

Application of Variance Ratio Criterion (VRC) by Calinski and Harabasz (1974)

Calinski and Harabasz (1974) introduced the variance ratio criterion (VRC), which can be used to determine the “correct” number of clusters in a cluster analysis and which has proven to work well in many situations (Milligan and Cooper 1985).

For a solution with N objects and K segments, the criterion is given by

$$VRC_k = (SS_B / (K - 1)) / (SS_W / (N - K)),$$

where SS_B is the overall between-segment variation and SS_W the overall within-segment variation with regard to all clustering variables. The criterion should look familiar, as this is actually the F-value of a one-way ANOVA with K representing the number of factor levels. Consequently, the VRC can easily be computed using SPSS, even though this is not readily available in the clustering procedures’ SPSS outputs

To finally determine the “correct” number of segments, we compute ω_K for each segment solution as follows:

$$\omega_k = (VRC_{k+1} - VRC_k) - (VRC_k - VRC_{k-1}).$$

In the next step we choose a value for K , which will minimize the value of ω_K . Owing to the term VRC_{k-1} , the minimum number of clusters that can be selected is three. This is a clear disadvantage of the criterion.

We want to illustrate the application of the VRC using the *cars.sav* dataset. Open the dataset and go to Analyze ► Classify ► K-Means Cluster. This displays a new dialog box (Figure A9.1).

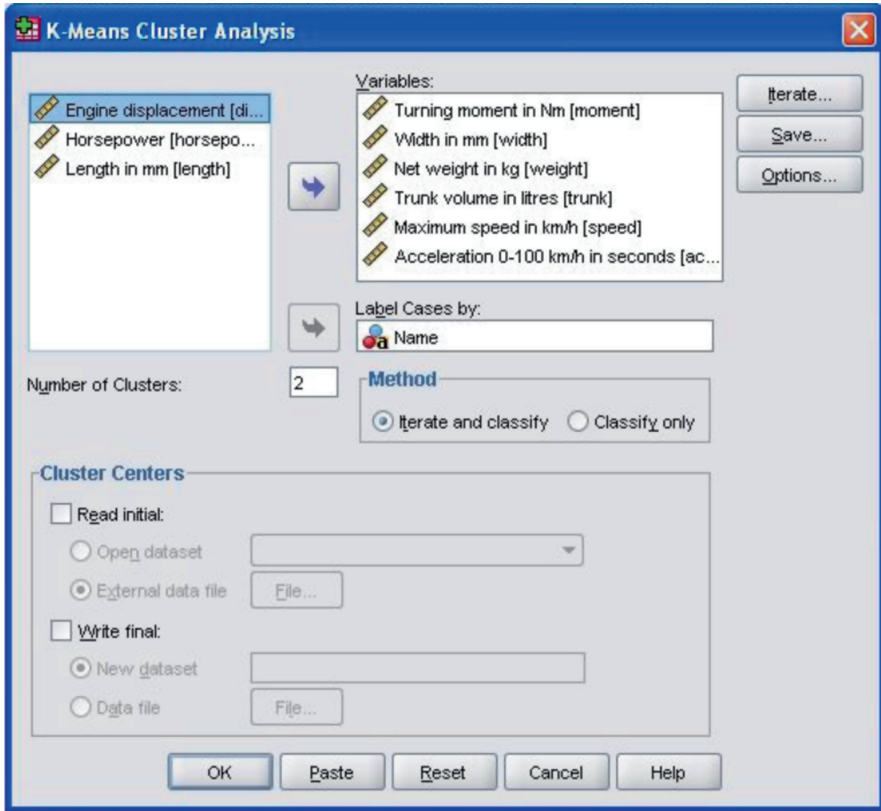


Figure A9.1 k-Means Cluster Analysis Dialog Box.

As with the examples discussed in Chapter 9, move the six clustering variables, i.e. *moment*, *width*, *weight*, *trunk*, *speed*, and *acceleration* into the **Variables** box and specify the case labels (variable: *name*). Instead of using the cluster centers from our previous hierarchical cluster analysis, we allow SPSS to randomly select the initial cluster centers.

As we are interested in comparing different segment solutions, we have to run the analysis for different numbers of segments. In this case, we want to compute VRC values for a three-, four-, and five-segment solution. Please note that this analysis is for illustrative purposes only, as it is not really meaningful to use these numbers of segments with such low numbers of objects in the dataset. Since determining a suitable number segments using VRC involves comparing the VRC values of solutions with one segment less than K and with one segment more than K , we need to run k-means for a three- through six-segment solution.

We start off by running k-means for a two-segment solution. Enter 2 in the **Number of Clusters** box (Figure A9.1). Before starting the analysis, you have to request an ANOVA table by clicking on **Options**, which will produce the dialog box shown in Figure A9.2. The resulting output provides the basis for the computation of VRC.

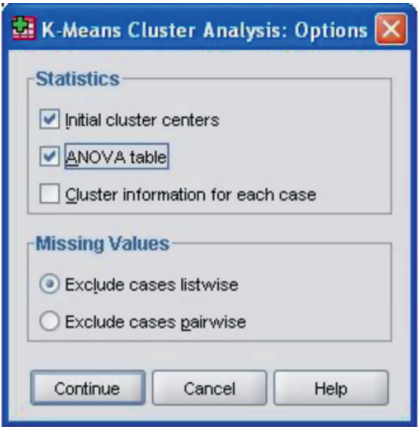


Figure A9.2 K-means Options Menu.

