# Ch.5 Classification and Clustering

In machine learning, there are two main types of learning problems, *supervised* and *unsupervised* learning.

An analogy for the former is a French class where the teacher demonstrates the correct French pronunciation.
An analogy for the latter is students working on a team project without supervision – i.e., the students are provided with learning rules, but must rely on self-organization to arrive at a solution, without a teacher.

In supervised learning, one is provided with the predictor data, $\mathbf{x}_1, \ldots, \mathbf{x}_n$, and the response data, $\mathbf{y}_1, \ldots, \mathbf{y}_n$. Given the predictor data as input, the model produces outputs, $\mathbf{y}'_1, \ldots, \mathbf{y}'_n$.

The model learning is "supervised" in that the model output $(\mathbf{y}'_1, \ldots, \mathbf{y}'_n)$ is guided towards the given response data $(\mathbf{y}_1, \ldots, \mathbf{y}_n)$, usually by minimizing an *objective function* (also called a *cost function* or *error function* or loss function).

Regression and classification involve supervised learning.

In contrast, for unsupervised learning, only input data are provided, and the model discovers the natural patterns or structure in the input data. Principal component analysis and clustering involve unsupervised learning.

With discrete variables, classification is supervised, while clustering is unsupervised.

## 5.4 K-means clustering [Book, Sec.1.7]

*Clustering* or cluster analysis is the unsupervised version of classification. The goal of clustering is to group the data into a number of subsets or 'clusters', such that the data within a cluster are more closely related to each other than data from other clusters.

A simple and widely used clustering method is *K-means clustering*: Start with initial guesses for the mean positions of the $K$ clusters in data space (to be referred to as the cluster centres), then iterates the following two steps till convergence:

(i) For each data point, find the closest cluster centre (based on Euclidean distance).

(ii) For each cluster, reassign the cluster centre to be the mean position of all the data belonging to that cluster.

