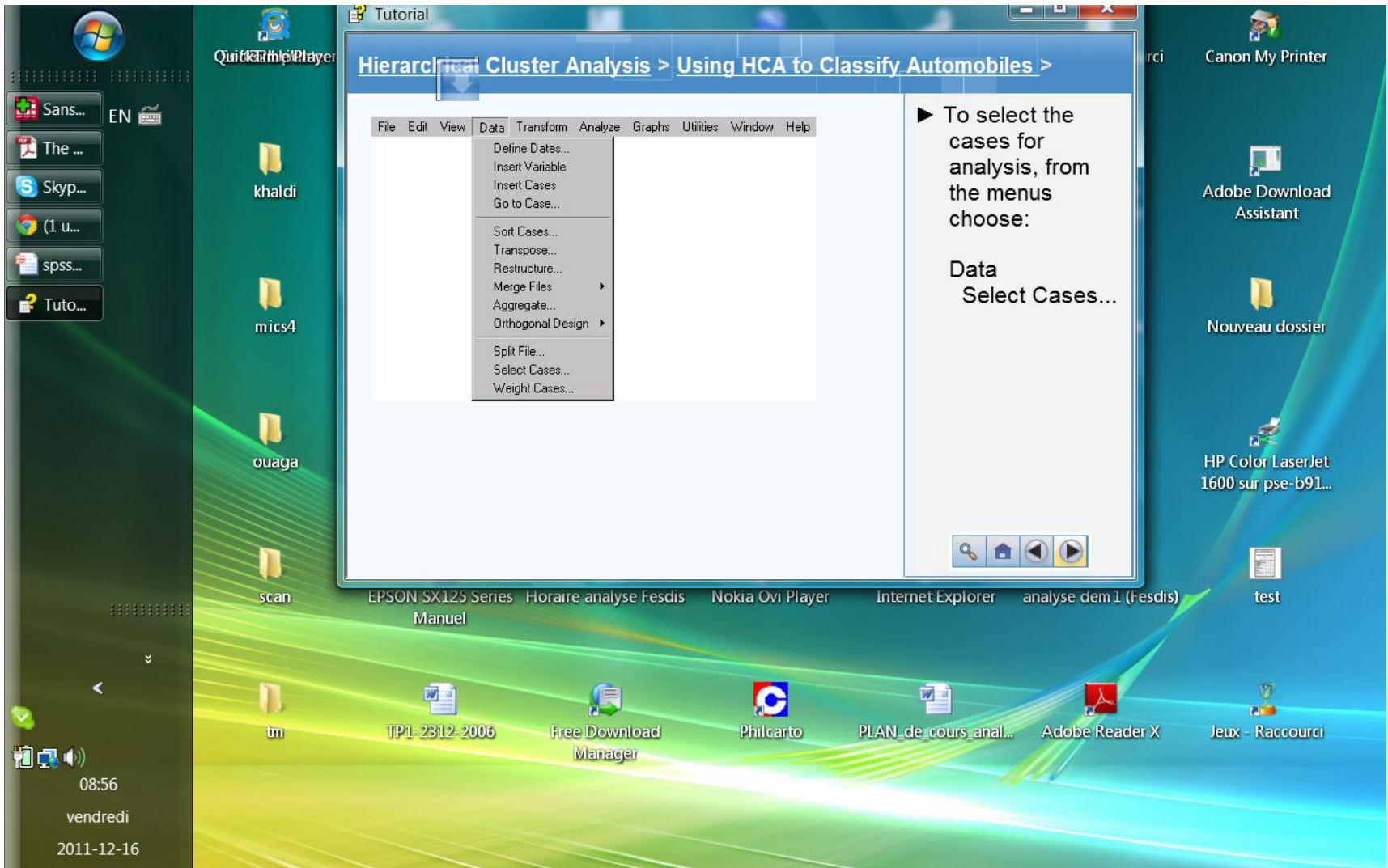
- **Cluster method** defines the rules for cluster formation. For example, when calculating the distance between two clusters, you can use the pair of nearest objects between clusters or the pair of furthest objects, or a compromise between these methods.
- **Measure** defines the formula for calculating distance. For example, the Euclidean distance measure calculates the distance as a "straight line" between two clusters. Interval measures assume that the variables are scale; count measures assume that they are discrete numeric; and binary measures assume that they take only two values.
- **Standardization** allows you to equalize the effect of variables measured on different scales.

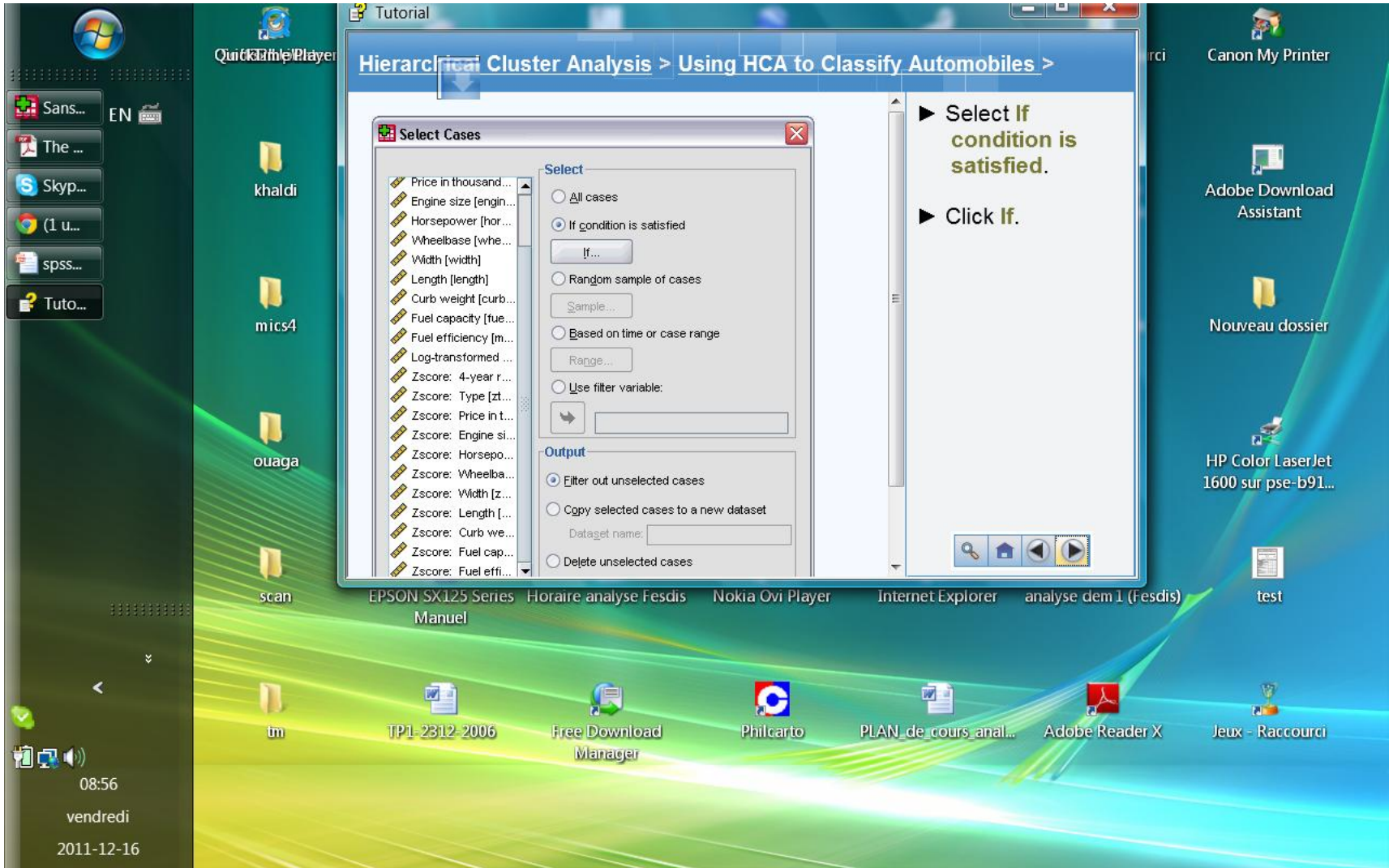


Hierarchical Cluster Analysis >

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 [Sample Files](#) for more information. Use the Hierarchical Cluster Analysis procedure to group the highest-selling automobiles according to their prices and physical properties.