This analysis produces a number of outputs such as the initial and final cluster centers. In addition to the outputs discussed in Chapter 9, SPSS generates an ANOVA table shown in Table A9.1. The ANOVA table indicates whether there are significant differences in the clustering variables across the segments retained from the data. As we can see, all p-values are rather low, indicating that, overall, each of the six clustering variables differs significantly across the clusters. Note that this does not mean that the clustering variables differ between all segments – this result merely indicates that at least one cluster is significantly different from the others with respect to each clustering variable. In order to evaluate whether all three clusters exhibit significant differences, we would have to carry out pairwise comparisons using post hoc tests (compare Chapter 6). Even though this analysis renders interesting results, we are primarily interested in the F-values (second column from the right) which partly correspond to the VRC statistic (Compare Chapter 9).

Table A9.2–A9.5 show the ANOVA outputs for a three- through six-segment solution using k-means.

Table A9.1 ANOVA output for k-means analysis with two segments

	ANOVA					
	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
moment	122331,868	1	1208,902	13	101,193	,000
width	47761,376	1	1818,643	13	26,262	,000
weight	668960,744	1	10833,584	13	61,749	,000
trunk	212487,619	1	9598,593	13	22,137	,000
speed	14994,525	1	415,990	13	36,045	,000
acceleration	83,160	1	3,326	13	25,004	,000

The F tests should only be used for descriptive purposes because the clusters have been chosen to maximize the differences between cases in different clusters. The observed significance levels are not corrected for this and cannot therefore be interpreted as tests of the hypothesis that the cluster means are equal.

Table A9.2 ANOVA output for k-means analysis with three segments

	ANOVA					
	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
moment	62902,371	2	1020,238	12	61,655	,000
width	23885,010	2	1969,476	12	12,128	,001
weight	337371,095	2	11254,595	12	29,976	,000
trunk	142820,381	2	4302,381	12	33,196	,000
speed	7655,200	2	424,333	12	18,041	,000
acceleration	44,975	2	3,037	12	14,808	,001

The F tests should only be used for descriptive purposes because the clusters have been chosen to maximize the differences between cases in different clusters. The observed significance levels are not corrected for this and cannot therefore be interpreted as tests of the hypothesis that the cluster means are equal.

Table A9.3 ANOVA output for k-means analysis with four segments

	ANOVA					
	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
moment	34064,533	3	3259,455	11	10,451	,002
width	16058,400	3	2111,685	11	7,605	,005
weight	244397,944	3	6963,955	11	35,095	,000
trunk	97282,878	3	4129,155	11	23,560	,000
speed	3854,400	3	803,564	11	4,797	,023
acceleration	23,753	3	5,012	11	4,739	,023

The F tests should only be used for descriptive purposes because the clusters have been chosen to maximize the differences between cases in different clusters. The observed significance levels are not corrected for this and cannot therefore be interpreted as tests of the hypothesis that the cluster means are equal.

Table A9.4 ANOVA output for k-means analysis with five segments

	ANOVA					
	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
moment	25989,233	4	3409,067	10	7,624	,004
width	12642,333	4	2083,440	10	6,068	,010
weight	188779,467	4	5467,947	10	34,525	,000
trunk	73369,167	4	4379,267	10	16,754	,000
speed	2974,133	4	850,587	10	3,497	,049
acceleration	17,917	4	5,473	10	3,274	,058

The F tests should only be used for descriptive purposes because the clusters have been chosen to maximize the differences between cases in different clusters. The observed significance levels are not corrected for this and cannot therefore be interpreted as tests of the hypothesis that the cluster means are equal.

Table A9.5 ANOVA output for k-means analysis with six segments

	ANOVA					
	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
moment	26402,460	5	670,589	9	39,372	,000
width	12188,247	5	1162,500	9	10,485	,002
weight	156543,607	5	3008,811	9	52,028	,000
trunk	65755,047	5	943,789	9	69,671	,000
speed	3675,120	5	225,200	9	16,319	,000
acceleration	22,554	5	1,514	9	14,895	,000

The F tests should only be used for descriptive purposes because the clusters have been chosen to maximize the differences between cases in different clusters. The observed significance levels are not corrected for this and cannot therefore be interpreted as tests of the hypothesis that the cluster means are equal.

To compute the VRC statistic for each case of number of segments, we simply sum up the F-values from Tables A9.1-A9.5, using a spreadsheet program such as Microsoft Excel. The results are shown in Table A9.6

Table A9.6 VRC values

Number of clusters	VRC
2	272.390
3	169.804
4	86.247
5	71.742
6	202.770

To determine the “correct” number of segments, we compute ω_K for each segment solution. For example, for $K = 3$, ω_K is given by

$$\omega_3 = (86.247 - 169.804) - (169.804 - 272.390) = 19.029$$

Similarly, we can compute ω_K for four and five segments resulting in $\omega_4 = 69.052$ and $\omega_5 = 145.533$, respectively.

Comparing the values for ω_K , we establish that the minimum is achieved for $K = 3$. Thus, we would choose a three-segment solution for our analysis.

References

- Calinski, T. and J. Harabasz (1974). “A Dendrite Method for Cluster Analysis,” *Communications in Statistics – Theory and Methods*, 3 (1), 1-27.
- Milligan, Glenn W. and Martha Cooper (1985). “An Examination of Procedures for Determining the Number of clusters in a Data Set,” *Psychometrika*, 50 (2), 159-179.

