Sequence analysis

Leonard Moulin

Université Paris 7 Diderot - LADYSS (CNRS UMR 7533)
leonard.moulin@ens-cachan.org

November 28, 2014
M2 EPOG - University Paris 13 - Econometrics & Economics of Education

Part 1. Theory

Introduction to sequence analysis

TraMineR

- TraMineR is an R package for sequence analysis.
- Specially designed for the social sciences.
- TraMineR: Trajectory Miner in R (Gewurztraminer wine ?)
- Freely available on the CRAN (Comprehensive R Archive Network)
  `http://cran.r-project.org/web/packages/TraMineR/`

Sequence analysis

- In the social sciences: survival analysis ⇒ using different types of modeling (mainly risk models and survival models) we focus on a particular event (marriage, first job, unemployment...).
- Development approach in terms of life course: whole trajectory (possibly with multiple transitions) as the unit of analysis

⇒ Holistic approach.

Sequence analysis

- In social sciences, sequences represent trajectories:
  - professional;
  - school;
  - sentimental;
  - residential;
  - union;
- With sequence analysis, we can study recurring patterns in the trajectories, taking into account the multiplicity of feasible states.
- A state is a situation in which an individual is at any given time (single, cohabiting, married, divorced, widowed).
- Sequences of states: a pathway is a sequence of states ordered along a time axis (single → married → widowed).

Sequence analysis

- Which characteristics of sequences are we interested in?
- What kind of indicators can we compute for a sequence set?
- What are suited plots for rendering sequences?
- How can we measure similarity between sequences?

Sequence analysis

With a more analytical or explanatory concern, we also consider issues such as:

- How can we identify groups with similar patterns and build typologies of sequences?
- How can we analyze the relationship of sequences with covariates?

Sequence analysis

Questions in social sciences:

- Do life courses obey some social norm? Which are the standard trajectories? What kind of departures do we observe from these standards ? How do life course patterns evolve over time ?
- Why are some people more at risk to follow a chaotic trajectory or to stay stuck in a state? How does the trajectory complexity evolve across birth cohorts?
- How is the life trajectory related to sex, social origin and other cultural factors?

Illustrative data set

- Study by McVicar and Anyadike-Danes (2002) on the transition from school to work in Northern Ireland.
- Dataset included in TraMineR.
- The aim of the study:
  - Describe the typical transition of young Irish.
  - Identified problematical trajectories. Distinguish between successful and unsuccessful transitions.
  - Understanding the factors that influences the trajectories.
  - Identified groups of young irish who have more problems to enter into the labor market.

Illustrative data set

- 712 individuals
- Follow-up starting at the end of the compulsory education (July 1993)
- Time series of 70 status variables: September 1993 to June 1999.
- The alphabet is made of the following statuses: EM (Employment), FE (Further Education), HE (Higher Education), JL (Joblessness), SC (School), TR (Training).
- Included in the TraMineR library.

Illustrative data set

Table: List of variables in MVAD

Variable	Description	Values
sex	Gender	female, male
region	Location of school in North Ireland	Belfast, North Eastern, South Eastern, Southern, Western
religion	Religion	Catholic, Protestant
funemp	Father unemployed at time of survey	yes, no
fmpr	Father has a professional, managerial or related job	yes, no
livboth	Living with both parents at time of first sweep of survey	yes, no
grammar	Grammar school secondary eduction	yes, no
gcse5eq	Qualifications gained by the end of compulsory education: 5 or more GCSEs at grades A–C, or equivalent	yes, no

Desciptive analysis of the sequences

State sequences

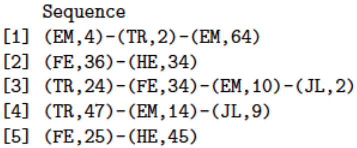
State sequences

We define a state sequence of length l as an ordered list of l elements successively chosen from a finite set A. We represent a sequence x by listing the successive elements that form the sequence $x = (x_1, x_2, ..., x_l)$, with $x_j \in A$.

- Two properties:
  - state sequences are formed by elements that are states;
  - the position of each element report information in terms of age, date or, more generally, past time or distance from the beginning of the sequence.

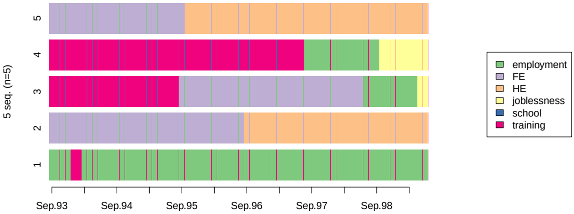
Visualizing individual state sequences

Individual state sequences



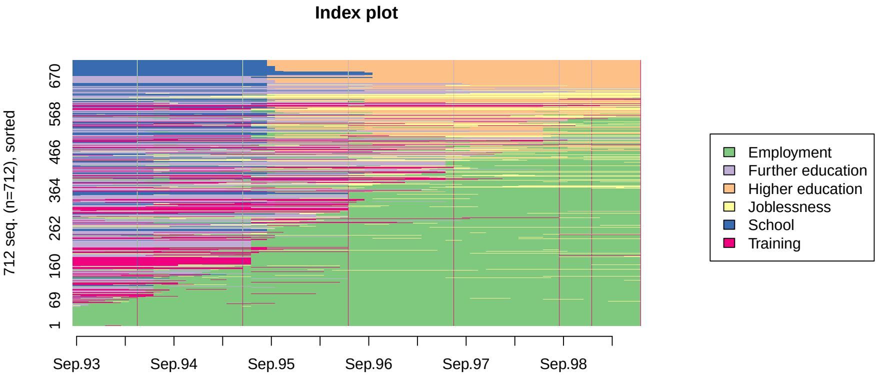
```
    Sequence
[1] (EM,4)-(TR,2)-(EM,64)
[2] (FE,36)-(HE,34)
[3] (TR,24)-(FE,34)-(EM,10)-(JL,2)
[4] (TR,47)-(EM,14)-(JL,9)
[5] (FE,25)-(HE,45)
```

Figure: Sequence index plot of sequences 1 to 5



Visualizing individual state sequences

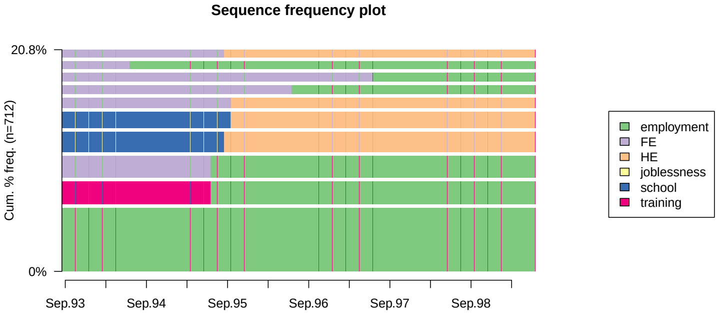
Figure: Individual state sequences



Visualizing individual state sequences

Sequences frequencies

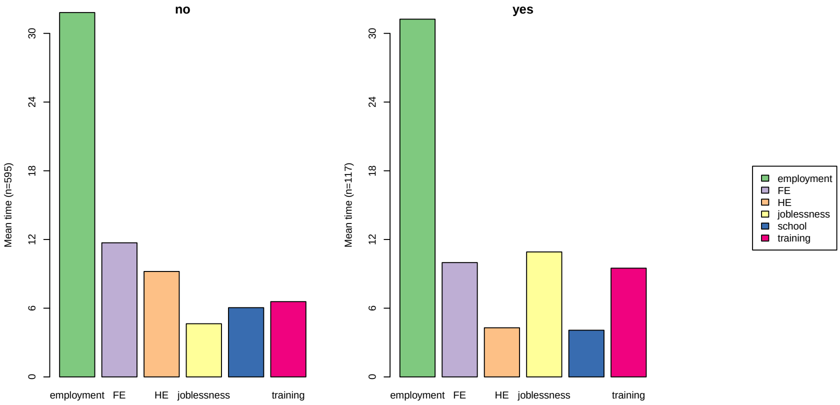
```
                   Freq Percent
...
FE/22-EM/48          14     2.5
TR/22-EM/48          18     2.5
EM/70                15     4.1
```



Overall and transversal statistics

Mean time

Figure: Mean time spent in each state by father's unemployment status

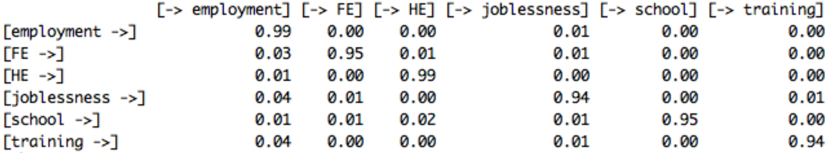


Overall and transversal statistics

Transition rates

The transition rate between (s_i, s_j) is the probability to switch at a given position from state s_i to state s_j . Let $n_t(s_i)$ be the number of sequences that do not end in t with state s_i at position t and let $n_{t,t+1}(s_i, s_j))$ be the number of sequences with state s_i at position t and states$_j$ at position $t + 1$. The transition rate $p(s_j|s_i)$ between states s_i and s_j is obtained as:

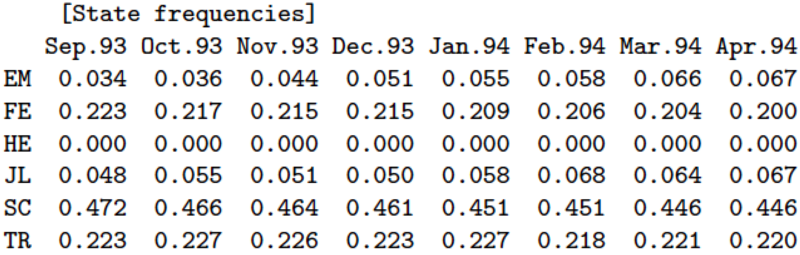
$$p(s_j|s_i) = \frac{\sum_{t=1}^{L-1} n_{t,t+1}(s_i, s_j)}{\sum_{t=1}^{L-1} n_t(s_i)}. \qquad (1)$$



```
                      [-> employment] [-> FE] [-> HE] [-> joblessness] [-> school] [-> training]
[employment ->]                  0.99    0.00    0.00             0.01        0.00          0.00
[FE ->]                          0.03    0.95    0.01             0.01        0.00          0.00
[HE ->]                          0.01    0.00    0.99             0.00        0.00          0.00
[joblessness ->]                 0.04    0.01    0.00             0.94        0.00          0.01
[school ->]                      0.01    0.01    0.02             0.01        0.95          0.00
[training ->]                    0.04    0.00    0.00             0.01        0.00          0.94
```

Overall and transversal statistics

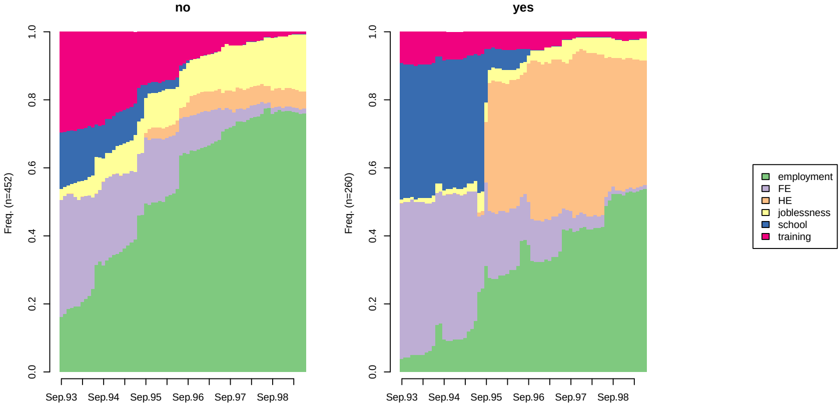
Transversal state distributions



```
     [State frequencies]
   Sep.93 Oct.93 Nov.93 Dec.93 Jan.94 Feb.94 Mar.94 Apr.94
EM  0.034  0.036  0.044  0.051  0.055  0.058  0.066  0.067
FE  0.223  0.217  0.215  0.215  0.209  0.206  0.204  0.200
HE  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000
JL  0.048  0.055  0.051  0.050  0.058  0.068  0.064  0.067
SC  0.472  0.466  0.464  0.461  0.451  0.451  0.446  0.446
TR  0.223  0.227  0.226  0.223  0.227  0.218  0.221  0.220
```

Overall and transversal statistics

Figure: State distribution plots by gcse5eq



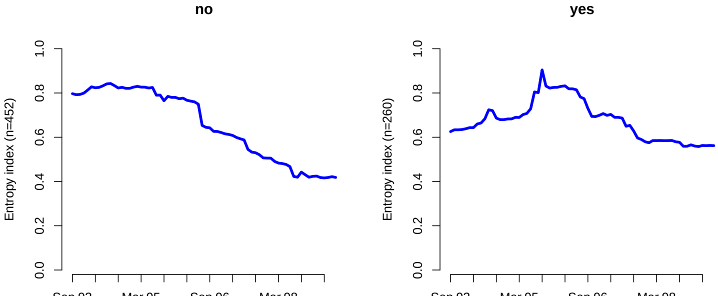
Overall and transversal statistics

Transversal entropy of state distributions

Entropy is a measure of the diversity of states observed at the considered position. Letting p_i denote the proportion of cases in state i at the considered position, the entropy is:

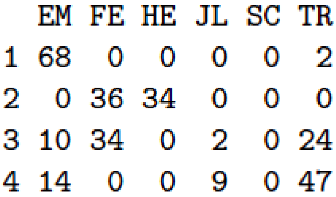
$$h(p_1, ..., p_a) = -\sum_{i=1}^{a} p_i \log(p_i). \tag{2}$$

Figure: Transversal entropy by gcse5eq



Individual sequence characteristics

Within sequence entropy



```
  EM FE HE JL SC TR
1 68  0  0  0  0  2
2  0 36 34  0  0  0
3 10 34  0  2  0 24
4 14  0  0  9  0 47
```

The total time spent in each state characterizes the state distribution within a sequence. The entropy of this distribution can be seen as a measure of the diversity of its states. Longitudinal entropy is:

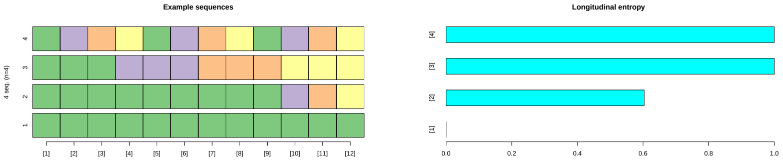
$$h(\pi_1, ..., \pi_s) = -\sum_{i=1}^{a} \pi_i \log(\pi_i). \tag{3}$$

Individual sequence characteristics

Within sequence entropy

- The entropy can be interpreted as the uncertainty in the prediction states of a given sequence:
  - if the state is the same throughout the sequence, the entropy is equal to 0;
  - entropy is maximum when the stay time in each of the states of the alphabet is the same.

Figure: Longitudinal entropy



Individual sequence characteristics

Turbulence

The turbulence $T(x)$ of a sequence x is a composite measure proposed by Elzinga and Liefbroer (2007) that accounts for:

1. the number $\phi(x)$ of distinct subsequences of the sequence;
2. the variance $s_t^2(x)$ of the consecutive times t_j spent in the $l_d(x)$ distinct states.

$$T(x) = \log_2 \left(\phi(x) \frac{s_{t,max}^2(x) + 1}{s_t^2(x) + 1} \right). \qquad (4)$$

where $s_{t,max}^2(x)$ is the maximum value that $s_t^2(x)$ can take given the total duration $l(x) = \sum_j t_j$ of that sequence.

Individual sequence characteristics

Complexity index

The complexity index, introduced in Gabadinho, Ritschard, Studer and Muller (2010), is a composite measure that combines:

1. the number of transitions in the sequence;
2. the longitudinal entropy of this sequence.

$$C(x) = \sqrt{\frac{l_d(x)}{l(x)}\frac{h(x)}{h_{max}}}. \tag{5}$$

where h_{max} is the theoretical maximum value of the entropy given the alphabet; i.e., $h_{max} = \log a$.

Individual sequence characteristics

Complexity index

- The minimum value of 0 can only be reached by a sequence with a single distinct state; i.e., with no transition and an entropy of 0.
- $C(x)$ reaches its maximum 1 if and only if the sequence x is such that:
  1. x contains each of the states in the alphabet;
  2. the same time $l(x)/a$ is spent in each state;
  3. the number of transitions is $l(x) - 1$.

Individual sequence characteristics

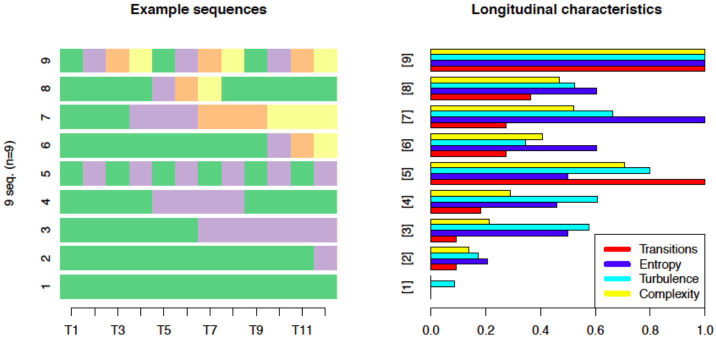


Figure: Different measures of complexity - comparison

Measuring sequences (dis)similarities

Measuring sequences (dis)similarities

- Methods from computer sciences (Hamming, 1950; Levenshtein,1966) and molecular biology. They appeared in the social sciences under the guidance of Abbott's seminal works.
- Methods based on the use of a measure of distance between sequences in order to observe similarities (and dissimilarities) between trajectories and build typology of sequences.
- Typology is used to identify and study the existing patterns in students pathways.

Measuring sequences (dis)similarities

What is distance ?

- Most of advanced sequence analysis methods rely on a dissimilarity measure.
- A dissimilarity is a quantification of how far two objects are.
- For instance, consider two incomes x and y:
  - $d(x, y) = |x - y|$.
  - $d(x, y) = log(1 + |x - y|)$.
  - $d(x, y) = (x - y)^2$.
- How to do it with categorical sequences?
- Depending on the issue, we want our dissimilarity measure to account for:
  - order of the states and transitions in each sequence;
  - temporality of the transitions;
  - duration of stay in each state;

Measuring sequences (dis)similarities

Basis

- Euclidean distance:

$$d(x,y) = \sqrt{\sum_i (x_i - y_i)^2}. \tag{6}$$

⇒ This distance measure is the shortest distance between two points.

- In geometry, the euclidean distance is a special case of that of Chebishev:

$$d(x,y) = \left(\sum_i |x_i - y_i|^p \right)^{\frac{1}{p}}. \tag{7}$$

⇒ This equation defines a distance family for Euclidean spaces.

