Tutorial PP: Clustering Fiber Length Histograms by Degree of Damage.

Guest Author: Mourad Krifa, Ph.D., Assistant Professor of Fiber Science, School of Human Ecology, University of Texas at Austin; Austin, TX.

Note to users: DATA set for this tutorial: The entire dataset illustrated above could not be provided for proprietary reasons; instead a sub-sample is provided, called: "*Fiber-Length DATA-subset_Elsevier.sta*". Please use this dataset to use in replicating the steps above; a similar, but not identical, set of results should be obtained.

Mechanical damage in fiber processing both shifts the fiber length distribution and alters its shape. This leads to length distributions with complex shapes (often bimodal) that cannot be classified based on simple summary statistics (Krifa, 2009). In such cases, the best way to compare and classify samples with varied degrees of fiber damage is to examine the empirical length histograms. The measurement instrument measures the length of individual fibers in samples ranging from 3000 to 10000 fibers (Krifa, 2006, 2008) and outputs histogram data (number of fibers detected in each of 40 length bins from 0.79 to 62.7 mm).

This tutorial illustrates the use of K-Means Clustering in STSATISTICA to classify samples based on histogram data. The dataset used in this tutorial consists of length histogram data from 356 cotton fiber samples forming a range of fiber damage. The data was organized in the spreadsheet such that each of the 356 samples is one case (row). The 40 variables (columns) correspond to the 40 length bins. The goal of the analysis is to quickly classify the samples into relatively homogenous groups of different damage levels.

I. Running K-Means Clustering:

- In the STATISTICA spreadsheet view, select *Generalized EM & k-Means Cluster Analysis* from the *Data Mining* menu (Fig. 1) to display the *Generalized Cluster Analysis* Startup Panel.

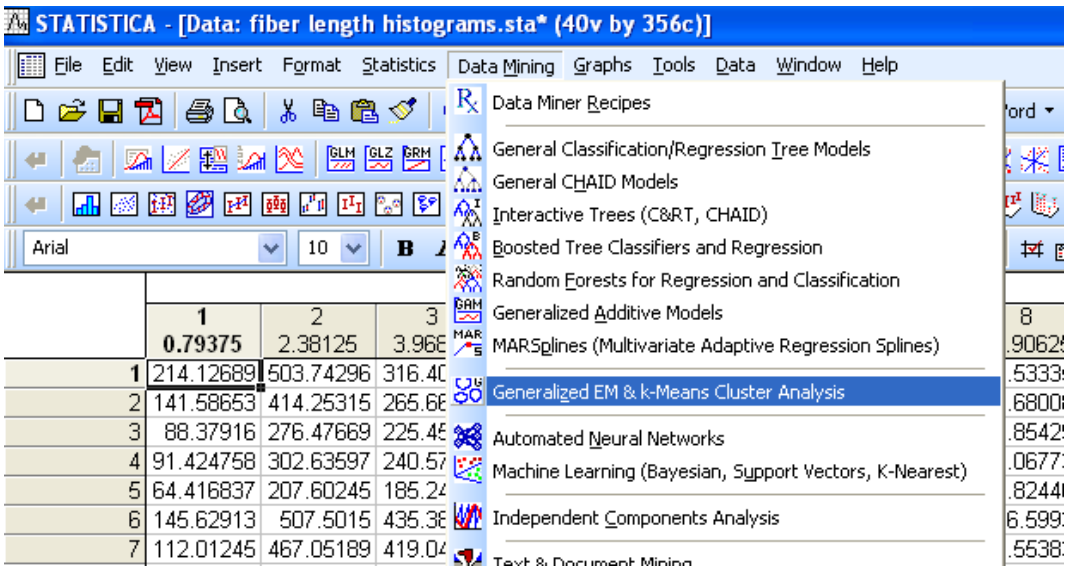


Fig. 1

- In the *Generalized Cluster Analysis* Startup Panel (Fig. 2), Click on *Variables* and select all variable for the analysis
- Click the *Validation* tab.

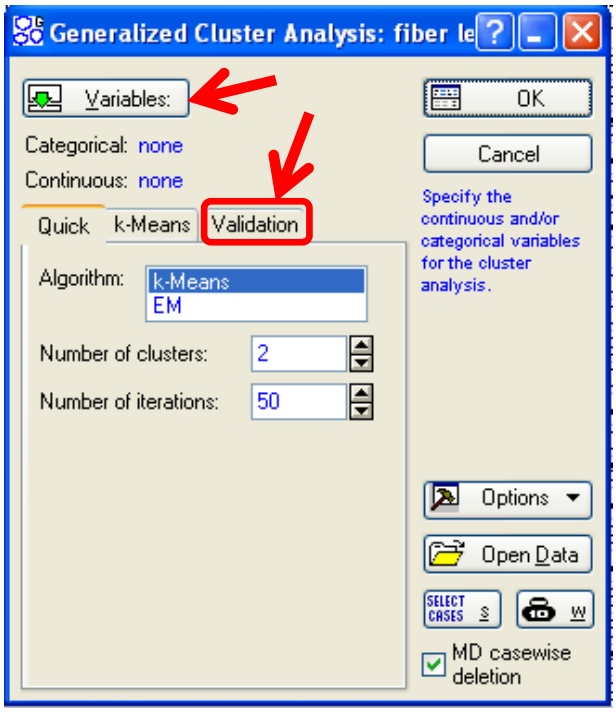


Fig. 2

- Select the *V-fold cross-validation* check box (Fig. 3) to let the algorithm determine the best number of clusters within the range specified in the *Minimum/Maximum number of cluster* fields also displayed on this tab. In this analysis, we specified a minimum number of 7 clusters based on knowledge about processing conditions. Click OK.

