Chapter 3
Hierarchical Clustering

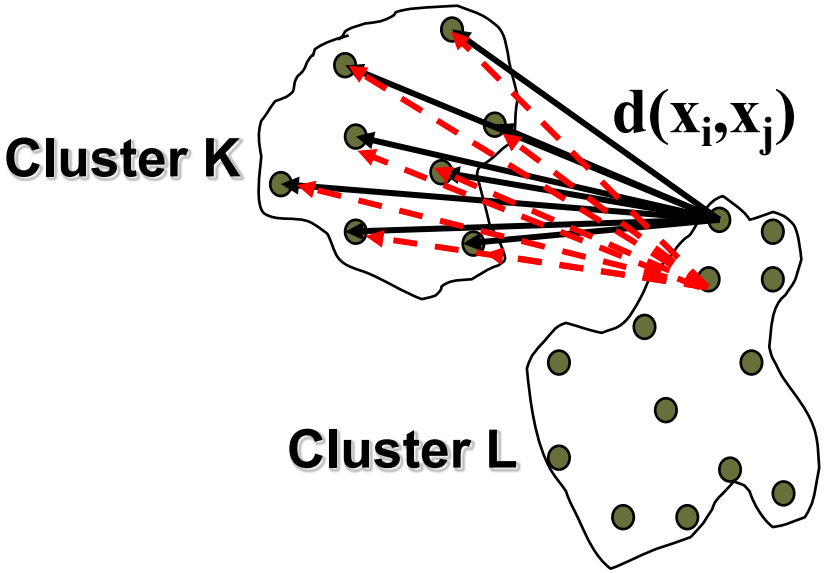
Hierarchical clustering methods

	Method	Method name (SAS)	Coordinate Data	Distance Data
1	**Average Linkage**	**AVERAGE**	√	√
2	Centroid Linkage	CENTROID	√	√
3	Complete Linkage	COMPLETE		√
4	Density Linkage	DENSITY		√
5	Equal variance maximum likelihood	EML	√	
6	Flexible-Beta	FLEXIBLE		√
7	McQuitty's	MCQUITTY		√
8	Median	MEDIAN		√
9	Single-Linkage	SINGLE		√
10	Two-Stage Linkage	TWOSTAGE		√
11	**Ward's**	**WARD**	√	√

(Different ways of computing cluster distance)

Average linkage (Sokal and Michener, 1958)

The distance between clusters is the average distance between pairs of observations.



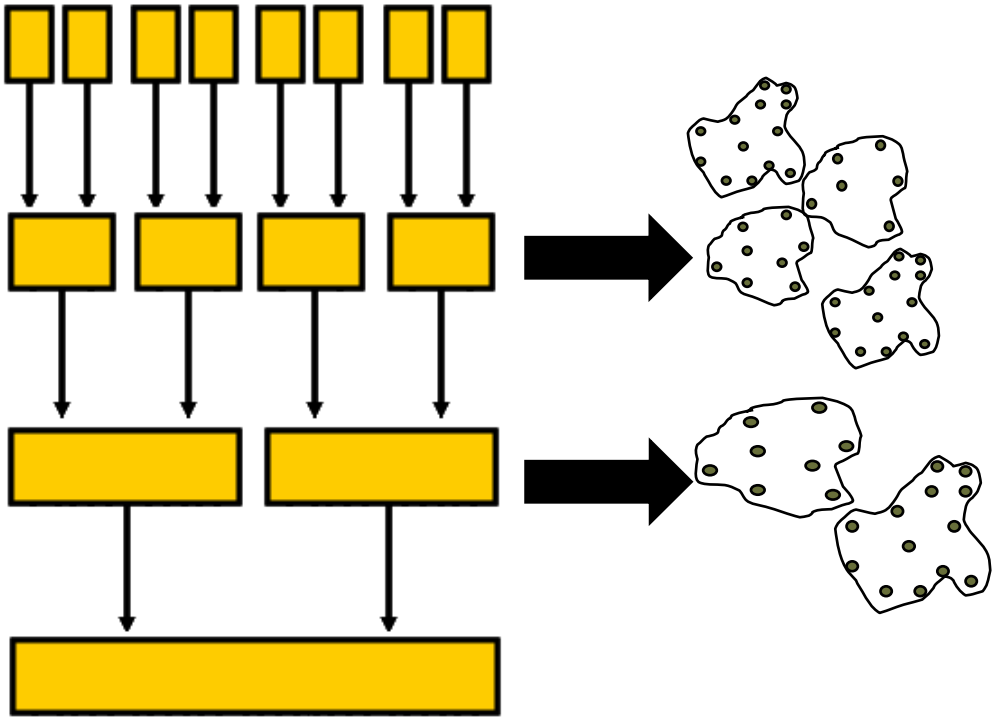
$$D_{KL} = \frac{1}{n_K n_L} \sum_{i \in C_K} \sum_{j \in C_L} d(x_i, x_j)$$

