Building Typologies of Individual Trajectories: An Overview of Statistical Methodologies

Yvette Grelet  Céreq-CMH Caen University (France)
Nicolas Robette – INED (France)
Plan

- 1. Reminders about clustering methods
- 2. Specificities of longitudinal data
- 3. Typologies of pathways
- 4. An example

- Conclusion
1- Reminders about clustering methods
Goals and general principles

- **Clustering** is the classification of objects into different groups (partitioning into clusters).
- A way to **explore** a data set, particularly adapted to the case of complex and numerous data.
- « Discover » (or confirm) some underlying pattern: the existence of groups (in case of heterogenous data); or create instrumental classes.

Methods may refer to a notion of proximity, to a model, to a notion of density. We will focus on the first approach: « **Gather what is alike** »
4 main steps

1. Data preparation: selection, coding and transformation, organisation
2. Dissimilarity measure between objects
3. Clustering method (measure of dissimilarity between sets of objects)
4. Consolidate the results; interpret them
1.1. Data preparation: selection, coding, transformation, organization

- Data selection
  - Individuals (elimination of outliers)
  - Variables (Homogeneity of data)
1.1.1. Data preparation: selection of variables. On which attribute are things alike? (1)
On which attribute are things alike? (2)
On which attribute are things alike? (3)
1.1.2. Data preparation: coding, transformation, organization (1)

- Coding, transformation
  - Quanti + quali → quali (homogeneisation)
  - Imputation of missing data...

- Organization
  - Table with I rows (individuals = observations) and J columns (variables)
  - If distance matrix → coordinates
1.1.2. Data preparation (2)

- Highly recommended, if possible (often) : to perform a factor analysis (CA or PCA) before Clustering, in order to:
  - Plot the data on the factorial planes and see the shape of the cloud (well separated clusters or continuum?)
  - Detect outliers
  - Perform the clustering on the first components (data transformation which eliminates noise, makes data homogeneous)
1.2. Measure of likeness between observations

- Quantitative data: a distance
  Euclidean distance: $D^2(i,i') = \sum_j w_j (X_{ij} - X_{i'j})^2$
  Where $w_j$ is the weight of variable $j$.

- Qualitative data: dissimilarity coefficient based for example on the number of concordances and discordances (a great variety of such indexes).
1.3. Choice of a clustering procedure

- Two groups:
  - Hierarchical clustering
  - Partitioning

- And neural networks
1.3.1. Hierarchical clustering

- At each step
  - Two clusters are agglomerated (hierarchical ascending classification - HAC)
  - Or one cluster is split into two smaller clusters (descending classification)

Clusters form a tree structure (dendrogram)
Examples of linkage rules

Single linkage, Complete linkage, Between centroids, Group average
A very usual agglomerative criterion: Ward’s

- Ward’s criterion: maximize variance between (minimize within) groups
A general frame: « Flexible » distances

- Lance and Williams formula

\[ d(C, A \cup B ) = \alpha_A d(C,A) + \alpha_B d(C,B) + \beta d(A,B) + \gamma | d(C,A) - d(C,B) | \]

with \( \alpha_A + \alpha_B = 1; \beta < 1; |\gamma| < 1 \)

Examples:
- Single linkage: \( \alpha_A = \alpha_B = 1/2 \)
- Complete, Centroid, Median, Group average, Ward,
- Flexible beta…
1.3.2. K-means partitioning

- Number k of clusters set *a priori*
- k seeds are selected as first guess of the means
- Observations agglomerated according to their *distance* to the seed, into k clusters
- Agglomeration around the centroid of the clusters
- Aso until convergence
- Possibly start over again with other sets of k seeds, to obtain several classifications (sets of k disjoint clusters). Crossed classif. gives some stable clusters.
1.3.3. Hybrid clustering (Wong-Lebart)

- Preliminary clustering with a k-means method to search stable groups, which may be numerous
- Hierarchical clustering of the previously obtained clusters, cutting of the dendrogram: determination of the final number of clusters k
- The centers of the previous clusters are taken as seeds for a partitioning procedure in order to consolidate the clusters
1.4 Interpretation of results