Measuring sequences (dis)similarities

Distance between sequences

- Distances are defined for euclidean spaces, that is to say when the observations are described using numerical coordinates in a number of predefined dimensions.
- Trajectories are not immediately be represented in such a space.
- The quantification of the distance of two objects d is a distance if it satisfies the following conditions $\forall x, y, z$ (see Kaufman and Rousseeuw, 1990):

$$\begin{aligned} d(x,y) &\geq 0. && (8) \\ d(x,y) &= 0 \Leftrightarrow x = y. && (9) \\ d(x,y) &= d(y,x). && (10) \\ d(x,z) &\leq d(x,y) + d(y,z). && (11) \end{aligned}$$

Dissimilarity measure between sequences

- The determination of the distance measurement is a crucial step ⇒ depends i) data and ii) research question.
- Classification of distance measurement in 2 groups (Gabadinho, Ritschard, Mueller and Studer 2011):
  1. measures based on common attributes between sequences, i.e. measures that do not allow to move a sequence or a part of it (HAM, LCP, LCS);
  2. editing measures, i.e. measures taking into for similar shifted patterns (OM, HAM, DHD).
- Example: without shift, $x = ABAB$ and $y = BABA$ are very distant, while they are quite similar if we shift y by just one position.

Dissimilarity measure between sequences

Dissimilarities based on counts of common attributes

Let $A(x, y)$ be a count of common attributes between sequences x and y. It is a proximity measure since the higher the counts, the closer the sequences. We derive a dissimilarity measure from it through the following general formula:

$$d(x, y) = A(x, x) + A(y, y) - 2A(x, y). \tag{12}$$

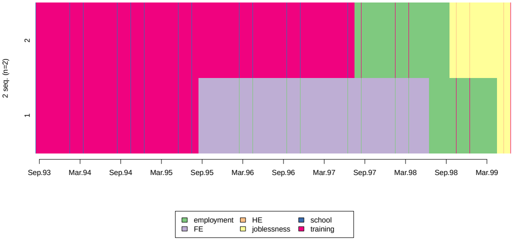
where $d(x, y)$ is the distance between objects x and y. The dissimilarity is maximal when $A(x, y) = 0$; i.e., when the two sequences have no common attribute. It is zero when the sequences are identical, in which case we have $A(x, y) = A(x, x) = |x| = A(y, y) = |y|$.

Dissimilarity measure between sequences

Hamming distance

The simple Hamming distance (Hamming, 1950) is the number of positions at which two sequences of equal length differ. It can equivalently be defined as $l - A_H(x, y)$, with $l = |x| = |y|$ the common sequence lenght and $A_H(x, y)$ the number of matching positions.

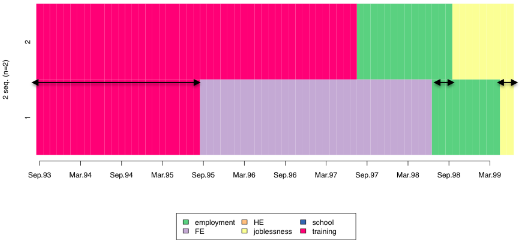
Figure: Sequences 3 and 4 of mvad



Dissimilarity measure between sequences

Hamming distance

Figure: Hamming distance - example on sequences 3 and 4 of mvad



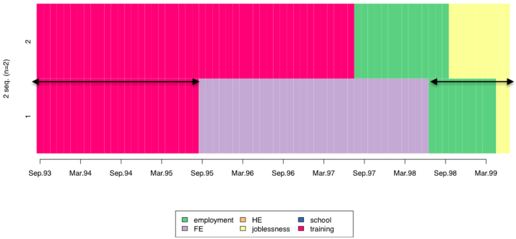
- Hamming distance do not allow us to use timing differences.
- With Hamming distance we focus on temporality: proximity = same state at same time.
- Take into account timing differences: count for common

Dissimilarity measure between sequences

LCS distance

$A_S(x,y)$ is the longest common subsequence (without time constraint): $A_{\mathcal{L}}(x,y) = \max\{|u| : u \in S(x,y)\}$, where $|u|$ is the lenght of the longest common subsequence for the pair of sequences (x,y) and $S(x,y)$ is the nonempty set of subsequences of sequences x and y. We get the LCS distance with equation 12 by using $A_S(x,y)$ as proximity measure.

Figure: LCS - example on sequences 3 and 4 of mvad

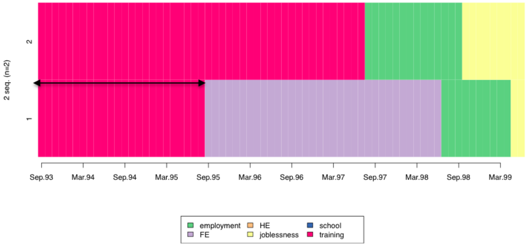


Dissimilarity measure between sequences

LCP distance

$A_P(x, y)$ of the longest common prefix (LCP) between two sequences. We get the LCP distance with equation 12 by using $A_P(x, y)$ as proximity measure.

Figure: LCP - example on sequences 3 and 4 of mvad



Edit distances

Distance

An edit distance is defined as the minimal cost of transforming one sequence into the other. Two types of operations are considered:

1. the substitution of one element by an other one (HAM, DHD, OM);
2. the insertion or deletion (indel) of an element, which generates a one position shift of all the elements on its right (OM).

Operations

- Operations of indel deform the time structure of sequences to allow common subsequences to emerge.
- Operations of substitution conserve the time structure of sequences so that elements can be compared at constant date.

Edit distances

Example: indel

- Consider the two following sequences:

1	SC	SC	SC	EM	EM	EM	JL
2	SC	SC	SC	EM	EM	JL	JL

- Insertion of “EM”, cost = 1:

1	SC	SC	SC	EM	EM	EM	JL	
2	SC	SC	SC	EM	EM	EM	JL	JL

- Insertion of “JL”, cost = 1:

1	SC	SC	SC	EM	EM	EM	JL
2	SC	SC	SC	EM	EM	EM	JL

⇒ Two sequences are now identical, total cost = 2.

Edit distances

Example: substitution

- Consider the two following sequences:

1	SC	SC	SC	EM	EM	EM	JL
2	SC	SC	SC	EM	EM	JL	JL

- Substitution of “EM” by “JL”, cost $= 2$:

1	SC	SC	SC	EM	EM	EM	JL
2	SC	SC	SC	EM	EM	EM	JL

$\Rightarrow$ Two sequences are now identical, total cost $= 2$.

Edit distances

HAM

HAM generalizes the basic Hamming distance by allowing for state dependent substitution costs. Indeed, the count of nonmatching positions is the cost of substituting a state at each position when all costs are set to 1.

Edit distances

DHD

- According to Lesnard (2010) indel operations (using in OM) leads to a distortion of the temporal sequences.
- Problem: in some analyzes the question of temporality is important (ex: schedules);
- The extension proposed by Lesnard (2006) to account for time-varying costs.

For each t we calculate the cots of substitution between states a and b of a same sequence, $S_t(a,b)$, using transition rates between states $P_t(a_t|b_{t_1}) = \frac{N(a_t|b_{t-1})}{N(b_{t-1})}$. We then calculate the substitution cost between states as follows

$$S_t(a,b) = 4 - P_t(a_t|b_{t-1}) - P_t(b_t|a_{t-1}) - P_t(a_{t+1}|b_t) - Pt(b_{t+1}|a_t). \tag{13}$$

Edit distances

Optimal matching

Let Σ the alphabet associated with sequences and λ the zero element. The different operations are:

- $a \to b$ an operation of substitution, with $a, b \in \Sigma \cup \{\lambda\}$ and $a \neq b$;
- $a \to \lambda$ an operation of delete; suppression;
- $\lambda \to a$ an operation of insertion.

$T_{x,y} = T_1 ... T_l$ denotes the set of l operations needed to transform a sequence x into a sequence y. We note γ the cost matrix associated with each operation and $\gamma(T_i)$ the cost associated with the operation T_i.
We obtain the distance $d_{OM}(x, y)$ by calculating the sequence of operations $T_{x,y}$ which minimizes the total cost.

$$d_{OM}(x, y) = \min \left\{ \sum_{i=1}^{l} \gamma(T_i) \right\}. \tag{14}$$

Edit distances

Example: OM

- Consider the two following sequences:

1	EM	EM	FE	FE	HE	HE
2	FE	FE	HE	HE	EM	EM

- With a substitution cost $= 2$ and a indel cost $= 1$, the OM distance $= 4$ (on a maximum of 12):

1	EM	EM	FE	FE	HE	HE	-	-
2	-	-	FE	FE	HE	HE	EM	EM
Coût	1	2	2	2	2	2	3	4

Edit distances

Costs

- Subsitution costs define the distance between two states *at the same time*.
- Indel costs defines the possibility of allowing a time lag in the comparison of sequences.
- If a substition cost is greater than twice indel cost, it will never be used.
- Ways to define subsitution costs:
  - **Theoretical costs:** costs are defined arbitrarily by the researcher. Some states may be closer to a state than others. Justification?
  - **Constant costs:** costs of substitution between states are constant.
  - **Estimated costs from the data:** costs estimated on transition rates between states.

