Solutions to Assignment #4 STA437H1S/2005H1S

1. (a) First of all,

$$\begin{aligned}\min\{d(A,B),d(A,C)\} &= \min\left\{\min_{a\in A,b\in B} d(a,b),\ \min_{a\in A,c\in C} d(a,c)\right\}\\ &= \min_{a\in A,b\in B\cup C} d(a,b)\\ &= d(A,B\cup C)\end{aligned}$$

The other equality follows since $\min(x,y) = (x+y)/2 - |x-y|/2$.
(b) This is exactly analogous to part (a):

$$\begin{aligned}\max\{d(A,B),d(A,C)\} &= \max\left\{\max_{a\in A,b\in B} d(a,b),\ \max_{a\in A,c\in C} d(a,c)\right\}\\ &= \max_{a\in A,b\in B\cup C} d(a,b)\\ &= d(A,B\cup C)\end{aligned}$$

The other equality follows since $\max(x,y) = (x+y)/2 + |x-y|/2$.
(c) For a, b, c in A, B, C, respectively, we have

$$d(a,c) \le d(a,b) + d(b,c).$$

Thus

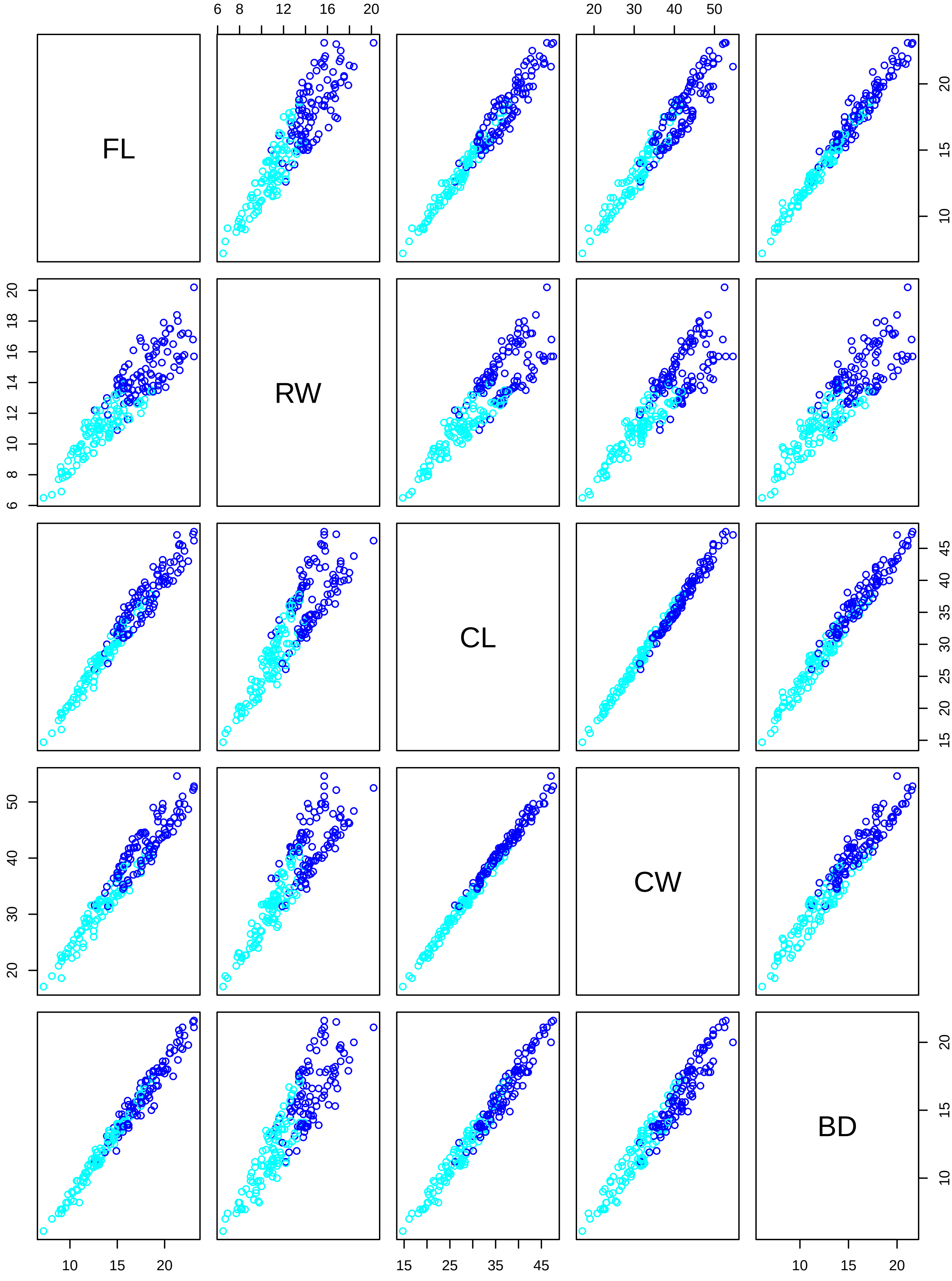
$$\max_{a\in A,b\in B,c\in C} d(a,c) = d(A,C) \le \max_{a\in A,b\in B,c\in C} \{d(a,b)+d(b,c)\}$$

