Cluster Analysis (Clustering)

COC131

Data Mining

19 March 2009

Richard Forsyth

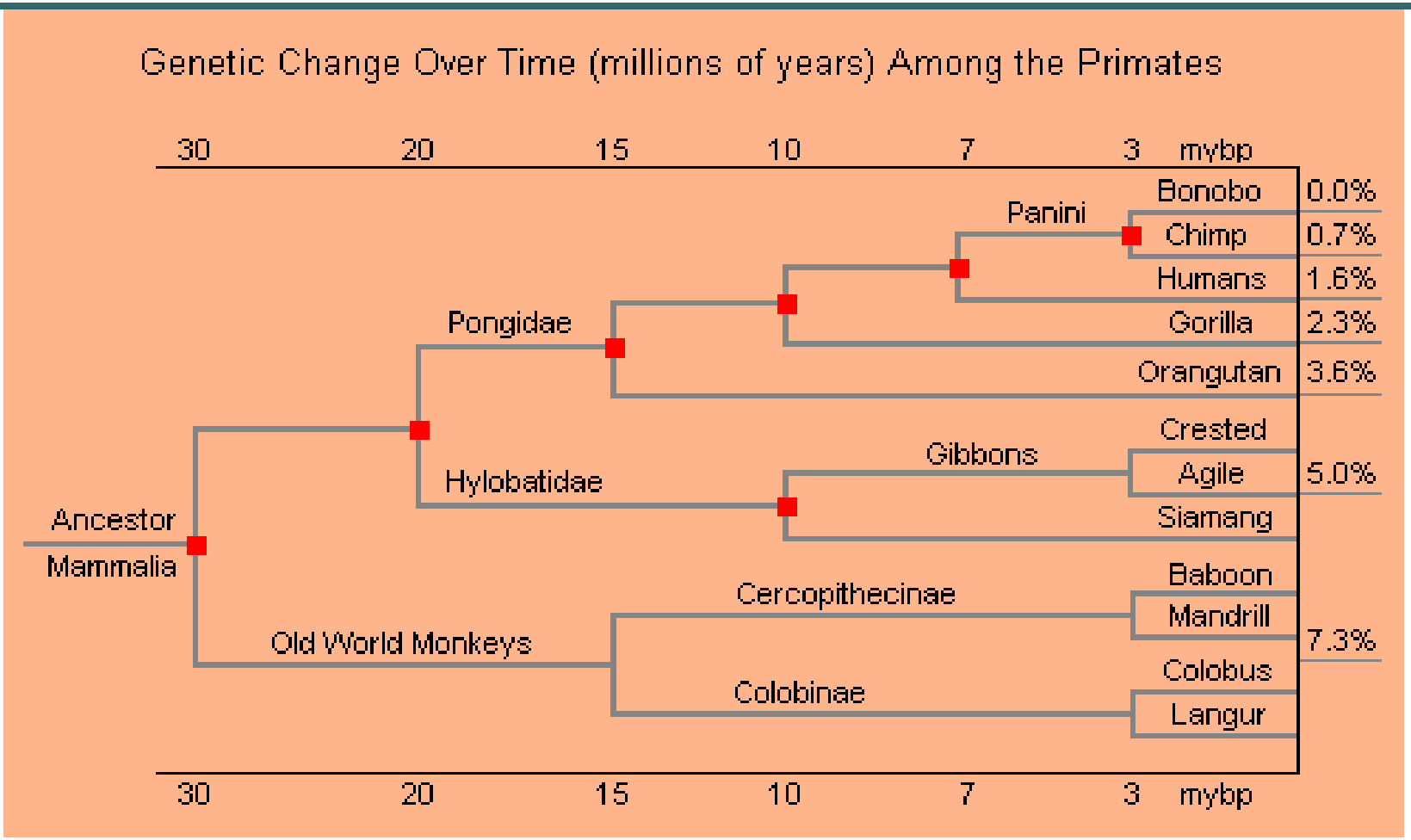
R.S.Forsyth@lboro.ac.uk

Cluster Analysis:

- What's it for?
  - to find (sub)groups in data
- Why?
  - fundamental to human thinking
  - discovery of "natural kinds" (Quine, 1969)
    - Carl Linnaeus : tree-structured taxonomy

A biological example

Data from *The Third Chimpanzee* by Jared Diamond, 1992



Who uses it?