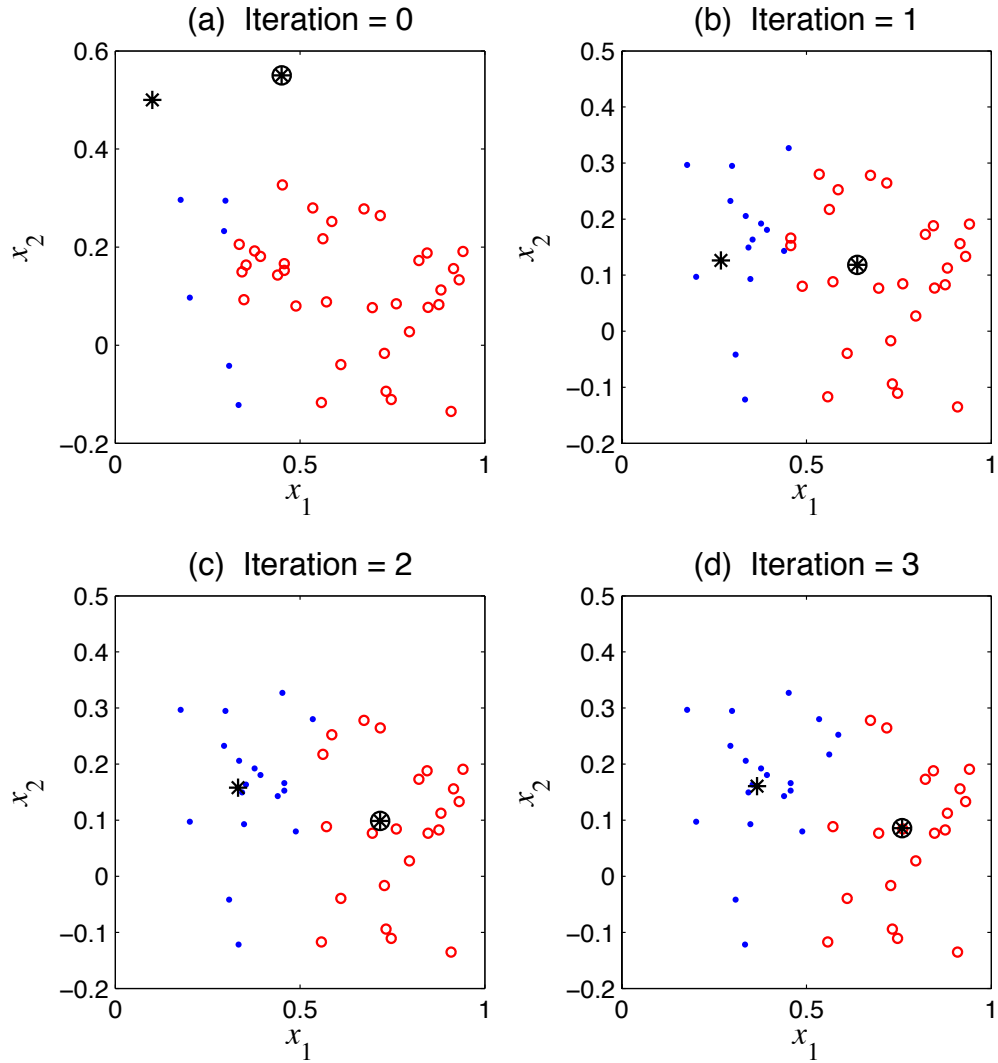


Figure : (a) The initial guesses for the two cluster centres are marked by the asterisk and the circled asterisk. The data points closest to the

asterisk are plotted as dots, while those closest to the circled asterisk are shown as circles. The location of the cluster centres and their associated clusters are shown after (b) one, (c) two, (d) three iterations.

As $K$, the number of clusters, is specified by the user, choosing a different $K$ will lead to very different clusters. Trouble is we don't know what the optimal $K$ value to use.

## 5.5 Hierarchical clustering

Hierarchical clustering allows clusters to be nested inside each other, i.e. one cluster branching into two smaller clusters, leading to an tree structure.
Two approaches to constructing the tree, bottom-up (agglomerative clustering) or top-down (divisive clustering). In the more common

bottom-up approach, the most similar clusters are merged at each step, while in the top-down approach, clusters are split up.

Given $N$ data points, agglomerative clustering starts with $N$ clusters, each containing one data point. At each step, it merges the two most "similar" clusters into one cluster. Many ways to define "similar", so many variants of this method.

Merging process can be plotted as an inverted tree, a dendrogram. Initial clusters are the leaves at the bottom, and the binary branching node in the tree indicates where two clusters are merged together. For a branching node, the vertical scale gives the distance or dissimilarity between the two clusters which are being merged. [The reason we can plot an inverted tree is because the distance between two merged clusters increases monotonically as the merging process continues.]

Example: Use satellite data to forecast Canadian Prairie crop yield, with yield data available at 40 Census Agricultural Regions (CAR). (Johnson, 2013)
Since data record is short (2000–2011) and there are 40 CARs, cluster the yield data from the 40 CARs, so build fewer models but with more training data for each model.

Since there are 12 annual values at each CAR, clustering is done for 40 data points in 12-dimensional space.

We can cut the tree at a selected height to choose the number of clusters – somewhat subjective.
Can probably cut when vertical link distances become relatively small further down in the inverted tree.