Tutorial

Hierarchical Cluster Analysis > Using HCA to Classify Automobiles >

Select Cases

Select

- All cases
- If condition is satisfied
- Random sample of cases
- Based on time or case range
- Use filter variable:

Output

- Filter out unselected cases
- Copy selected cases to a new dataset
- Delete unselected cases

► Select If condition is satisfied.

► Click If.

Windows XP desktop sidebar with icons for:
- Sans...
- The ...
- Skyp...
- (1 u...
- spss...
- Tuto...
- System tray: 08:56, vendredi, 2011-12-16

Desktop icons:
- QuickTime Player
- khaldi
- mics4
- ouaga
- scan
- tm

Taskbar icons:
- EPSON SX125 Series
- Manuel
- Horaire analyse Fesdis
- Nokia Ovi Player
- Internet Explorer
- analyse dem 1 (Fesdis)
- test
- TP1-2312-2006
- Free Download Manager
- Phillcarto
- PLAN_de_cours_anal...
- Adobe Reader X

Desktop icons:
- Canon My Printer
- Adobe Download Assistant
- Nouveau dossier
- HP Color LaserJet 1600 sur pse-b91...
- Jeux - Raccourci

Tutorial

Hierarchical Cluster Analysis > Using HCA to Classify Automobiles >

Select Cases: If

- Manufacturer [man...]
- Model [model]
- Sales in thousand...
- 4-year resale valu...
- Vehicle type [type]
- Price in thousand...
- Engine size [engin...]
- Horsepower [hors...]
- Wheelbase [wheel...]
- Width [width]
- Length [length]
- Curb weight [curb...]
- Fuel capacity [fuel...]
- Fuel efficiency [mpg]
- Log-transformed s...
- Zscore: 4-year res...
- Zscore: Type [ztype]
- Zscore: Price in th...
- Zscore: Engine si...
- Zscore: Horsepo...

(type=0) & (sales>100)

Next

- ▶ In the text field, type **(type=0) & (sales>100)**.
- ▶ Click **Continue**.
- ▶ Click **OK** in the Select Cases dialog box.

Further analysis of the data set will be based on automobiles that sold at least 100,000 units.

Sans... EN

The ...

Skyp...

(1 u...

spss...

Tuto...

08:56

vendredi

2011-12-16

QuickTime Player

khaldi

mics4

ouaga

scan

Manuel

tm

EPSON SX125 Series

Horaire analyse Fesdis

Nokia Ovi Player

Internet Explorer

analyse de cours (Fesdis)

TP1-2312-2006

Free Download Manager

Philcarto

PLAN_de_cours_anal...

Adobe Reader X

Canon My Printer

Adobe Download Assistant

Nouveau dossier

HP Color LaserJet 1600 sur pse-b91...

test

Jeux - Raccourci

Tutorial
Hierarchical Cluster Analysis > Using HCA to Classify Automobiles >

File Edit View Data Transform **Analyze** Graphs Utilities Window Help

- Reports
- Descriptive Statistics
- Tables
- Compare Means
- General Linear Model
- Generalized Linear Models
- Mixed Models
- Correlate
- Regression
- Loglinear
- Neural Networks
- Classify**
 - TwoStep C...
 - K-Means C...
 - Hierarchic...**
 - Tree...
 - Discrimina...
 - Nearest N...
- Dimension Reduction
- Scale
- Nonparametric Tests
- Forecasting
- Survival

▶ To run the cluster analysis, from the menus choose:
**Analyze
Classify
Hierarchical
Cluster...**

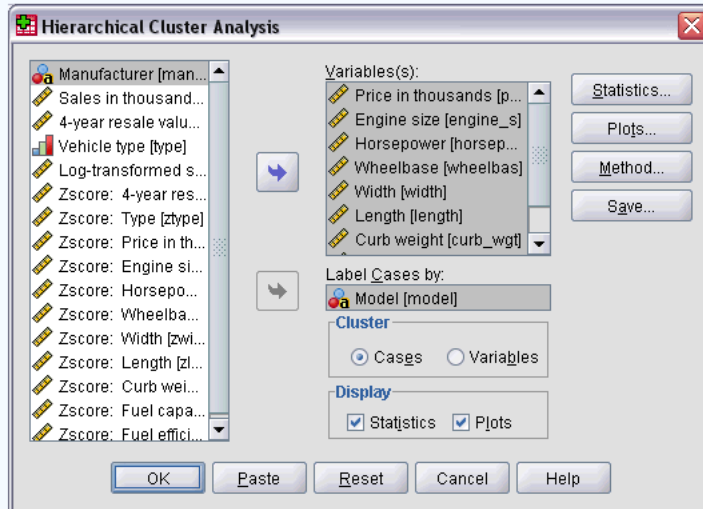
Next

EPSON SX125 Series Horaire analyse Fesdis Nokia Ovi Player Internet Explorer analyse uem 1 (Fesdis) Manuel

tm TP1-2312-2006 Free Download Manager Phillcarto PLAN_de_cours_anal... Adobe Reader X Jeux - Raccourci

08:56
vendredi
2011-12-16

Hierarchical Cluster Analysis > Using HCA to Classify Automobiles >



► 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**.

► Select *Price in thousands* through *Fuel efficiency* as analysis variables.

► Select *Model* as the case labeling variable.



Tutorial

Hierarchical Cluster Analysis > Using HCA to Classify Automobiles >

Hierarchical Cluster Analysis: Plots

Dendrogram

Icicle

All clusters

Specified range of clusters

Start cluster:

Stop cluster:

By:

None

Orientation

Vertical

Horizontal

Continue Cancel Help

- ▶ Select **Dendrogram**.
- ▶ Select **None** in the Icicle group.
- ▶ Click **Continue**.
- ▶ Click **Method** in the Hierarchical Cluster Analysis dialog box.

Windows 7 desktop sidebar with Start menu and taskbar.

Start menu items: Sans..., The ..., Skyp..., (1 u..., spss..., Tuto...

Taskbar: 08:57, vendredi, 2011-12-16

Desktop icons on the left side:

- QuickTime Player
- khaldi
- mics4
- ouaga
- scan
- tm

Taskbar and desktop icons at the bottom:

Taskbar: EPSON SX125 Series Manuel, Horaire analyse Fesdis, Nokia Ovi Player, Internet Explorer, analyse dem 1 (Fesdis)

Desktop icons: TP1-2312-2006, Free Download Manager, Phillcarto, PLAN_de_cours_anal..., Adobe Reader X, Jeux - Raccourci

Desktop icons on the right side:

- Canon My Printer
- Adobe Download Assistant
- Nouveau dossier
- HP Color LaserJet 1600 sur pse-b91...
- test