Typologie of trajectories

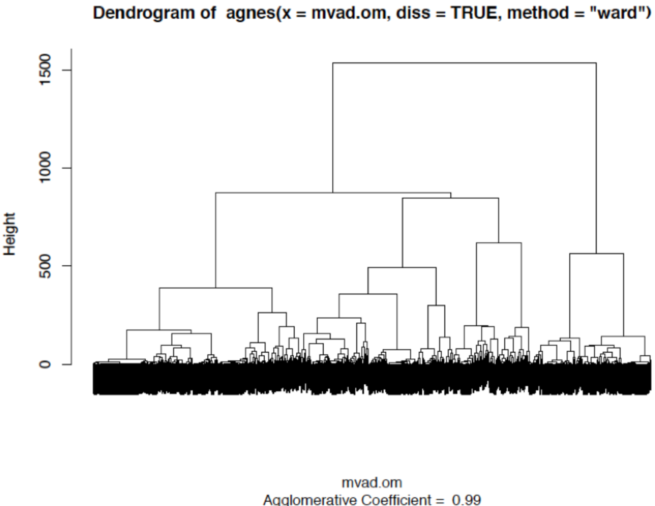
Typologie of trajectories

- Classification methods are used to construct a typology of sequences, i.e regroup the population of the sample into groups based on common characteristics between sequences.
- This grouping procedure is based on a simplification of the data.
- Describe reality?
- Two main types of clustering procedures:
    1. hierarchical clustering (ascending and descending);
    2. non hierarchical clustering (partitioning).

Clustering

Hierarchical clustering

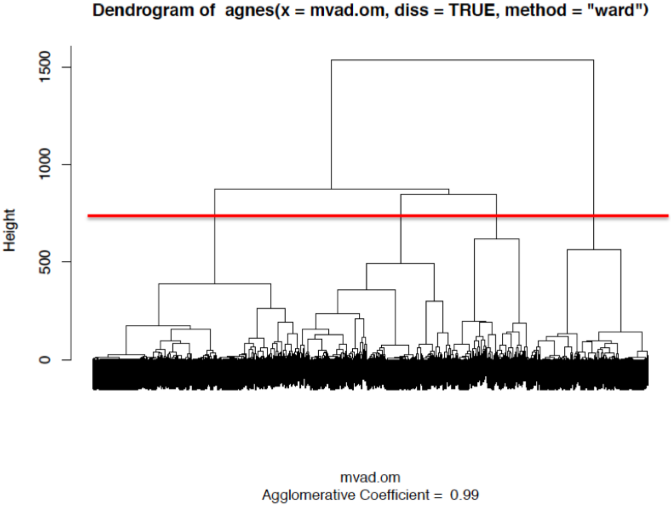
Figure: Dendrogram



Clustering

Hierarchical clustering

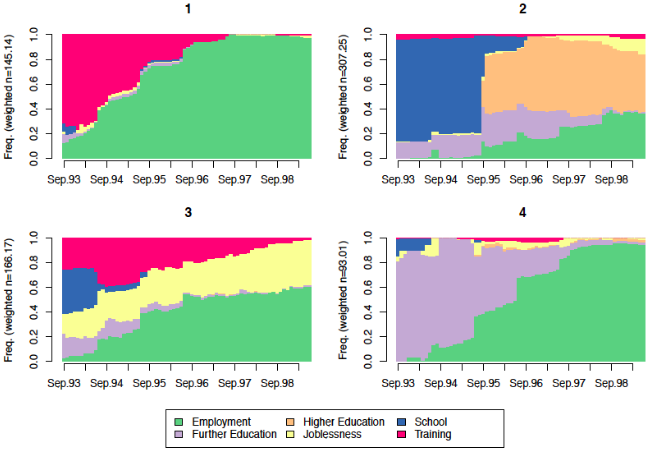
Figure: Dendrogram



Clustering

Hierarchical clustering

Figure: Typical trajectories - ward



Clustering

Hierarchical clustering

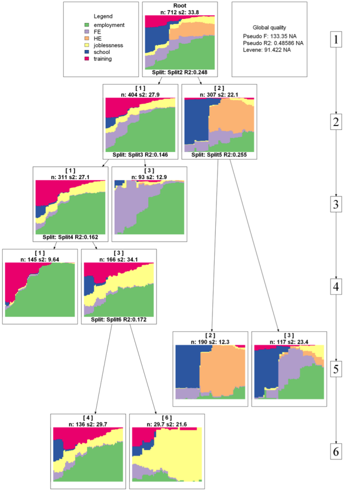
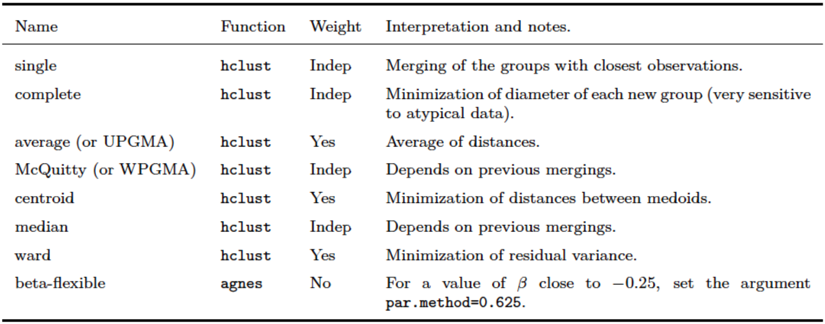


Figure: WardTree

Clustering

Hierarchical clustering

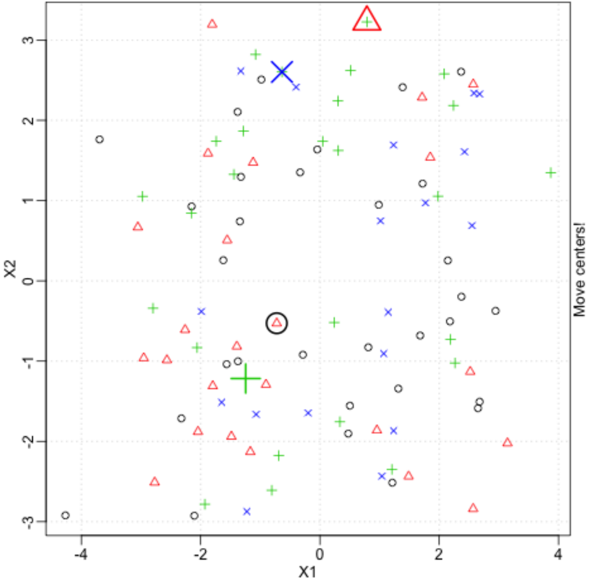
Table: Hierarchical clustering algorithms



Name	Function	Weight	Interpretation and notes.
single	`hclust`	Indep	Merging of the groups with closest observations.
complete	`hclust`	Indep	Minimization of diameter of each new group (very sensitive to atypical data).
average (or UPGMA)	`hclust`	Yes	Average of distances.
McQuitty (or WPGMA)	`hclust`	Indep	Depends on previous mergings.
centroid	`hclust`	Yes	Minimization of distances between medoids.
median	`hclust`	Indep	Depends on previous mergings.
ward	`hclust`	Yes	Minimization of residual variance.
beta-flexible	`agnes`	No	For a value of β close to -0.25, set the argument `par.method=0.625`.

