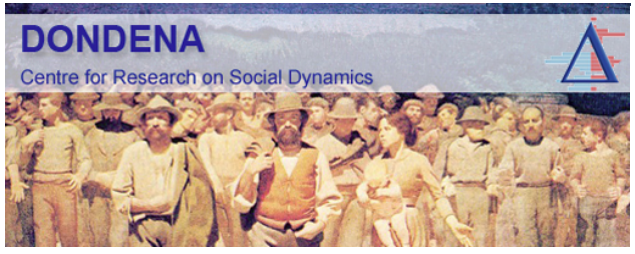
Clustering and analysis of typical trjectories

Nicola Barban

Carlo F. Dondena Centre for Research on Social Dynamics
Università Bocconi

Madrid. March 7-8, 2011



Università Commerciale
Luigi Bocconi

dissimilarity measures

- Distance between sequences Different metrics (LCP, LCS, OM, HAM, DHD, ...)
- A dissimilarity is a quantification of how far two objects are. For instance, consider two incomes x and y:
  - $d(x, y) = (x - y)^2$
  - $d(x, y) = |x - y|$
  - $d(A_{x_1,y_1}, B_{x_1,y_2}) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$
- Optimal Matching, or LCS, DHD, ... compute distances for categorical trajectories?

Cluster

- Cluster analysis automatically classify different objects in a reduced number of categories.
- It simplifies the large number of distinct sequences in a few different types of trajectories.
- It is used to build a typology of the trajectories. It offers a descriptive approach to analyze the sequences.

Cluster

- Clustering always start from a distance matrix. Usually euclidean distances between variables
- But clustering may be done using a dissimilarity matrix.
- Several methods for agglomerating observations in cluster procedures
- Usually iterative procedure. At every step the most “similar” observations are grouped

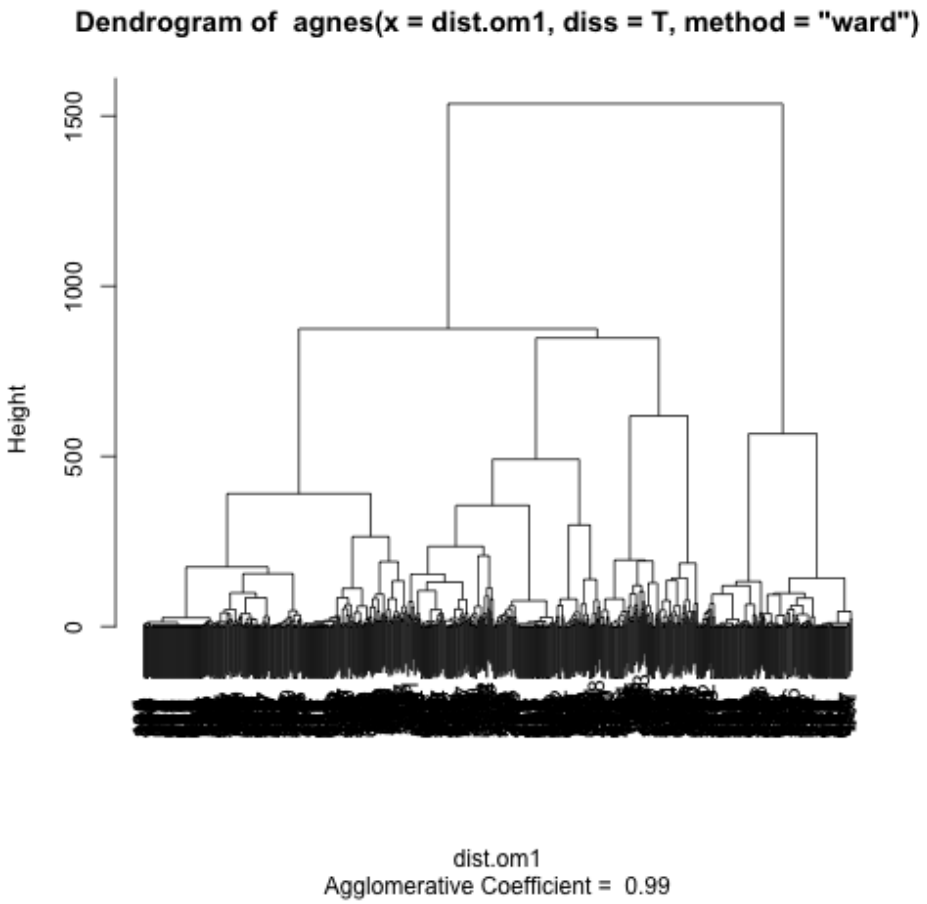
Ward clustering

- Ward is a hierarchical clustering algorithm.
- At each step, it joins together the two less distant groups.
- Ward aims at minimizing the within cluster discrepancy.