Hierarchical Cluster Analysis > Using HCA to Classify Automobiles >

Hierarchical Cluster Analysis: Method

Cluster Method: Nearest neighbor

Measure

Interval: Squared Euclidean distance
 Power: 2 Root: 2

Counts: Chi-squared measure

Binary: Euclidean distance
 Present: 1 Absent: 0

Transform Values

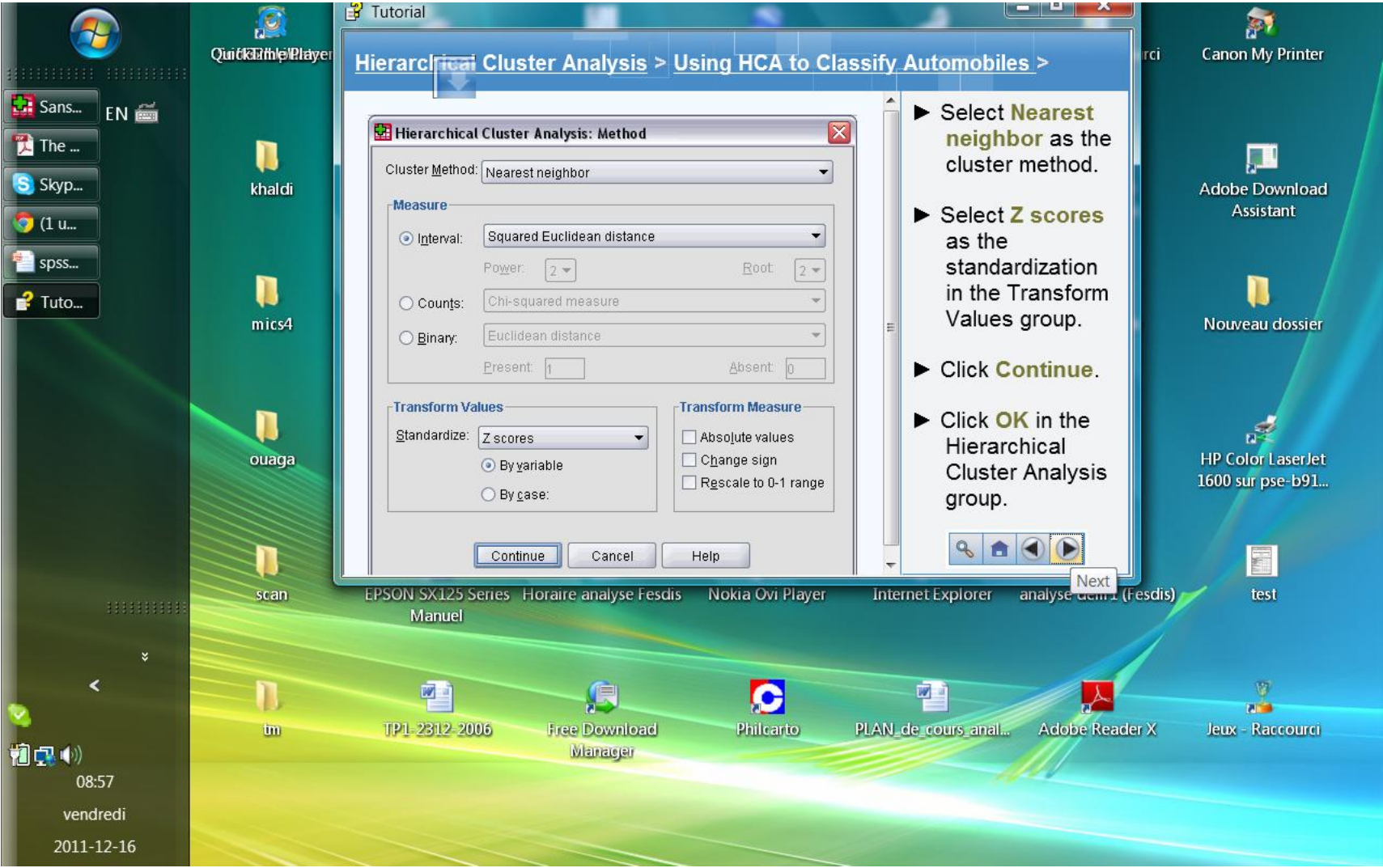
Standardize: Z scores
 By variable
 By case:

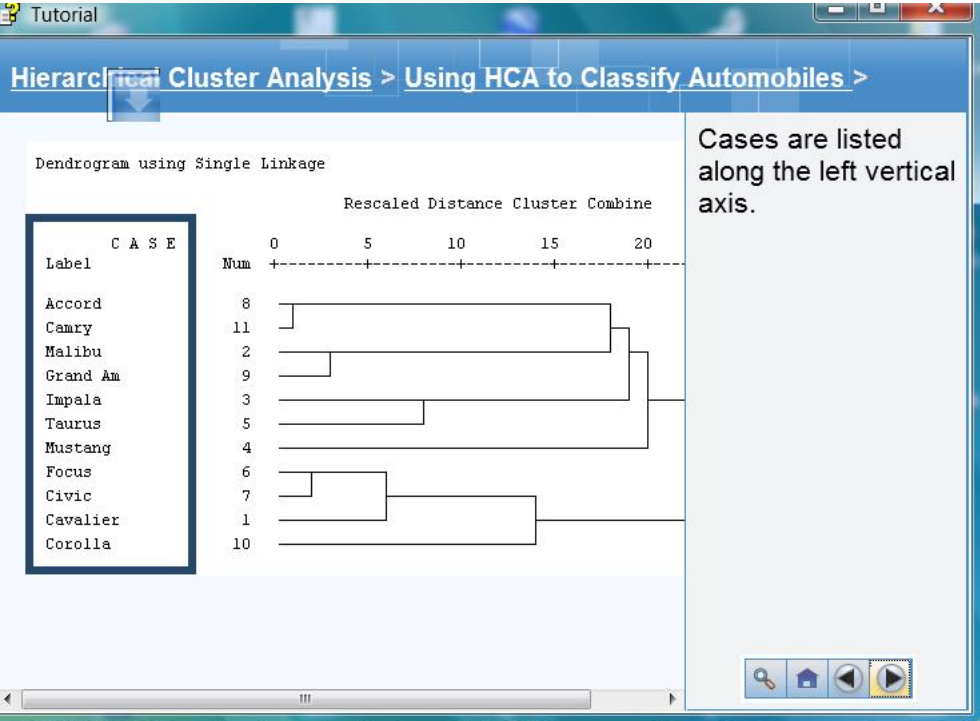
Transform Measure

Absolute values
 Change sign
 Rescale to 0-1 range

Continue Cancel Help

- ▶ Select **Nearest neighbor** as the cluster method.
- ▶ Select **Z scores** as the standardization in the Transform Values group.
- ▶ Click **Continue**.
- ▶ Click **OK** in the Hierarchical Cluster Analysis group.





EPSON SX125 Series Manuel Horaire analyse Fesdis Nokia Ovi Player Internet Explorer analyse dem 1 (Fesdis) test

tm TP1-2312-2006 Free Download Manager Phillcarto PLAN_de_cours_anal... Adobe Reader X Jeux - Raccourci

08:58
vendredi
2011-12-16

Windows 7 desktop environment with a tutorial window open.

Tutorial: Hierarchical Cluster Analysis > Using HCA to Classify Automobiles >

Program using Single Linkage

Rescaled Distance Cluster Combine

CASE	Number	0	5	10	15	20	25
1	8						
2	11						
3	2						
4	9						
5	3						
6	5						
7	4						
8	6						
9	7						
10	1						
11	10						

The horizontal axis shows the distance between clusters when they are joined.

Taskbar: EPSON SX125 Series, Horaire analyse Fesdis, Nokia Ovi Player, Internet Explorer, analyse, Next (Fesdis), Manuel

Taskbar: tm, TP1-2312-2006, Free Download Manager, Phillcarto, PLAN_de_cours_anal..., Adobe Reader X, Jeux - Raccourci

System tray: 08:58, vendredi, 2011-12-16

Windows 7 desktop environment with a tutorial window open.