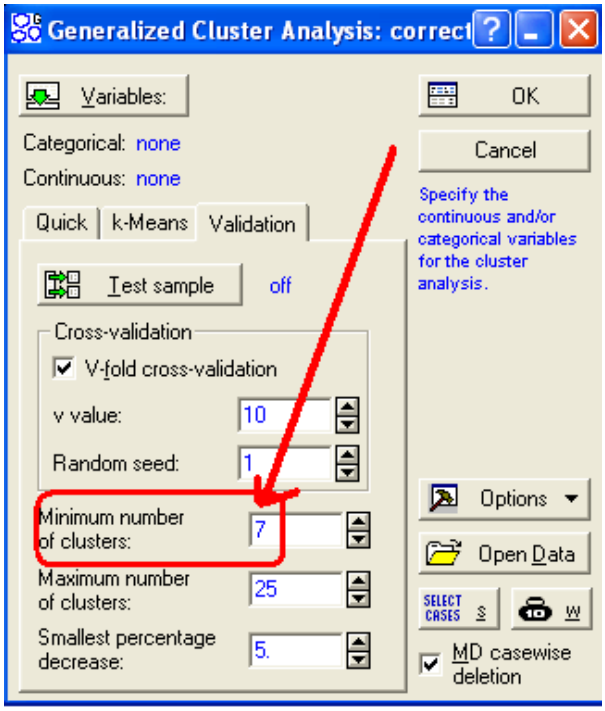


Fig. 3

II. RESULS:

The *Quick* tab of the *Generalized Cluster Analysis* results dialog (Fig. 4) gives access to multiple options (Cluster means, Cluster distances...). In this analysis we are interested in cluster memberships of the 356 observations.

- Click on Members & distances to display a spreadsheet with cluster memberships of the observations (Fig. 5). The drop-down list allows generating seperate spreadsheets for spcific clusters.

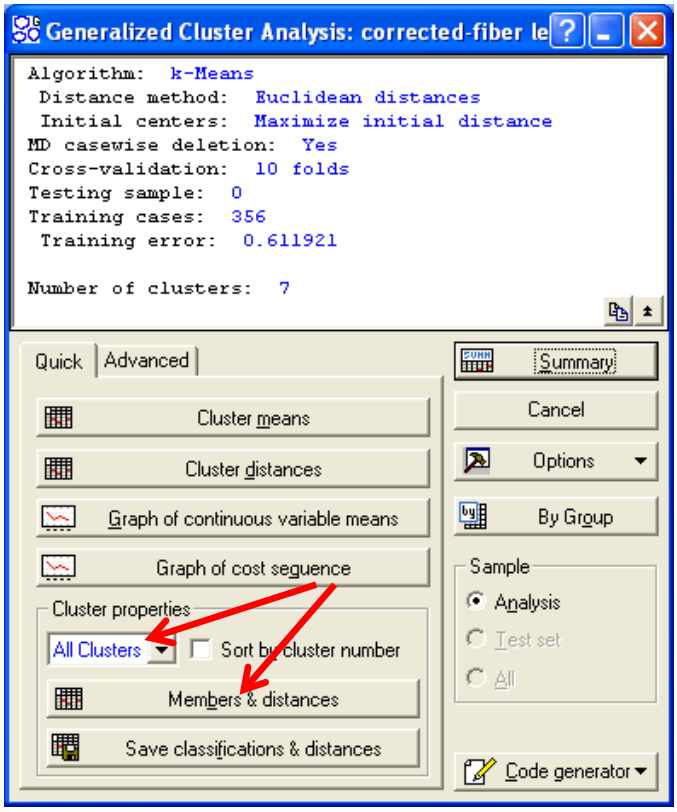
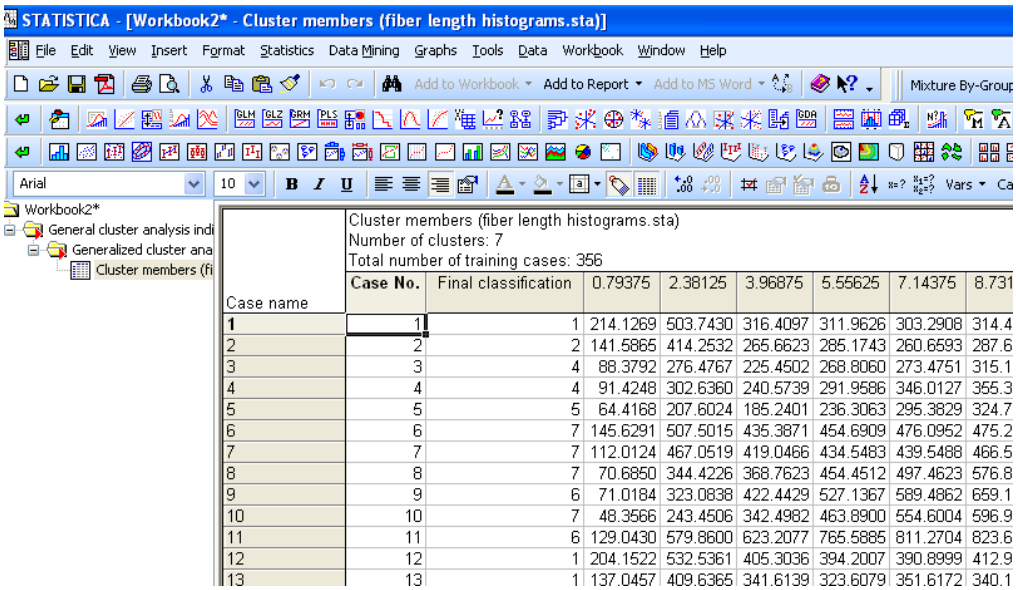