n_K, n_L: numbers of observations in clusters C_K and C_L

$d(x_i, x_j)$: distance between observations i and j

Ward's (minimum-variance)

A fusion results in minimum loss of information.



$$D_{KL} = \frac{\left\| \bar{x}_K - \bar{x}_L \right\|^2}{\left(\frac{1}{n_K} + \frac{1}{n_L} \right)}$$

$\bar{\mathbf{x}}_K$, $\bar{\mathbf{x}}_L$: mean vectors of cluster C_K and C_L

n_K, n_L: number of observations in clusters C_K and C_L

What is the best method?

Is there a way to determine this?

Comparing hierarchical clustering methods

Ordered by Number of Misclassifications

Obs	method	misclassified	chisq	pchisq	cramers
1	TWOSTAGE	8	255.840	3.5923E-54	0.92347
2	AVERAGE	17	223.881	2.7403E-47	0.86387
3	WARD	17	223.881	2.7403E-47	0.86387
4	COMPLETE	18	218.449	4.0424E-46	0.85333
5	EML	18	218.449	4.0424E-46	0.85333
6	FLEXIBLE	18	218.449	4.0424E-46	0.85333
7	MCQUITTY	18	218.449	4.0424E-46	0.85333
8	MEDIAN	18	218.449	4.0424E-46	0.85333
9	SINGLE	20	149.390	2.7504E-31	0.73839
10	CENTROID	21	212.857	6.4533E-45	0.84515
11	DENSITY	40	210.536	2.0373E-44	0.9694