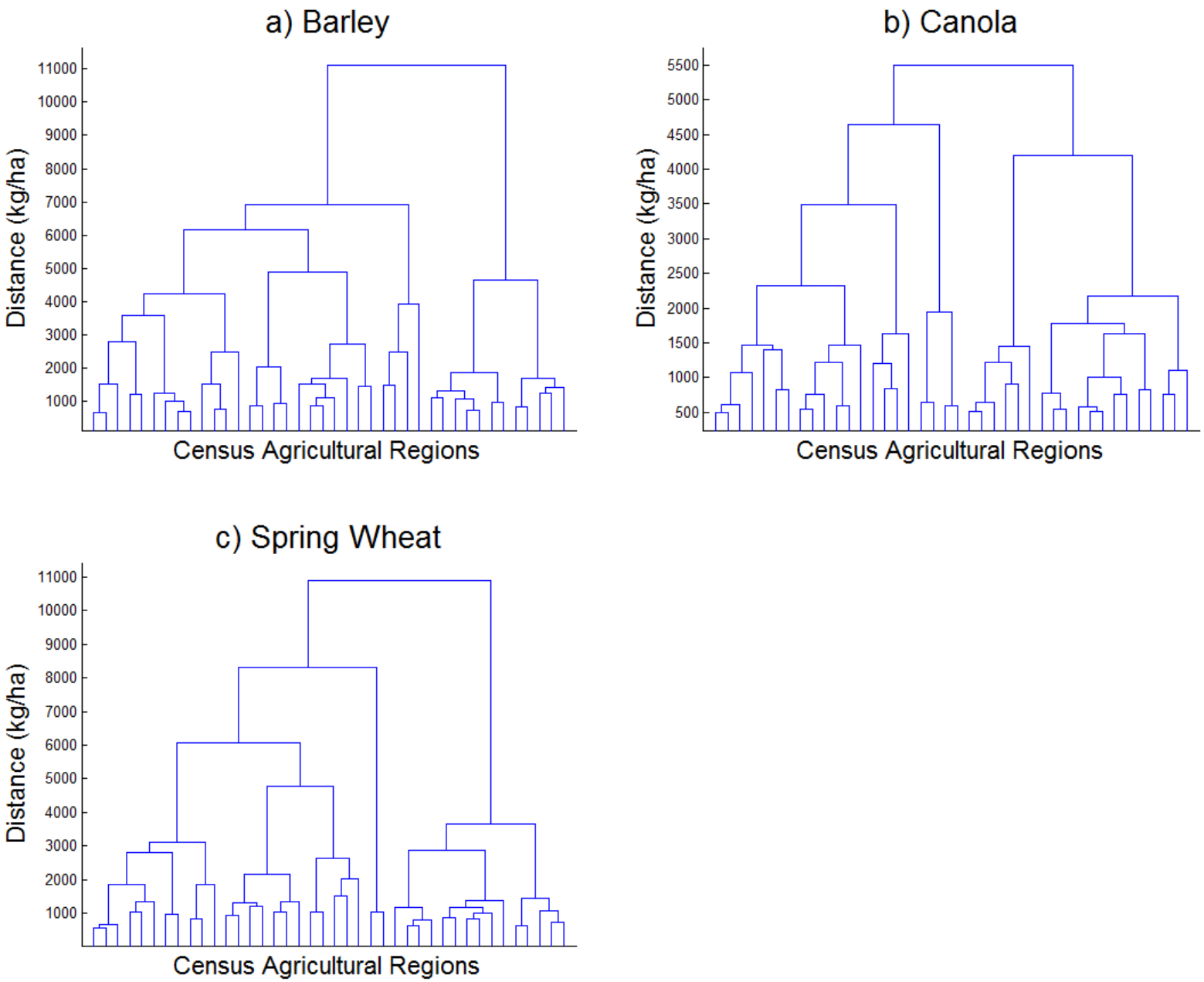


Figure : Dendrogram for Canadian prairie (a) barley, (b) canola and (c) spring wheat from various regions (with Ward's method used).

(a) Barley: vertical link distances becomes much smaller after 2 clusters and again after 4 clusters, so choose 2-4 clusters.
(b) Canola: link distances become small after 5 clusters, so choose 5 clusters.
(c) Spring wheat: smaller distances after 3 clusters and again after 5 clusters, so choose 3-5 clusters.