Fig. 4



Case name	Case No.	Final classification	0.79375	2.38125	3.96875	5.55625	7.14375	8.731
1	1	1	214.1269	503.7430	316.4097	311.9626	303.2908	314.4
2	2	2	141.5865	414.2532	265.6623	285.1743	260.6593	287.6
3	3	4	88.3792	276.4767	225.4502	268.8060	273.4751	315.1
4	4	4	91.4248	302.6360	240.5739	291.9586	346.0127	355.3
5	5	5	64.4168	207.6024	185.2401	236.3063	295.3829	324.7
6	6	7	145.6291	507.5015	435.3871	454.6909	476.0952	475.2
7	7	7	112.0124	467.0519	419.0466	434.5483	439.5488	466.5
8	8	7	70.6850	344.4226	368.7623	454.4512	497.4623	576.8
9	9	6	71.0184	323.0838	422.4429	527.1367	589.4862	659.1
10	10	7	48.3566	243.4506	342.4982	463.8900	554.6004	596.9
11	11	6	129.0430	579.8600	623.2077	765.5885	811.2704	823.6
12	12	1	204.1522	532.5361	405.3036	394.2007	390.8999	412.9
13	13	1	137.0457	409.6365	341.6139	323.6079	351.6172	340.1

Fig. 5

- From the SEPARATE SPREADSHEETS created for EACH of the 7 CLUSTERS, we can draw a 2-D-LINE PLOT {Case Profiles}, as shown in Fig. 6

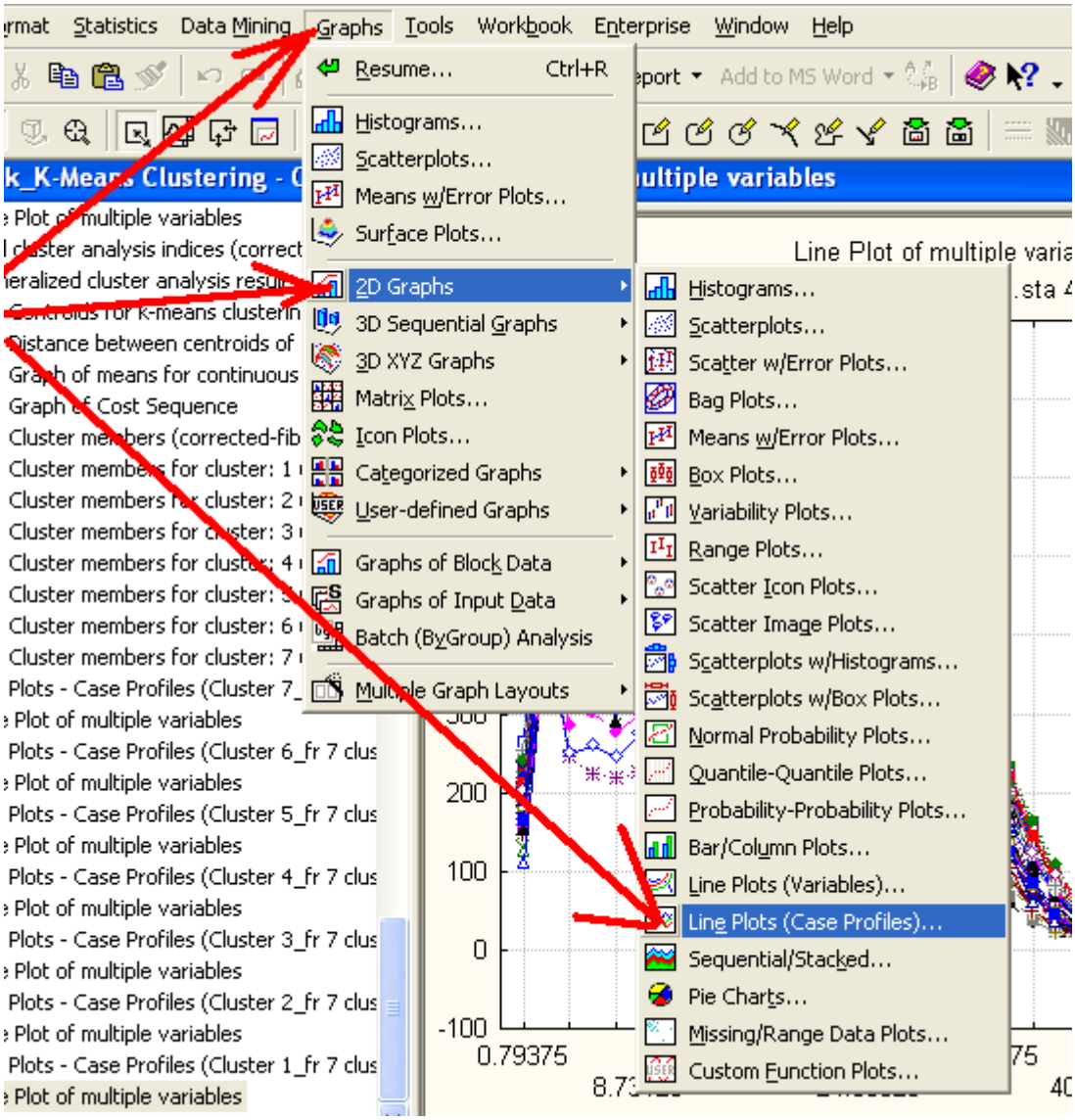
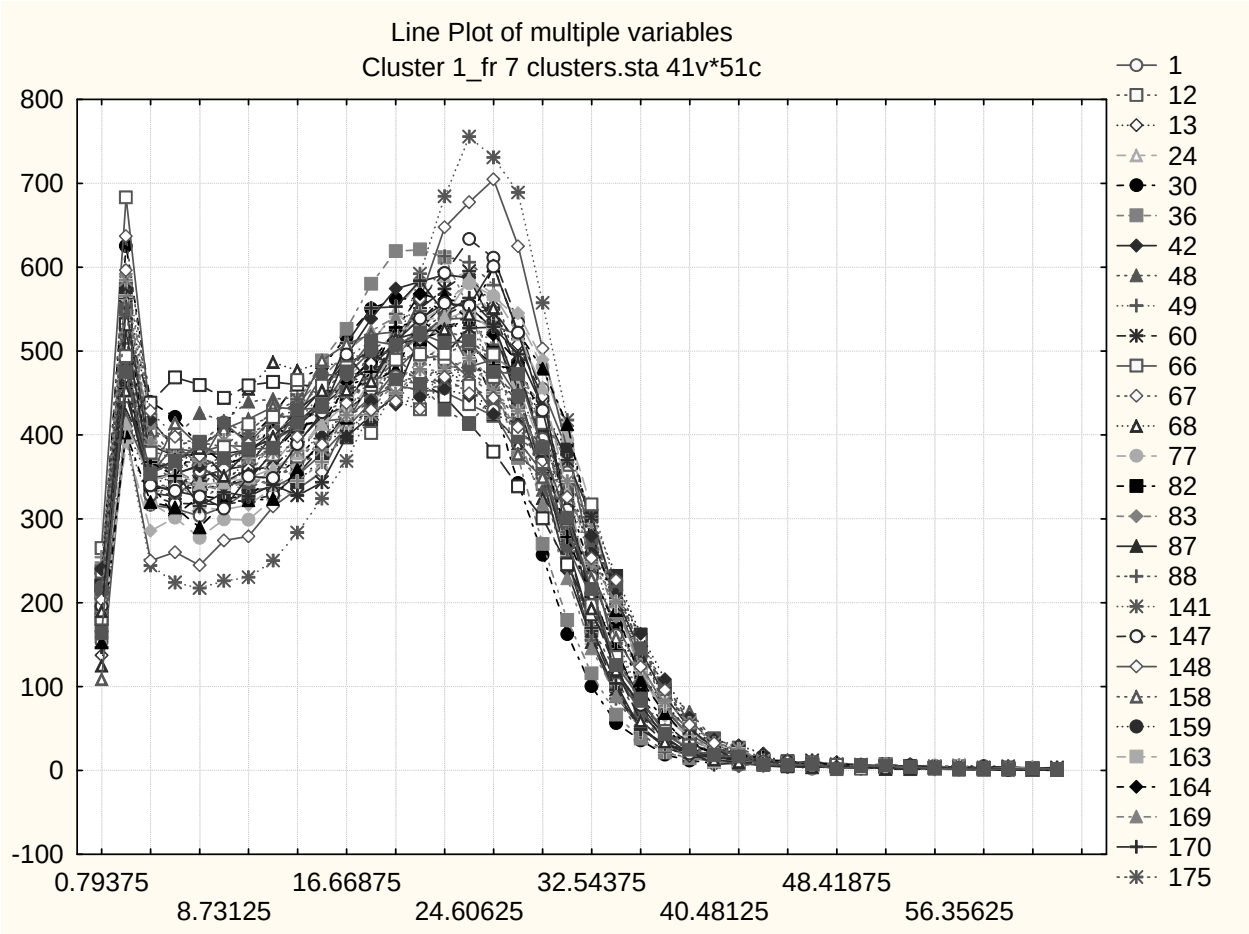


