Python seminar Homework for Chap. 7.3 No.5

Please conduct cluster analyses on wet weight time series of fishes around Japan.

The standardized wet weight time series of 6 fish stocks are in "standardized_weight.csv".
The data is from 1978 to 2018.
The header has xxxY_Z format, where xxx is fish species, Y is regional stock and Z is life stage.
For species,

maiwashi: Japanese sardine
masaba: chub mackerel
urumeiwashi: round herring
katakuchi: Japanese anchovy

For regional stock,

P: Pacific
T: Tsushima

For life stage,

J: juvenile
m: juvenile to mature mixed
M: mature

For cluster analysis, there are two types of clustering: hierarchical clustering and non-hierarchical clustering.
For big data, non-hierarchical is effective. For simple data, hierarchical clustering is easy to understand.
In this case, since the data set is not large, we will try hierarchical clustering.

For hierarchical clustering, there are several methods including single (nearest), complete (farthest), average, weighted, centroid, median, and ward methods.
Please see the detail options in
https://docs.scipy.org/doc/scipy/reference/cluster.hierarchy.html

Each method needs metric to evaluate the distance between the data, including 'braycurtis', 'canberra', 'chebyshev', 'cityblock', 'correlation', 'cosine', 'dice', 'euclidean', 'hamming', 'jaccard', 'jensenshannon', 'kulsinski', 'mahalanobis', 'matching', 'minkowski', 'rogerstanimoto', 'russellrao', 'seuclidean',

‘sokalmichener’, ‘sokalsneath’, ‘sqeuclidean’, ‘yule’.
For, normal data, standardized Euclid distance is feasible. For diversity based on presence/absence data, Jaccard metric is effective. For diversity including abundance ratio, Bray-Curtis metric is effective.

In this study, please use standardized Euclid distance (but the data is already standardized, therefore the results should be the same even if Euclid distance is used).

Spicy has cluster analyses tools. To use hierarchical cluster analyses, please import linkage & dendrogram by

from scipy.cluster.hierarchy import linkage, dendrogram

Please compare the results of cluster analyses based on single, complete, average and ward methods.

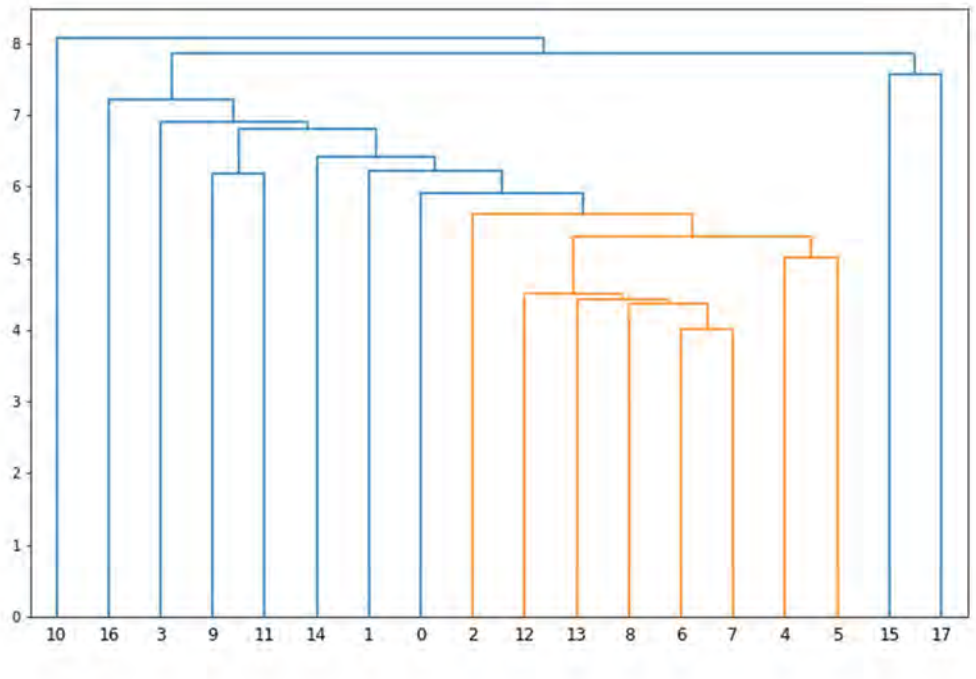

Figure 1. Dendrogram from the single method. The single method gathers the nearest data. It is not suitable for grouping but for identification of resembling data (orange stocks).