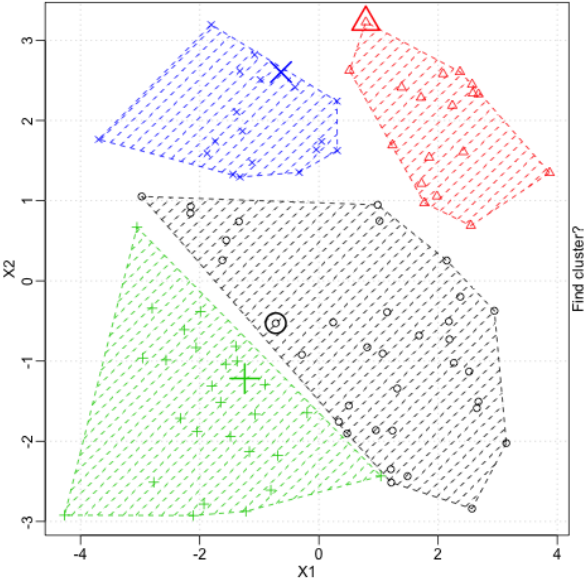
Clustering

Partitionning around medoids



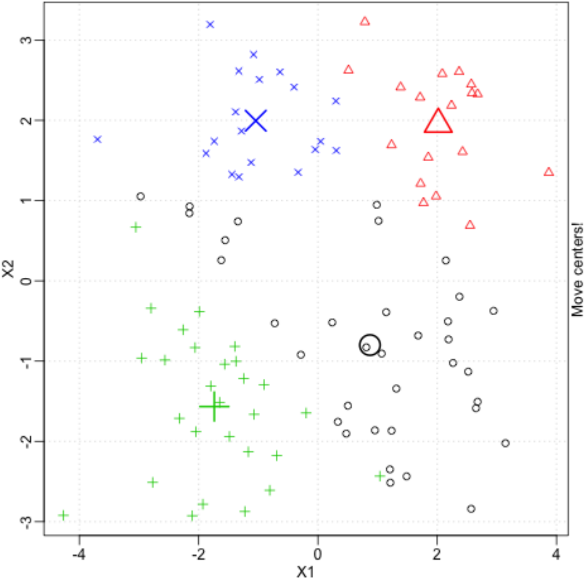
Clustering

Partitionning around medoids



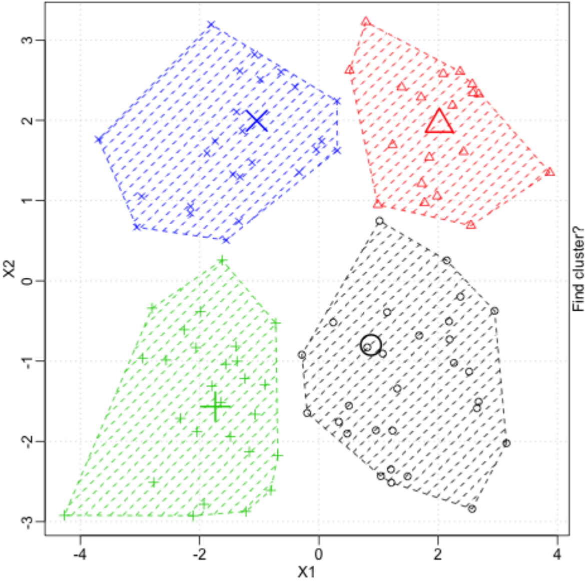
Clustering

Partitionning around medoids



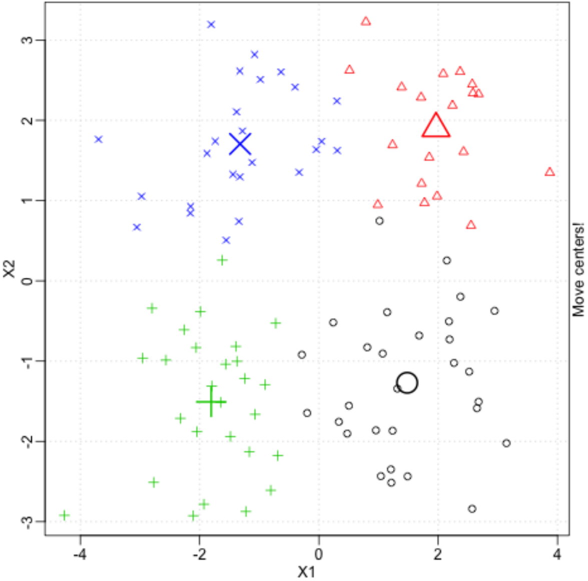
Clustering

Partitionning around medoids



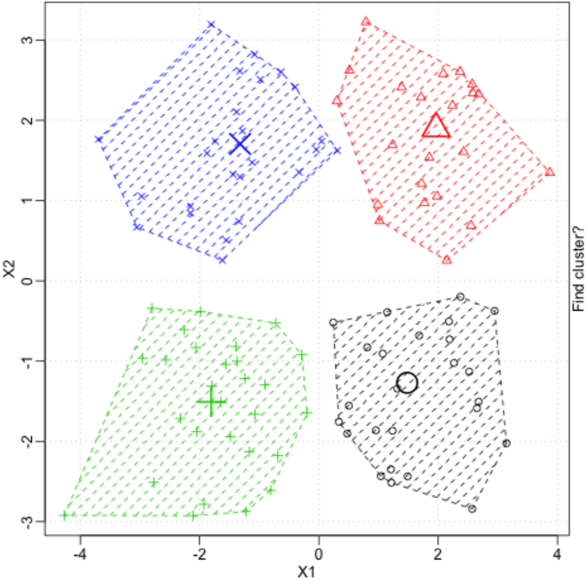
Clustering

Partitionning around medoids



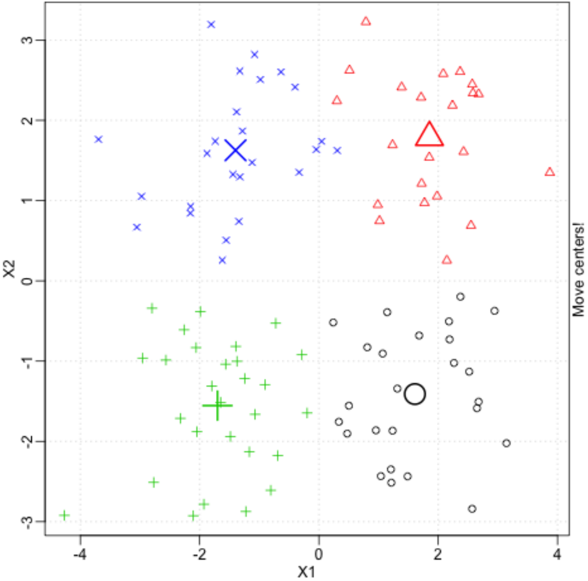
Clustering

Partitionning around medoids



Clustering

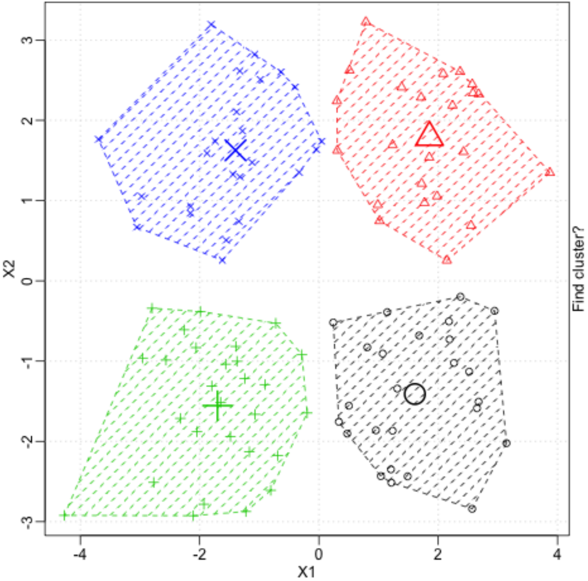
Partitionning around medoids



!

Clustering

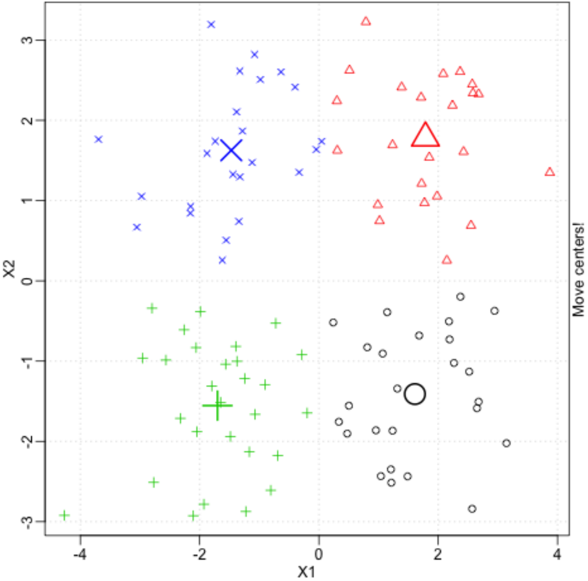
Partitionning around medoids



!

Clustering

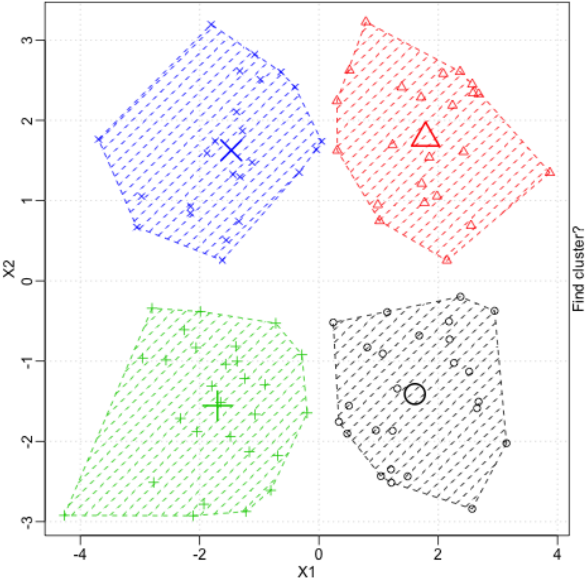
Partitionning around medoids



!

Clustering

Partitionning around medoids



!

Clustering

Partitionning around medoids

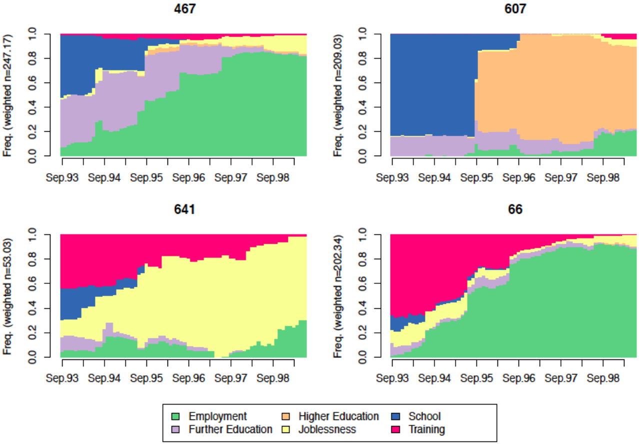
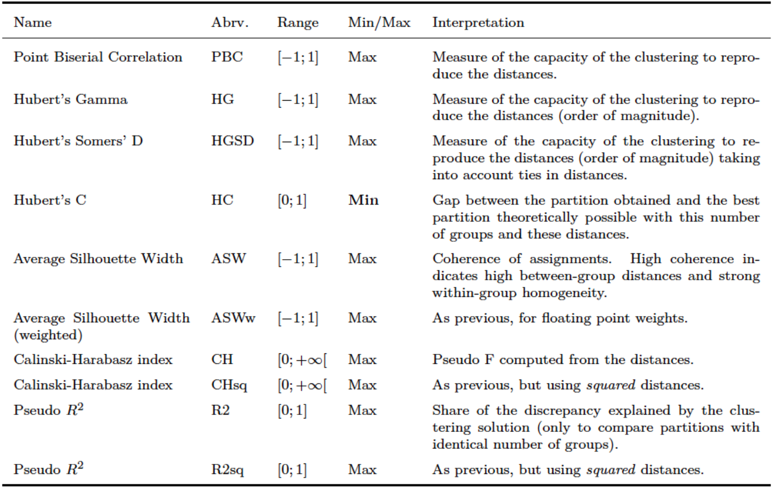