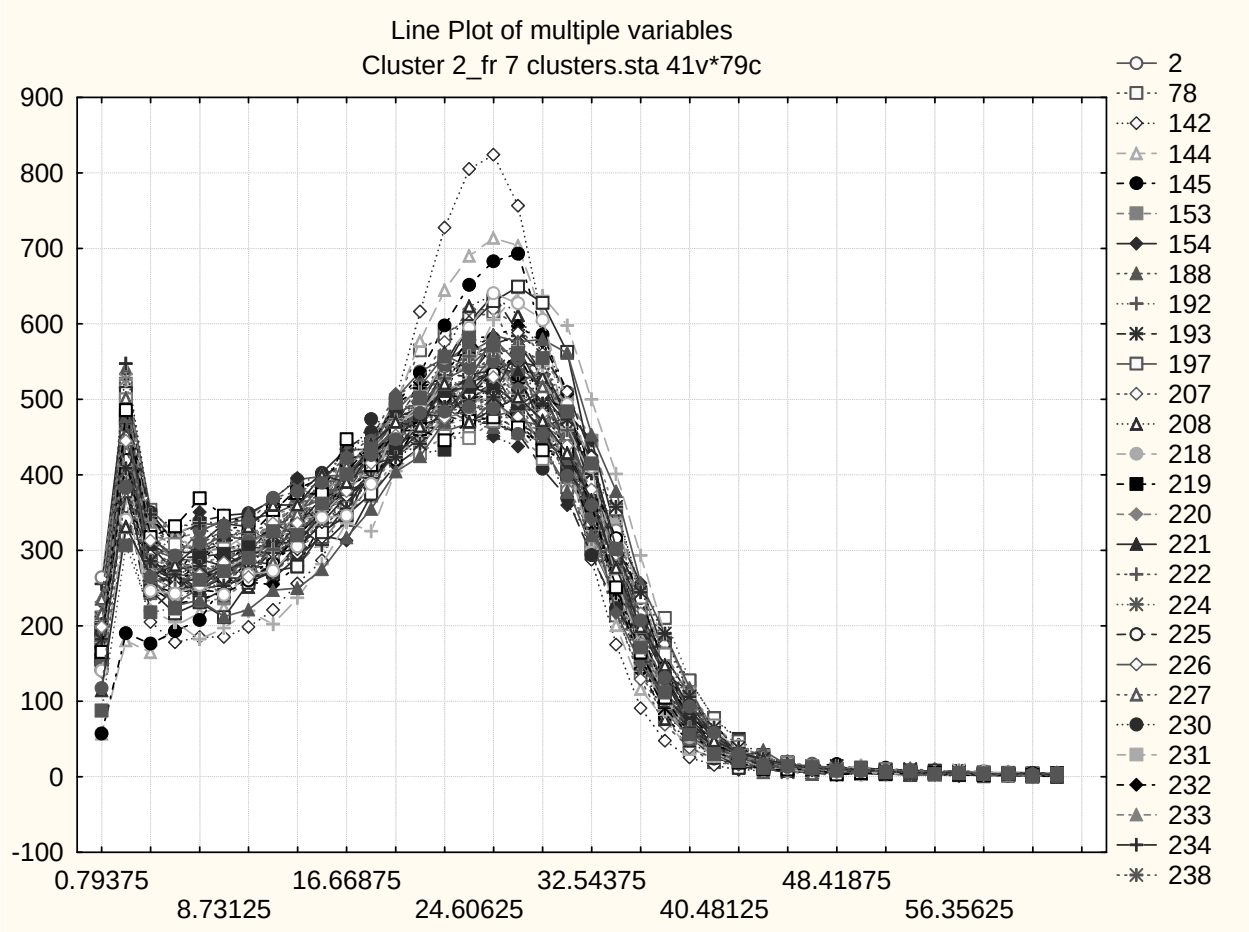
Fig. 6

The case profile plots show empirical density traces (histograms) of the length distributions classified in each of the 7 clusters:

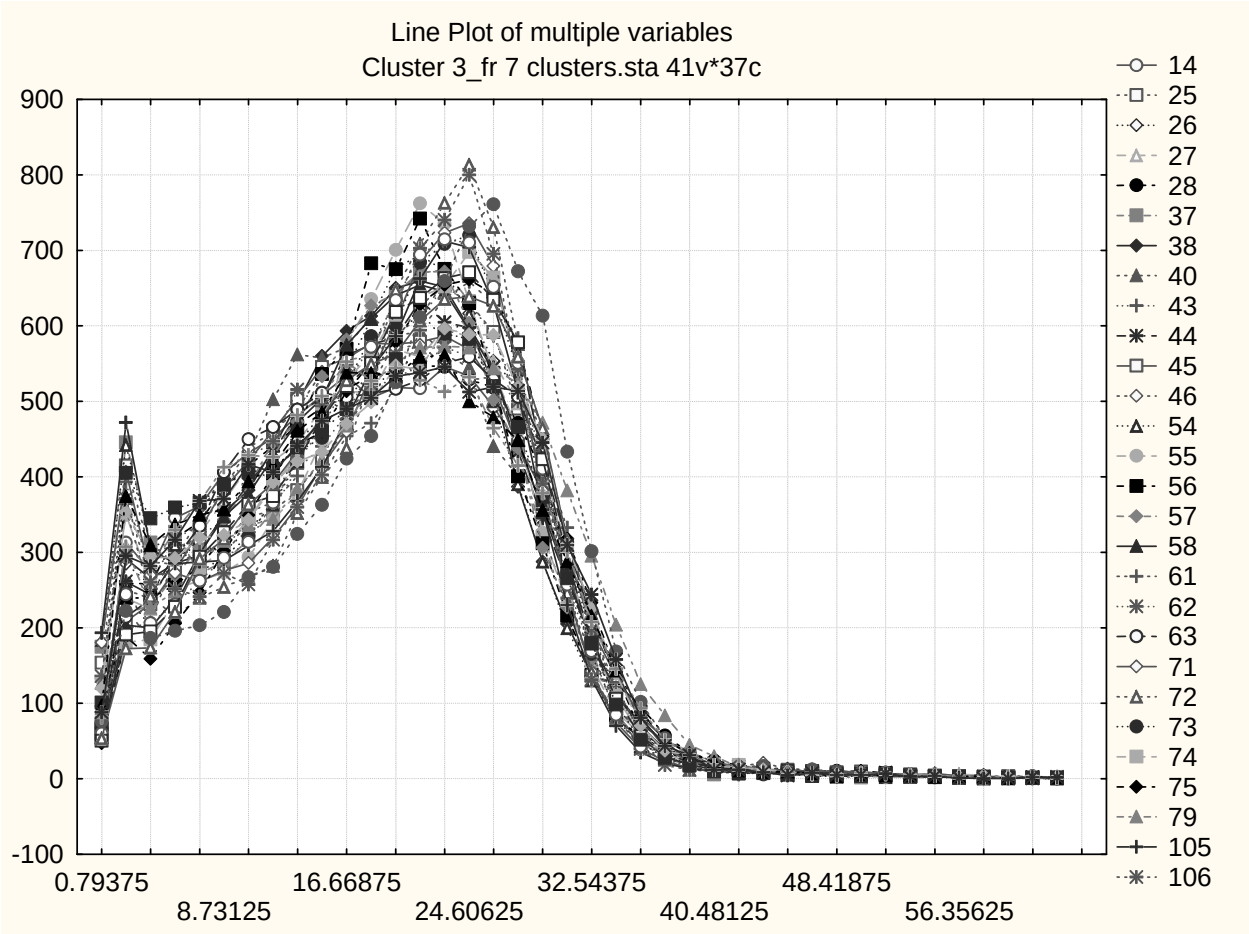
Cluster 1:


Line Plot of multiple variables
Cluster 1_fr 7 clusters.sta 41v*51c
800
700
600
500
400
300
200
100
0
-100
0.79375
8.73125
16.66875
24.60625
32.54375
40.48125
48.41875
56.35625
1
12
13
24
30
36
42
48
49
60
66
67
68
77
82
83
87
88
141
147
148
158
159
163
164
169
170
175