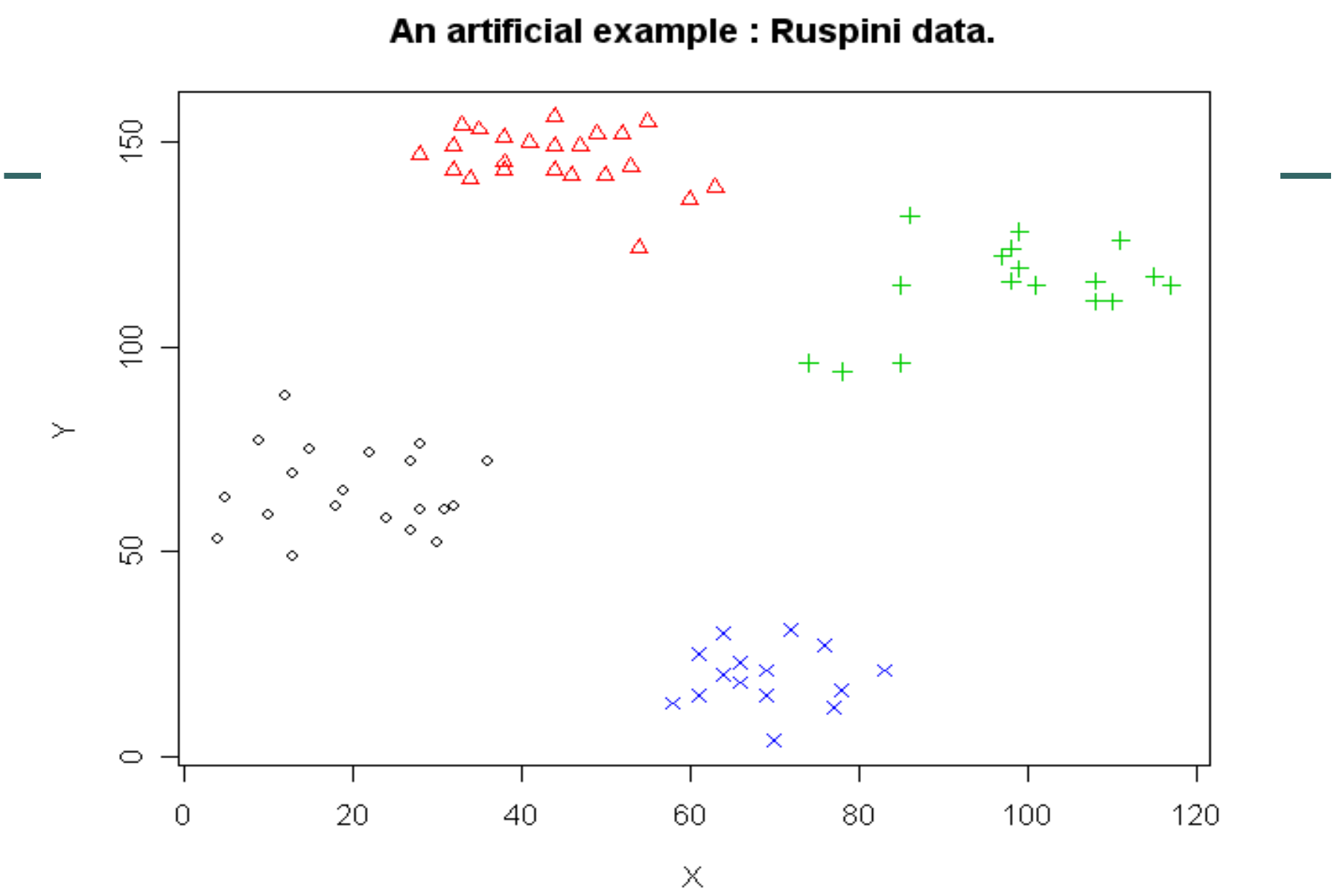
- Astronomers
  - Types of galaxies, planets, stars
- Biologists
  - Families, genera, species
- Marketing people
  - Market segmentation
- Psychiatrists
  - Different types of suicide attempts
- Etc.

What is it?

- a form of unsupervised learning
  - related to k-NNC
- difference between clustering and classification
  - group membership must be known for classification
    - supervised learning
  - group membership is unknown prior to clustering
    - unsupervised learning = discovery procedure
- Not just a question of getting out the class structure you knew about

Basic idea:

- Clusters are
  - "continuous regions of [feature] space containing a relatively high density of points, separated from other such [clusters] by regions containing a relatively low density of points." -- Everitt (1980).


An artificial example : Ruspini data.
Y
X
0
50
100
150
0
20
40
60
80
100
120

Most important point!

- There is NO "right answer"
- Danger of evaluating by correspondence to pre-existing grouping
  - more valuable to find unexpected groupings
- (more on evaluating clusterings later)

Outline:

1. Basic ideas
2. Clustering methods
3. Clustering problems
4. Odds & ends

2. Clustering Methods:

- Major design choices:
  - what sort of clustering strategy?
  - which variables are to be used?
  - how is distance/similarity to be measured?
  - what criteria will be used to link cases into clusters & clusters into bigger clusters?

2.1 Clustering Strategies:

- Incremental
  - Agglomerative
  - Divisive
- Iterative

- [Weka rather offbeat]

