1. 概述

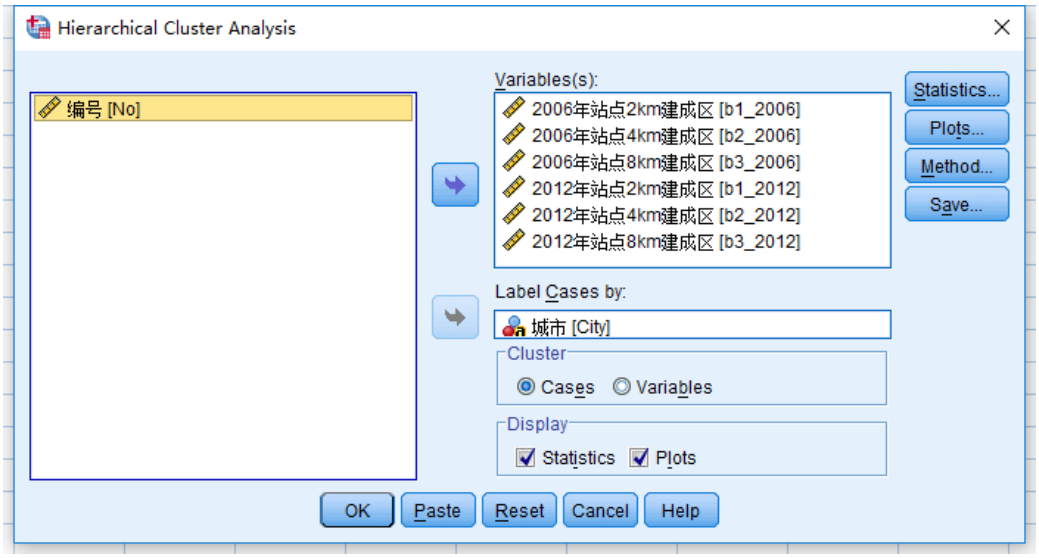
本节以 SPSS 为软件，介绍两种聚类分析的操作过程。首先，以高铁站点周边地区的城镇化数据解说层次聚类法的应用，数据文件为“HST_urbanization.sav”；然后，以商业空间客流统计数据解说 k-means 聚类法的应用，数据文件为“consumers.sav”。另外，附件中还提供了这两份数据对应的 csv 文件。

2. 层次聚类法

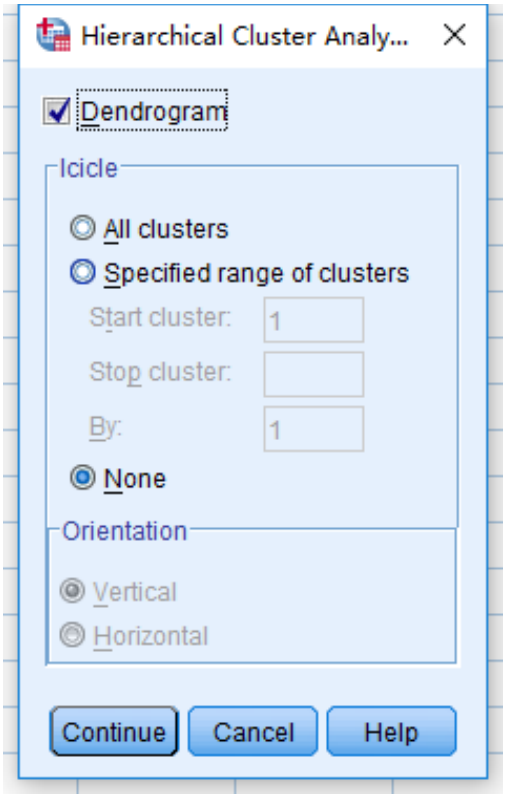
用 SPSS 打开“HST_urbanization.sav”，数据如下图所示。

No	City	b1_2006	b2_2006	b3_2006	b1_2012	b2_2012	b3_2012
1	北京	100.00	100.00	100.00	100.00	100.00	100.00
2	南京	100.00	94.99	81.04	100.00	97.91	92.33
3	上海	85.20	82.76	65.57	100.00	98.96	91.62
4	天津	5.85	8.07	19.11	15.90	28.47	40.35
5	常州	48.60	58.40	41.11	64.22	68.56	58.42
6	昆山	57.71	69.37	51.86	79.45	85.07	73.09
7	济南	5.28	21.19	30.17	97.79	79.76	51.34
8	泰安	28.76	21.18	20.28	80.13	49.75	28.93
9	苏州	54.35	28.66	19.40	73.97	37.69	21.52
10	廊坊	89.51	62.68	10.58	100.00	86.96	31.26

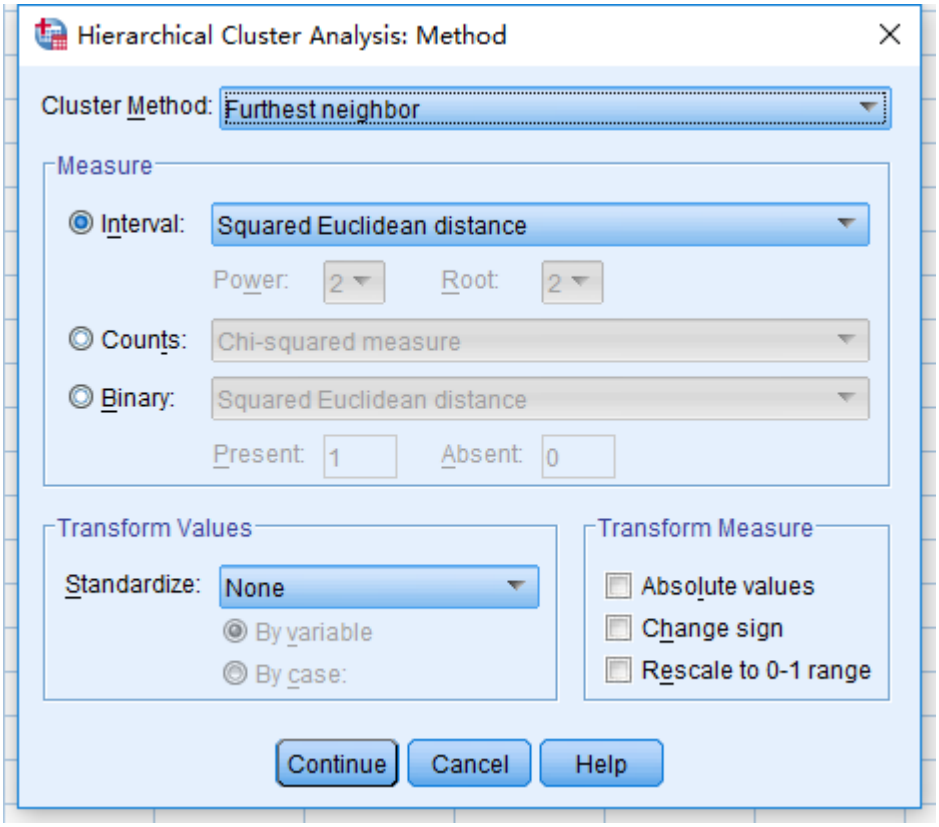
在菜单栏中选择“Analyze – Classify – Hierarchical Cluster”，进入层次聚类对话框。将“b1_2006”、“b2_2006”、“b3_2006”、“b1_2012”、“b2_2012”、“b3_2012”这 6 个变量选入“Variable(s)”中，作为聚类分析所使用的变量；将“城市 [City]”选入“Label Cases by”中，以城市名对样本进行标注，如下图所示。