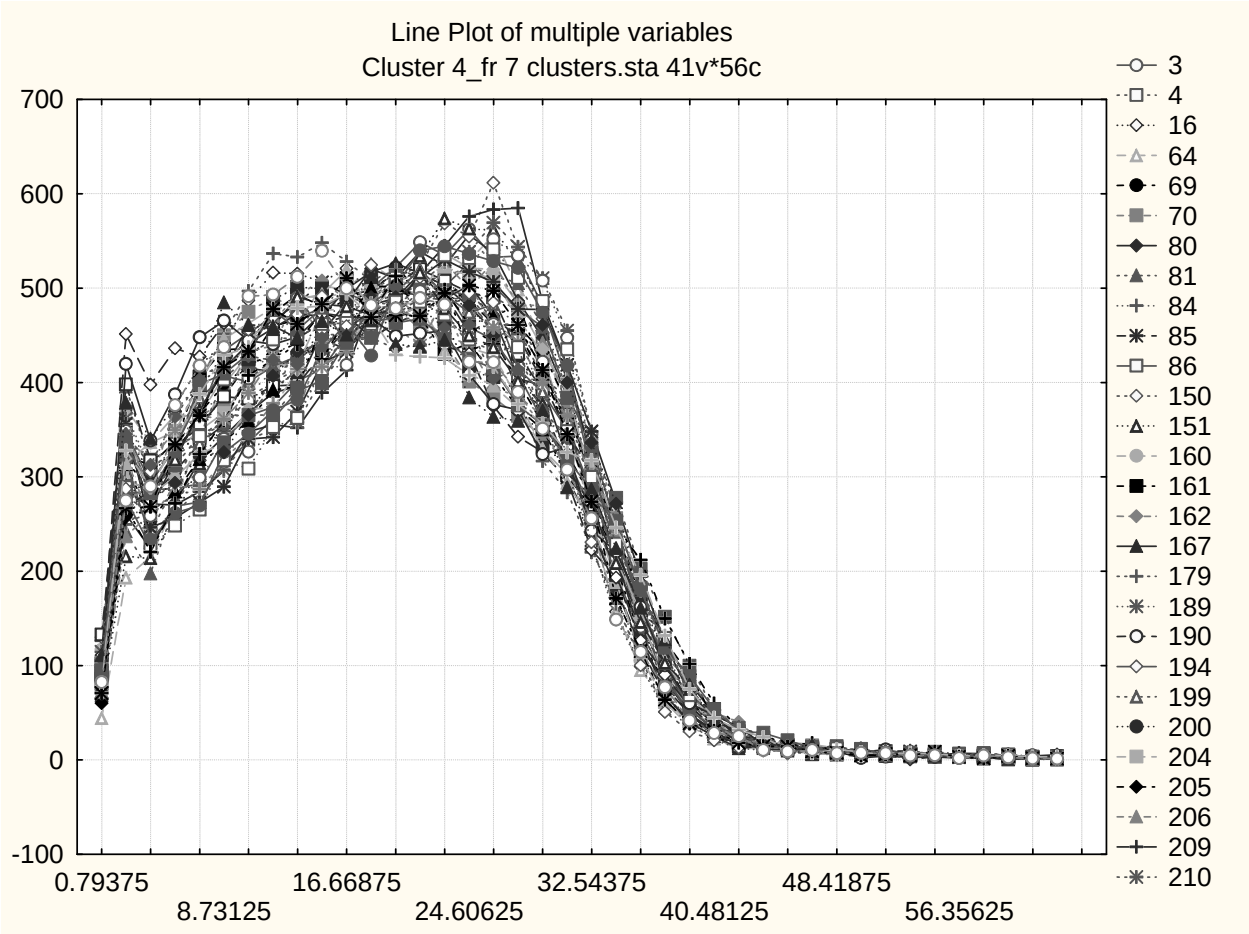
Cluster 2:


Line Plot of multiple variables
Cluster 2_fr 7 clusters.sta 41v*79c
900
800
700
600
500
400
300
200
100
0
-100
0.79375
8.73125
16.66875
24.60625
32.54375
40.48125
48.41875
56.35625
2
78
142
144
145
153
154
188
192
193
197
207
208
218
219
220
221
222
224
225
226
227
230
231
232
233
234
238

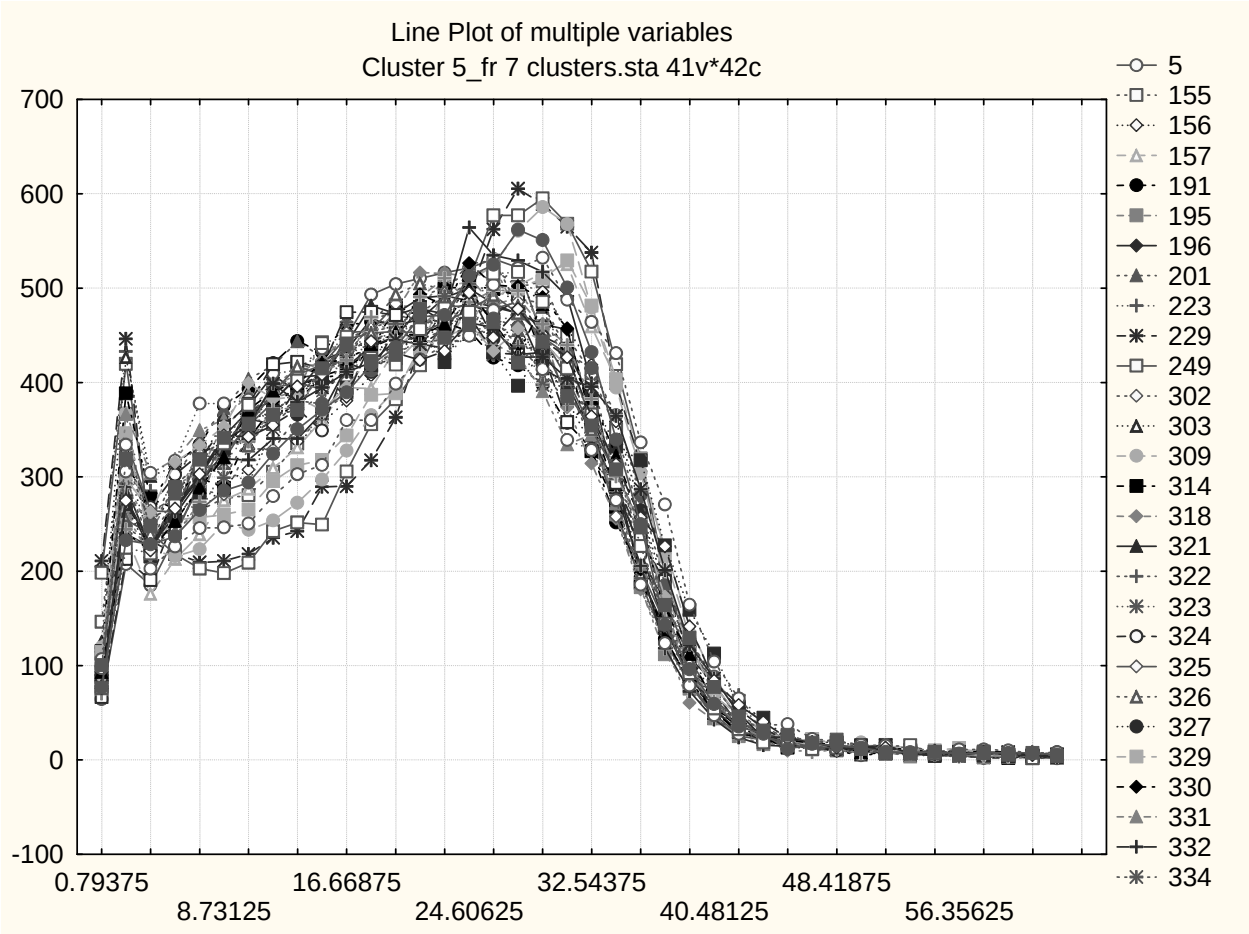
Cluster 3:


Line Plot of multiple variables
Cluster 3_fr 7 clusters.sta 41v*37c
900
800
700
600
500
400
300
200
100
0
-100
0.79375
8.73125
16.66875
24.60625
32.54375
40.48125
48.41875
56.35625
14
25
26
27
28
37
38
40
43
44
45
46
54
55
56
57
58
61
62
63
71
72
73
74
75
79
105
106

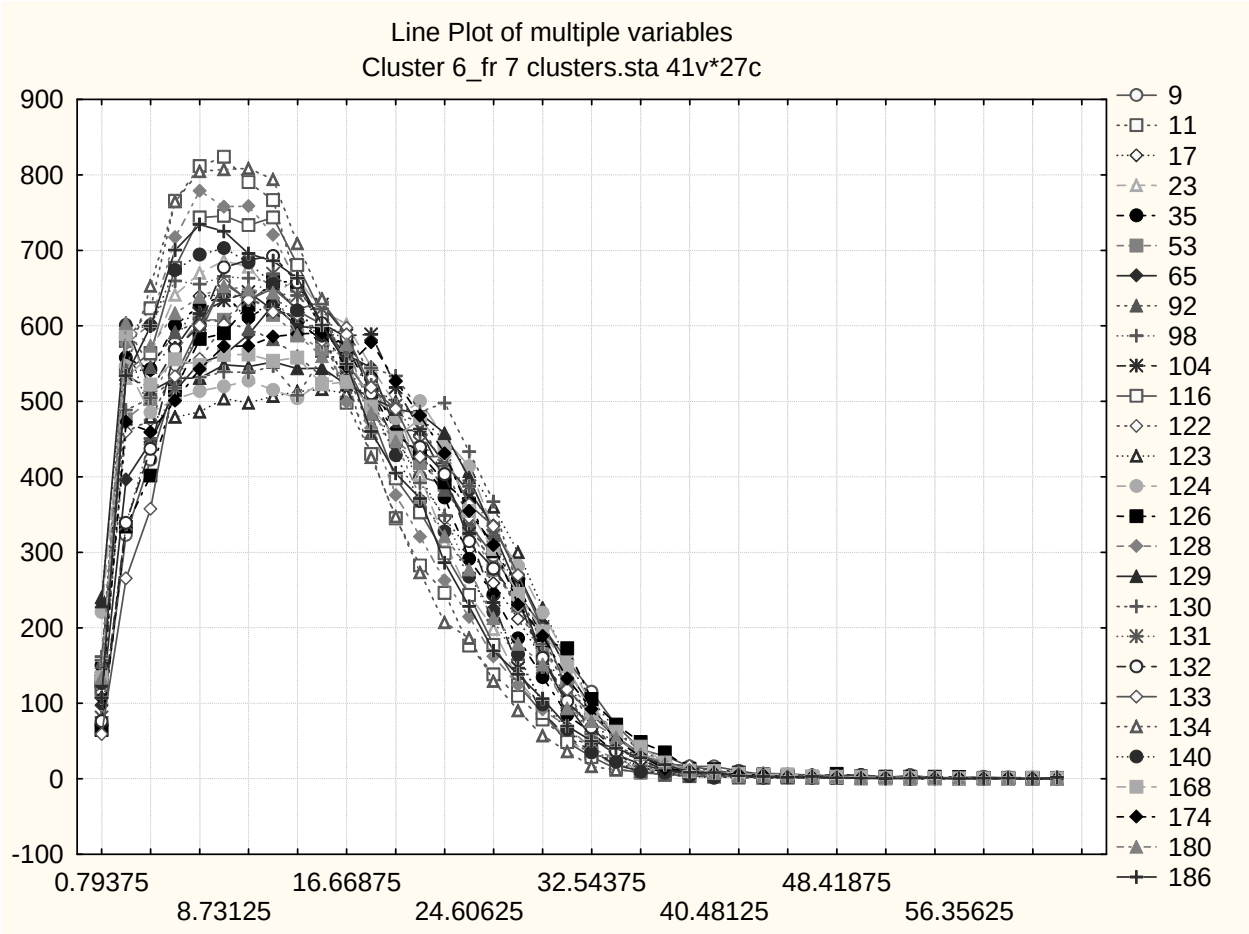
Cluster 4:


Line Plot of multiple variables
Cluster 4_fr 7 clusters.sta 41v*56c
700
600
500
400
300
200
100
0
-100
0.79375
8.73125
16.66875
24.60625
32.54375
40.48125
48.41875
56.35625
3
4
16
64
69
70
80
81
84
85
86
150
151
160
161
162
167
179
189
190
194
199
200
204
205
206
209
210