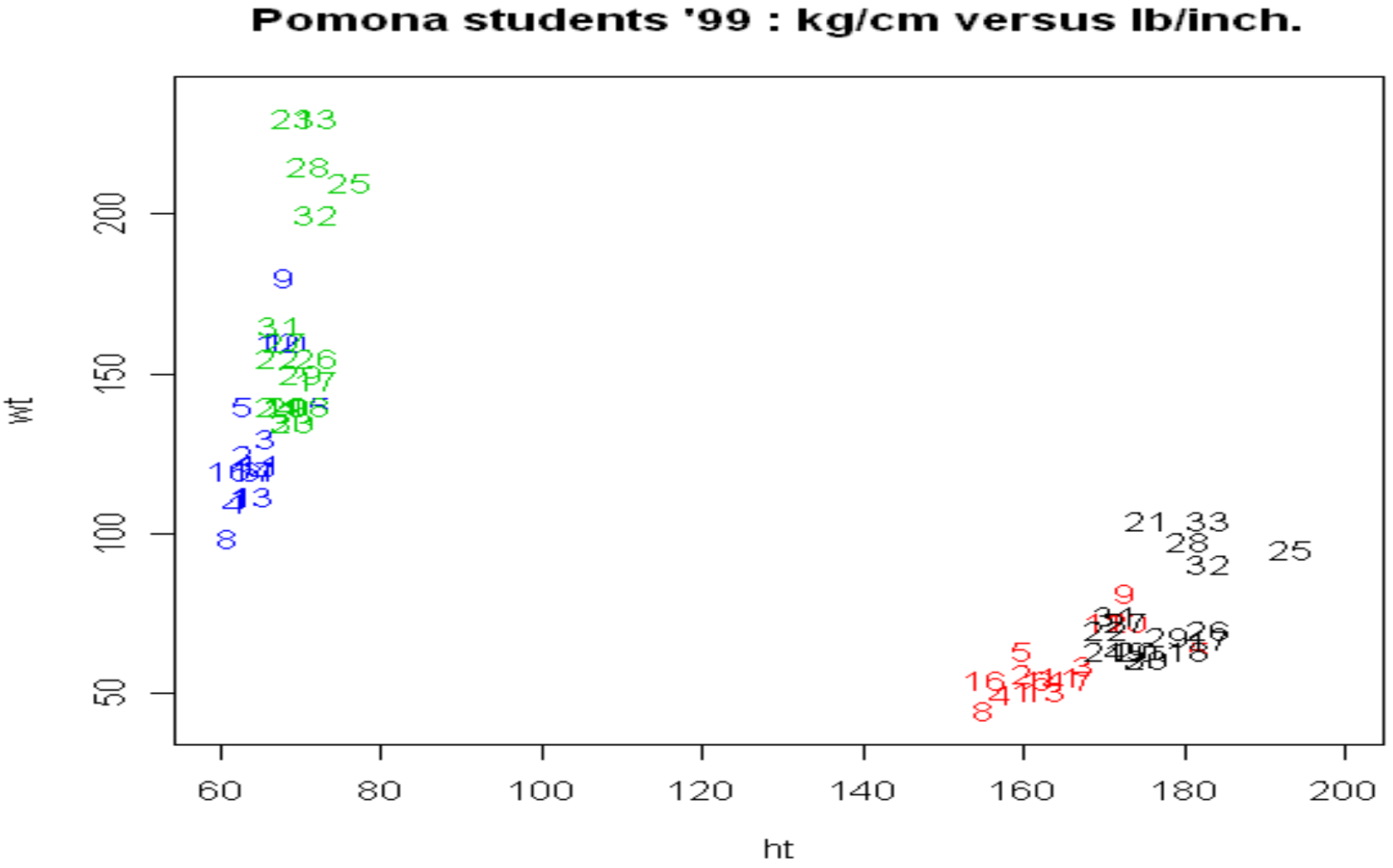
2.1 K-means pseudocode:

- Input k = number of clusters sought
- Pick k random cases as initial cluster centroids
- Repeat
  - Assign each case to nearest cluster centroid
  - Measure total distance of each case to its cluster centroid
  - Calculate new cluster centroids
  - Until convergence ## no change
- Display results
- ## basically just hill-climbing (local optimum).
- ## best to have several re-starts.

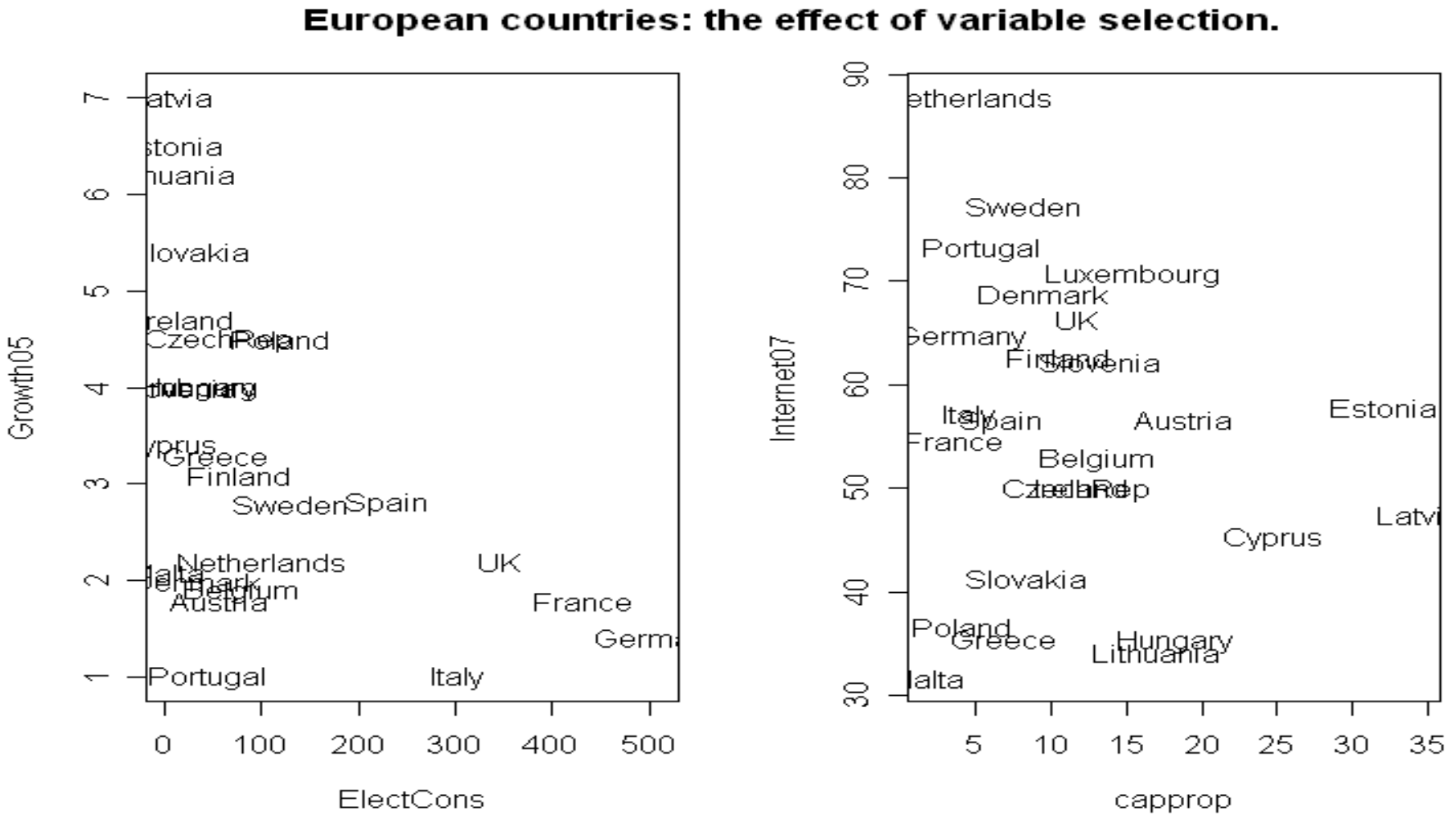
2.2 Variable Selection & Scaling:

- Selection of variables:
  - crucial to result of clustering
  - often helpful to do Principal Components Analysis first (Witten & Frank, 7.3)
- Variable Scaling:
  - customary to standardize using z-scores
    - $z_{ij} = (x_{ij} - m_j) / s_j$
    - each variable relative
    - avoids arbitrary variable weighting

2.2 Variable Scaling:



2.2 Variable Selection:



2.3 Distance measures:

- Euclidean distance
  - $d(i,j) = sqrt(\sum(z_{ik} - z_{jk})^2)$
- City-block distance
  - $d(i,j) = \sum abs(z_{ik} - z_{jk})$
- Canberra metric
  - $d(i,j) = \Sigma(abs(z_{ik} - z_{jk})/(z_{ik} + z_{jk}))$
- What about other data types?
  - nominal, ordinal, interval, ratio
  - string?
    - String similarity metrics (computationally expensive)

2.4 Methods of Combining Clusters:

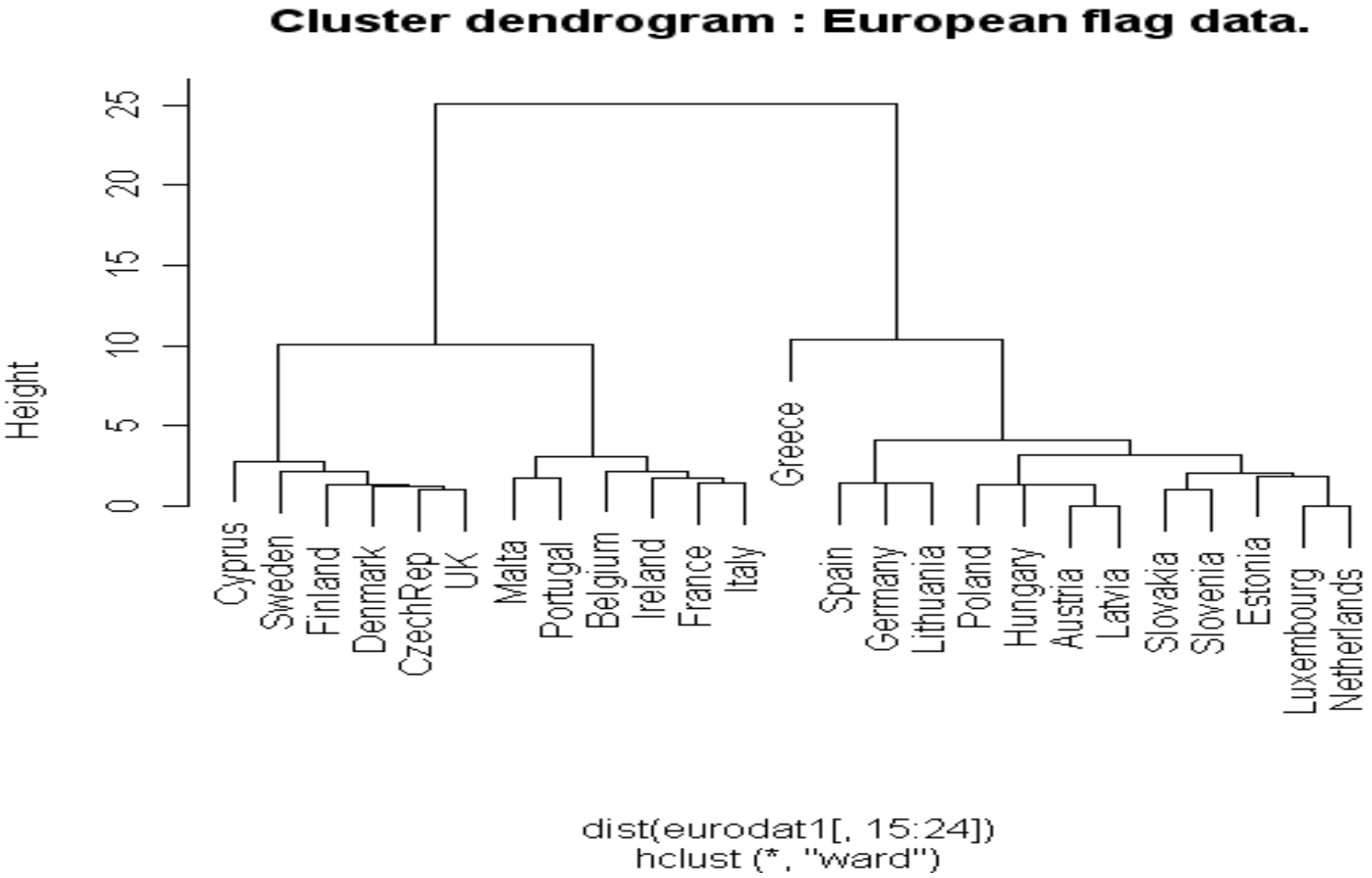
- Single linkage (nearest neighbour)
- Complete linkage (furthest neighbour)
- Average linkage
- Centroid linkage

- (RF recommends Ward's method as default)