- Compute the average profile of clusters (active variables, individual characteristics)
- Factor analysis combined with cluster analysis after clustering, in order to:
  - See the position of the clusters and their centroid on the factorial planes → interpretation of clusters/factors
  - Characterize the clusters
- Display the representative elements of each cluster (the type)
  - Fictitious modal element of each cluster
  - The real element(s) which are near from the centre of gravity of the cluster
2- Longitudinal data
2.1. General frame: Three-way data table
The 3-way table and its (un)foldings

\[ \sum_{t} \]

\[ i \quad j_1 \quad j_2 \quad \ldots \quad j_T \]
Factor Analysis of 2-way tables

\[
\sum_{t} i_j
\]
On the factorial planes: Trajectories of individuals (or groups of individuals)
On the factorial planes: Time-evolution of variables
2.2. Example

- Céreq’s survey on transition from school-to-work
Céreq’s « Generation 98 »’ monthly calendar

State  At school  Employed  Unempl  National Service  Other

MOIS  1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39


PAULE

PIERRE

JEAN

RENÉE
Display of individual transition pathways

Two types of graphs:

- Chronograms
- « Carpets »
Monthly histogram / Individual pathways

Generation 98’ cohort

Generation 98’ pathways
3- Typologies of pathways
Goals

Explore the diversity of pathways:
  ex) 5 states, 12 months → $5^{12}$ (>200M) possible pathways

Identify clusters of « similar » observations, types of transition pathways

Summarize pathways:
Compute a new variable, which may be used in a model
  □ Factor analysis → interval variable
  □ Cluster analysis → nominal variable
A series of decisions

- **Upstream** from clustering procedure:
  - How describe pathways?
    - Selection of the relevant information (which states?)
    - Coding of the information (monthly calendar, indicators, …)
  - How measure dissimilarity (choice of a distance)

- **Choice of a clustering method** (agglomerative procedure)

- **Downstream**: understand why observations are in the same cluster (characterize clusters)
Coding Monthly Calendars

- Dummy coding restoring the whole information:
  \( S = \text{nbr of states}, \ M = \text{nbr of Months} \rightarrow \text{SM variables } x_{sm} \)
  with value 1 if observed state \( s \) on month \( m \), 0 otherwise.

- Transitions between states (\( S^2 \) values)

- Both

- Summarized calendar = Qualitative Harmonic Analysis

- More summarized: Indicators
  ex) time spent (or %) in each state, nbr of spells for each state, duration before access to a particular state, etc.
Choice of a dissimilarity index

- Depends on the nature of variables resulting from the coding (quali/quanti)

- Examples with monthly calendars:
  - Euclidean distance (discordance)
    \[ D_L^2(i,i') = \sum_m \delta_{i'm} \text{ where } \delta_{i'm} = 1 \text{ if } i \text{ and } i' \text{ are in different states during month } m, 0 \text{ otherwise} \]
  
  - Chi-square distance
    (row i, column sm, dummy coding)
    \[ D_C^2(i,i') = \frac{1}{M} \sum_{sm} \left( \frac{1}{f_{sm}} \right) (k_{ism} - k_{i'sm})^2 \]
To weight or not to weight distances

Why? Take account of:

- Proximity between states
- Frequency of states’ occurrence
- Time (proximity and frequency may vary over time)
Qualitative Harmonic Analysis

- Deville et Saporta (1980)

- Method to describe « a set of individuals characterized by a career, that is (to say) by an evolution in a finite set of states » (Deville, 1982)

QHA: methodology

- Choice of a common period of observation and division of the period into intervals
- Measuring, for each interval, the proportion of time spent in each state
- Factor analysis of the created variables
- Cluster Analysis on the factors

→ Typology of trajectories
Split of the period

- Choice of the observation period:
  - The same for all individuals
  - Censoring is not taken into account

- How to split the period:
  - Split into intervals of equal durations,
  - Split according to events quantiles,
  - Arbitrary split…

- Number of intervals:
  - If too high: many variables equal 0 ➞ sparse matrix and bad quality factor analysis
  - If too low: important reduction of information ➞ bad quality description
Coding the variables

- Choice of states nature and number:
  - Same trade-off as for number of intervals

- Computation of the proportion of interval duration spent in a state
  \[ \text{nb of variables} = \text{nb of states} \times \text{nb of intervals} \]

- Possible refinement (Degenne et al, 1995):
  - Addition of series of variables calculating the proportion of each transition between two states
    \[ \text{nb of variables} = (\text{nb of states})^2 \]
Example

Trajectory = A B B B A

- 2 states: A and B
- Period: t1 to t5, split into 2 intervals: [t1; t2] and [t3; t5]

⇒ 4 + 4 variables:

<table>
<thead>
<tr>
<th>interval [t1; t2]</th>
<th>interval [t3; t5]</th>
<th>transitions</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>A</td>
<td>AB</td>
</tr>
<tr>
<td>B</td>
<td>B</td>
<td>AA</td>
</tr>
<tr>
<td>0,5</td>
<td>0,33</td>
<td>0,25</td>
</tr>
<tr>
<td>0,5</td>
<td>0,66</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0,5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0,25</td>
</tr>
</tbody>
</table>
Factor analysis

- Correspondence Analysis on the variables

- Only a part of the factors are retained, so as to discard « noise » while keeping the major part of information (70% to 80% of inertia)

- Hierarchical Cluster Analysis on selected factors
Description of trajectories by means of QHA

3 dimensions compose a trajectory:

- DURATION spent in each state: measured by duration variables
- MOMENT when one’s in each state: introduced by the split into intervals
- TRANSITIONS from a state to another: measured by transition variables
QHA limits

- Problem of the censorship not taken into account
- Sequence of states not taken into account
Sequence analysis

- Individual trajectories are built as sequences of states.

Example: a school-to-work trajectory (8 months):

- S = student; U = unemployed; W = wage-earner

<table>
<thead>
<tr>
<th>m1</th>
<th>m2</th>
<th>m3</th>
<th>m4</th>
<th>m5</th>
<th>m6</th>
<th>m7</th>
<th>m8</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>S</td>
<td>S</td>
<td>U</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>W</td>
</tr>
</tbody>
</table>

- Then they are grouped together according to their degree of similarity technique = optimal matching analysis (OMA), ...

⇒ Typology of trajectories
Optimal Matching Analysis (1)

- Method born from molecular biology (DNA strings)
- Introduced in social sciences by Andrew Abbott in the 80’s
- **Principle**: measuring dissimilarity between pairs of sequences by calculating the cost of the transformation of one sequence into the other

*see e.g. (Abbott & Tsay, 2000)*
Optimal Matching Analysis (2)

- 3 basic operations:
  - insertion
  - deletion
  - substitution

- Each operation is assigned a cost

- The distance between two sequences is equal to the minimal cost needed to transform one sequence into the other

- A cluster analysis of the distance matrix (comparison of all pairs of sequences) gives typologies
Optimal Matching Analysis (3)

Example:

X: B B A B A B A B

Y: B A B A B B

→ 4 substitutions

→ 1 insertion, 1 deletion
The choice of costs (1)

- Crucial issue of OMA: sequence = event + time

- **Substitution**: retains the temporal structure (moment) but distorts events structure (order)

  **Insertion/deletion**: distorts time but retains the order of events

- A common choice: indel=1, subst=2
The choice of costs (2)

- Substitution costs:
  - According to theoretical assumptions:
    - ex) stratification
  - Data driven:
    - ex) based on transition likelihoods
Criticism

- Different lengths: censoring or process?
  - Variable indel costs (Stovel, Bolan, 2004)

- Substitution does not take order into account
  - Common sub-sequences (Dijkstra, Taris, 1995; Elzinga, 2003)

- Trajectories are time-dependent
  - No indel operations, substitution costs computed for each moment (Lesnard, 2004)
Conclusion

- exploration of complex trajectories
- complementarity with stochastic approaches
- robustness
- flexibility

- reflection about data and research question: 
  *through coding; costs (OMA)*...

- OMA: possible varying length of pathways (but… the sense of it?)
4- EXAMPLE
2400 young people having exited school in 1992, surveyed in 1997

64 Months’ calendar

8 States:

- Unlimited term contract (CDI)
- Limited term contract (CDD)
- Stabilisation from unstable to stable contract (CDD-DI)
- Employment and training scheme (CQ)
- Part-time Employment scheme without training (CES)
- Unemployment (RE)
- Inactivity (INA)
- Return to education, training (FOR)

MCA of the table: 2400 rows x 512 columns (8x64)

Total time spent in each state as supplementary variables
MCA – Plot F1xF2, individuals
MCA – Plot F1xF2, 8 states (nbr of months)
MCA – Plot F3xF4, individuals
MCA – Plot F5xF6, individuals
MCA – Plot F7xF8, individuals