Chapter 4 Assessing Results: Determining the Number of Clusters

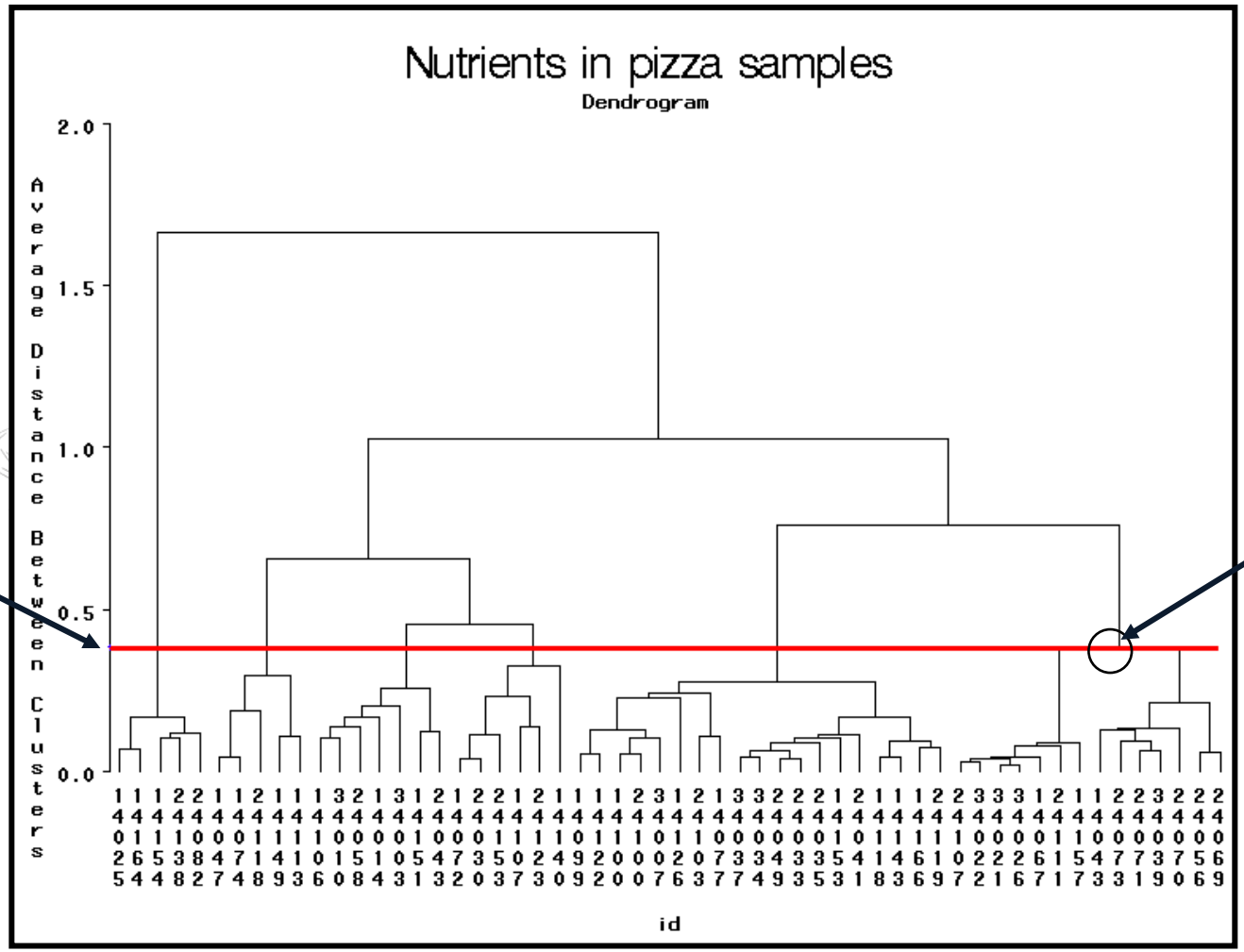
Objectives

To answer the question "How many clusters?" using the following methods:

- Dendrograms
- Cubic clustering criterion (CCC)
- Pseudo *F* statistic (PSF)
- Pseudo T^2 test (PST2)

(With some assumptions that never/hard to hold)

Interpreting Dendrograms



Fusion Level

Fusion Point

Business (domain) knowledge is useful.

Cubic Clustering Criterion (CCC)

- CCC >2: good clustering
- 0 ≤ CCC ≤ 2: possible cluster structure
- Large negative CCC values: possible outliers

Important Limitations of CCC

1. CCC is computed under the assumption that the sample size is at least 20, and that the variables are uncorrelated.
2. CCC is not appropriate for highly elongated or irregularly shaped clusters.
3. CCC is only calculated for coordinate data.
4. CCC is appropriate for clustering techniques in which the within-group sum of squares is minimized, for example, Ward's method (PROC CLUSTER).

CCC: No. of Clusters

Cluster History

NCL	Clusters Joined		FREQ	SPRSQ	RSQ	ERSQ	CCC	Tie
10	CL12	CL44	25	0.0099	.911	.891	2.84	
9	CL11	CL15	45	0.0129	.898	.882	2.16	
8	CL18	CL27	18	0.0130	.885	.870	1.86	
7	CL17	CL22	29	0.0137	.872	.854	1.87	
6	CL8	CL20	26	0.0149	.857	.834	2.24	
5	CL10	CL16	30	0.0150	.842	.806	3.24	
4	CL19	CL7	49	0.0364	.805	.764	3.28	
3	CL9					.694	4.00	
2	CL3	CL5	101	0.1331	.619	.552	2.86	
1	CL4	CL2	150	0.6188	.000	.000	0.00	