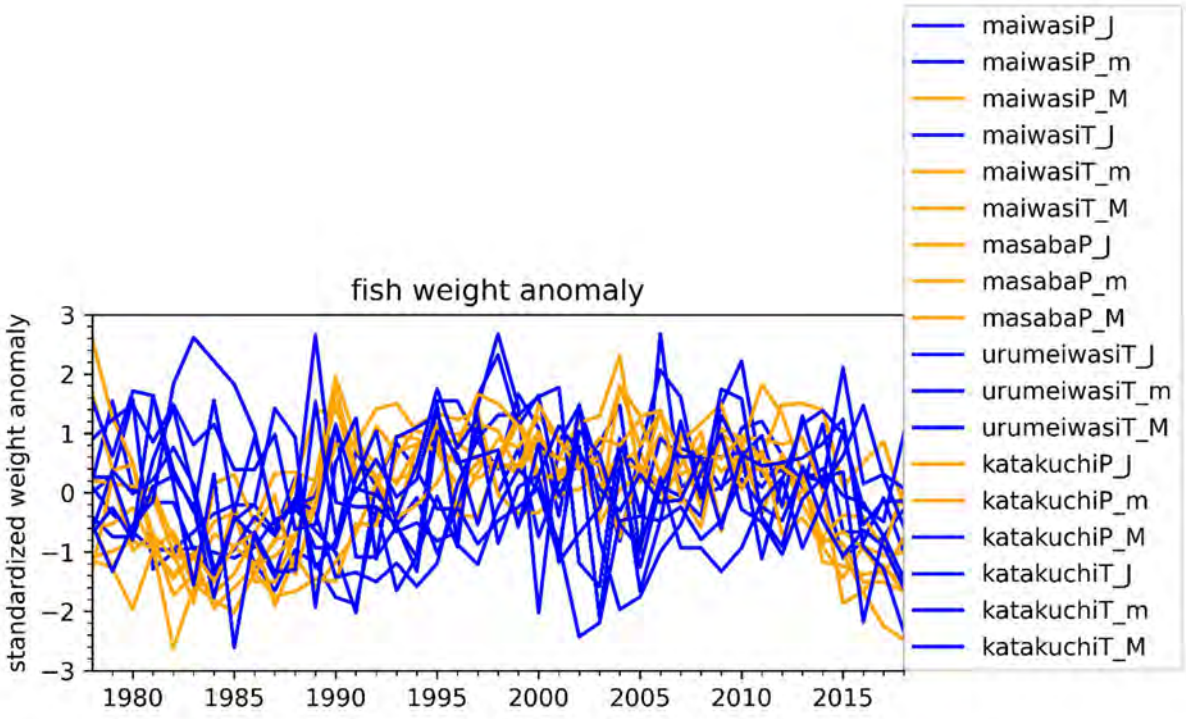

Figure 2. Time series of standardized fish wet weight anomaly. The colors correspond to that in Figure 1. The orange color stocks show similar fluctuations.