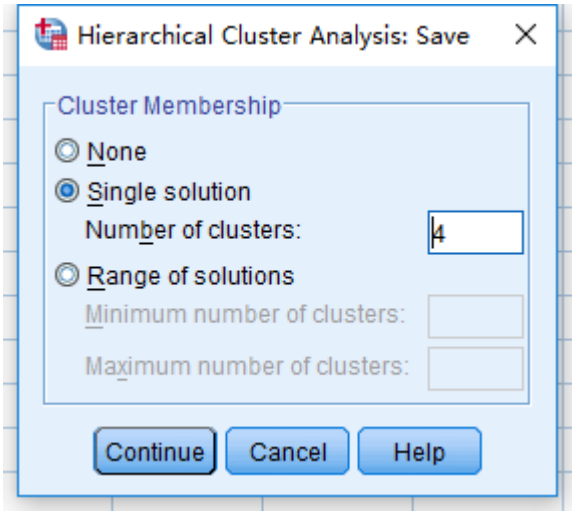
点击“Plot”进入绘图子对话框。勾选“Dendrogram”，要求显示谱系图。同时，下方的“Icicle”默认选中“All clusters”，将其改选至“None”，要求不显示冰柱图。一般而言，谱系图比冰柱图更加清晰直观，是实务中的首选。



点击“Continue”回到主对话框，点击“Method”进入方法子对话框。在“Cluster Method”下拉菜单中有不同的测度类间距离的方法，默认为“Between-groups linkage”，这里将其更改为“Furthest neighbor”，使用最长距离法。在“Measure”中不同的测度每两个样本点之间距离的方法，这里使用默认的“Interval：Squared Euclidean distance”，即把所有变量作为连续变量，使用平方欧氏距离。



点击“Continue”回到主对话框，点击“Save”进入保存子对话框。选中“Single solution”，并在“Number of clusters”中输入 4，要求 SPSS 将样本分成 4 类，并保存类别变量。如果希望尝试不同数量的类别，可以选择“Range of solutions”，然后在下面的“Minimum number of clusters”和“Maximum number of clusters”中设置类别数量的上下限。



点击“Continue”回到主对话框，点击“OK”运行分析。在 SPSS 的结果窗口中首先报告了样本情况，本案例共有 21 个样本。下方出现“Squared Euclidean Distance”的注释，表明样本点之间距离的测度方式是平方欧氏距离，同时，“Complete Linkage”的注释表明在计算类别之间的距离时使用了最长距离法。

Case Processing Summary[a,b]

Cases					
Valid		Missing		Total	
N	Percent	N	Percent	N	Percent
21	100.0	0	.0	21	100.0

a. Squared Euclidean Distance used

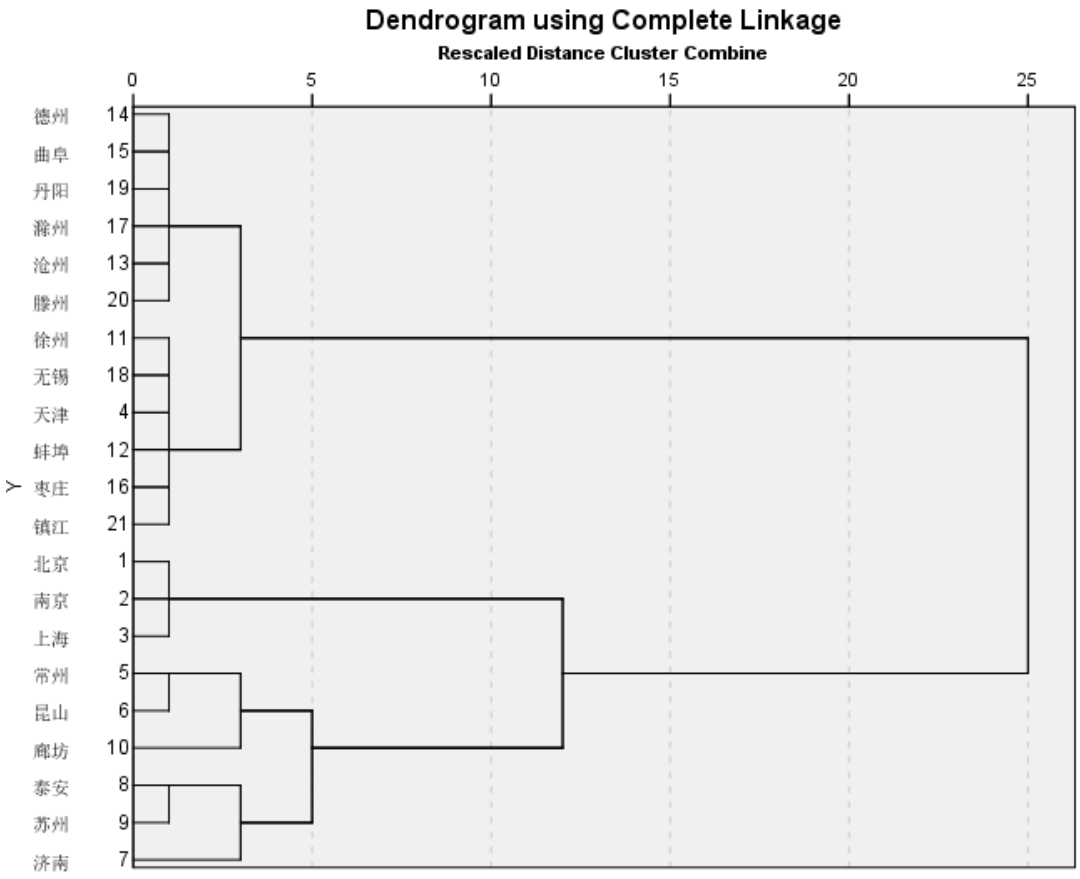
b. Complete Linkage

然后，SPSS 汇报的是层次聚类的聚合过程。

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	14	15	36.959	0	0	2
2	14	19	75.742	1	0	4
3	11	18	123.605	0	0	8
4	14	17	127.042	2	0	6
5	13	20	132.632	0	0	6
6	13	14	362.704	5	4	16
7	1	2	447.779	0	0	14
8	4	11	466.697	0	3	10
9	8	9	949.870	0	0	17
10	4	12	975.962	8	0	13
11	5	6	1038.637	0	0	15
12	16	21	1114.468	0	0	13
13	4	16	1717.824	10	12	16
14	1	3	1772.989	7	0	19
15	5	10	4980.461	11	0	18
16	4	13	5246.277	13	6	20
17	7	8	5806.168	0	9	18
18	5	7	9659.812	15	17	19
19	1	5	25613.833	14	18	20
20	1	4	57916.813	19	16	0