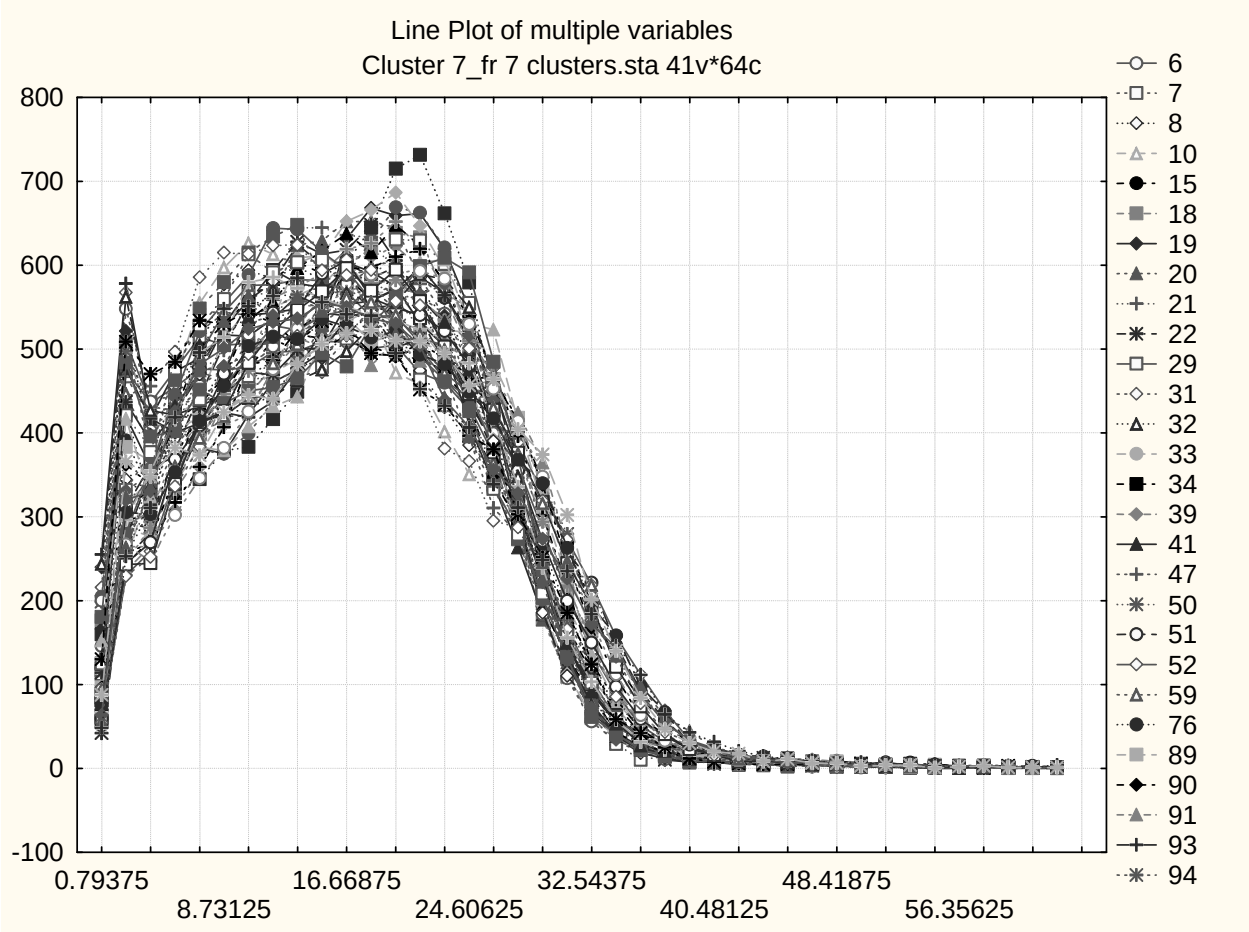
Cluster 5:


Line Plot of multiple variables
Cluster 5_fr 7 clusters.sta 41v*42c
700
600
500
400
300
200
100
0
-100
0.79375
8.73125
16.66875
24.60625
32.54375
40.48125
48.41875
56.35625
5
155
156
157
191
195
196
201
223
229
249
302
303
309
314
318
321
322
323
324
325
326
327
329
330
331
332
334

Cluster 6:


Line Plot of multiple variables
Cluster 6_fr 7 clusters.sta 41v*27c
900
800
700
600
500
400
300
200
100
0
-100
0.79375
8.73125
16.66875
24.60625
32.54375
40.48125
48.41875
56.35625
9
11
17
23
35
53
65
92
98
104
116
122
123
124
126
128
129
130
131
132
133
134
140
168
174
180
186

Cluster 7:



III. Conclusion:

We used K-Means Clustering to classify fiber length histograms of 356 samples into 7 groups. Line plots generated for the 7 clusters show distinct patterns across clusters and relatively homogenous distribution shapes within clusters. The various patterns correspond to different degrees of fiber damage.

Note to users: DATA set for this tutorial: The entire dataset illustrated above could not be provided for proprietary reasons; instead a sub-sample is provided, called: "*Fiber-Length DATA-subset_Elsevier.sta*". Please use this dataset to use in replicating the steps above; a similar, but not identical, set of results should be obtained.

IV. References:

Krifa M., 2006. Fiber Length Distribution in Cotton Processing: Dominant Features and Interaction Effects. *Textile Res. J.*, 76 (5): 426-435.
Krifa M., 2008. Fiber Length Distribution in Cotton Processing: A Finite Mixture Distribution Model. *Textile Res. J.*, 78 (8): 688-698.
Krifa M., 2009. A Mixed Weibull Model for Size Reduction of Particulate and Fibrous Materials. *Powder Technology*, 194 (3): 233-238.