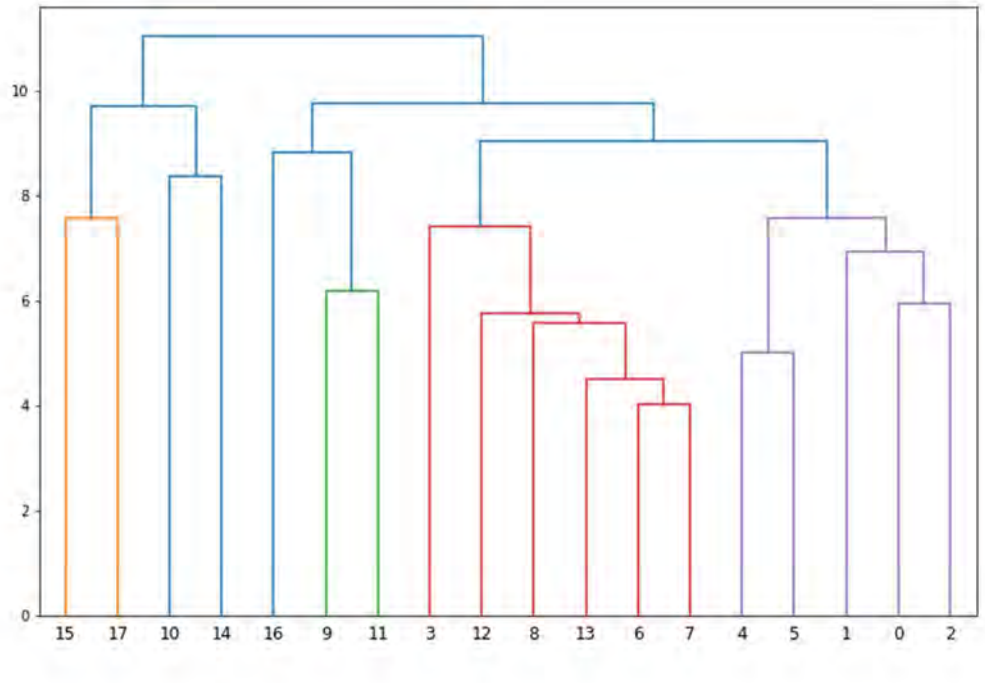

Figure 3. Dendrogram from the complete method. The complete method separates the farthest data. It is suitable for grouping but not complicated

structure data.