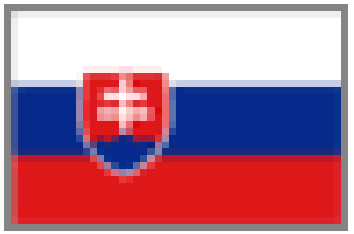
2.5 Displaying a Clustering:

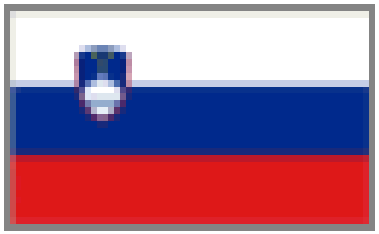
- Hierarchic methods:
  - Dendrogram
  - Icicle plot
    - (Ugly to my eyes)
- Iterative methods:
  - No standard graphical mode
    - (Will show k-means plot later)

2.6 Example Dendrogram:

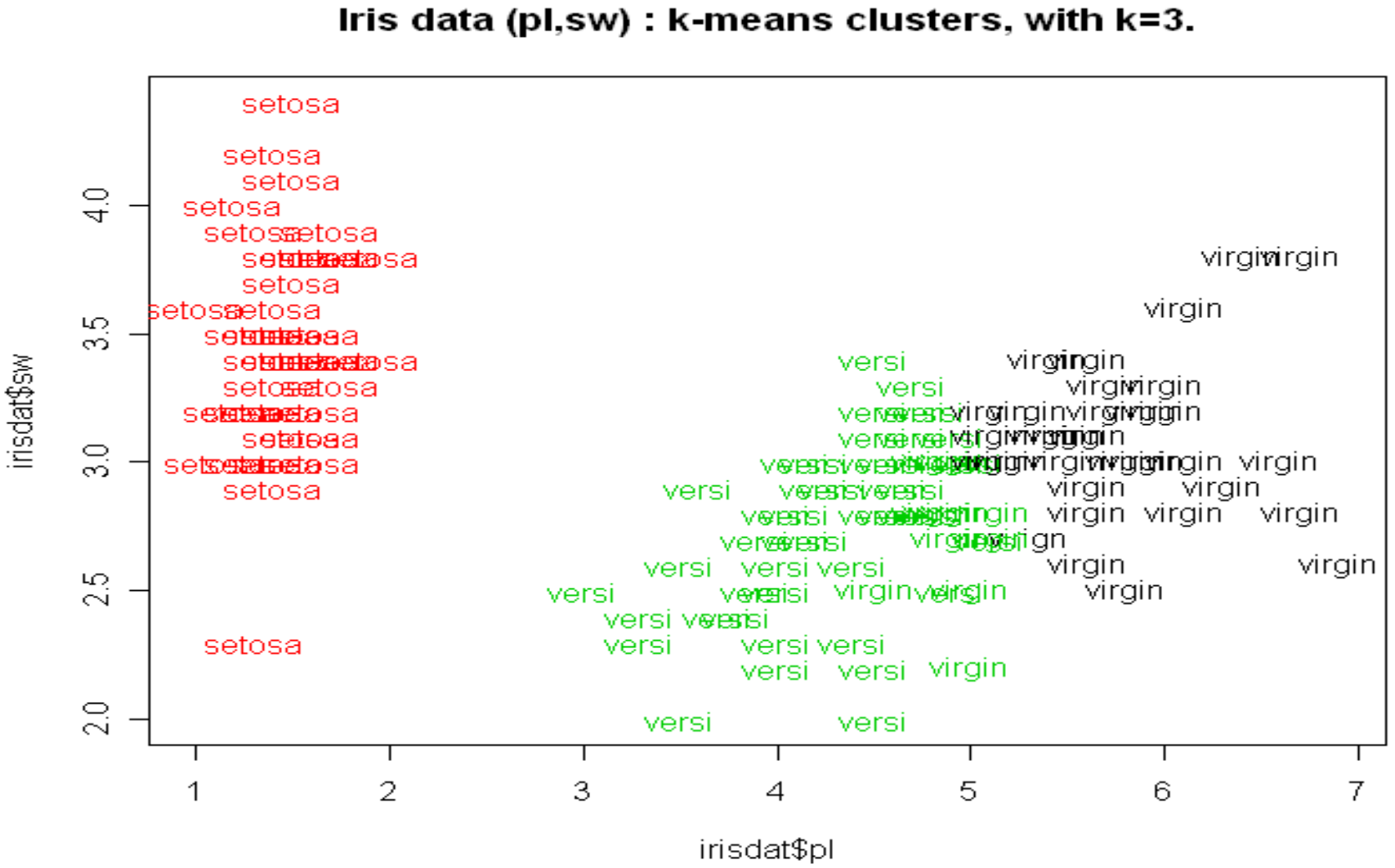


Example Flags (Lux, NL, Slovakia, Slovenia)