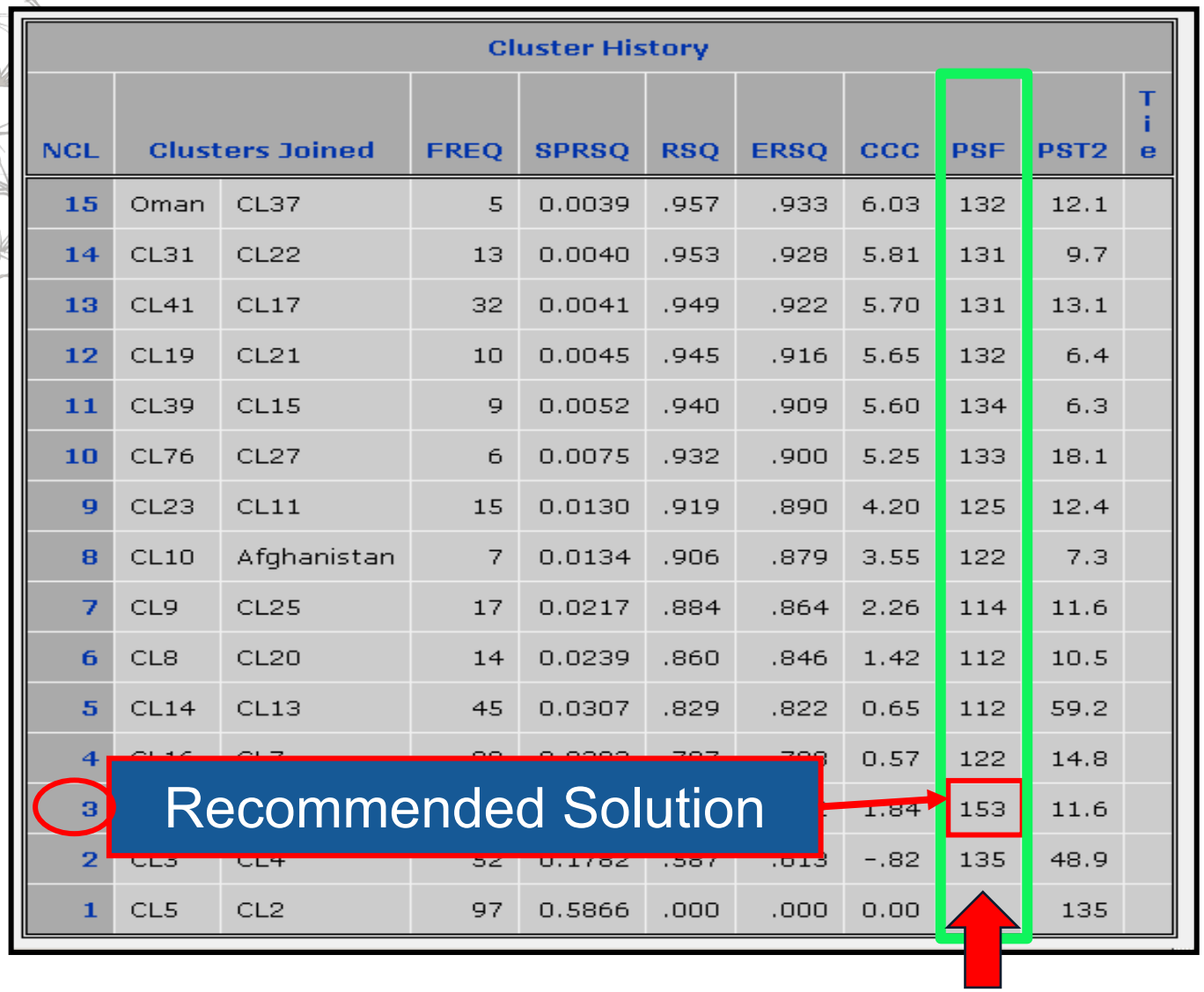
Recommended Solution

From the bottom look for the 1st **peak** of CCC (normally>2)

The Pseudo *F* Statistic (PSF)

- PSF measures the separation among the clusters.
- It is **not** distributed as an *F* random variable.
- The maximum PSF value indicates a strong cluster solution.
- The "pseudo" option must be specified in the PROC CLUSTER statement and only when AVERAGE, CENTRIOD, or WARD method is used.