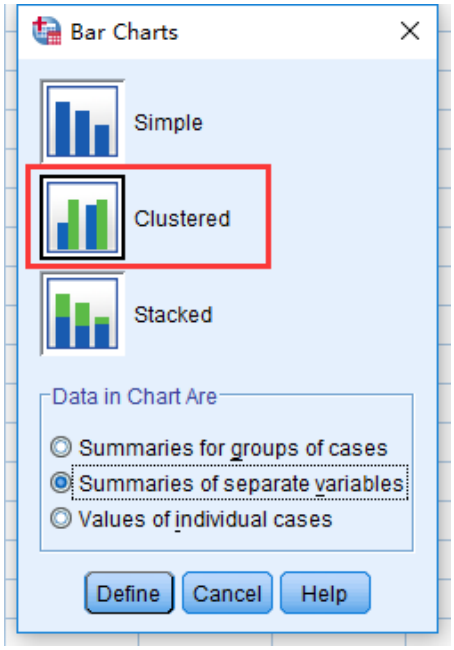
接下来，SPSS 呈现了聚类谱系图。



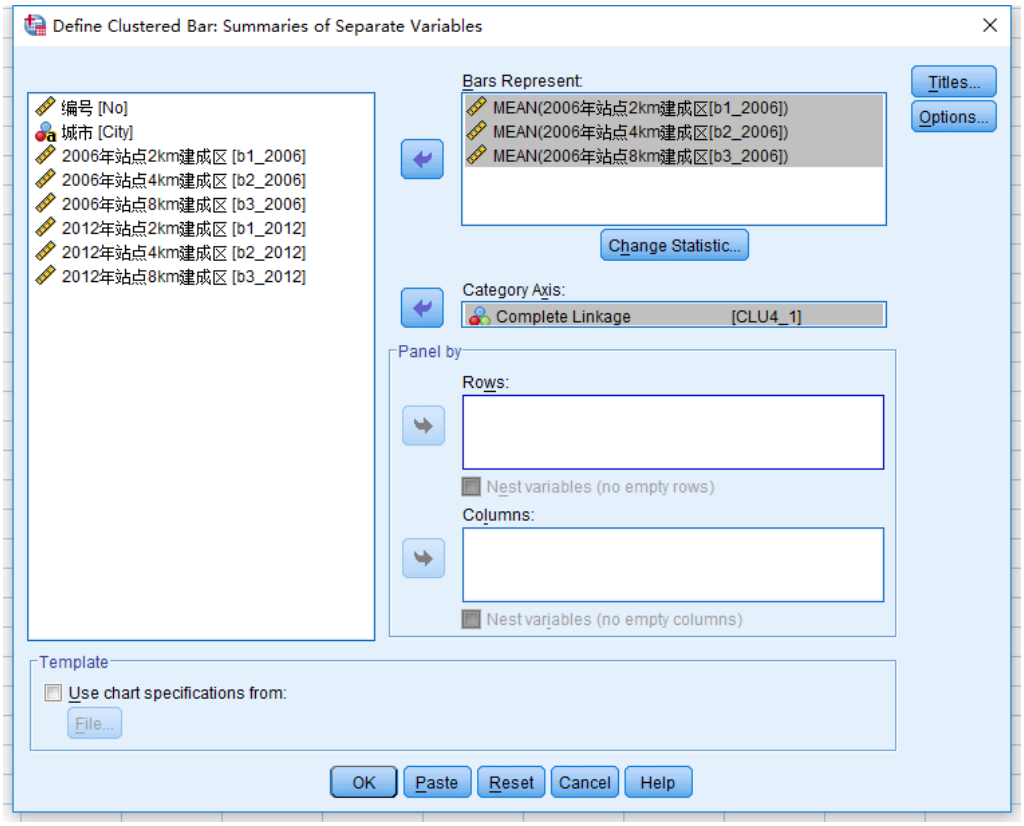
回到数据视图中，可以看到，SPSS 已经在最后一列保存了类别变量，其取值为 1、2、3、4，分别代表样本所属的四个类别。

No	City	b1_2006	b2_2006	b3_2006	b1_2012	b2_2012	b3_2012	CLU4_1
1	北京	100.00	100.00	100.00	100.00	100.00	100.00	1
2	南京	100.00	94.99	81.04	100.00	97.91	92.33	1
3	上海	85.20	82.76	65.57	100.00	98.96	91.62	1
4	天津	5.85	8.07	19.11	15.90	28.47	40.35	2
5	常州	48.60	58.40	41.11	64.22	68.56	58.42	3
6	昆山	57.71	69.37	51.86	79.45	85.07	73.09	3
7	济南	5.28	21.19	30.17	97.79	79.76	51.34	4
8	泰安	28.76	21.18	20.28	80.13	49.75	28.93	4
9	苏州	54.35	28.66	19.40	73.97	37.69	21.52	4
10	廊坊	89.51	62.68	10.58	100.00	86.96	31.26	3
11	徐州	1.38	5.05	16.18	25.48	33.15	36.79	2
12	蚌埠	14.71	12.88	12.55	22.58	21.60	14.44	2
13	沧州	.00	.84	13.21	2.85	8.87	18.72	2
14	德州	.00	.00	3.25	.00	.00	7.50	2
15	曲阜	.00	.39	7.84	.00	.61	11.42	2
16	枣庄	3.21	30.44	8.46	31.16	44.82	15.92	2
17	滕州	.56	1.65	.65	1.08	6.76	6.87	2
18	无锡	.00	11.08	16.62	33.90	36.12	34.46	2
19	丹阳	.00	.31	.86	.00	.21	15.86	2
20	滕州	2.39	5.65	11.18	12.15	6.95	15.64	2
21	镇江	13.68	18.00	24.44	44.83	38.84	35.21	2

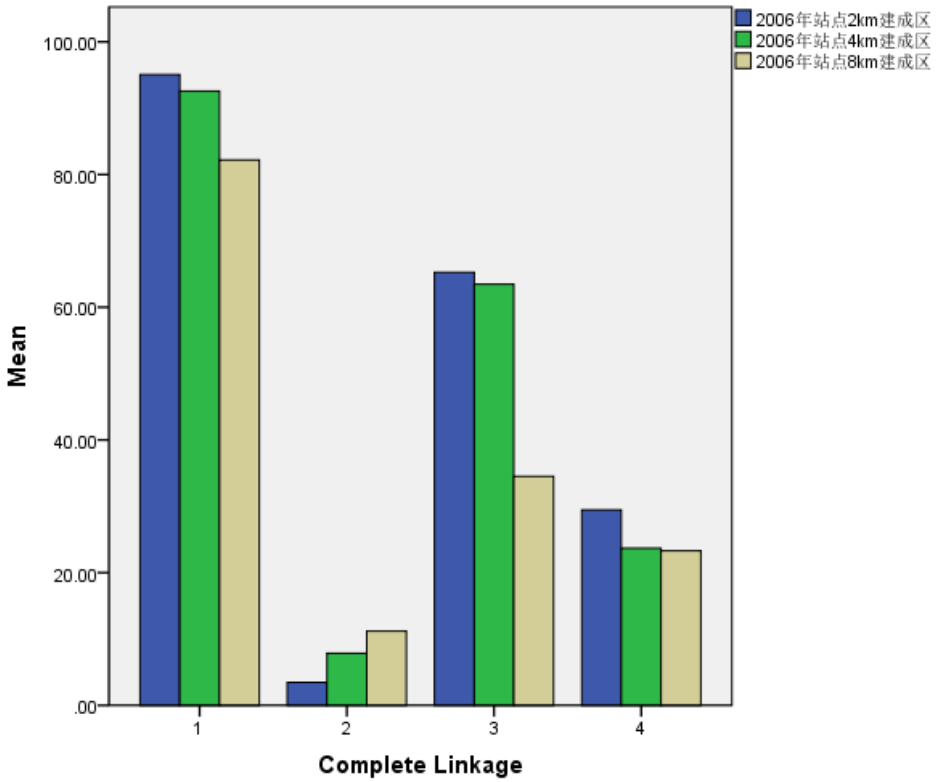
在聚类分析完成之后，通常需要对比各类别的样本在原始变量上呈现的不同特征，从而对各类别的意义进行解读，这里我们采用柱状图的方式。在菜单栏中选择“Graphs – Legacy Dialogs - Bar”，弹出柱状图对话框，然后选择“Clustered”，使用集群柱状图，并将下方的“Data in Chart Are”改选为“Summaries of separate variables”，即图中的各个柱子代表不同的变量，如右图所示。