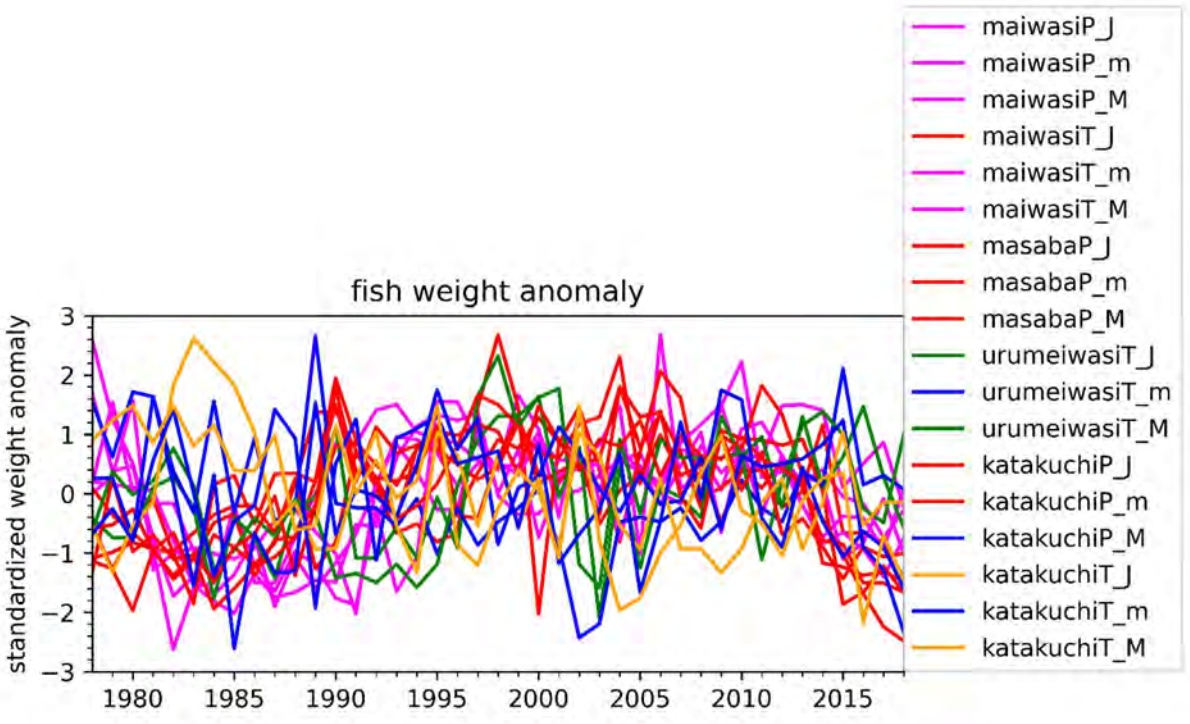

Figure 4. Time series of standardized fish wet weight anomaly. The colors correspond to that in Figure 3.

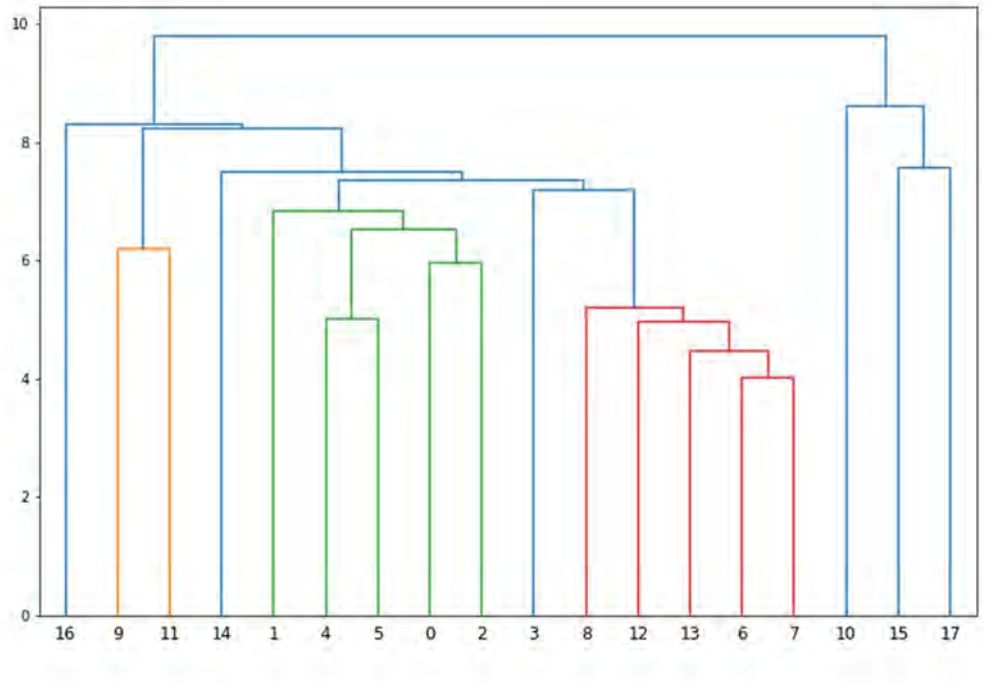

Figure 5. Dendrogram from the average method. The average method separates the groups based on averaged distance. It is suitable for grouping but sometime

it shows many single cluster.