2.5 Displaying the results of a non-hierarchical method:



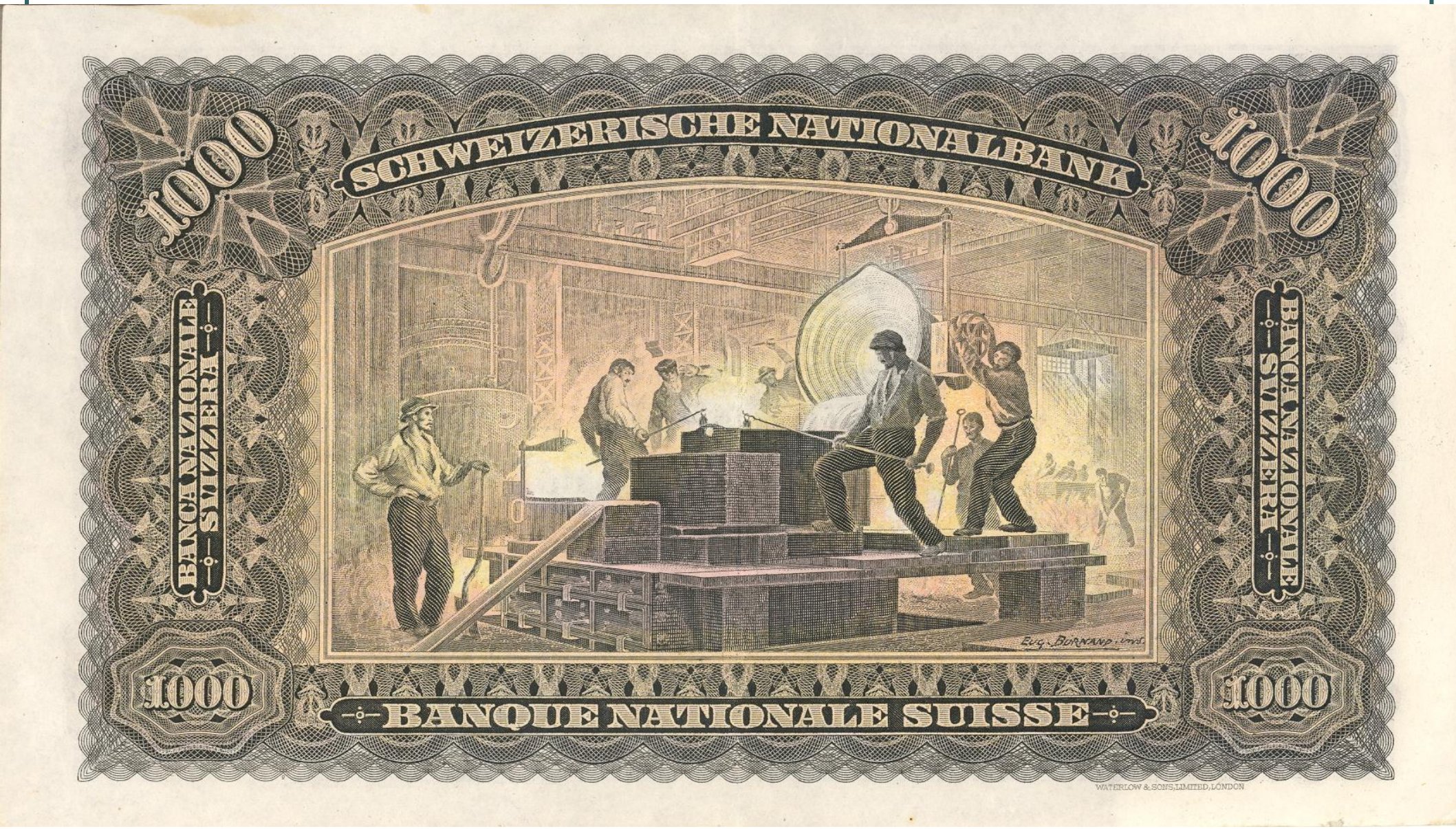
3. Problems with Cluster Analysis:

- Choosing an appropriate number of clusters
- Validating a clustering

3.1 How many clusters ?

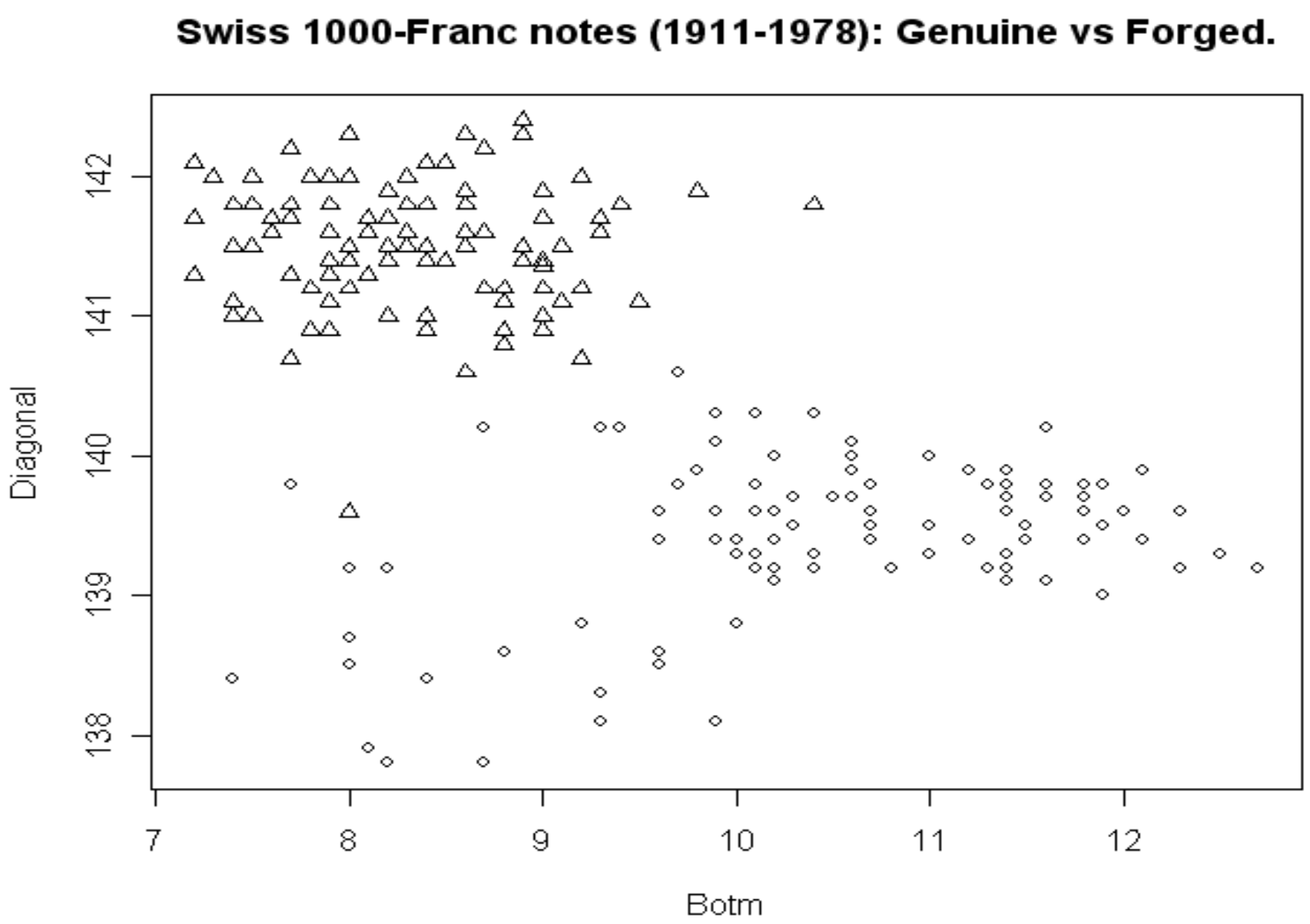
- Add a wrapper round k-means
  - try g-1 to g+1
    - pick best on quality index
  - rule of thumb for guess of g:
    - g = round(ln(n+1))