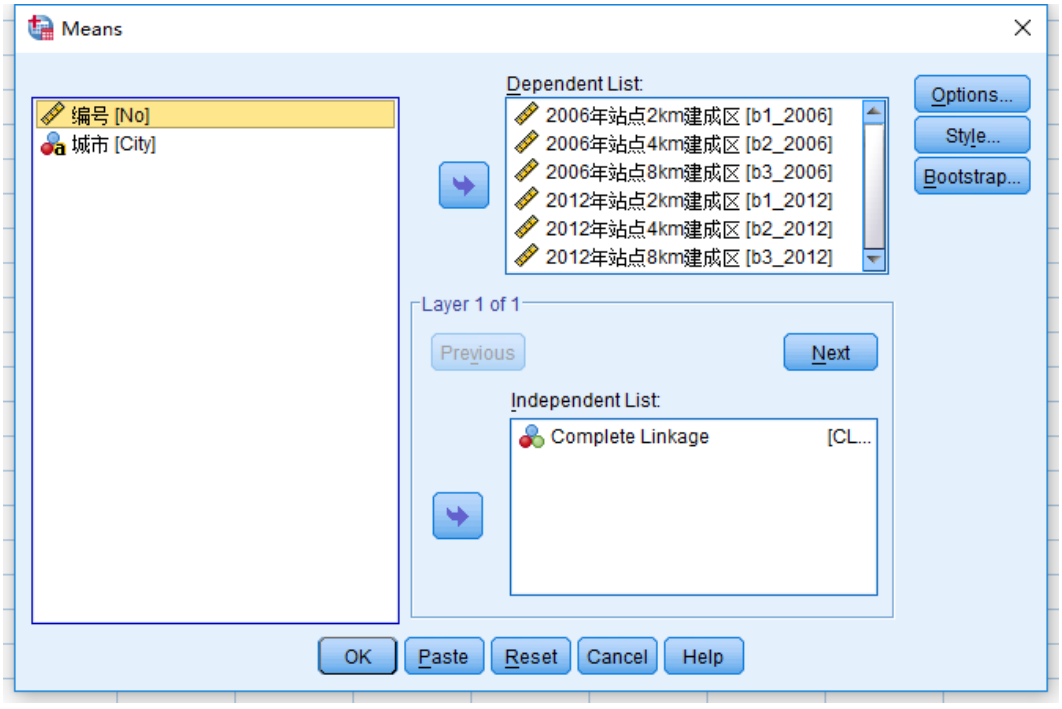
点击“Define”，在弹出的对话框中将“b1_2006”、“b2_2006”、“b3_2006”、这 3 个变量选入“Bars Represent”中，并将“Complete Linkage [CLU4_1]”选入“Category Axis”。



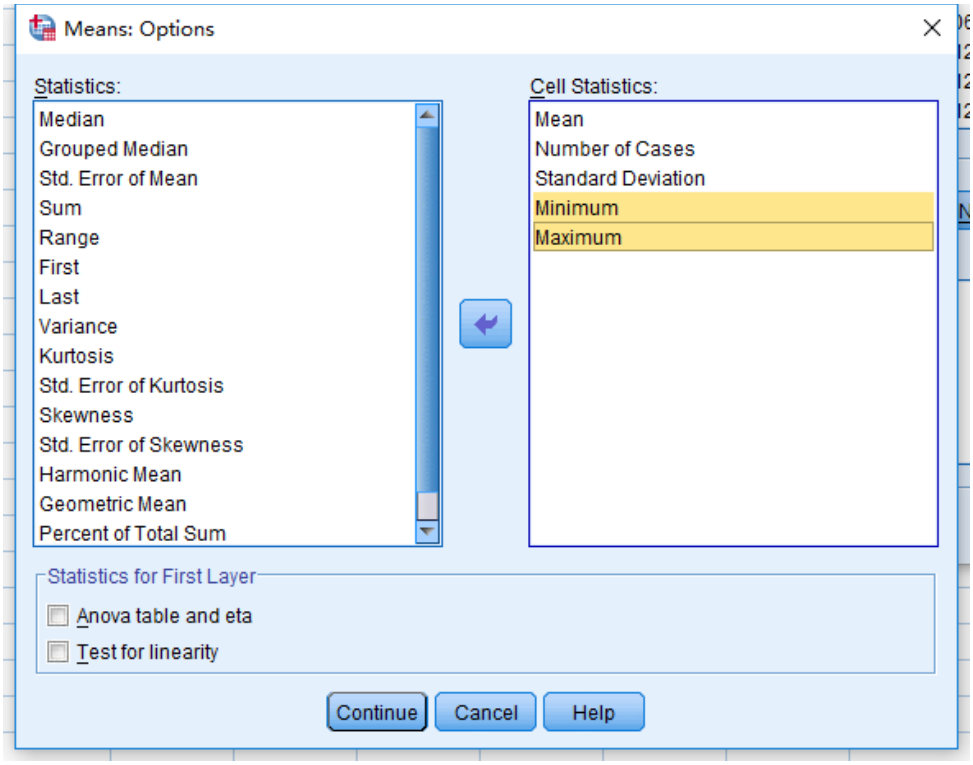
点击“OK”，则 SPSS 的结果窗口中出现以下柱状图。



另外，我们也可以采用表格的形式统计各类别的指标特征。一种简单的实现方式是 SPSS 中的均值比较功能。在菜单栏中选择“Analyze – Compare Means - Means”，打开均值比较对话框。将“b1_2006”、“b2_2006”、“b3_2006”、“b1_2012”、“b2_2012”、“b3_2012”这 6 个变量选入“Dependent List”中，将“Complete Linkage [CLU4_1]”选入“Independent List”中，如下所示。



点击“Options”，在弹出的选项对话框中，将左侧“Statistics”中的“Minimum”、“Maximum”选入右侧的“Cell Statistics”。SPSS 将会比较各类别在“Cell Statistics”中的 5 个统计指标：均值、样本量、标准差、最小值、最大值。