Figure: Typical trajectories - PAM

Measuring the quality of a partition

Table: Measures of the quality of a partition



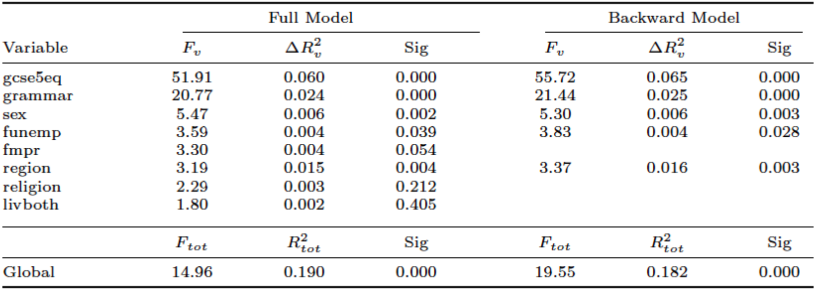
Name	Abrv.	Range	Min/Max	Interpretation
Point Biserial Correlation	PBC	$[-1;1]$	Max	Measure of the capacity of the clustering to reproduce the distances.
Hubert's Gamma	HG	$[-1;1]$	Max	Measure of the capacity of the clustering to reproduce the distances (order of magnitude).
Hubert's Somers' D	HGSD	$[-1;1]$	Max	Measure of the capacity of the clustering to reproduce the distances (order of magnitude) taking into account ties in distances.
Hubert's C	HC	$[0;1]$	**Min**	Gap between the partition obtained and the best partition theoretically possible with this number of groups and these distances.
Average Silhouette Width	ASW	$[-1;1]$	Max	Coherence of assignments. High coherence indicates high between-group distances and strong within-group homogeneity.
Average Silhouette Width (weighted)	ASWw	$[-1;1]$	Max	As previous, for floating point weights.
Calinski-Harabasz index	CH	$[0;+\infty[$	Max	Pseudo F computed from the distances.
Calinski-Harabasz index	CHsq	$[0;+\infty[$	Max	As previous, but using *squared* distances.
Pseudo R^2	R2	$[0;1]$	Max	Share of the discrepancy explained by the clustering solution (only to compare partitions with identical number of groups).
Pseudo R^2	R2sq	$[0;1]$	Max	As previous, but using *squared* distances.

Linking trajectory types and explanatory factors

- Econometric models.
- Discrepancy analysis of state sequences (Studer et al., 2011).
- Tree-structured analysis of sequences (Studer et al., 2009).

Multi-factor discrepancy analysis

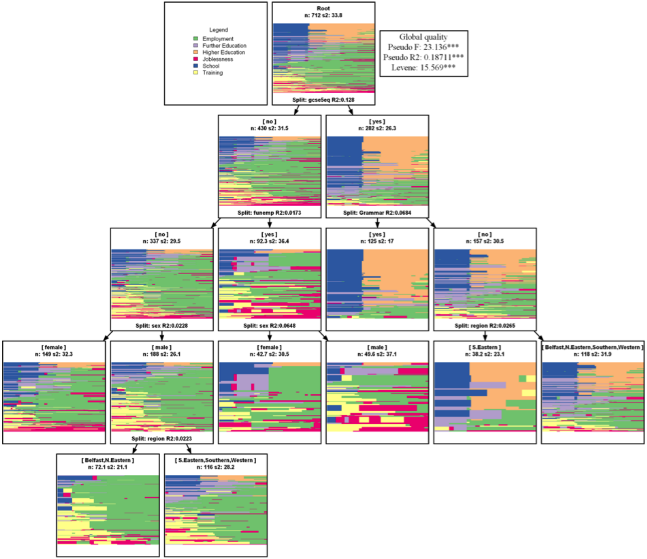
Table: Multi-factor discrepancy analysis



	Full Model			Backward Model		
Variable	F_v	ΔR_v^2	Sig	F_v	ΔR_v^2	Sig
gcse5eq	51.91	0.060	0.000	55.72	0.065	0.000
grammar	20.77	0.024	0.000	21.44	0.025	0.000
sex	5.47	0.006	0.002	5.30	0.006	0.003
funemp	3.59	0.004	0.039	3.83	0.004	0.028
fmpr	3.30	0.004	0.054			
region	3.19	0.015	0.004	3.37	0.016	0.003
religion	2.29	0.003	0.212			
livboth	1.80	0.002	0.405			
	F_{tot}	R_{tot}^2	Sig	F_{tot}	R_{tot}^2	Sig
Global	14.96	0.190	0.000	19.55	0.182	0.000

Sequence regression tree

Figure: Tree regression analysis



Part 2. Application

Tuition fees and social segregation

Introduction

- The increase of tuition fees is a “political” solution adopted by many countries.
- France: universities autonomy (LRU) and statutes derogatory in some institutions (Dauphine, Science Po, Mines Telecom...).
- Analysis of the effects of the introduction of tuition fees at the University Paris Dauphine 9 on:
  1. students pathways selection ;
  2. student achievement.
- Methodology:
  1. Optimal matching to construct a typology of students trajectories in HE.
  2. Multinomial logit to assess the effect of tuition fees on the academic pathways selected by this university.
  3. Difference-in-differences in a non-linear model Puhani (2012).

Litterature review

- The relationship between tuition fees and students' decisions has been extensively studied in the literature, in terms of:
    1. access to higher education (Cameron and Heckman, 2001; Coelli, 2009; Hubner, 2012...);
    2. choice of curriculum (Callender and Jackson, 2008; Dietrich and Gerner, 2012; Field, 2009...);

⇒ Persistent controversy of on the impact of tuition fees.

- Small number of empirical works on student achievement:
    1. effects of tuition fees on the time necessary to graduation (Garibaldi, Giavazzi, Ichino and Rettore, 2012);
    2. effects of tuition fees on student achievement (as endogenous variable, see Heineck, Kifmann and Lorenz, 2006).

⇒ The various impacts of the introduction of tuition fees put forward a debate on the existence, nature and extent of segregation phenomena.

Data

SISE

- The SISE database collects data on students.
- Data concerning enrollment and success of students in French universities;
- Data (more than 70 variables) about:
    - Sociodemographic characteristics of students: sex, age, social category, nationality, geographic origin, etc.
    - Schooling: establishment, diploma course, school attended, registration type, etc.
    - Previous schooling: bachelor academy, graduation year, baccalaureat, year of first registration in the French university system, success Diploma etc.
- Matching: SISE *universites - inscriptions*, SISE *universites - resultats* and ALGAE.

Data

Dauphine

- 1st university to implement tuition fees in France .
- This allows to analyze the effect induced by the increase of tuition fees on the students trajectories.

Gross income (per year)	Tuition fees
< 40 000	1 500
40 000 - 50 000	2 000
50 000 - 60 000	2 500
60 000 - 70 000	3 000
70 000 - 80 000	3 500
> 80 000	4 000

- Two cohorts:
  1. 2009-2010: all second year master in economy and management are without tuition fees;
  2. 2010-2011: a part of them increase tuition fees.

Table: Socio-economic characteristics of students in Master 2 economics-management at Dauphine by year and by type of Masters

Caracteristiques	2009 (n=1139)			2010 (n=1192)			Total
socio-economiques	DN	DGE	Total	DN	DGE	Total	(n=2331)
Social class							
Very well-off	56.03	60.17	58.21	59.34	60.48	59.98	59.12
Well-off	10.02	10.00	10.01	10.40	11.14	10.82	10.42
Average	14.84	14.33	14.57	17.34	16.64	16.95	15.79
Disadvantaged	19.11	15.50	17.21	12.91	11.74	12.25	14.67
Scholarships							
No	90.54	84.67	87.45	86.51	82.47	84.23	85.80
Yes	9.46	15.33	12.55	13.49	17.53	15.77	14.20
Sex							
Female	53.25	56.33	54.87	52.99	53.64	53.36	54.10
Male	46.75	43.67	45.13	47.01	46.36	46.64	45.90
Nationality							
French	69.20	79.33	74.54	68.40	81.43	75.76	75.16
Foreign	30.80	20.67	25.46	31.60	18.57	24.24	24.84
Location							
Paris	40.45	39.83	40.12	40.66	43.54	42.28	41.23
Seine et Marne	2.97	2.17	2.55	1.93	1.78	1.85	2.19
Yvelines	3.90	8.00	6.06	4.82	6.54	5.79	5.92
Essonne	2.04	3.83	2.99	4.05	3.27	3.61	3.30
Hauts de Seine	14.66	14.17	14.40	16.38	13.97	15.02	14.71
Seine Saint Denis	4.27	2.33	3.25	4.24	2.53	3.27	3.26
Val de Marne	7.42	6.50	6.94	7.13	5.35	6.12	6.52
Val d'Oise	2.41	2.00	2.19	3.85	3.86	3.86	3.05
Outside Ile-de-France	21.89	21.17	21.51	16.96	19.17	18.20	19.82
Age							
≤ 22	24.12	32.17	28.36	24.66	25.41	25.08	26.68
[23; 24]	48.24	46.33	47.23	46.24	57.80	50.06	50.06
≥ 25	27.64	21.50	24.41	29.09	16.79	22.15	23.25