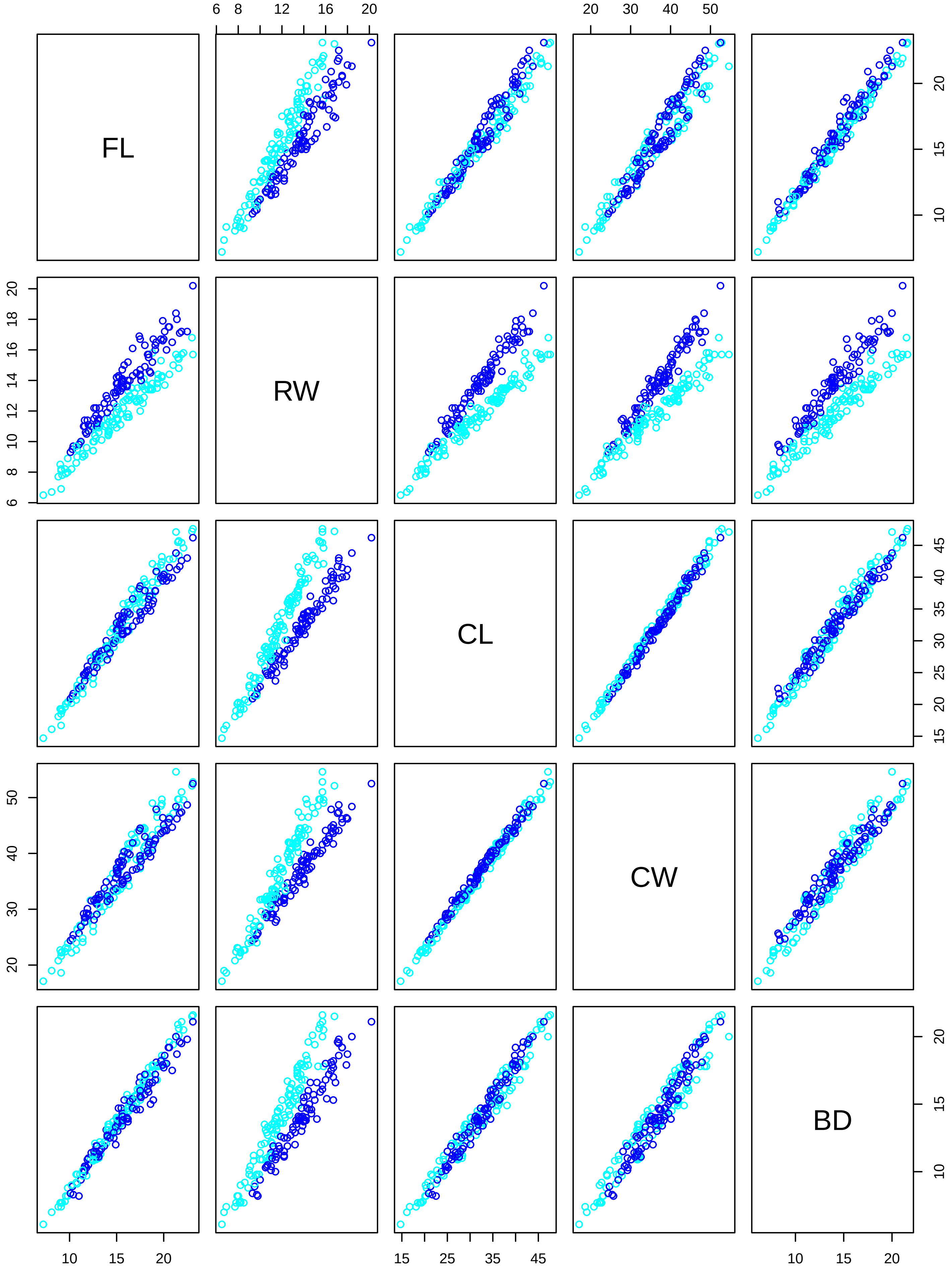
Likewise

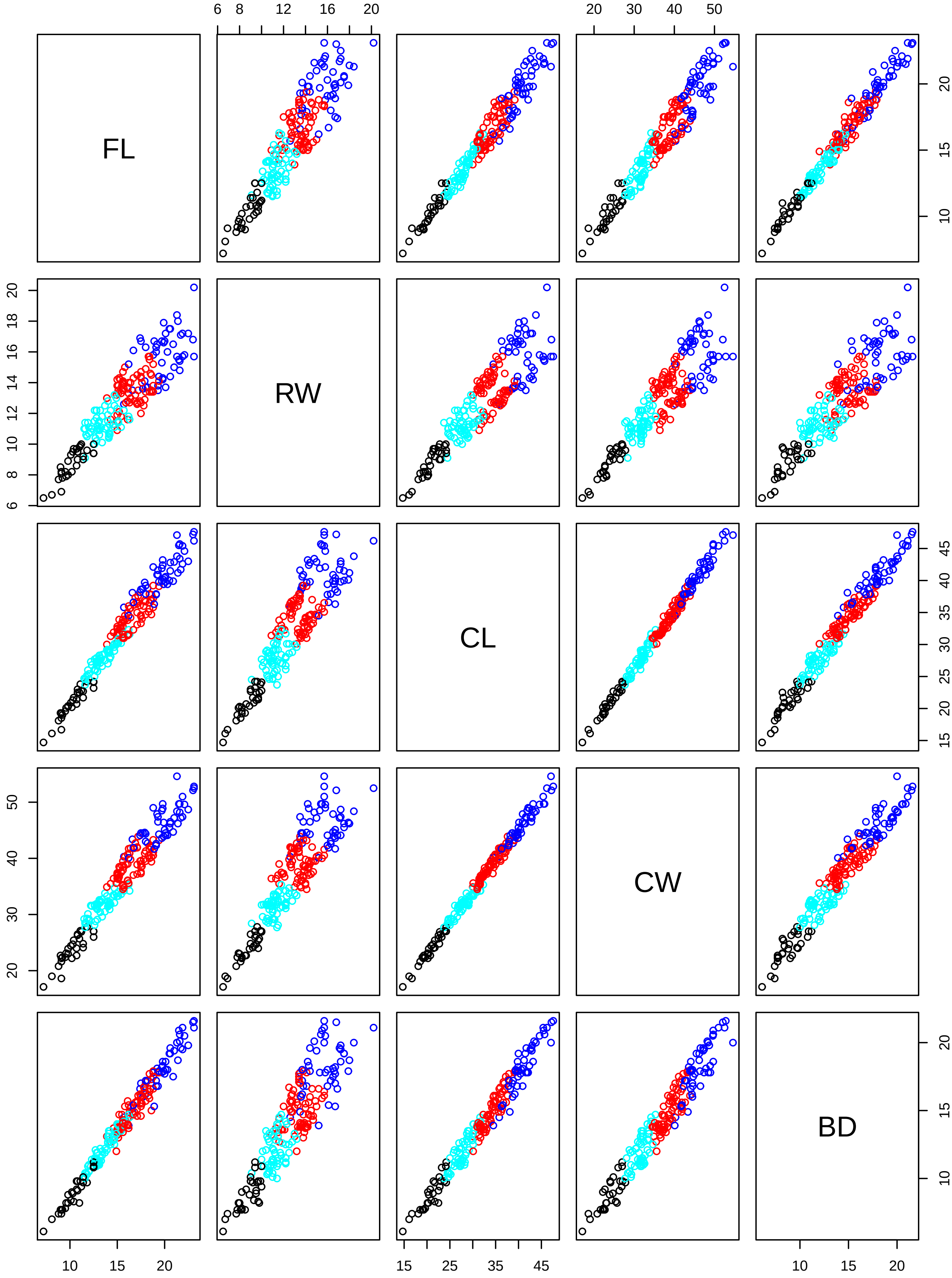
$$\max_{a\in A,b\in B,c\in C} \{d(a,b)+d(b,c)\} \le \max_{a\in A,b\in B,c\in C} d(a,b) + \max_{a\in A,b\in B,c\in C} d(b,c) = d(A,B)+d(B,C).$$

Note that the argument given above will not work for single linkage clustering as the second inequality above will not hold in generaly if max is replaced by min.

2. (a) The clusters after 30 iterations are shown on pairwise scatterplots. Essentially, these clusters divide each variable according to size, reflecting the initial clusters given by the k-means procedure.
(b) After 100 iterations, the two clusters are markedly different from those after 30 iterations. In particular, the clusters seem to recover the two sex groups – that is, one cluster consists mainly of females, the other mainly of males.
(c) The four clusters shown in the pairwise scatterplots are those after 500 iterations of the EM algorithm. The four clusters essentially divide each variable into 4 groups from smallest to largest, and do not recover the 4 sex/species groups. Whether it is possible to recover the sex/species groups using this mixture model is not clear; it is possible that the initial clusters provided by k-means are not "good enough" for the EM algorithm to isolate the sex/species groups.


FL
RW
CL
CW
BD
6
8
12
16
20
20
30
40
50
20
15
10
20
18
16
14
12
10
8
6
45
40
35
30
25
20
15
50
40
30
20
20
15
10
10
15
20
15
25
35
45
10
15
20


FL
RW
CL
CW
BD
6
8
12
16
20
20
30
40
50
20
15
10
20
18
16
14
12
10
8
6
45
40
35
30
25
20
15
50
40
30
20
20
15
10
10
15
20
15
25
35
45
10
15
20


FL
RW
CL
CW
BD
6
8
12
16
20
20
30
40
50
20
15
10
20
18
16
14
12
10
8
6
45
40
35
30
25
20
15
50
40
30
20
20
15
10
10
15
20
15
25
35
45
10
15
20