点击“Continue”回到主菜单，点击“OK”运行分析，统计结果如下：

Report

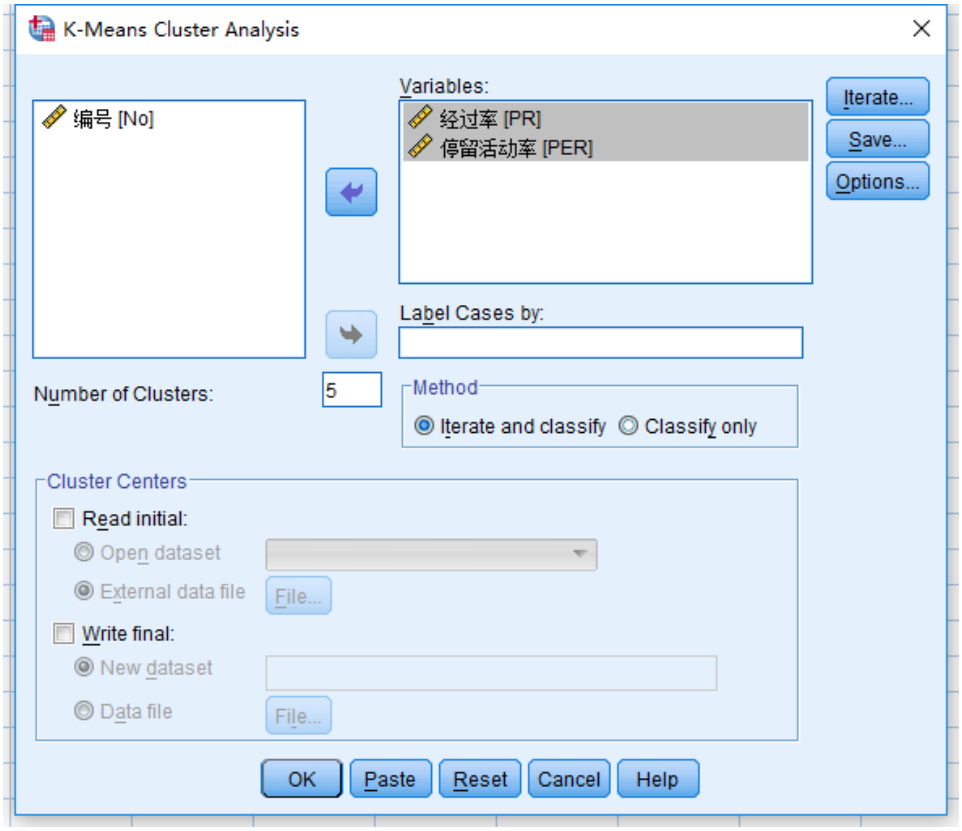
Complete Linkage		2006年站点2km建成区	2006年站点4km建成区	2006年站点8km建成区	2012年站点2km建成区	2012年站点4km建成区	2012年站点8km建成区
1	Mean	95.0667	92.5833	82.2033	100.0000	98.9567	94.6500
	N	3	3	3	3	3	3
	Std. Deviation	8.54478	8.86840	17.24446	.00000	1.04500	4.64682
	Minimum	85.20	82.76	65.57	100.00	97.91	91.62
	Maximum	100.00	100.00	100.00	100.00	100.00	100.00
2	Mean	3.4817	7.8633	11.1958	15.8275	18.8667	21.0983
	N	12	12	12	12	12	12
	Std. Deviation	5.31310	9.16265	7.37393	15.64543	16.78673	12.09657
	Minimum	.00	.00	.65	.00	.00	6.87
	Maximum	14.71	30.44	24.44	44.83	44.82	40.35
3	Mean	65.2733	63.4833	34.5167	81.2233	80.1967	54.2567
	N	3	3	3	3	3	3
	Std. Deviation	21.47813	5.52895	21.41527	17.95580	10.12186	21.22351
	Minimum	48.60	58.40	10.58	64.22	68.56	31.26
	Maximum	89.51	69.37	51.86	100.00	86.96	73.09
4	Mean	29.4633	23.6767	23.2833	83.9633	55.7333	33.9300
	N	3	3	3	3	3	3
	Std. Deviation	24.54256	4.31570	5.98024	12.36402	21.66383	15.52605
	Minimum	5.28	21.18	19.40	73.97	37.69	21.52
	Maximum	54.35	28.66	30.17	97.79	79.76	51.34
Total	Mean	29.1043	30.1710	26.3981	46.9281	44.3362	38.1757
	N	21	21	21	21	21	21
	Std. Deviation	37.04062	33.38587	26.94850	39.58743	35.51593	29.09753
	Minimum	.00	.00	.65	.00	.00	6.87
	Maximum	100.00	100.00	100.00	100.00	100.00	100.00

3. k-means 聚类法

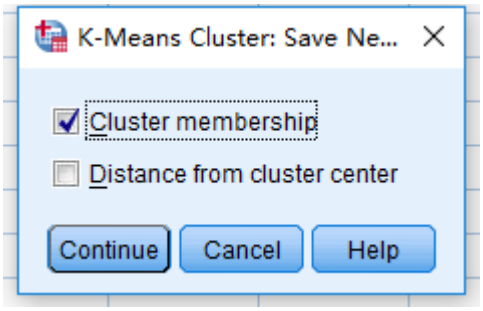
在 SPSS 中打开“consumers.sav”，数据如下图所示。其中，“PR”为经过率，“PER”为停留活动率。

No	PR	PER
1	.7761	.0479
2	.7361	.0279
3	.7000	.1700
4	.4203	.2431
5	.4000	.2300
6	.4500	.1900
7	.4739	.1241
8	.4500	.1200
9	.4400	.1500

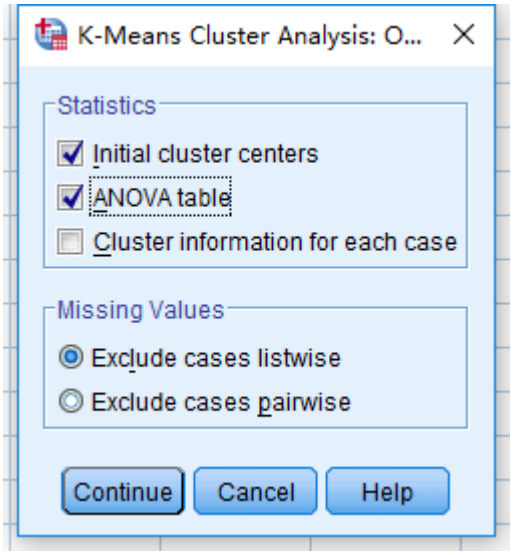
在菜单栏中选择“Analyze – Classify – K-Means Cluster”，打开 k-means 聚类分析对话框。将“经过率 [PR]”和“停留活动率 [PER]”选入“Variables”，作为聚类分析的原始变量。在“Number of Clusters”中输入“5”，将样本分为 5 类。



点击“Save”进入保存子对话框，勾选“Cluster membership”，将分类变量保存到数据中。



点击“Continue”回到主对话框，然后点击“Options”进入选项子对话框。勾选“ANOVA table”，要求汇报方差分析表。



点击“Continue”回到主对话框，再点击“OK”运行分析。在 SPSS 的结果窗口中首先显示了 5 个类别的初始中心的位置。请注意，k-means 聚类法对初始中心的位置敏感，不同的初始中心可能导致不同的聚类结果。

Initial Cluster Centers

	Cluster				
	1	2	3	4	5
经过率	.7281	.0248	.0093	.4225	.5000
停留活动率	.3189	.2500	1.0000	.0769	.7100

然后，SPSS 报告了 k-means 的迭代过程，即中心位置不断更新的过程，其原理与课件中所展示的相同。可以看到，在第 4 次更新时，5 个中心的位置已经几乎不发生移动，算法收敛。

Iteration History[a]

Iteration	Change in Cluster Centers				
	1	2	3	4	5
1	.155	.095	.163	.080	.110
2	.000	.034	.022	.000	.000
3	.000	.018	.011	.014	.000
4	.000	.000	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 4. The minimum distance between initial centers is .390.

接下来，SPSS 报告了 5 个聚类中心的最终位置，此时即可根据这些中心的位置，确定每个样本点的类别。

Final Cluster Centers

	Cluster				
	1	2	3	4	5
经过率	.7321	.1254	.0702	.3895	.3968
停留活动率	.1642	.2548	.8193	.1351	.6719

下面报告的是方差分析表，用于确定各个原始变量在不同类别上是否存在显著差异。从表中可以看到，经过率和停留活动率均存在显著差异。但是，下方的注释提示，该结果仅用于描述性目的，因为聚类分析的过程本身就是使类间差异最大化，因此这里的方差分析不能作为统计推断的依据。

ANOVA

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
经过率	.855	4	.006	82	147.837	.000
停留活动率	1.730	4	.012	82	139.137	.000