<table>
<thead>
<tr>
<th>P.0</th>
<th>1680</th>
<th>529</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>879</td>
<td>240</td>
</tr>
<tr>
<td></td>
<td>1200</td>
<td>1444</td>
</tr>
<tr>
<td></td>
<td>&gt; 764</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>1.6</th>
<th>21300</th>
<th>0</th>
<th>0</th>
<th>13998</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>8160</td>
<td></td>
<td></td>
<td>1460</td>
</tr>
<tr>
<td></td>
<td>6640</td>
<td></td>
<td></td>
<td>1050</td>
</tr>
<tr>
<td></td>
<td>1020</td>
<td></td>
<td></td>
<td>1360</td>
</tr>
<tr>
<td></td>
<td>1190</td>
<td></td>
<td></td>
<td>1930</td>
</tr>
<tr>
<td></td>
<td>1310</td>
<td></td>
<td></td>
<td>2170</td>
</tr>
<tr>
<td></td>
<td>1380</td>
<td></td>
<td></td>
<td>2310</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>1.0</th>
<th>&gt; 1689</th>
<th>23360</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1410</td>
<td></td>
<td></td>
<td>2030</td>
</tr>
<tr>
<td></td>
<td>1690</td>
<td></td>
<td></td>
<td>2410</td>
</tr>
<tr>
<td></td>
<td>1770</td>
<td></td>
<td></td>
<td>2550</td>
</tr>
<tr>
<td></td>
<td>1790</td>
<td></td>
<td></td>
<td>2570</td>
</tr>
<tr>
<td></td>
<td>1830</td>
<td></td>
<td></td>
<td>2630</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>0.8</th>
<th>&gt; 1297</th>
<th>11945</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1100</td>
<td></td>
<td></td>
<td>1400</td>
</tr>
<tr>
<td></td>
<td>1190</td>
<td></td>
<td></td>
<td>1500</td>
</tr>
<tr>
<td></td>
<td>2190</td>
<td></td>
<td></td>
<td>2660</td>
</tr>
<tr>
<td></td>
<td>2210</td>
<td></td>
<td></td>
<td>2750</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>0.6</th>
<th>&gt; 1218</th>
<th>11895</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1110</td>
<td></td>
<td></td>
<td>1420</td>
</tr>
<tr>
<td></td>
<td>1230</td>
<td></td>
<td></td>
<td>1540</td>
</tr>
<tr>
<td></td>
<td>2190</td>
<td></td>
<td></td>
<td>2680</td>
</tr>
<tr>
<td></td>
<td>2220</td>
<td></td>
<td></td>
<td>2780</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>0.4</th>
<th>&gt; 1203</th>
<th>11895</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1100</td>
<td></td>
<td></td>
<td>1400</td>
</tr>
<tr>
<td></td>
<td>1200</td>
<td></td>
<td></td>
<td>1500</td>
</tr>
<tr>
<td></td>
<td>2190</td>
<td></td>
<td></td>
<td>2680</td>
</tr>
<tr>
<td></td>
<td>2220</td>
<td></td>
<td></td>
<td>2780</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>0.2</th>
<th>&gt; 1094</th>
<th>11895</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1100</td>
<td></td>
<td></td>
<td>1400</td>
</tr>
<tr>
<td></td>
<td>1200</td>
<td></td>
<td></td>
<td>1500</td>
</tr>
<tr>
<td></td>
<td>2190</td>
<td></td>
<td></td>
<td>2680</td>
</tr>
<tr>
<td></td>
<td>2220</td>
<td></td>
<td></td>
<td>2780</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>0.0</th>
<th>&gt; 1036</th>
<th>11895</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1100</td>
<td></td>
<td></td>
<td>1400</td>
</tr>
<tr>
<td></td>
<td>1200</td>
<td></td>
<td></td>
<td>1500</td>
</tr>
<tr>
<td></td>
<td>2190</td>
<td></td>
<td></td>
<td>2680</td>
</tr>
</tbody>
</table>
MCA Outcomes

- The successive 2-dimensional plots show that:
  - There is no evident clump of individuals, but rather a continuum, at least on the first axis.
  - The main differentiation between pathways is due to occurrence of states: did youngsters experience unemployment -stable contract, inactivity, etc. - or not?
  - Time variation is minor compared to « synchronic pattern » (appears only after the 6th factor).
HAC of 2400 pathways (50 factors) → 9 clusters
One among 9 clusters: youngsters who went through a CES (employment scheme)
Split into 4 clusters: time of occurrence
5- Conclusion - discussion

- One cluster analysis for one question
- The classification used to distinguish between states (events) has a decisive effect on the result
- Also the way data are transformed (monthly calendar, set of periods, indicators …)
- If possible Cluster analysis and Factor analysis (complementary methods)
- How many clusters?
Thank you for your attention
References (1)


References (2)