Tutorial: Hierarchical Cluster Analysis > Using HCA to Classify Automobiles >

Program using Single Linkage

Rescaled Distance Cluster Combine

C A S E	Num	0	5	10	15	20	25
1							
ord	8						
y	11						
bu	2						
d Am	9						
la	3						
us	5						
ang	4						
s	6						
c	7						
lier	1						
lla	10						

Parsing the classification tree to determine the number of clusters is a subjective process. Generally, you begin by looking for "gaps" between joinings along the horizontal axis.

Taskbar: EPSON SX125 Series Manuel, Horaire analyse Fesdis, Nokia Ovi Player, Internet Explorer, analyse dem 1 (Fesdis), test

Taskbar: tm, TP1-2312-2006, Free Download Manager, Phillcarto, PLAN_de_cours_anal..., Adobe Reader X, Jeux - Raccourci

System tray: 08:58, vendredi, 2011-12-16

Windows 7 desktop environment with a tutorial window open.

Tutorial: Hierarchical Cluster Analysis > Using HCA to Classify Automobiles >

Program using Single Linkage

Rescaled Distance Cluster Combine

Starting from the right, there is a gap between 20 and 25, which splits the automobiles into two clusters.

C A S E	Num
1	8
2	11
3	2
4	9
5	3
6	5
7	4
8	6
9	7
10	1
11	10

Taskbar: EPSON SX125 Series, Horaire analyse Fesdis, Nokia Ovi Player, Internet Explorer, analyse dem 1 (Fesdis), Manuel

Desktop icons: khalidi, mics4, ouaga, scan, tm, TP1-2312-2006, Free Download Manager, Phillcarto, PLAN_de_cours_anal..., Adobe Reader X, Jeux - Raccourci, Canon My Printer, Adobe Download Assistant, Nouveau dossier, HP Color LaserJet 1600 sur pse-b91..., test

System tray: 08:59, vendredi, 2011-12-16

Windows 7 desktop environment with a tutorial window open.

Tutorial Window: Hierarchical Cluster Analysis > Using HCA to Classify Automobiles >

Program using Single Linkage

Rescaled Distance Cluster Combine

There is another gap from approximately 10 to 15, which suggests 6 clusters.

CASE	Num	0	5	10	15	20	25
1	8						
2	11						
3	2						
4	9						
5	3						
6	5						
7	4						
8	6						
9	7						
10	1						
11	10						

Next

Taskbar: EPSON SX125 Series, Horaire analyse Fesdis, Nokia Ovi Player, Internet Explorer, analyse de cours (Fesdis), Manuel

Desktop icons: tm, TP1-2312-2006, Free Download Manager, Phillcarto, PLAN_de_cours_anal..., Adobe Reader X, Jeux - Raccourci, test

System tray: 08:59, vendredi, 2011-12-16

Tutorial

Hierarchical Cluster Analysis > Using HCA to Classify Automobiles >

The agglomeration schedule is a numerical summary of the cluster solution.

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	8	11	1.260	0	0	7
2	6	7	1.579	0	0	4
3	2	9	1.625	0	0	7
4	1	6	2.318	0	2	8
5	3	5	2.619	0	0	8
6	1	10	3.670	4	0	10
7	2	8	4.420	3	1	8
8	2	3	4.505	7	5	9
9	2	4	4.774	8	0	10
10	1	2	5.718	6	9	0

Next

Windows 7 taskbar area:

- Start button
- Language bar: Sans... EN
- Taskbar buttons: The..., Skyp..., (1 u..., spss..., Tuto...
- System tray: Volume, Network, Date/Time (08:59, vendredi, 2011-12-16)

Left sidebar of desktop:

- QuickTime Player
- Folder: khaldi
- Folder: mics4
- Folder: ouaga
- Folder: scan
- Folder: tm

Desktop icons:

- EPSON SX125 Series
- Manuel
- Horaire analyse Fesdis
- Nokia Ovi Player
- Internet Explorer
- analyse dem 1 (Fesdis)
- TP1-2312-2006
- Free Download Manager
- Philcarto
- PLAN_de_cours_anal...
- Adobe Reader X

Right sidebar of desktop:

- Canon My Printer
- Adobe Download Assistant
- Nouveau dossier
- HP Color LaserJet 1600 sur pse-b91...
- test
- Jeux - Raccourci

Hierarchical Cluster Analysis > Using HCA to Classify Automobiles >

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	8	11	1.260	0	0	7
2	6	7	1.579	0	0	4
3	2	9	1.625	0	0	7
4	1	6	2.318	0	2	8
5	3	5	2.619	0	0	8
6	1	10	3.670	4	0	10
7	2	8	4.420	3	1	8
8	2	3	4.505	7	5	9
9	2	4	4.774	8	0	10
10	1	2	5.718	6	9	0

At the first stage, cases 8 and 11 are combined because they have the smallest distance.

EPSON SX125 Series Manuel Horaire analyse Fesdis Nokia Ovi Player Internet Explorer analyse dem 1 (Fesdis) test

tm TP1-2312-2006 Free Download Manager Phillcarto PLAN_de_cours_anal... Adobe Reader X Jeux - Raccourci