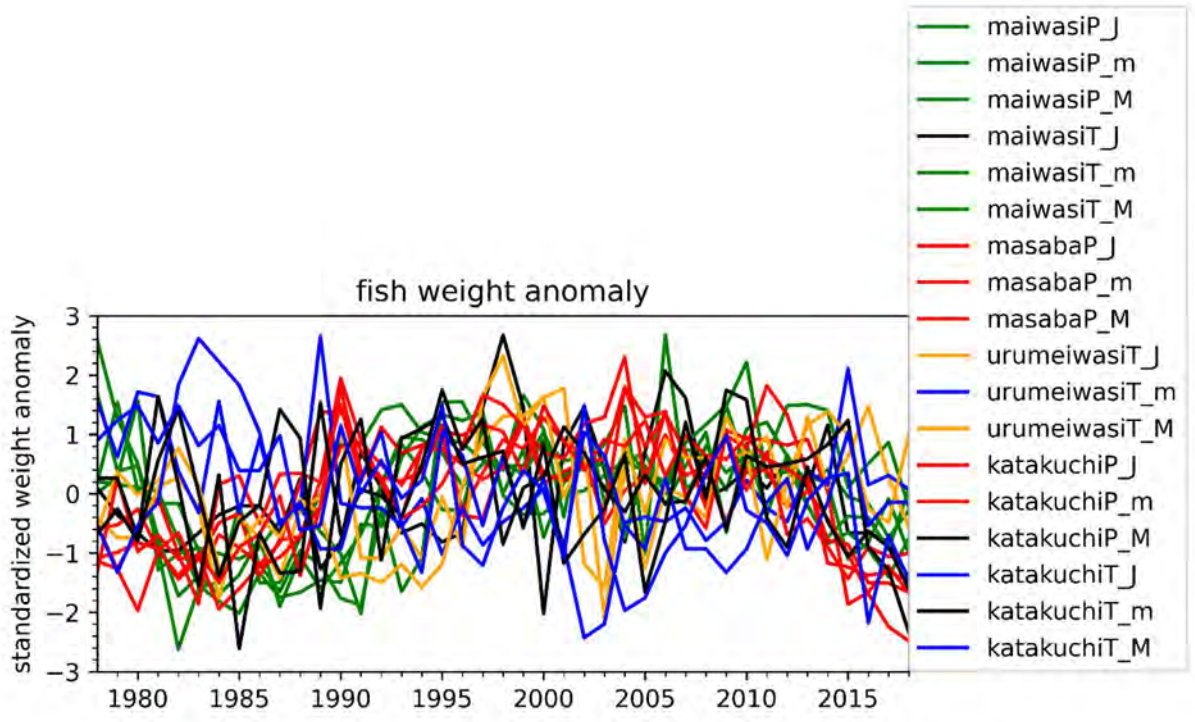

Figure 6. Time series of standardized fish wet weight anomaly. The colors correspond to that in Figure 5.