Socio-economic	2009 (n=23 807)			2010 (n=25 910)			Total
characteristics	Univ.	Dauphine	Total	Univ.	Dauphine	Total	(n=49 717)
Social class							
Very well-off	30,7	58,2	32,0	30,7	60,0	32,1	32,0
Well-off	12,3	10,0	12,1	11,7	10,8	11,7	11,9
Average	18,5	14,6	18,3	18,0	17,0	18,0	18,1
Disadvantaged	38,6	17,2	37,6	39,5	12,3	38,3	37,9
Scholarship							
No	78,4	87,5	78,8	77,4	84,2	77,7	78,2
Yes	21,6	12,5	21,2	22,6	15,8	22,3	21,8
Sex							
Female	53,1	54,9	53,2	53,4	53,4	53,4	53,3
Male	46,9	45,1	46,8	46,6	46,6	46,6	46,7
Nationality							
French	61,8	74,5	62,4	59,6	75,8	60,4	61,3
Foreign	38,2	25,5	37,6	40,4	24,2	39,6	38,7
Location							
Paris	6,3	40,1	7,9	5,7	42,3	7,3	7,6
Île de France	13,2	38,4	14,4	12,4	39,5	13,7	14,0
Province	62,5	16,1	60,3	63,2	14,4	61,0	60,6
Out of France	18,0	5,4	17,4	18,7	3,8	18,0	17,7
Baccalauréat							
Littéraire	4,9	3,4	4,8	4,9	3,4	4,9	4,8
Économique	28,3	26,9	28,2	28,0	27,6	28,0	28,1
Scientifique	22,8	50,8	24,2	21,5	51,0	22,9	23,5
Technologique	6,7	1,1	6,4	6,5	0,9	6,3	6,3
Others technol.	1,8	0,1	1,7	1,6	0,2	1,5	1,6
Professionnel	0,7	0,3	0,6	0,5	0,1	0,5	0,6
Dispensé	34,9	17,6	34,1	36,9	16,8	36,0	35,1
Age							
≤ 22	22,5	28,4	22,8	22,9	25,1	23,0	22,9
$[23; 24]$	40,4	47,2	40,8	41,3	52,8	41,8	41,3
≥ 25	37,1	24,4	36,5	35,8	22,2	35,2	35,8

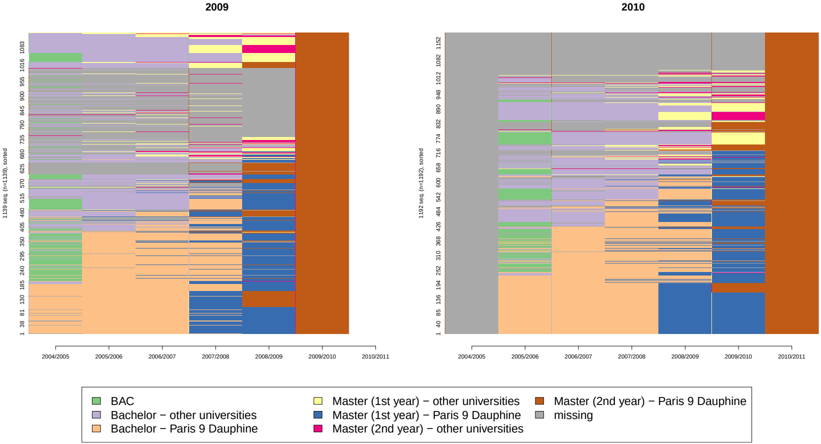
Methodology

Optimal Matching Analysis

- OM with indel=max(costs)/2.
- Costs based on transition rates.

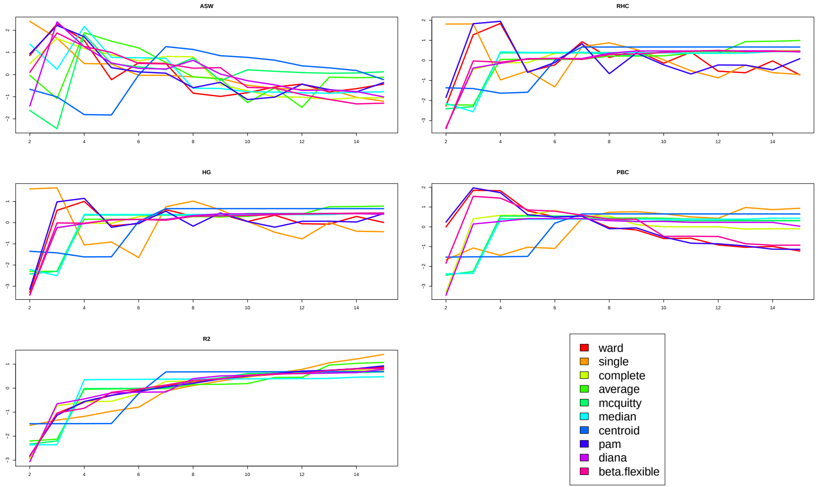
Methodology

Figure: Student trajectories depending on the year of enrollment in second year of master in economics and management at University Paris 9 Dauphine



Methodology

Figure: Evaluation of quality measures depending number of groupes and clustering procedures



Logit multinomial and marginal effects

- Logit multinomial to determine the relationships between typical trajectories and socio-economic variables (x_i).
- The dependent variable in our model is the typical trajectory taken by the student. We note $J+1$ the number of modalities. The probability that a student with characteristics x_i borrows the trajectory j is given by:

$$\Pr\left(Y_i = j \mid x_i\right) = \frac{e^{\beta_j' x_i}}{\sum_{k=0}^{J} e^{\beta_k' x_i}}, \text{ for } j = 0, ..., J. \tag{15}$$

- In the following we estimate the marginal effects of different variables on the probability that the student borrows the trajectory j:

$$\delta_i = \frac{\partial P_j}{\partial x_i} = P_j \left(\beta_j - \sum_{k=0}^{J} P_k \beta_k \right) = P_j \left(\beta_j - \overline{\beta} \right). \tag{16}$$

Difference in differences and non linear model

- In the case of a non-linear model the model to estimate can be written:

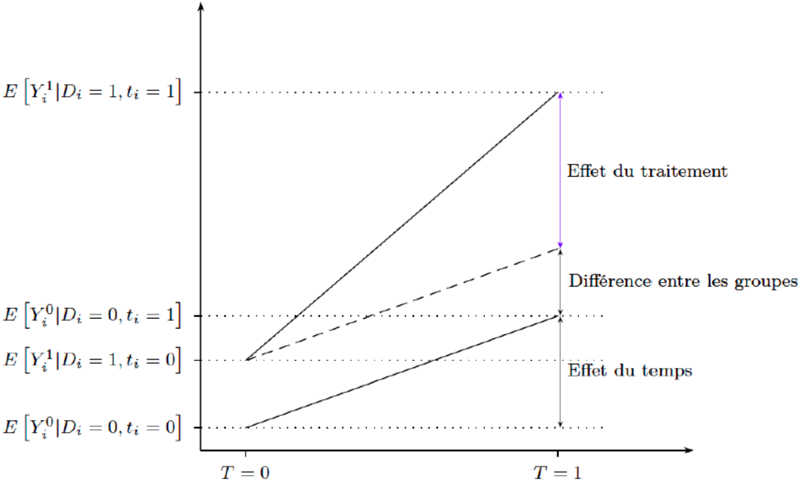
$$\Pr\left(Y_i = 1\right) = \Phi\left(\alpha + \beta D_i + \delta t_i + \tau D_i t_i + \theta X_i + \epsilon_i\right). \quad (17)$$

- Formally, Puhani (2012) shows that the effect of the treatment is equal to:

$$\begin{aligned}\gamma &= \frac{\Delta^2 E\left[Y_i | D_i, t_i, X_i\right]}{\Delta D_i \Delta t_i} - \frac{\Delta^2 E\left[Y_i^0 | D_i, t_i, X_i\right]}{\Delta D_i \Delta t_i}. && (18)\\ &= \Phi\left(\alpha + \beta + \delta + \tau + \theta\right) - \Phi\left(\alpha + \beta + \delta + \theta\right). && (19)\end{aligned}$$

Methodology

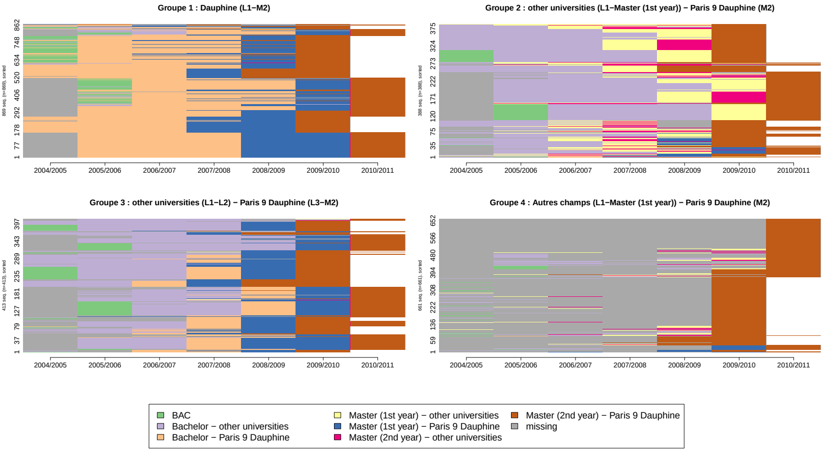
Figure: Difference-in-differences



$E\left[Y_i^1|D_i=1,t_i=1\right]$
$E\left[Y_i^0|D_i=0,t_i=1\right]$
$E\left[Y_i^1|D_i=1,t_i=0\right]$
$E\left[Y_i^0|D_i=0,t_i=0\right]$
Effet du traitement
Différence entre les groupes
Effet du temps
$T=0$
$T=1$