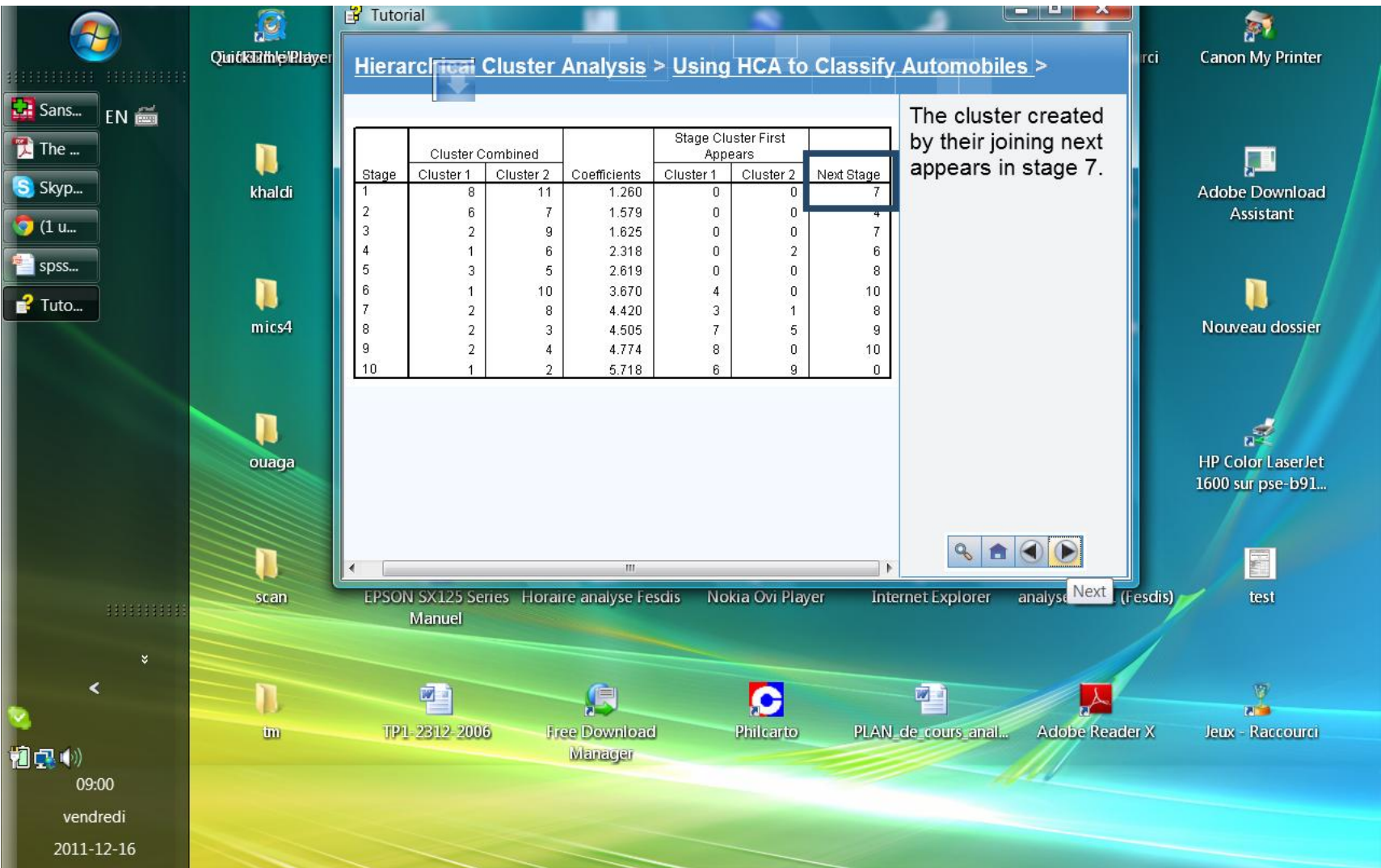
09:00
vendredi
2011-12-16

Tutorial

Hierarchical Cluster Analysis > Using HCA to Classify Automobiles >

The cluster created by their joining next appears in stage 7.

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	8	11	1.260	0	0	7
2	6	7	1.579	0	0	4
3	2	9	1.625	0	0	7
4	1	6	2.318	0	2	6
5	3	5	2.619	0	0	8
6	1	10	3.670	4	0	10
7	2	8	4.420	3	1	8
8	2	3	4.505	7	5	9
9	2	4	4.774	8	0	10
10	1	2	5.718	6	9	0



Tutorial

Hierarchical Cluster Analysis > Using HCA to Classify Automobiles >

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	8	11	1.260	0	0	7
2	6	7	1.579	0	0	4
3	2	9	1.625	0	0	7
4	1	6	2.318	0	2	6
5	3	5	2.619	0	0	8
6	1	12	3.375	1	2	10
7	2	8	4.420	3	1	8
8	2	2	4.585	7	5	9
9	2	4	4.774	8	0	10
10	1	2	5.718	6	9	0

In stage 7, the clusters created in stages 1 and 3 are joined. The resulting cluster next appears in stage 8.

EPSON SX125 Series Manuel Horaire analyse Fesdis Nokia Ovi Player Internet Explorer analyse dem 1 (Fesdis) test

tm TP1-2312-2006 Free Download Manager Phillcarto PLAN_de_cours_anal... Adobe Reader X Jeux - Raccourci

09:01
vendredi
2011-12-16

Tutorial

Hierarchical Cluster Analysis > Using HCA to Classify Automobiles >

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	8	11	1.260	0	0	7
2	6	7	1.579	0	0	4
3	2	9	1.625	0	0	7
4	1	6	2.318	0	2	6
5	3	5	2.619	0	0	8
6	1	10	3.670	4	0	10
7	2	8	4.420	3	1	8
8	2	3	4.505	7	5	9
9	2	4	4.774	8	0	10
10	1	2	5.718	6	9	0

When there are many cases, this table becomes rather long, but it may be easier to scan the coefficients column for large gaps rather than scan the dendrogram.

Next

Sans... EN
The ...
Skyp...
1 u...
spss...
Tuto...

09:01
vendredi
2011-12-16

QuickTime Player
khaldi
mics4
ouaga
scan
tm

EPSON SX125 Series
Manuel
Horaire analyse Fesdis
Nokia Ovi Player
Internet Explorer
analyse... (Fesdis)
TP1-2312-2006
Free Download Manager
Phillcarto
PLAN_de_cours_anal...
Adobe Reader X
Jeux - Raccourci

Canon My Printer
Adobe Download Assistant
Nouveau dossier
HP Color LaserJet 1600 sur pse-b91...
test

Tutorial

Hierarchical Cluster Analysis > Using HCA to Classify Automobiles >

A good cluster solution sees a sudden jump (gap) in the distance coefficient. The solution before the gap indicates the good solution.

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	8	11	1.260	0	0	7
2	6	7	1.579	0	0	4
3	2	9	1.625	0	0	7
4	1	6	2.318	0	2	6
5	3	5	2.619	0	0	8
6	1	10	3.670	4	0	10
7	2	8	4.420	3	1	8
8	2	3	4.505	7	5	9
9	2	4	4.774	8	0	10
10	1	2	5.718	6	9	0

Next

Sans... EN

The ...

Skyp...

(1 u...

spss...

Tuto...

09:01

vendredi

2011-12-16

QuickTime Player

khaldi

mics4

ouaga

scan

tm

EPSON SX125 Series

Manuel

Horaire analyse Fesdis

Nokia Ovi Player

Internet Explorer

analyse ue... (Fesdis)

TP1-2312-2006

Free Download Manager

Phillcarto

PLAN_de_cours_anal...

Adobe Reader X

Canon My Printer

Adobe Download Assistant

Nouveau dossier

HP Color LaserJet 1600 sur pse-b91...

test

Jeux - Raccourci

Hierarchical Cluster Analysis > Using HCA to Classify Automobiles >

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	8	11	1.260	0	0	7
2	6	7	1.579	0	0	4
3	2	9	1.625	0	0	7
4	1	6	2.312	0	2	6
5	3	5	2.619	0	0	8
6	1	10	3.670	4	0	10
7	2	8	4.420	3	1	8
8	2	3	4.505	7	5	9
9	2	4	4.774	8	0	10
10	1	2	5.718	6	9	0

The largest gaps in the coefficients column occur between stages 5 and 6, indicating a 6-cluster solution, and stages 9 and 10, indicating a 2-cluster solution. These are the same as your findings from the dendrogram.

EPSON SX125 Series Horaire analyse Fesdis Nokia Ovi Player Internet Explorer analys Next L (Fesdis) Manuel

tm TP1-2312-2006 Free Download Manager Phillcarto PLAN_de_cours_anal... Adobe Reader X Jeux - Raccourci