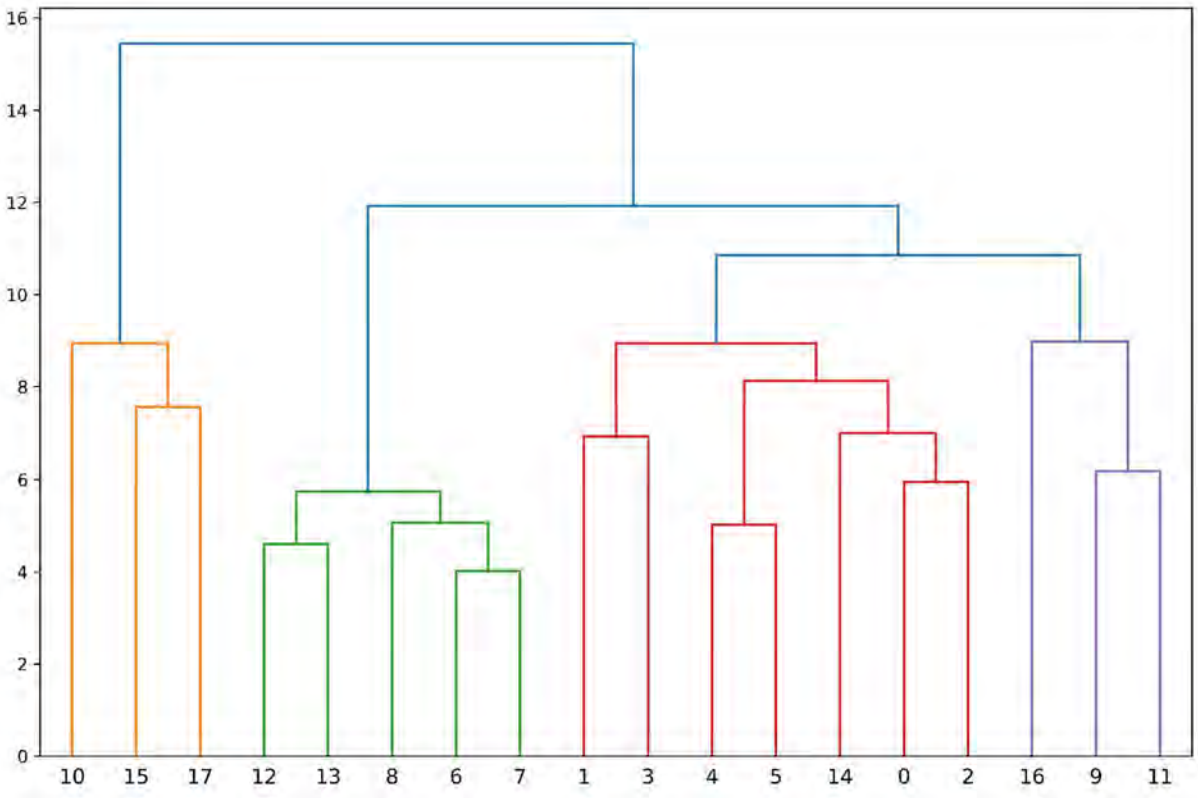

Figure 7. Dendrogram from the Ward method. The Ward method minimizes the variance. It is suitable for grouping in general.

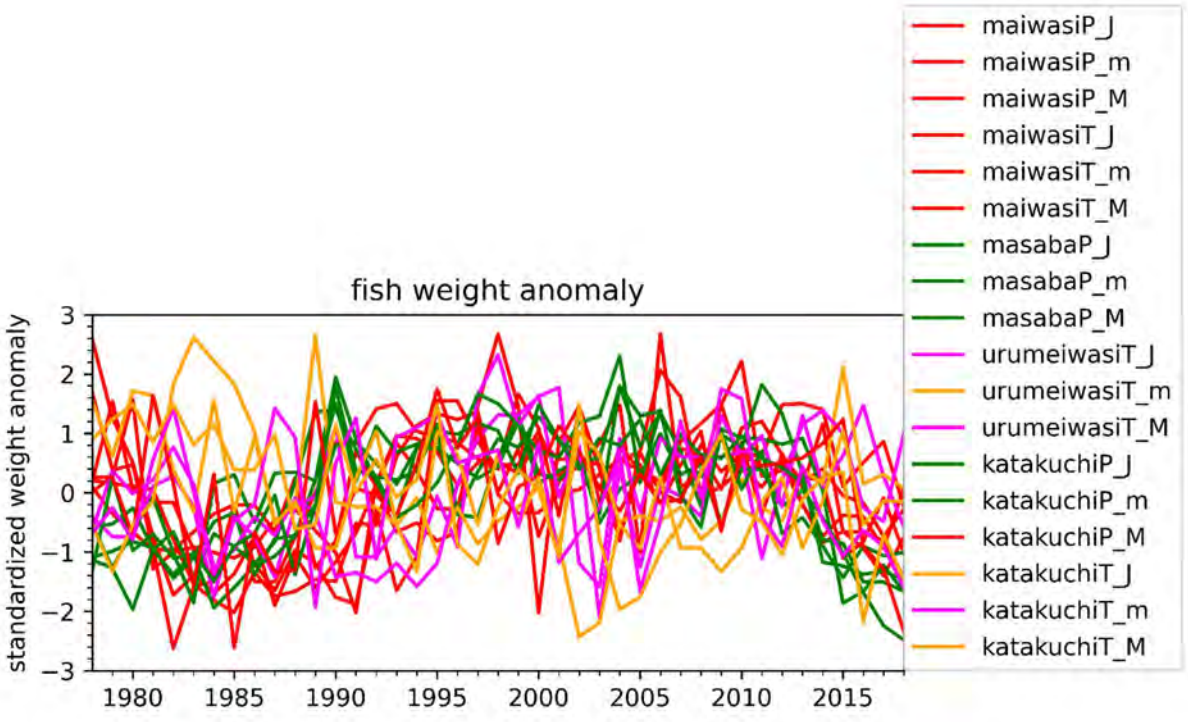

Figure 6. Time series of standardized fish wet weight anomaly. The colors correspond to that in Figure 7.