Swiss 1000-Franc note (1911-1978):

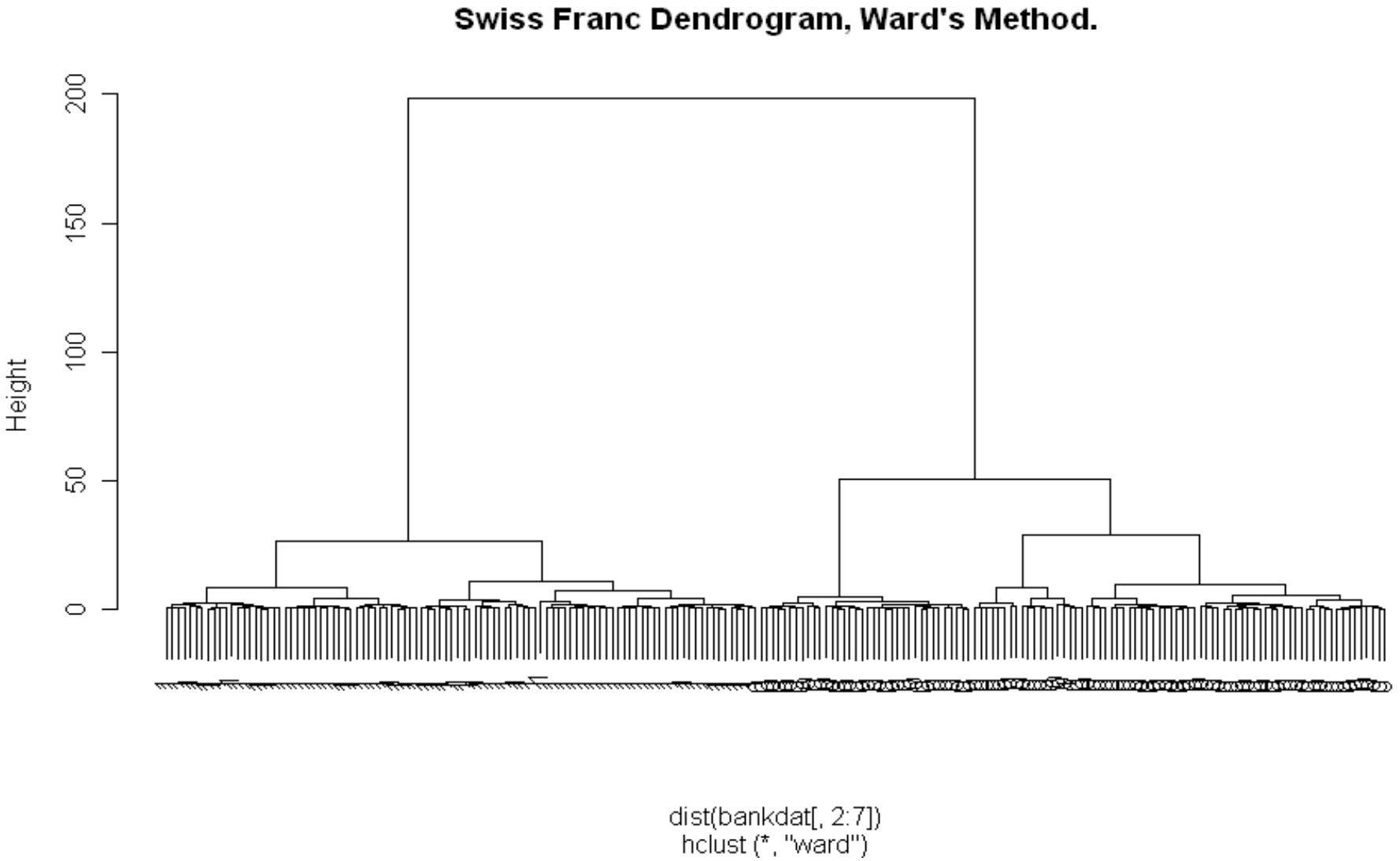


Swiss Franc Data:

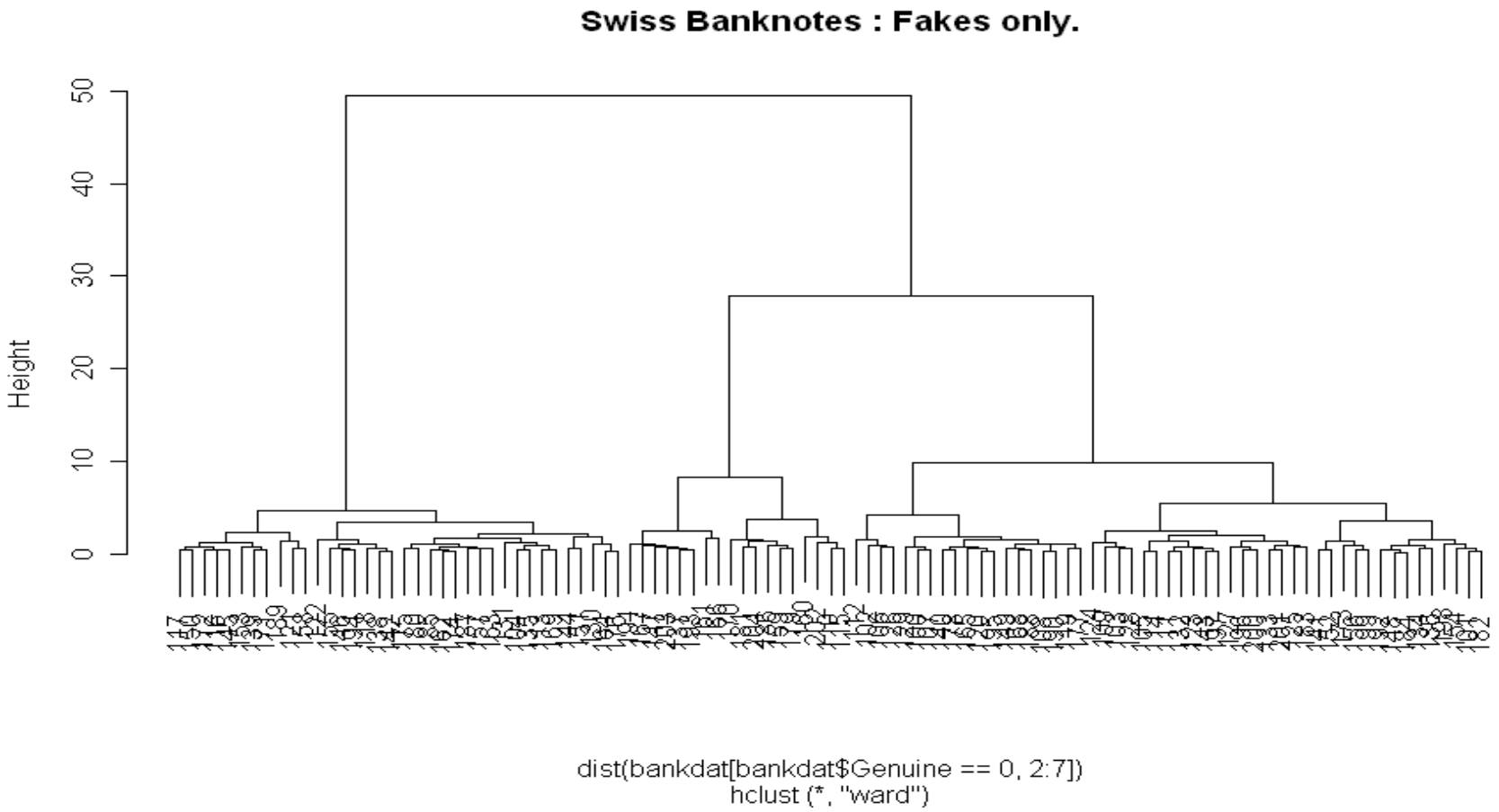
- Flury & Riedwyl (1988)
- 205 1000-Franc notes
- Six measures (in mm) :
  - Breadth
  - Lheight
  - Rheight
  - Botm
  - Topm
  - Diagonal


Swiss 1000-Franc notes (1911-1978): Genuine vs Forged.
Diagonal
142
141
140
139
138
7
8
9
10
11
12
Botm

Swiss Francs dendrogram:



Swiss Banknotes : Take Fakes.



3.2 How to validate a clustering ?

- Quality measures
  - C = (trace(B)/(g-1)) / (trace(W)/(n-g))
  - CU : Category Utility
  - "Silhouette coefficient“
  - MDL principle
- Re-sampling approach
  - quasi-cross-validation

4. Future directions:

- Long-term answer must be evolutionary optimization
- Fitness function ~ cluster quality:
- Problem:
  - Impurity measures always reduce as number of clusters increases
  - Therefore need an approach based on the MDL principle:
    - Cost = cost(cluster_spec) + cost(data | cluster_spec)
    - See Witten & Frank, p. 181-184
  - Still a research issue

Further reading:

- Dunham, M. (2003). Data Mining. Pearson Educational.
- Everitt, B.S. (1993). Cluster Analysis, 3e. London: Arnold.
- Flury, B. & Riedwyl, H. (1988). Multivariate Statistics: a Practical Approach. C.U.P.
- Murtagh, F. & Heck, A. (1987). Multivariate Data Analysis. Dordrecht: Kluver.

- Witten & Frank: 4.8, 6.6, 10.6 (7.3)

Relevant websites:

- http://astro.u-strasbrg.fr/~fmurtagh/mda-sw/
- http://darwiniana.org/intro1.htm
- http://lib.statlib.cmu.edu/
- http://thames.cs.rhul.uk/~fionn/strule/books/holdall.pdf
- http://www.r-project.org/
- http://www-users.cs.umn.edu/~kumar/dmbook/ch8.pdf

Practicalities:

- Martin Sykora has Eurodata
- Tomorrow: practice using Weka
- Will transfer ppt to pdf & then Sameer will put onto course website
- Next week: text mining case study