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

在结果窗口的最后，SPSS 报告了各类别的样本量。

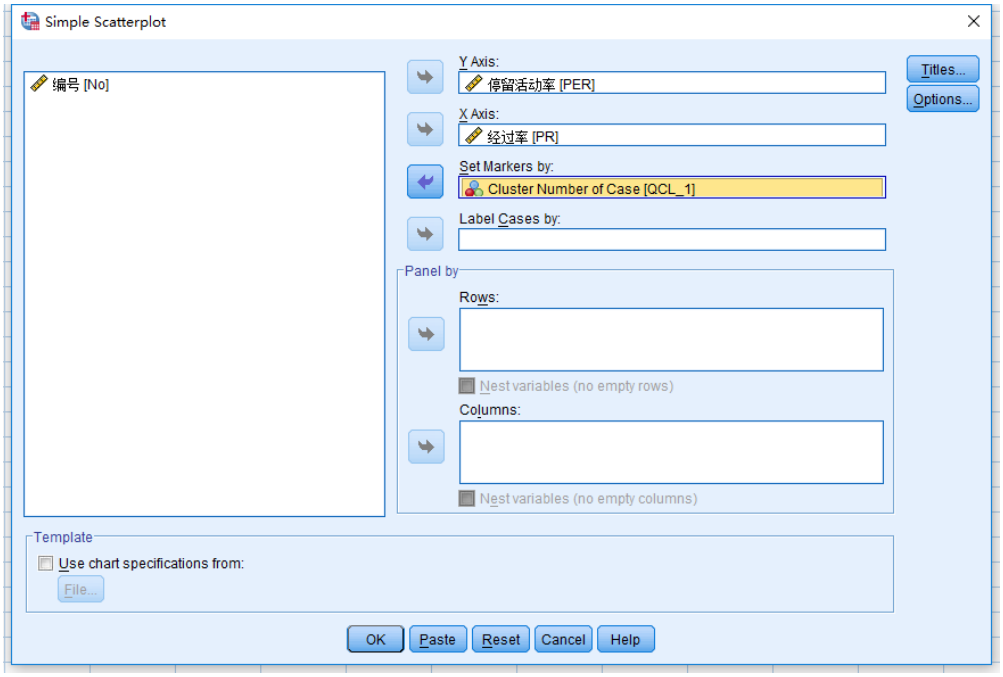
Number of Cases in each Cluster

Cluster	1	7.000
	2	19.000
	3	24.000
	4	14.000
	5	23.000
Valid		87.000
Missing		.000

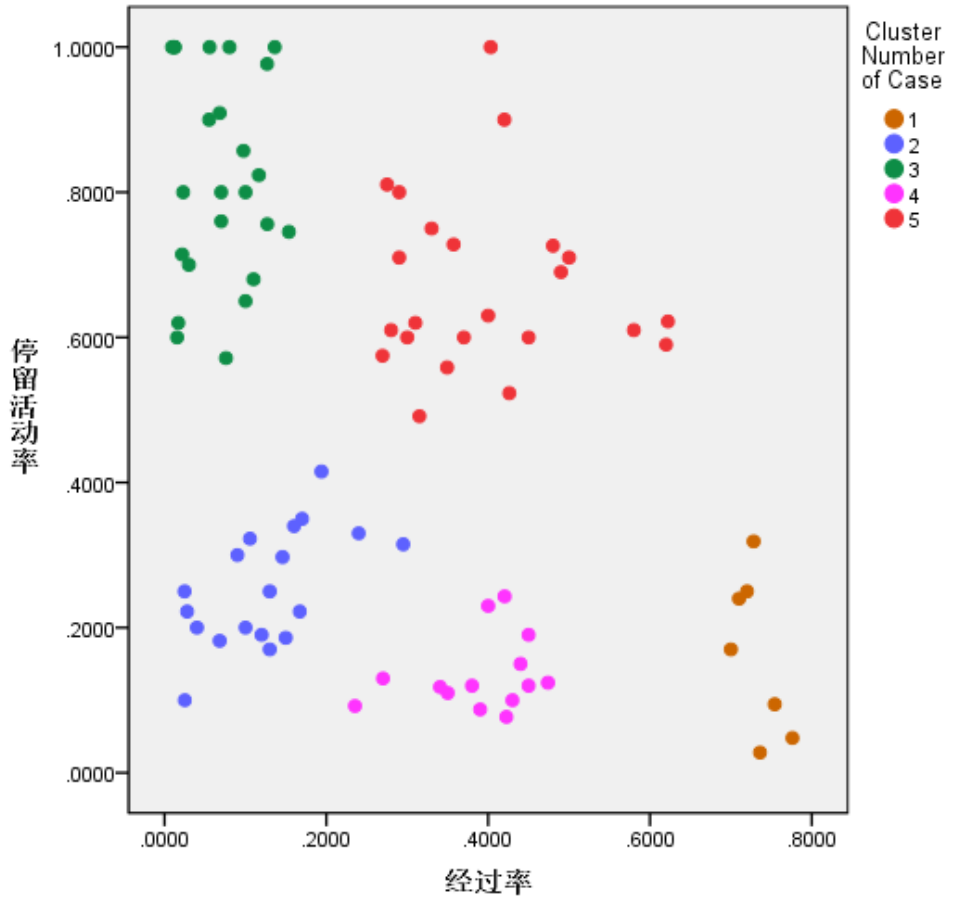
回到数据窗口，我们可以看到 SPSS 已经将分类变量保存到最后一列。

No	PR	PER	QCL_1
1	.7761	.0479	1
2	.7361	.0279	1
3	.7000	.1700	1
4	.4203	.2431	4
5	.4000	.2300	4
6	.4500	.1900	4
7	.4739	.1241	4
8	.4500	.1200	4
9	.4400	.1500	4

我们可以继续使用前面介绍的方法，通过图表分析各类别的特征。另外，由于本案例只有 2 个变量，我们也可以绘制散点图。在菜单栏中选择“Graph – Legacy Dialogs – Scatter/Dot”，确认当前选中的是“Simple Scatter”后点击“Define”，进入散点图定义对话框。将“停留活动率 [PER]”选入“Y Axis”，将“经过率 [PR]”选入“X Axis”，将“Cluster Number of Case [QCL_1]”选入“Set Marker by”。



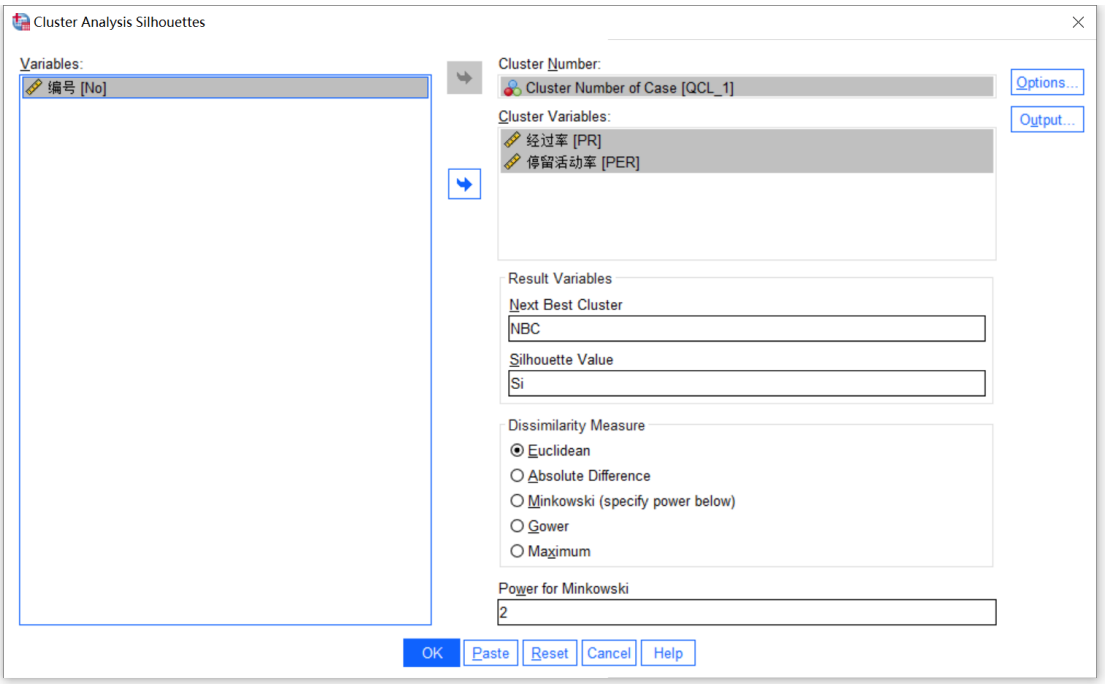
点击“OK”，则 SPSS 的结果窗口将显示“经过率—停留活动率”散点图，且以不同颜色表示 5 种类别。在 SPSS 的图形编辑窗口中对该图的颜色进行微调，可以得到课件中所示的散点图效果。