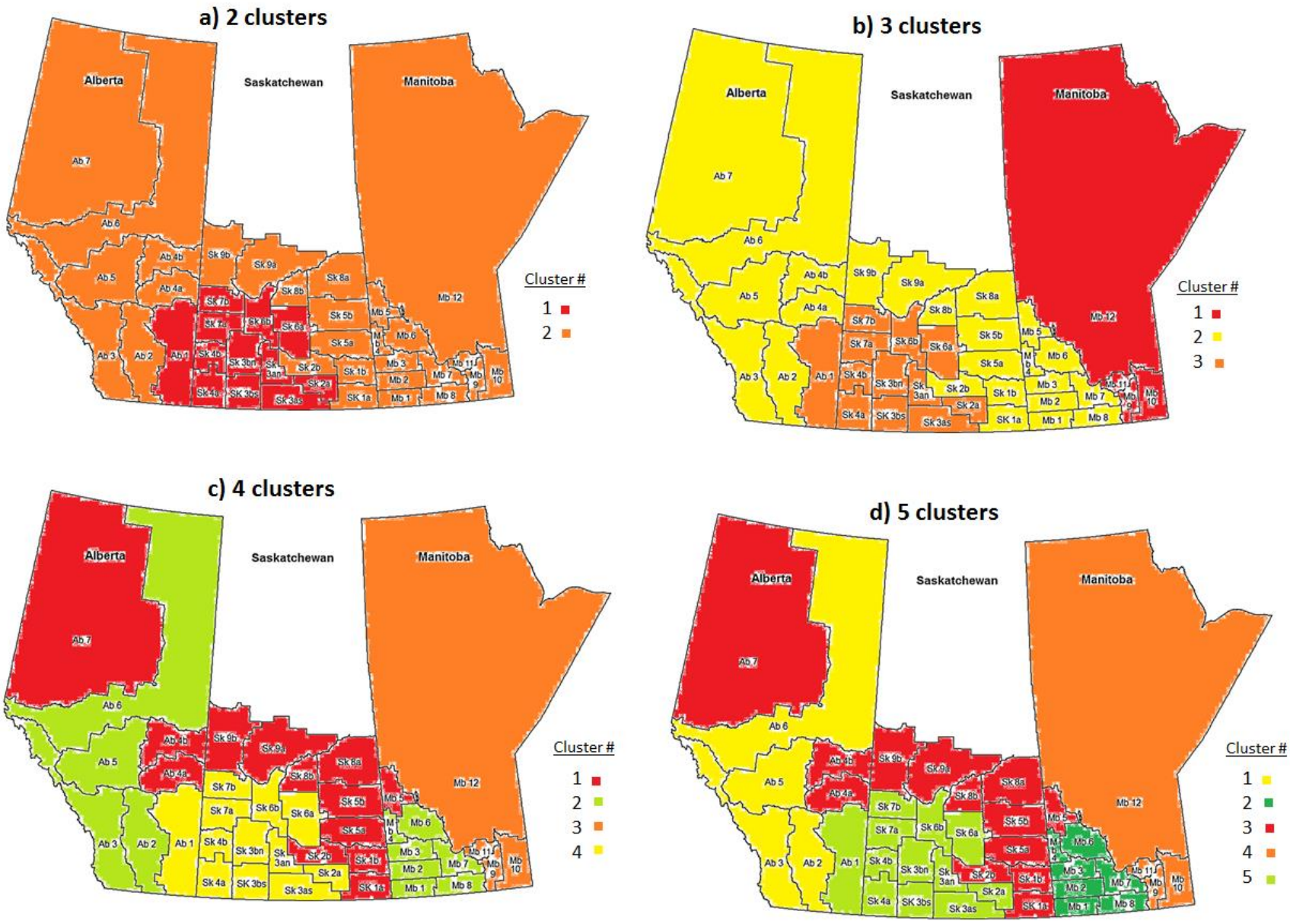


Figure : Barley yield clustered in (a) 2, (b) 3, (c) 4 and (d) 5 clusters.

Many ways to define "distance" or dissimilarity between two clusters. The average link clustering method measures the average distance between all pairs of data points in the two clusters $A$ and $B$:

$$d_{\text{ave}}(A, B) = \frac{1}{n_A n_B} \sum_{i \in A} \sum_{j \in B} d_{ij}, \tag{1}$$

where $n_A$ and $n_B$ are the number of data points in cluster $A$ and $B$, and $d_{ij}$ is the distance between the data points $i$ and $j$.
A variety of choices for $d_{ij}$, e.g. Eucidean distance, etc.

Other link approaches, e.g. single link, complete link, are generally not as good as average link.

Gong and Richman (1995): Best hierarchical clustering method is the Ward's method, though some non-hierarchical methods can be better.

Ward's method: At each step find the pair of clusters that leads to minimum increase in total within-cluster variance after merging.

Q4: Hierarchical clustering analysis (using Ward's method) was applied to a dataset containing the concentration of 11 ions in the honey from 44 locations in Europe (Fermo et al., 2013). How many clusters for the locations appear optimal?