PSF: No. of Clusters



Cluster History

NCL	Clusters Joined		FREQ	SPRSQ	RSQ	ERSQ	CCC	PSF	PST2	Tie
15	Oman	CL37	5	0.0039	.957	.933	6.03	132	12.1	
14	CL31	CL22	13	0.0040	.953	.928	5.81	131	9.7	
13	CL41	CL17	32	0.0041	.949	.922	5.70	131	13.1	
12	CL19	CL21	10	0.0045	.945	.916	5.65	132	6.4	
11	CL39	CL15	9	0.0052	.940	.909	5.60	134	6.3	
10	CL76	CL27	6	0.0075	.932	.900	5.25	133	18.1	
9	CL23	CL11	15	0.0130	.919	.890	4.20	125	12.4	
8	CL10	Afghanistan	7	0.0134	.906	.879	3.55	122	7.3	
7	CL9	CL25	17	0.0217	.884	.864	2.26	114	11.6	
6	CL8	CL20	14	0.0239	.860	.846	1.42	112	10.5	
5	CL14	CL13	45	0.0307	.829	.822	0.65	112	59.2	
4							0.57	122	14.8	
3							1.84	153	11.6	
2							-.82	135	48.9	
1	CL5	CL2	97	0.5866	.000	.000	0.00		135	

- From the bottom look for peaks of PSF values.
- Select the peak that produces the smallest number of clusters.

The Pseudo t^2 Statistic (PST2)

The "pseudo" option must be specified in the PROC CLUSTER statement and only when AVERAGE, CENTRIOD, or WARD method is used.

PST2: No. of Clusters

Cluster History

NCL	Clusters Joined		FREQ	SPRSQ	RSQ	ERSQ	CCC	PSF	PST2	Tie
15	Oman	CL37	5	0.0039	.957	.933	6.03	132	12.1	
14	CL31	CL22	13	0.0040	.953	.928	5.81	131	9.7	
13	CL41	CL17	32	0.0041	.949	.922	5.70	131	13.1	
12	CL19	CL21	10	0.0045	.945	.916	5.65	132	6.4	
11	CL39	CL15	9	0.0052	.940	.909	5.60	134	6.3	
10	CL76	CL27	6	0.0075	.932	.900	5.25	133	18.1	
9	CL23	CL11	15	0.0138	.918	.888	4.89	125	12.4	
8	CL10	Afg						122	7.3	
7	CL9	CL25	17	0.0217	.884	.864	2.26	114	11.6	
6	CL8	CL20	14	0.0239	.860	.846	1.42	112	10.5	
5	CL14	CL13	45	0.0307	.829	.822	0.65	112	59.2	
4	CL16	CL7	28	0.0323	.797	.788	0.57	122	14.8	
3	CL12	CL6	24	0.0323	.765	.732	1.84	153	11.6	
2	CL3	CL4	52	0.1782	.587	.613	-.82	135	48.9	
1	CL5	CL2	97	0.5866	.000	.000	0.00	.	135	

Potential Solutions

- Move up from the bottom and look for significant **drops** (see the values in the red boxes) of the PST2 values.
- Choose the one with smallest No. of clusters.

Determining the Optimal Number of Clusters

```
proc format;
   value specfmt
      1='Bream'
      2='Roach'
      3='Whitefish'
      4='Parkki'
      5='Perch'
      6='Pike'
      7='Smelt';
run;

%let inputs=newlength1
logLengthRatio height width
     weight3;
%let group=species;
proc stdize data=teaching.fish
method=range out=temp1;
   var &inputs;
run;
```

```
title 'Ward''s Method';
proc cluster data=temp1
method=ward ccc pseudo
     outtree=tree;
   var &inputs;
   copy &inputs &group;
run;

title 'Dendrogram';
proc tree data=tree out=treeout
level=.0098;
   copy &inputs &group;

run;
title 'Confusion Matrix';
proc freq data=treeout;
   format species specfmt.;
   tables &group*cluster / norow
    nocol nopercent chisq;
run;
```