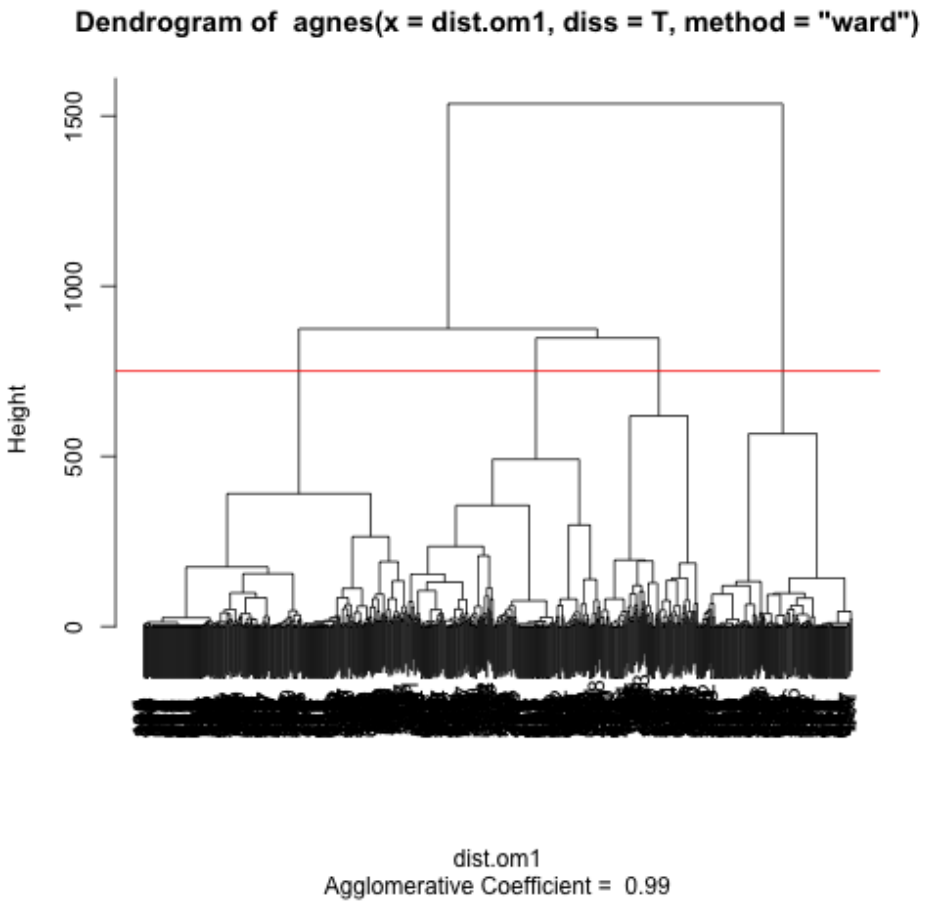
Number of clusters

- The number of clusters needs to be chosen by the researcher
- Several way to do that. No best method
  1. Theory driven. You have some reason to believe that the best number of group is . . .
  2. Description of the clusters. Try different solutions
  3. Dendogram

Dendogram


Dendrogram of agnes(x = dist.om1, diss = T, method = "ward")
Height
0
500
1000
1500
dist.om1
Agglomerative Coefficient = 0.99

Dendogram pruning



Analysis of cluster

- Check the sample size of each cluster. You don't want to have too small clusters
- Check the distribution of clusters. Do you have "residual" clusters
- Try one less clusters. Check distribution
- Be parsimonious.

Medoid

- Clusters can be described by their “center”
- This is called centroid sequence or **medoid**
- What is the sequence that is more “central”?
- “centrality” is equivalent less distance.
- The medoid distance is the sequence that is less distant in average to all the other sequences in the cluster

Medoid 2

- Medoid are **real** sequence
- Easy to describe!
- (S-12)-(C-6)-(M-24)
- (S-6)-(C-03)-(S-09)-(M-12)-(S-12)

Exploring clusters

Three types of graphics:

- Transversal distribution with `seqdplot()`
- Frequency plots with `seqfplot()`
- Individual index-plots `seqiplot()`

Use `group = cluster.membership.factor` to get plots by clusters

Determinants of trajectories

- It is possible to estimate the influence of independent covariates on the probability of belonging to a given cluster (i.e. type of trajectory) rather than another.
- We can fit, for instance, a logistic (multinomial) regression model
- Class membership can be used for further analysis

logistic regression

```
> summary(jobless.reglog)

Call:
glm(formula = jobless ~ male + funemp + gcse5eq, family = bino
    data = mvad)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8116  -0.5948  -0.5813  -0.3565   2.3613

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.64230    0.19297  -8.510  < 2e-16 ***
maleMen     -0.05032    0.22333  -0.225 0.821748
funempyes    0.70083    0.25466   2.752 0.005923 **
gcse5eqyes  -1.03169    0.27872  -3.702 0.000214 ***
---
Signif. codes:  0 ‚Äò***‚Äô 0.001 ‚Äò**‚Äô 0.01 ‚Äò*‚Äô 0.05 ‚Ä

(Dispersion parameter for binomial family taken to be 1)
```