4. 聚类评价指标——Silhouette 值

本小节延续第 3 节中商业空间客流数据的 K-means 聚类分析，通过计算聚类 Silhouette 值（轮廓值），评价聚类质量，为类别数量的确定提供参考依据。

在菜单栏中选择“Analyze – Classify – Cluster Silhouettes”，打开聚类轮廓分析对话框。在“Cluster Number”中选入“Cluster Number of Case [QCL_1]”， 即刚刚由 K-means 分析所生成的分类变量（指示每个样本属于哪一类），在“Cluster Variables”中选入用于聚类分析的所有变量，这里为“经过率 [PR]”和“停留活动率 [PER]”。然后，在“Result Variables”中指定要保存的变量名称：在“Next Best Cluster”中输入“NBC”，用以保存每个样本次优的所属类别；在“Silhouette Value”中输入“Si”，用以保存每个样本的 Silhouette 值，这两个变量名是任意设定的。计算 Silhouette 值需要测度距离，在“Dissimilarity Measure”中，我们使用默认的“Euclidean”，即欧氏距离，所有设定如下图所示。



点击“OK”运行分析，首先回到数据窗口，我们会发现 SPSS 新生成了两列变量，其中，“NBC”列指示了每个样本在最有可能的分类之外，次优的分类类别；“Si”列则是每个样本的 Silhouette 值。

再查看结果窗口中，首先报告的是 Silhouette 值的统计指标，如下图所示。指标包括第 1—5 类每一类样本以及全体样本 Silhouette 值的平均值、最大值、最小值。其中，全体样本的平均 Silhouette 值为 0.552。

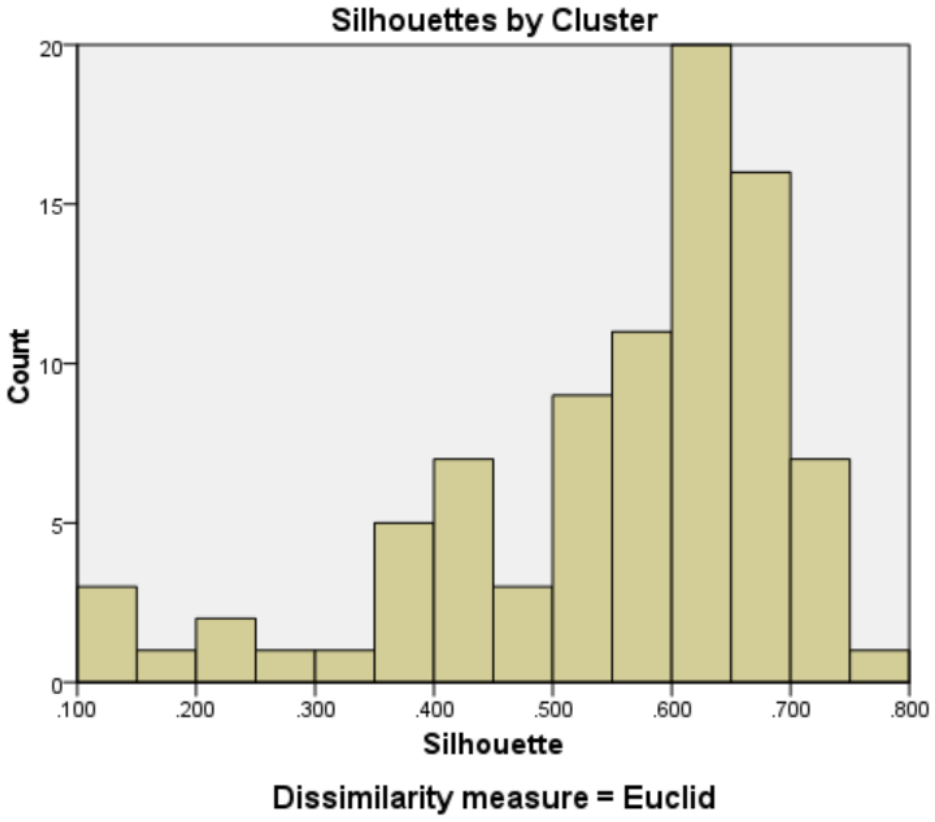
Silhouette Statistics

Statistics

Cluster	Case Count	Mean	Minimum	Maximum
1	7.000	.667	.597	.711
2	19.000	.577	.110	.690
3	24.000	.545	.230	.657
4	14.000	.637	.234	.753
5	23.000	.453	.101	.643
Total	87.000	.552	.101	.753