The CLUSTER Procedure
Ward's Minimum Variance Cluster Analysis

Cluster History

NCL	--Clusters Joined---		FREQ	SPRSQ	RSQ	ERSQ	CCC	PSF	PST2	Tie
36	CL54	CL65	12	0.0010	.968	.	.	104	4.7	
35	CL45	CL73	16	0.0011	.967	.	.	104	6.1	
34	CL51	CL115	8	0.0011	.966	.	.	104	4.9	
33	OB71	OB89	2	0.0012	.964	.	.	105	.	
32	OB53	CL74	5	0.0012	.963	.	.	105	6.9	
31	CL44	OB95	4	0.0013	.962	.908	17.6	106	2.5	
30	CL57	OB156	3	0.0013	.961	.905	17.5	107	2.6	
29	CL37	CL68	11	0.0014	.959	.902	17.5	107	5.9	
28	OB35	CL42	8	0.0015	.958	.899	17.4	108	5.1	
27	CL55	CL59	11	0.0016	.956	.896	17.3	109	7.8	
26	CL64	CL39	12	0.0017	.954	.893	17.3	109	5.2	
25	CL41	CL53	10	0.0018	.952	.889	17.2	110	4.8	
24	CL35	CL85	19	0.0019	.951	.886	17.2	111	8.1	
23	CL46	CL47	9	0.0021	.949	.882	17.1	112	7.2	
22	CL43	CL60	11	0.0021	.946	.878	17.1	114	9.1	
21	CL36	CL58	15	0.0024	.944	.874	17.0	115	8.3	
20	CL52	CL24	25	0.0024	.942	.869	16.9	116	7.5	
19	CL32	CL61	11	0.0026	.939	.864	16.9	118	8.9	
18	CL28	CL34	16	0.0033	.936	.859	16.7	119	8.0	
17	CL40	CL26	14	0.0033	.932	.853	14.4	121	6.9	
16	CL62	CL154	5	0.0034	.929	.847	14.3	123	16.1	
15	CL25	CL95	12	0.0041	.925	.841	14.2	125	8.2	
14	CL20	CL31	29	0.0042	.921	.833	14.2	128	9.3	
13	CL29	CL38	19	0.0052	.916	.825	14.1	130	16.2	
12	CL13	CL21	34	0.0064	.909	.816	13.8	132	12.3	
11	CL14	CL33	31	0.0064	.903	.806	13.7	135	10.7	
10	CL17	CL23	23	0.0068	.896	.794	13.8	141	10.7	
9	CL22	CL30	14	0.0085	.887	.780	13.8	146	16.9	
8	CL11	CL19	42	0.0088	.879	.764	14.1	154	12.1	
7	CL15	CL16	17	0.0098	.869	.744	14.7	166	11.2	
6	CL18	CL8	58	0.0227	.846	.718	13.8	166	27.1	
5	CL12	CL27	45	0.0302	.816	.686	12.9	168	48.6	
4	CL7	CL9	31	0.0627	.753	.640	9.86	156	48.6	
3	CL6	CL10	81	0.0744	.679	.570	7.20	163	65.6	
2	CL3	CL4	112	0.2307	.448	.444	0.16	126	96.1	
1	CL5	CL2	157	0.4482	.000	.000	0.00	.	126	

Confusion Matrix

The FREQ Procedure

Table of species by CLUSTER

species	CLUSTER							
Frequency	1	2	3	4	5	6	7	Total
Bream	0	0	0	0	34	0	0	34
Roach	6	0	0	0	0	13	0	19
Whitefish	0	0	0	0	0	3	3	6
Parkki	0	0	0	11	0	0	0	11
Perch	36	0	0	0	0	0	20	56
Pike	0	17	0	0	0	0	0	17
Smelt	0	0	14	0	0	0	0	14
Total	42	17	14	11	34	16	23	157

Statistics for Table of species by CLUSTER

Statistic	DF	Value	Prob
Chi-Square	36	725.5882	<.0001
Likelihood Ratio Chi-Square	36	472.9589	<.0001
Mantel-Haenszel Chi-Square	1	28.1276	<.0001
Phi Coefficient		2.1498	
Contingency Coefficient		0.9067	
Cramer's V		0.8776	

WARNING: 84% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

Cluster Assignment and Associated Probability for all the Species

Species	*Assigned Cluster*	*Probability*
Bream	5	34/34 = 1
Roach	6	13/19 = 0.6842
Whitefish	6 or 7	3/6 = 0.5
Parkki	4	11/11 = 1
Perch	1	36/56 = 0.6429
Pike	2	17/17 = 1
Smelt	3	14/14 = 1