09:02
vendredi
2011-12-16

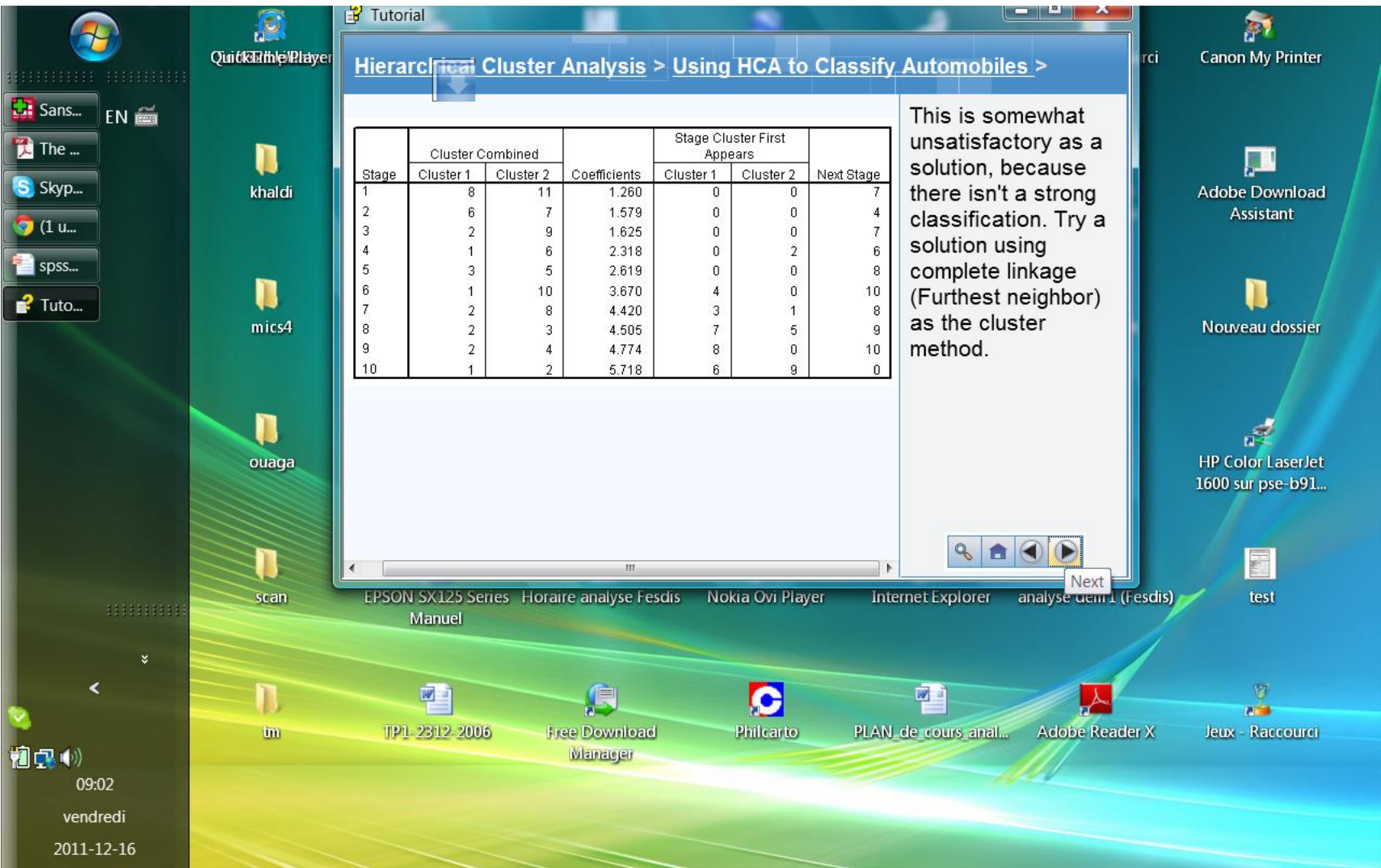
Tutorial

Hierarchical Cluster Analysis > Using HCA to Classify Automobiles >

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	8	11	1.260	0	0	7
2	6	7	1.579	0	0	4
3	2	9	1.625	0	0	7
4	1	6	2.318	0	2	6
5	3	5	2.619	0	0	8
6	1	10	3.670	4	0	10
7	2	8	4.420	3	1	8
8	2	3	4.505	7	5	9
9	2	4	4.774	8	0	10
10	1	2	5.718	6	9	0

This is somewhat unsatisfactory as a solution, because there isn't a strong classification. Try a solution using complete linkage (Furthest neighbor) as the cluster method.

Next



- Sans...
- The ...
- Skyp...
- (1 u...
- spss...
- Tuto...

- khaldi
- mics4
- ouaga
- scan
- tm

- EPSON SX125 Series
- Manuel
- Horaire analyse Fesdis
- Nokia Ovi Player
- Internet Explorer
- analyse ueim1 (Fesdis)
- test
- TP1-2312-2006
- Free Download Manager
- Phillcarto
- PLAN_de_cours_anal...
- Adobe Reader X
- Jeux - Raccourci

- Canon My Printer
- Adobe Download Assistant
- Nouveau dossier
- HP Color LaserJet 1600 sur pse-b91...

09:02
vendredi
2011-12-16

Tutorial

Hierarchical Cluster Analysis > Using HCA to Classify Automobiles >

Hierarchical Cluster Analysis

Manufacturer [man...]
Sales in thousand...
4-year resale valu...
Vehicle type [type]
Log-transformed s...
Zscore: 4-year res...
Zscore: Type [ztype]
Zscore: Price in th...
Zscore: Engine si...
Zscore: Horsepo...
Zscore: Wheelba...
Zscore: Width [zwi...
Zscore: Length [zl...
Zscore: Curb wel...
Zscore: Fuel capa...
Zscore: Fuel effici...

Variables(s):
Price in thousands [p...
Engine size [engine_s...
Horsepower [horsep...
Wheelbase [wheelbas...
Width [width]
Length [length]
Curb weight [curb_wgt]

Label Cases by:
Model [model]

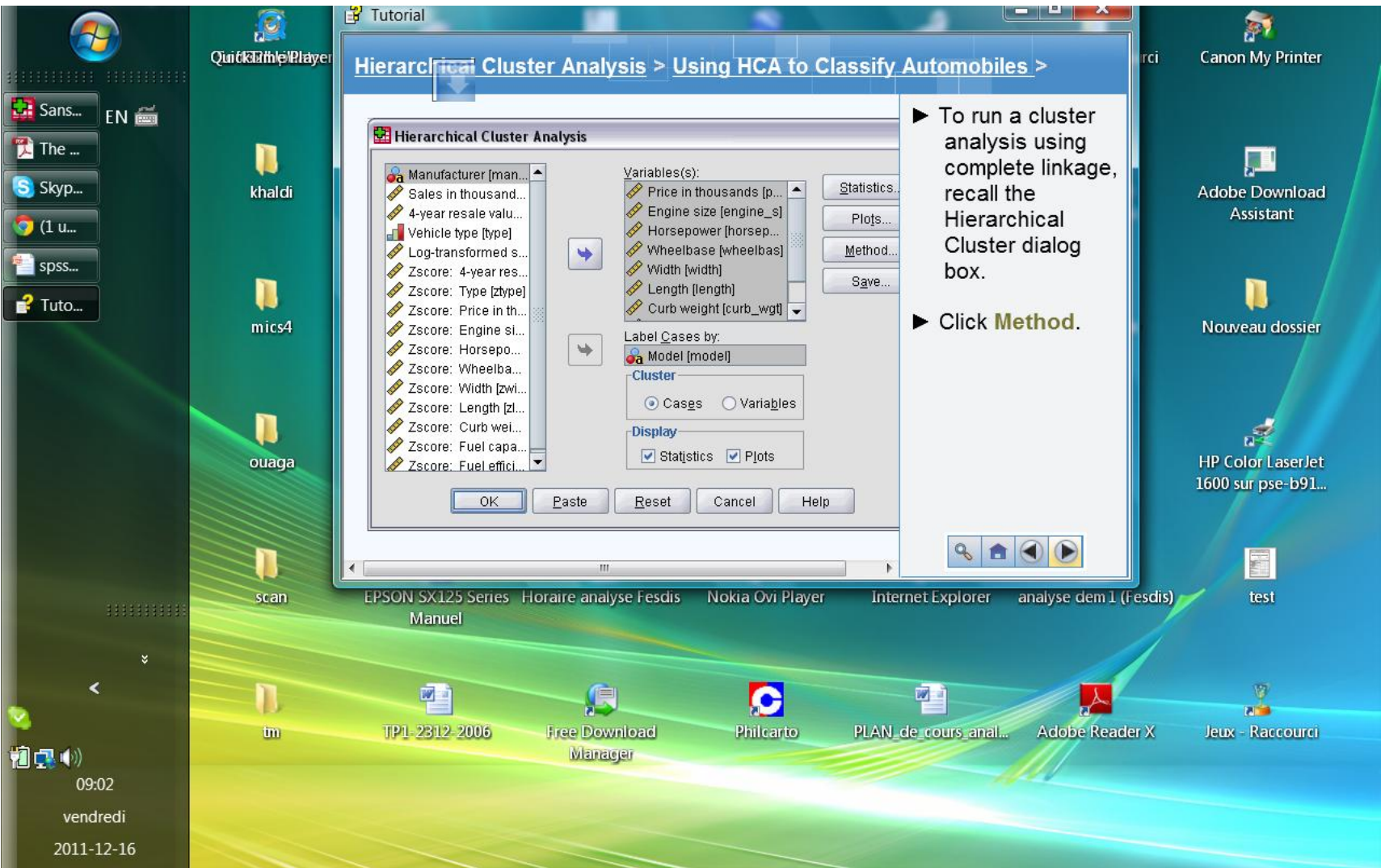
Cluster
 Casgs Variables

Display
 Statistics Plots

Statistics...
Plots...
Method...
Save...