Results

Figure: Typical pathways of students in Master 2 at Dauphine



Results

Table: Situation in the previous year for the students in Group 4

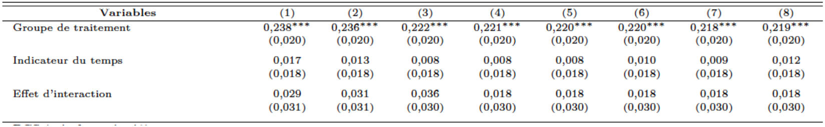
Situation in	2009 (n=346)			2010 (n=315)			Total
previous year	DN	DGE	Total	DN	DGE	Total	(n=661)
University	38.55	17.22	27.46	43.06	22.81	32.06	29.65
Maneagement school	10.84	17.78	14.45	12.50	22.81	18.10	16.19
Engineering school	10.24	21.11	15.90	9.03	16.96	13.33	14.67
Foreign establishment	17.47	17.22	18.69	17.34	13.19	14.62	15.73
Other establishment in France	4.82	4.44	4.62	6.25	0	2.86	3.78
Other SISE establishment	0	0	0	0.69	0.58	0.63	0.30
Return to studies	18.07	22.22	20.23	15.28	22.22	19.05	19.67

Table: Socio-economic characteristics of students in the Master 2 economics-management at Dauphine by typical pathway

Variables	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total
Social class					
Very well-off	67.55	50.26	56.90	54.61	59.12
Well-off	9.67	11.86	11.38	9.98	10.42
Average	13.46	18.81	16.46	16.64	15.79
Disadvantaged	9.32	19.07	15.25	18.76	14.67
Scholarship					
No	86.19	79.38	77.48	94.25	85.80
Yes	13.81	20.62	22.52	5.75	14.20
Sex					
Female	58.11	51.55	48.67	53.71	54.10
Male	41.89	48.45	51.33	46.29	45.90
Nationality					
French	87.11	83.51	80.87	50.98	75.16
Foreign	12.89	16.49	19.13	49.02	24.84
Location					
Paris	45.91	34.79	35.11	42.66	41.23
Seine et Marne	2.88	3.87	1.21	0.91	2.19
Yvelines	9.78	3.35	6.30	2.12	5.92
Essonne	3.57	2.84	4.12	2.72	3.30
Hauts de Seine	16.00	7.22	17.19	15.89	14.71
Seine Saint Denis	3.45	4.12	2.66	2.87	3.26
Val de Marne	5.75	8.25	6.54	6.51	6.52
Val d'Oise	3.68	4.38	2.66	1.66	3.05
Outside Ile-de-France	8.98	31.19	24.21	24.66	19.82
Age					
≤ 22	36.02	21.65	32.93	13.46	26.68
[23; 24]	53.97	52.32	55.21	40.39	50.06
≥ 25	10.01	26.03	11.86	46.14	23.25

Variables	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Very well-off (ref.)	-	-	-	-
Well-of	-0.064†	0.056†	0.016	-0.008
	(0.036)	(0.029)	(0.029)	(0.036)
Average	-0.090**	0.055*	0.000	0.035
	(0.031)	(0.024)	(0.025)	(0.030)
Disadvantaged	-0.155***	0.084***	0.015	0.056†
	(0.036)	(0.026)	(0.027)	(0.032)
Scholarship				
No (ref.)	-	-	-	-
Yes	0.023	0.052*	0.096***	-0.172***
	(0.034)	(0.025)	(0.024)	(0.038)
Sex				
Female (ref.)	-	-	-	-
Male	-0.057**	0.003	0.045**	0.009
	(0.022)	(0.018)	(0.017)	(0.021)
Nationality				
French (ref.)	-	-	-	-
Foreign	-0.192***	-0.082***	-0.000	0.275***
	(0.024)	(0.024)	(0.023)	(0.023)
Localisation				
Paris (ref.)	-	-	-	-
Ile de France	-0.025	-0.004	0.025	0.003
	(0.023)	(0.021)	(0.020)	(0.024)
Outside Ile de France	-0.349***	0.129***	0.053*	0.167***
	(0.033)	(0.023)	(0.024)	(0.028)
Age	-0.081***	0.025***	-0.027***	0.082***
	(0.008)	(0.005)	(0.006)	(0.006)
Type of Master's				
National (ref.)	-	-	-	-
Fee paying	0.039	-0.066***	-0.056*	0.083**
	(0.030)	(0.025)	(0.024)	(0.031)
Year				
2009 (ref.)	-	-	-	-
2010	0.002	0.0247	0.048*	-0.074**
	(0.028)	(0.022)	(0.022)	(0.028)

Figure: Difference-in-differences results



Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Groupe de traitement	0,238*** (0,020)	0,236*** (0,020)	0,222*** (0,020)	0,221*** (0,020)	0,220*** (0,020)	0,220*** (0,020)	0,218*** (0,020)	0,219*** (0,020)
Indicateur du temps	0,017 (0,018)	0,013 (0,018)	0,008 (0,018)	0,008 (0,018)	0,008 (0,018)	0,010 (0,018)	0,009 (0,018)	0,012 (0,018)
Effet d'interaction	0,029 (0,031)	0,031 (0,031)	0,036 (0,030)	0,018 (0,030)	0,018 (0,030)	0,018 (0,030)	0,018 (0,030)	0,018 (0,030)

(1) D&D

(2) D&D + social class

(3) D&D + social class + scholarship

(4) D&D + social class + scholarship + sex

(5) D&D + social class + scholarship + sex + nationality

(6) D&D + social class + scholarship + sex + nationality + location

(7) D&D + social class + scholarship + sex + nationality + location + baccalaureat

(8) D&D + social class + scholarship + sex + nationality + location + baccalaureat + age

Socio-economic category (Ref. very well-off)								
Well-off	-	0.001	-0.016	-0.017	-0.015	-0.016	-0.018	-0.011
		(0.025)	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)
Average	-	-0.014	-0.037†	-0.037†	-0.035†	-0.040†	-0.044*	-0.042*
		(0.021)	(0.021)	(0.021)	(0.021)	(0.021)	(0.021)	(0.021)
Disadvantaged		-0.056**	-0.085***	-0.086***	-0.076***	-0.076***	-0.081***	-0.068***
		(0.020)	(0.020)	(0.020)	(0.021)	(0.021)	(0.021)	(0.021)
Scholarships (ref. no scholarship)								
Grade 0	-	-	0.163**	0.163**	0.148*	0.136*	0.137*	0.127*
			(0.059)	(0.059)	(0.059)	(0.059)	(0.059)	(0.058)
Grade 1	-	-	0.011†	0.012†	0.095†	0.086	0.085	0.075
			(0.057)	(0.058)	(0.057)	(0.057)	(0.058)	(0.057)
Grade 2	-	-	0.185*	0.187*	0.174*	0.172*	0.170*	0.157†
			(0.084)	(0.084)	(0.084)	(0.085)	(0.086)	(0.085)
Grade 3	-	-	0.190**	0.191*	0.178*	0.169*	0.166*	0.156*
			(0.080)	(0.080)	(0.079)	(0.077)	(0.078)	(0.077)
Grade 4	-	-	0.208†	0.211†	0.196	0.203	0.197	0.180
			(0.123)	(0.124)	(0.122)	(0.125)	(0.123)	(0.122)
Grade 5	-	-	0.241***	0.240***	0.225***	0.217***	0.217***	0.207***
			(0.063)	(0.063)	(0.062)	(0.062)	(0.062)	(0.062)
Grade 6	-	-	0.193***	0.194***	0.185**	0.186**	0.193**	0.181**
			(0.060)	(0.060)	(0.060)	(0.060)	(0.061)	(0.061)
Sex (Ref. female)								
Male	-	-	-	-0.019	-0.023	-0.023	-0.020	-0.017
				(0.014)	(0.014)	(0.014)	(0.014)	(0.014)
Nationality (Ref. French)								
Foreign	-	-	-	-	-0.043**	-0.042**	-0.062**	-0.061**
					(0.016)	(0.016)	(0.024)	(0.024)

Geographical origin (Ref. Paris)								
Seine et Marne	-	-	-	-	-	-0.035	-0.039	-0.047
						(0.045)	(0.045)	(0.045)
Yvelines	-	-	-	-	-	0.034	0.031	0.022
						(0.034)	(0.034)	(0.034)
Essonne	-	-	-	-	-	0.039	0.039	0.036
						(0.044)	(0.044)	(0.044)
Hauts de Seine	-	-	-	-	-	-0.007	-0.010	-0.011
						(0.021)	(0.021)	(0.021)
Seine Saint Denis	-	-	-	-	-	-0.010	-0.011	-0.015
						(0.039)	(0.039)	(0.039)
Val de Marne	-	-	-	-	-	0.025	0.027	0.025
						(0.030)	(0.030)	(0.030)
Val d'Oise	-	-	-	-	-	0.018	0.010	0.004
						(0.044)	(0.044)	(0.044)
Outside the Île de France region	-	-	-	-	-	0.055**	0.051*	0.047*
						(0.020)	(0.021)	(0.021)
Baccalauréat stream (Ref. sciences)								
Literature	-	-	-	-	-	-	0.044	0.051
							(0.044)	(0.044)
Economics	-	-	-	-	-	-	0.031†	0.031†
							(0.018)	(0.018)
Technology STT	-	-	-	-	-	-	0.042	0.078
							(0.078)	(0.078)
Other technology	-	-	-	-	-	-	-0.205	-0.165
							(0.145)	(0.145)
Vocational	-	-	-	-	-	-	-0.079	-0.043
							(0.139)	(0.139)
Exemption	-	-	-	-	-	-	0.042	0.055*
							(0.028)	(0.028)
Age	-	-	-	-	-	-	-	-0.009***
								(0.003)

Conclusion

- ▶ Paris 9 Dauphine is the first university wich introduces tuition fees in France.
- ⇒ Tuition fees have changed the types of students pathways allowing access to second year of master 2 and therefore the nature of people admitted in these curriculum ⇒ Cumulative mechanism.
- ⇒ Unlike theoretical requirements, tuition fees have no effect on student achievement.
- ▶ Generalization of our results? Specificity of Dauphine in the landscape of French universities?