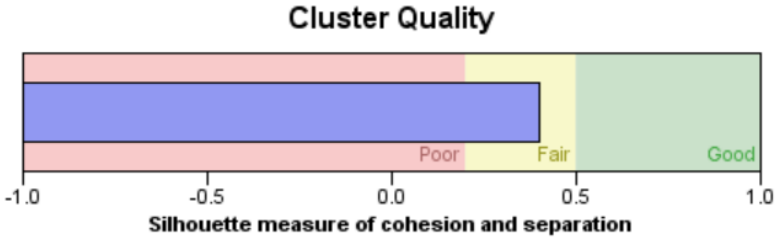
Evaluating Goodness of Clustering

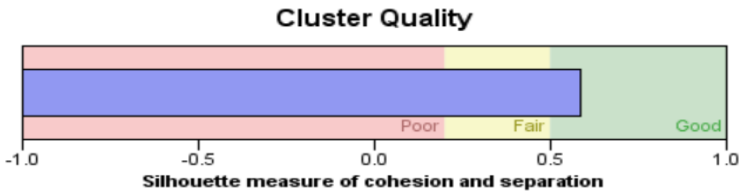
Look at the following SPSS documentation as well
http://www.norusis.com/pdf/SPC_v19.pdf

1. Silhouette Measure of Cohesion and Separation

Once you make some choices or do nothing and go with the defaults, the clusters are formed. At this point, you can consider whether the number of clusters is "good." If you use automated cluster selection, IBM SPSS Statistics finds the number of clusters at which the Schwarz Bayesian Criterion, abbreviated BIC (the **I** stands for Information), becomes small and the change in BIC between adjacent number of clusters is small. Various measures are used to quantify the "goodness" of a cluster solution. In a good cluster solution, the elements within a cluster are similar to one another (cohesive) while dissimilar or distant from points in other clusters (separated).

A popular measure is the **silhouette coefficient**, which is a measure of both cohesion and separation. For each element in a cluster, you calculate the average distance to all other elements in its cluster and the average distance to all elements in each of the other clusters. For each element, the **silhouette measure** is the difference between the smallest average between cluster distance and the average within cluster distance, divided by the larger of the two distances. In a good solution, the within-cluster distances are small and the between cluster distances are large, resulting in a silhouette measure close to the maximum value of 1. If the silhouette measure is negative, the average distance of a case to members of its own cluster is larger than the average distance to cases in other clusters, an undesirable feature. The silhouette measure for a cluster is just the average of the silhouette measures for the cases within the cluster. The silhouette measure ranges from –1 to +1. In the examples below, taken from SPSS Modeler, the first image shows an average Silhouette coefficient of **0.4** (average across the clusters), which is not the best you can hope for. When comparing two clusterings, Average Silhouette Coefficient is a measure that you should take into account. The second image below is of better quality.





SPSS Modeler Algorithm Guide: Chapter 11. Cluster Evaluation Algorithms

Calculating the Silhoutte coefficient

The average Silhouette coefficient is simply the average over all cases of the following calculation for each individual case:

$$(B - A) / \max(A, B)$$

where A is the average distance from the case to every other case assigned to the same cluster and B is the minimal average distance from the case to cases of a different cluster across all clusters.

Unfortunately, this coefficient is computationally expensive. In order to ease this burden, we use the following definitions of A and B:

- A is the distance from the case to the centroid of the cluster which the case belongs to;
- B is the minimal distance from the case to the centroid of every other cluster.

Distances may be calculated using Euclidean distances. The Silhouette coefficient and its average range between −1, indicating a very poor model, and 1, indicating an excellent model. As found by Kaufman and Rousseeuw (1990), an average silhouette greater than 0.5 indicates reasonable partitioning of data; less than 0.2 means that the data do not exhibit cluster structure.

2. Sum of Squared Error Measurement

The Most common measure for evaluating the Goodness of a clustering is the Sum of Squared Error (SSE), where error for each point (record) is defined as the distance to the centroid of the cluster it belongs to, once cluster labels are assigned. To get SSE, we square these errors and sum them.

- Cohesion is measured by the within cluster sum of squares (SSE)

$$WSS = \sum_{i} \sum_{x \in C_i} (x - m_i)^2$$

- The smaller the SSE, as defined above, more cohesion there is it. i.e; points within a cluster are close to each other.
- In the diagram below, distance of each point to its cluster centroid are shown. You need to get each of these distances (SPSS will provide it to you), square them and add them. You can use EXCEL for it.
- Given different clusterings, we can choose the one with the smallest SSE. Or some may prefer to calculate the Average SSE, by dividing the SSE by the number of clusters.

- One easy way to reduce SSE is to increase K, the number of clusters. SO you need to be careful
- A good clustering with a smaller number of clusters can have a lower SSE than a poor clustering with higher number of clusters.