OK Paste Reset Cancel Help

- ▶ To run a cluster analysis using complete linkage, recall the Hierarchical Cluster dialog box.
- ▶ Click **Method**.



09:02
vendredi
2011-12-16

Tutorial

Hierarchical Cluster Analysis > Using HCA to Classify Automobiles >

Hierarchical Cluster Analysis: Method

Cluster Method: Furthest neighbor

Measure

Interval: Squared Euclidean distance
Power: 2 Root: 2

Counts: Chi-squared measure

Binary: Euclidean distance
Present: 1 Absent: 0

Transform Values

Standardize: Z scores
 By variable
 By case:

Transform Measure

Absolute values
 Change sign
 Rescale to 0-1 range

Continue Cancel Help

- ▶ Select **Furthest neighbor** as the cluster method.
- ▶ Click **Continue**.
- ▶ Click **OK** in the Hierarchical Cluster Analysis dialog box.

Next

Windows 7 desktop environment. Taskbar includes: Sans..., The..., Skyp..., (1 u..., spss..., Tuto... (Tutorial). Start menu is open showing: Sans..., The..., Skyp..., (1 u..., spss..., Tuto... (Tutorial). Desktop icons include: QuickTime Player, khaldi, mics4, ouaga, scan, tm. System tray shows: 09:02, vendredi, 2011-12-16.

Windows 7 desktop environment. Taskbar includes: EPSON SX125 Series, Manuel, Horaire analyse Fesdis, Nokia Ovi Player, Internet Explorer, analyse ueim1 (Fesdis). Desktop icons include: Canon My Printer, Adobe Download Assistant, Nouveau dossier, HP Color LaserJet 1600 sur pse-b91..., test, TP1-2312-2006, Free Download Manager, Phillcarto, PLAN_de_cours_anal..., Adobe Reader X, Jeux - Raccourci.

Tutorial

Hierarchical Cluster Analysis > Using HCA to Classify Automobiles >
 Complete Linkage Solution >

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	8	11	1.260	0	0	7
2	6	7	1.579	0	0	5
3	2	9	1.625	0	0	6
4	3	5	2.619	0	0	6
5	6	10	4.012	2	0	9
6	2	3	7.333	3	4	8
7	1	8	9.183	0	1	9
8	2	4	12.440	6	0	10
9	1	6	25.486	7	5	10
10	1	2	54.607	9	8	0

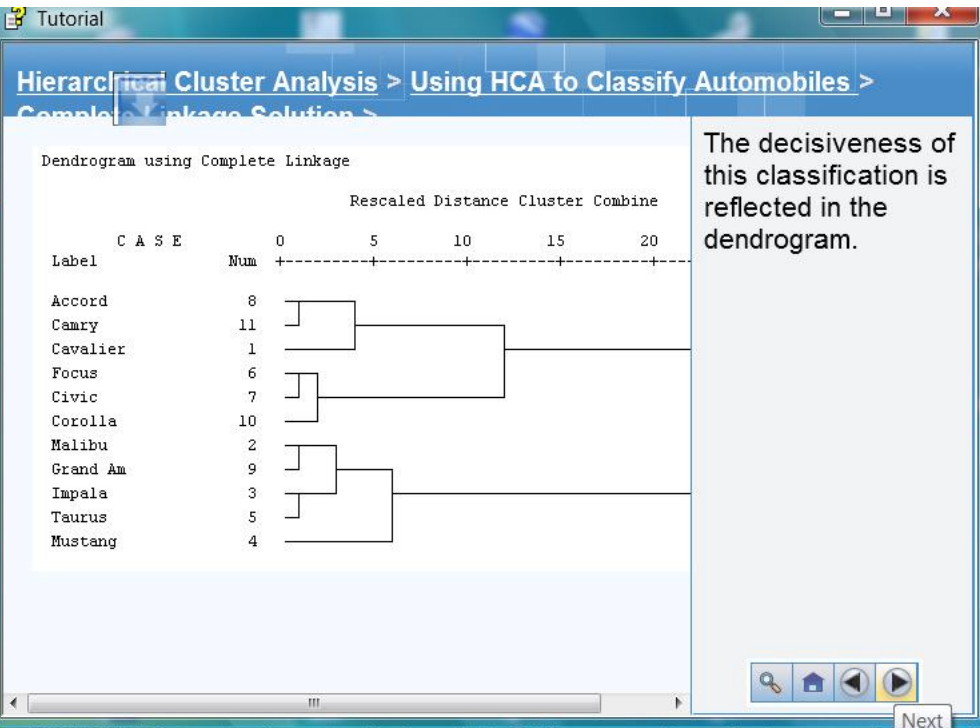
For the first few stages, the schedule for the complete linkage solution is similar to that for the single linkage solution. In the final few stages, they are quite different as the complete linkage solution makes a strong classification of two or three clusters.

Windows taskbar area showing system tray icons (volume, network, clock) and system information: 09:03, vendredi, 2011-12-16.

Windows desktop area with folders: khalidi, mics4, ouaga, scan, tm.

Windows taskbar showing open applications: EPSON SX125 Series Manuel, Horaire analyse Fesdis, Nokia Ovi Player, Internet Explorer, analyse dem 1 (Fesdis).

Windows desktop area with icons: Canon My Printer, Adobe Download Assistant, Nouveau dossier, HP Color LaserJet 1600 sur pse-b91..., test, Jeux - Raccourci.



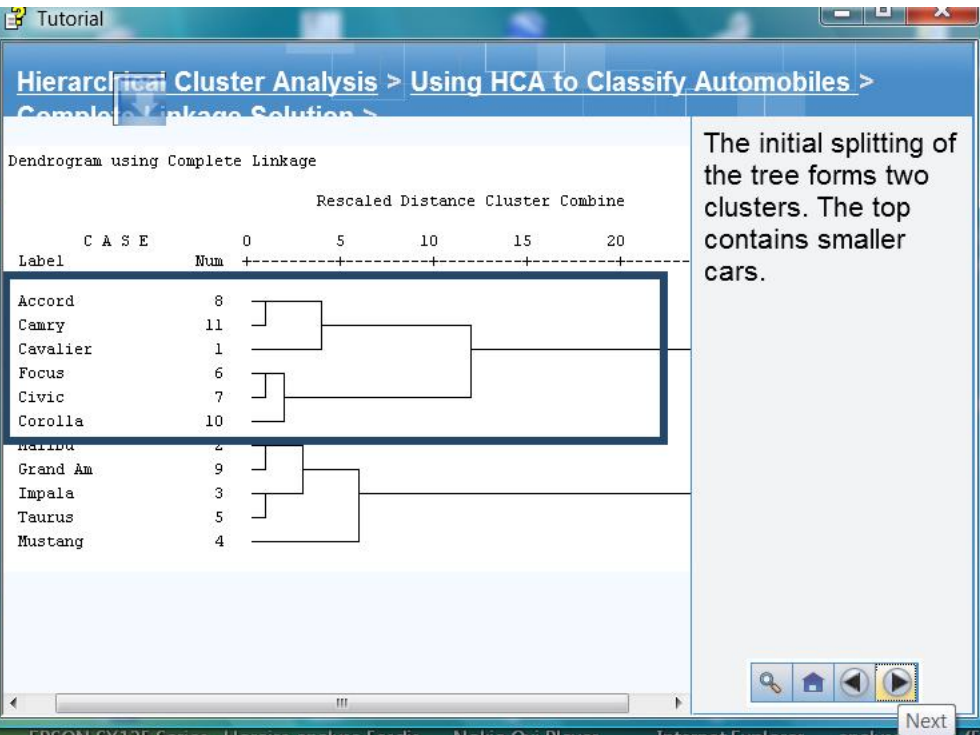
EPSON SX125 Series Manuel Horaire analyse Fesdis Nokia Ovi Player Internet Explorer analyse de m... (Fesdis)

tm TP1-2312-2006 Free Download Manager Phillcarto PLAN_de_cours_anal... Adobe Reader X Jeux - Raccourci

Windows taskbar area including Start button, system tray (clock: 09:03, date: vendredi, 2011-12-16), and application shortcuts (Sans..., The..., Skyp..., spss..., Tuto...).