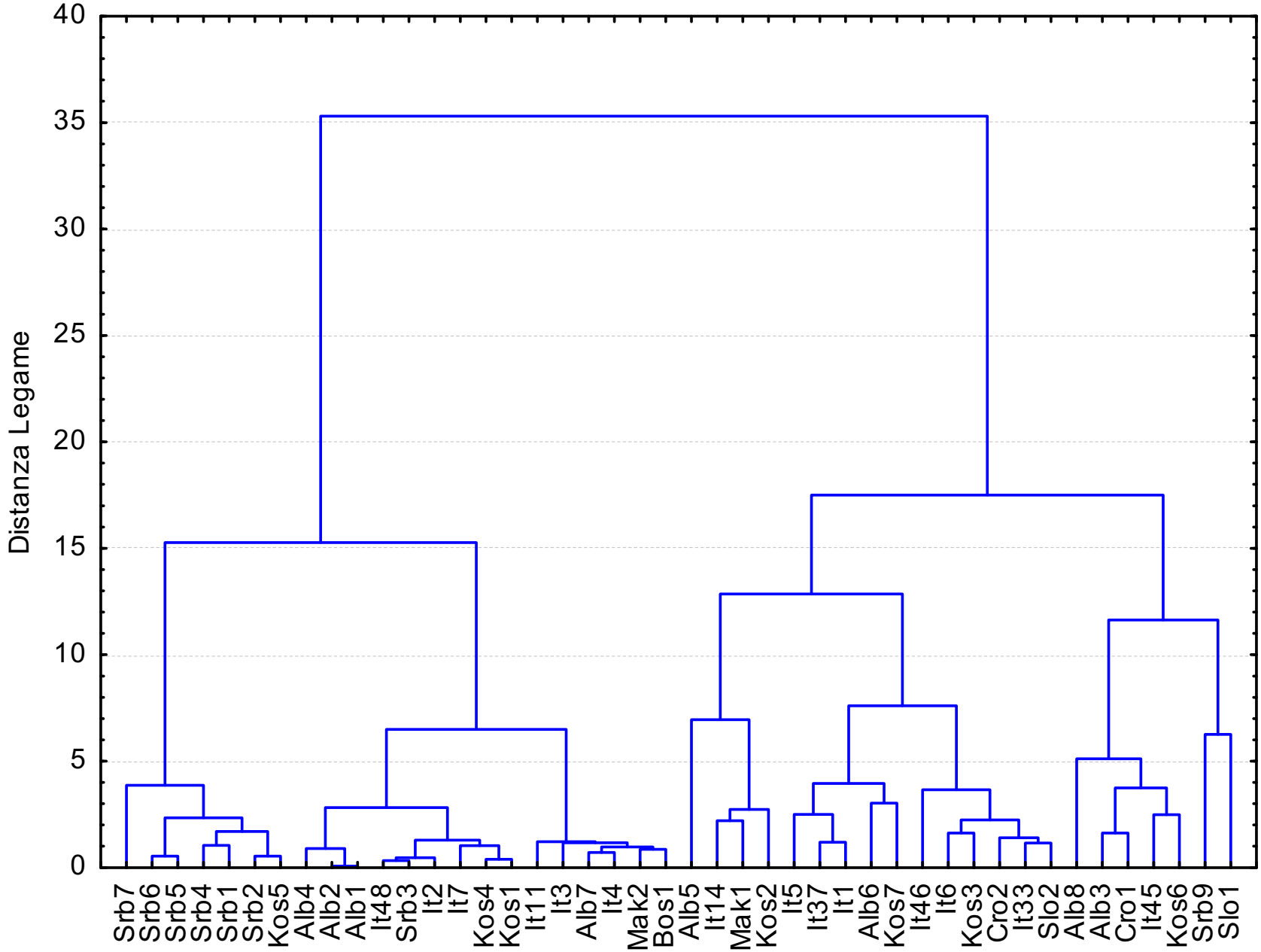


**Fig. 4.** Dendrogram obtained from PCA (Ward's method) of honey samples from Italy and from W. Balkans.

## Matlab functions for clustering:

*K*-means clustering:
`http://www.mathworks.com/help/stats/kmeans.html`

Hierarchical clustering:
`http://www.mathworks.com/help/stats/hierarchical-clustering.html#bq_679x-10`

`http://www.mathworks.com/help/stats/linkage.html`
For method, choose 'ward' (best) or 'average'. The default 'single' is poor. I.e.
Y = pdist(X); % calculates pairs of distances from data matrix X
Z = linkage(Y, 'ward'); % cluster using Ward's method
dendrogram(Z) % plots dendrogram (up to 30 nodes)
dendrogram(Z,0) % plots dendrogram (all nodes)

## References:


Fermo, P., Beretta, G., Facino, R. M., Gelmini, F., and Piazzalunga, A. (2013). Ionic profile of honey as a potential indicator of botanical origin and global environmental pollution. *Environmental Pollution*, 178:173–81.

Gong, X. F. and Richman, M. B. (1995). On the application of cluster-analysis to growing-season precipitation data in North-America east of the Rockies. *J. Climate*, 8(4):897–931.

Johnson, M. D. (2013). *Crop Yield Forecasting on the Canadian Prairies by Satellite Data and Machine Learning*. M.Sc. thesis, Univ. of British Columbia.