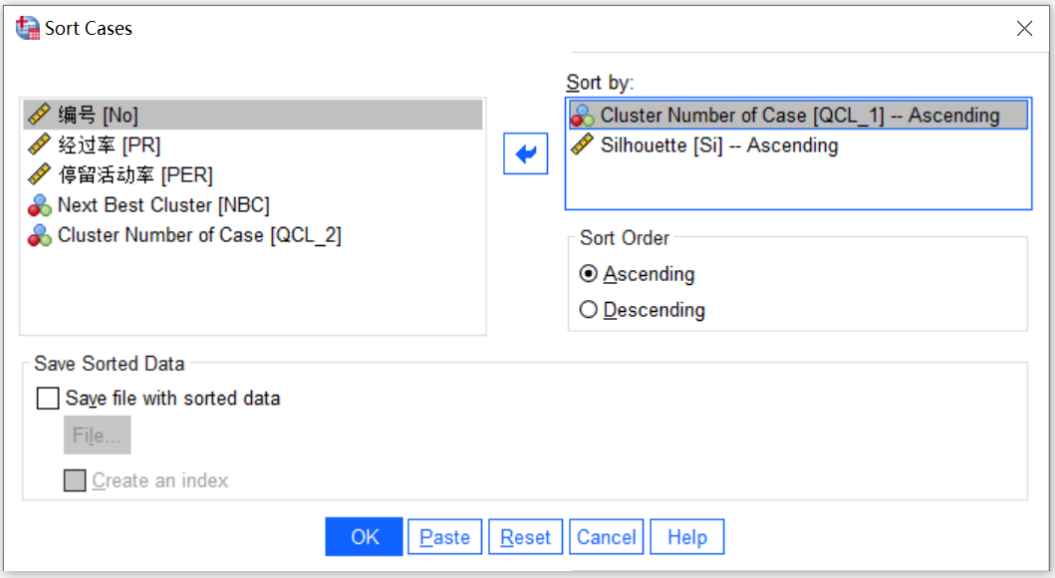
Dissimilarity measure = Euclid

接下来，SPSS 以直方图的形式呈现了各样本 Silhouette 值的分布情况，如下图所示。

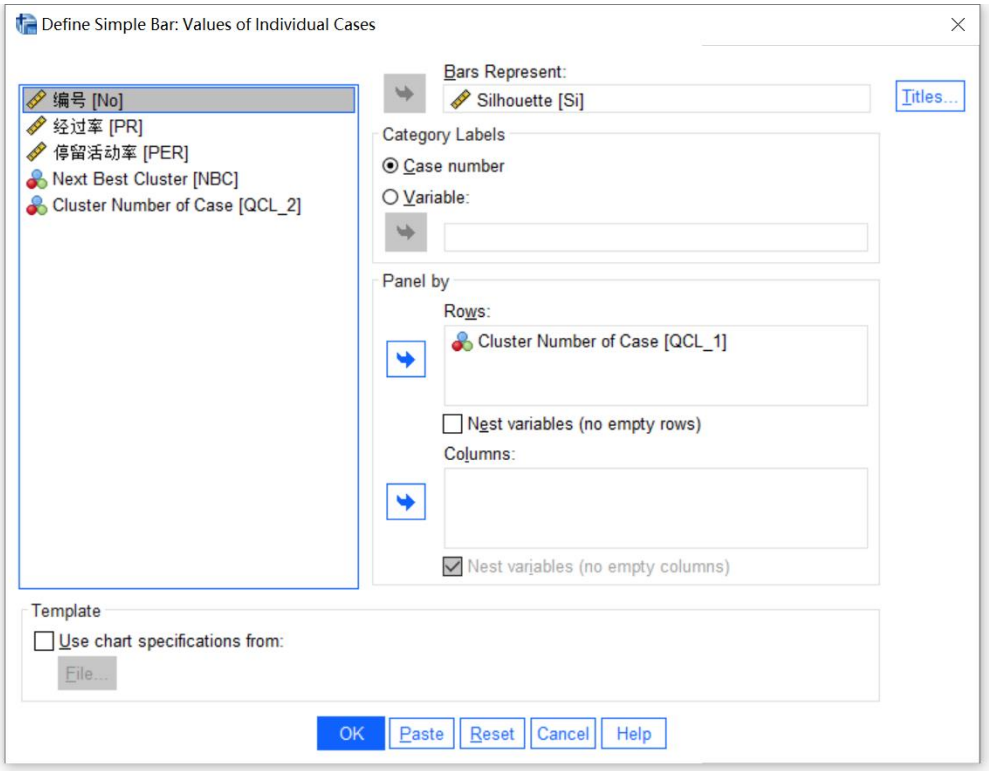


除了这种方式以外，我们还可以按不同的组别具体呈现每一个样本的 Silhouette 值。为此，我们先对数据进行一些排序处理。在菜单栏中依次选择“Data — Sort Cases”，在弹出的样本排序对话框中，首先把“Cluster Number of Case [QCL_1]”选入，并在“Sort Order”中选择“Ascending”，然后，再把“Silhouette

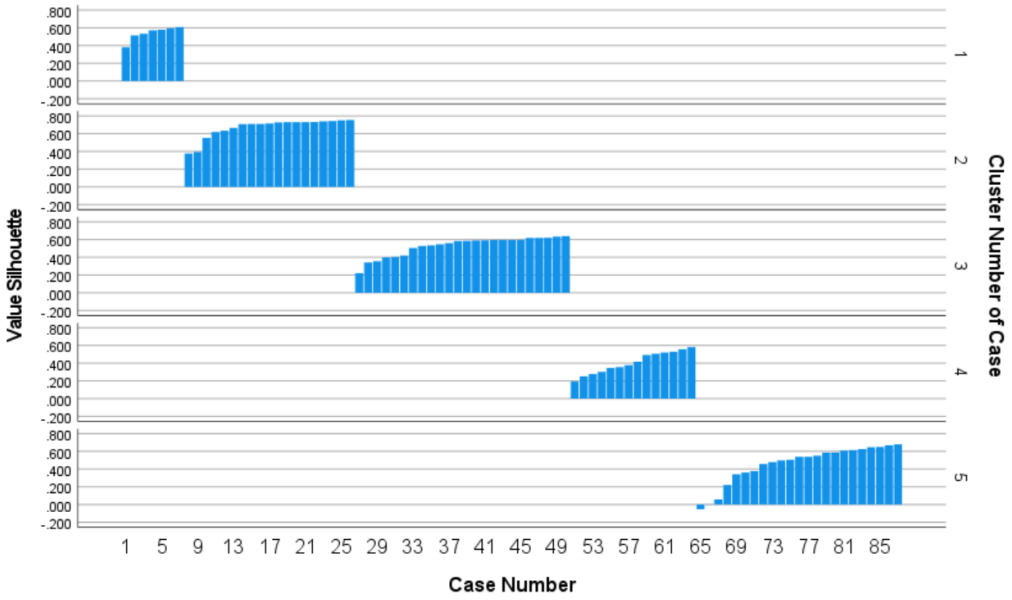
[Si]”选入，并在“Sort Order”中依旧选择“Ascending”， 如下图所示。点击“OK”后，SPSS 将对数据按类别（第一顺序）和 Silhouette（第二顺序）升序排序。



然后，在菜单栏中依次选择“Graphs — Legacy Dialogs — Bar”，在弹出的对话框中选择“Simple”，在“Data in Chat Are”中选择“Values of individual cases”，点击“Define”，进入条形图定义对话框。在“Bars Represent”中选入“Silhouette [Si]”，在“Panel by”的“Rows：”中选入“Cluster Number of Case [QCL_1]”，如下图所示。

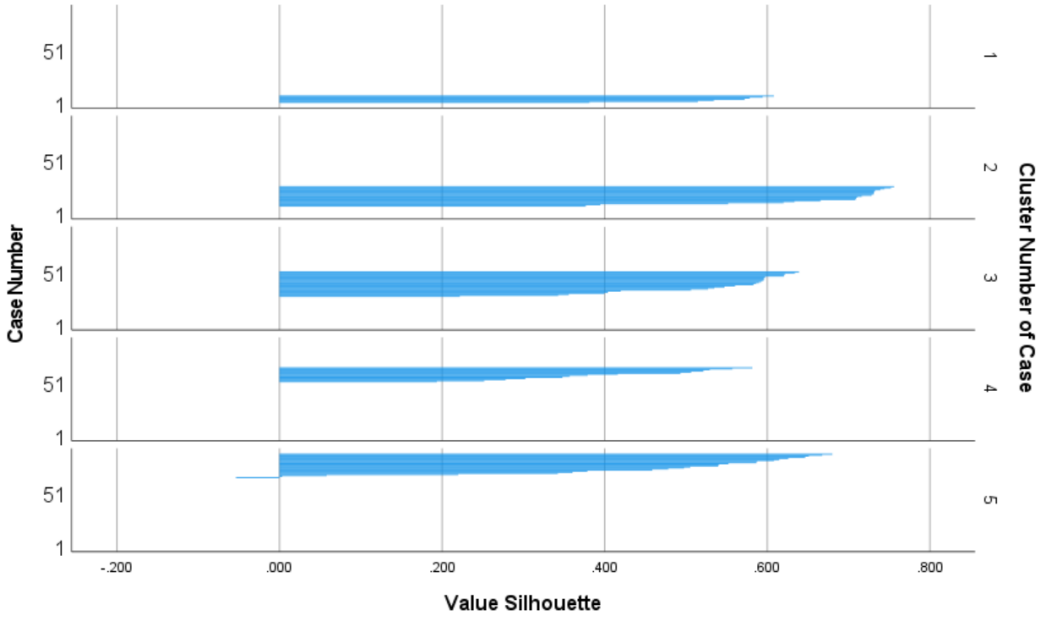


点击“OK”运行分析，结果窗口中绘制出如下条形图。



为了获得更好的图面效果，在条形图上双击，打开图形编辑器，在工具栏中点击“Transpose chart coordinate system”，把竖向条形图转换为横向条形图，如下图所示。

SPSS 最终生成如下所示的 Silhouette 分组分布图，其中，右侧纵轴为组别。



如果需要以 Silhouette 值为参照，确定最佳类别数量，我们可以在 K-means 分析中依次设定不同的类别数量，保存相应的分类变量，然后再依次进行 Silhouette 分析，得到每一种聚类结果的平均 Silhouette 值，指出 Silhouette 均值最高的聚类方案即可，具体过程不再重复演示。