Left sidebar of the desktop with folders: khaldi, mics4, ouaga, scan, and tm.

Right sidebar of the desktop with icons: Canon My Printer, Adobe Download Assistant, Nouveau dossier, HP Color LaserJet 1600 sur pse-b91..., and test.

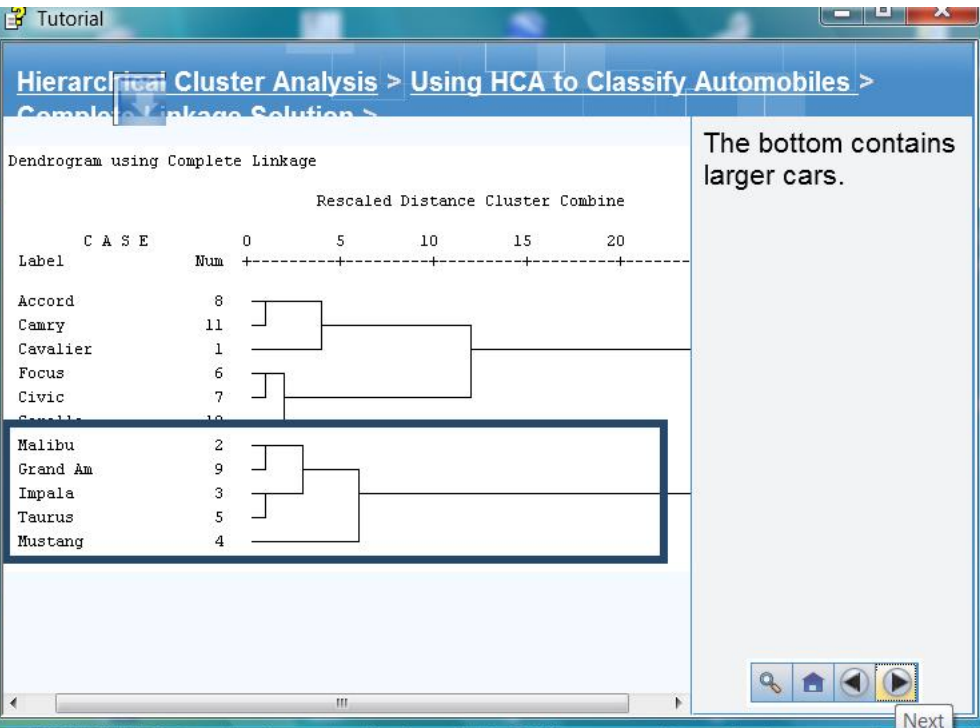


Windows taskbar and desktop icons on the left side of the screen.

- Start button
- System tray: Network, Volume, Date (09:03), Day (vendredi), Date (2011-12-16)
- Taskbar: Sans..., The..., Skyp..., (1 u..., spss..., Tuto...
- Desktop icons: QuickTime Player, khaldi, mics4, ouaga, scan, tm

Windows desktop background and taskbar on the right side of the screen.

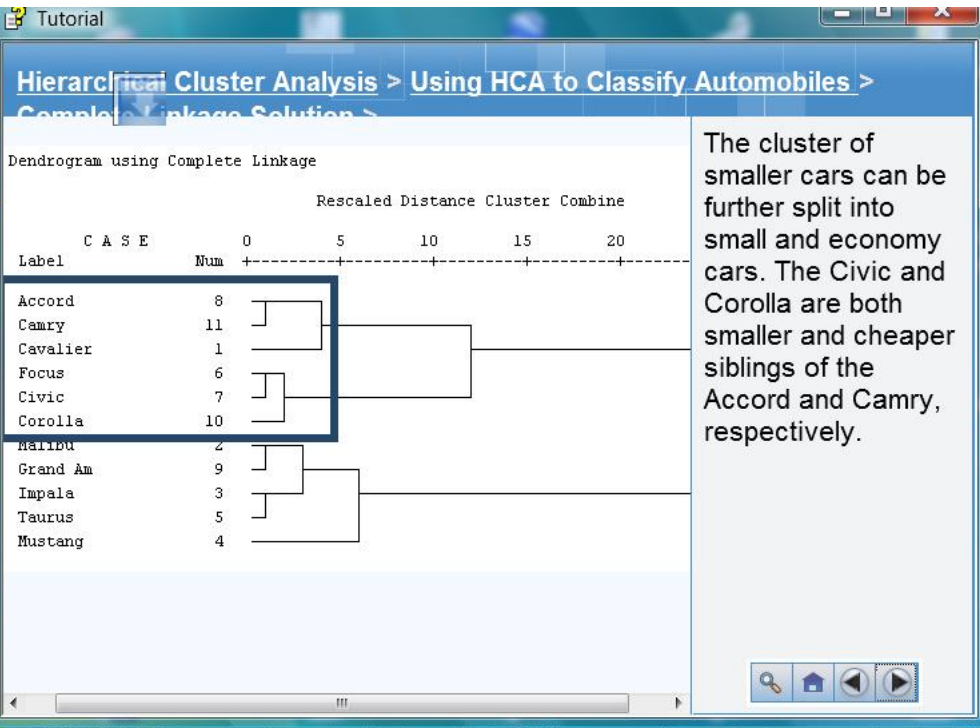
- Taskbar: EPSON SX125 Series, Manuel, Horaire analyse Fesdis, Nokia Ovi Player, Internet Explorer, analyse ueen1 (Fesdis), Next
- Desktop icons: Canon My Printer, Adobe Download Assistant, Nouveau dossier, HP Color LaserJet 1600 sur pse-b91..., test, Jeux - Raccourci



EPSON SX125 Series Horaire analyse Fesdis Nokia Ovi Player Internet Explorer analyse dem 1 (Fesdis) Manuel

tm TP1-2312-2006 Free Download Manager Philcarto PLAN_de_cours_anal... Adobe Reader X Jeux - Raccourci

09:03
vendredi
2011-12-16



The cluster of smaller cars can be further split into small and economy cars. The Civic and Corolla are both smaller and cheaper siblings of the Accord and Camry, respectively.

Windows Desktop Environment

Taskbar: Sans... EN, Skype™ - kouaouciali, Skyp..., (1 u..., spss..., Tuto...

Start Menu: QuickTime Player, Canon My Printer, Adobe Download Assistant, Nouveau dossier, HP Color LaserJet 1600 sur pse-b91...

Taskbar Icons: EPSON SX125 Series, Horaire analyse Fesdis, Nokia Ovi Player, Internet Explorer, analyse dem 1 (Fesdis) Manuel, tm, TP1-2312-2006, Free Download Manager, Phillcarto, PLAN_de_cours_anal..., Adobe Reader X, Jeux - Raccourci

System Tray: 09:04, vendredi, 2011-12-16

[Hierarchical Cluster Analysis](#) > [Using HCA to Classify Automobiles](#) >

The complete linkage solution is satisfying because its clusters are distinct, while the single linkage solution is less conclusive. Using complete linkage clustering, you can determine the competition for vehicles in the design phase by entering their specifications as new cases in the data set and rerunning the analysis.

