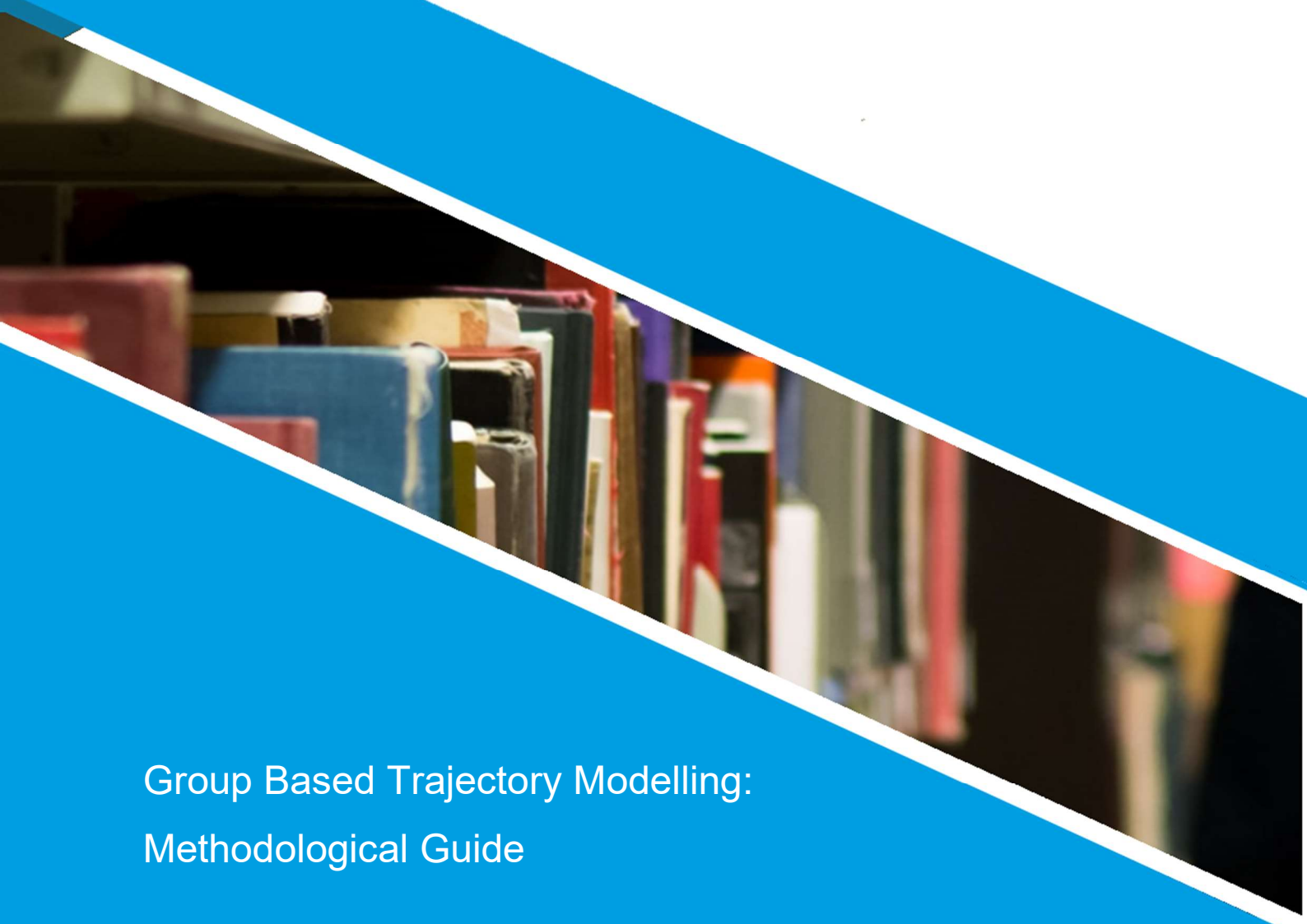
COMPLEX TRAJECTORIES



Group Based Trajectory Modelling: Methodological Guide

Co-funded by the Erasmus+ Programme of the European Union

Author:

Albert Sánchez-Gelabert

Project:

Erasmus+

The Complex Trajectories project is funded by European Union Erasmus+ grant agreement No. KA203-082842, 2020-2023.

Consortium members:

Universitat Oberta de Catalunya

Universitat Autònoma de Barcelona

Generalitat de Catalunya
Departament d'Empresa i Coneixement
Secretaria d'Universitats i Recerca

Universitat de València

Université de Bourgogne

The Open University

U.PORTO
FEUP Faculdade de Engenharia
Universidade do Porto

Universidade Aberta
www.uab.pt

Co-funded by the Erasmus+ Programme of the European Union

Contents

Co-funded by the Erasmus+ Programme of the European Union

1. Group-based Trajectory Modelling (GBTM)

Group-based trajectory modelling (GBTM) is a statistical methodology for analysing developmental trajectories - the evolution of an outcome over age or time (Nagin, 2014). GBTM is a specialized application of finite mixture models using maximum likelihood estimation and allows us to analyse developmental trajectories and to describe and explain changes over a relatively long period of time (Nagin, 2005). At the same time, it allows us to summarize individual differences in the developmental progression of a variable.

According to Joengbloed's working paper (Jongbloed, 2021)[1] to perform the GBTM analysis, we will follow the next steps:

1. Create a hypothesis of a plausible number of groups based on theory/literature.
2. Refine the model from step 1 to determine:
   a) the optimal number of groups, typically testing K=1–7 groups.
   b) the optimal shape of the trajectories, typically testing linear, quadratic, and cubic functions of each trajectory.
3. Assess model fit using Bayesian information criteria (BIC) values, average posterior probability of assignments (APPA), and odds of correct classification (OCC).
4. Investigate graphical presentations and assess for substantive interpretation.

In the following pages an example of a GBTM analysis is performed using the "traj" module in Stata with data from the Catalan university system. It is worth mentioning that there are also alternatives to perform GBTM with free software such as R and packages like "crimCV", "LCMM" (Latent Class Mixed Models) or flexmix.

In order to perform the analysis, different documents, guides and tutorials have been followed. One of the most useful and easy to follow documents is the article "Latent Class Growth Modelling: A Tutorial" (Andruff, Carraro, Thompson, Gaudreau, & Louvet, 2009), in which the steps to follow with SAS are explained.

Also the following documents can be very useful to perform the GBTM analysis. In them you can find examples of analysis and syntax that can solve some doubts:

- Jones, B. L., & Nagin, D. S. (2012). *A Stata Plugin for Estimating Group-Based Trajectory Models*.

[1] Also referred to on "video 1 – GBTM fundamentals" of this same Unit 4.

Co-funded by the Erasmus+ Programme of the European Union

- van der Nest, G., Lima Passos, V., Candel, M. J. J. M., & van Breukelen, G. J. P. (2020). An overview of mixture modelling for latent evolutions in longitudinal data: Modelling approaches, fit statistics and software. Advances in Life Course Research, 43(January), 100323. https://doi.org/10.1016/j.alcr.2019.100323

1.1.1. Preparing the database for the analysis

The variable to be analysed with the GBTM must be a continuous variable, measured repeatedly at different time points. In addition, it is necessary to include a variable that will indicate how many times the observation is made.

A decision must be taken on how to consider the cases that drop out of the sample. In our case, students who drop out of the system are considered, at the points in time when they are not enrolled, as zeros (a student does not obtain credits during these courses).

It is likewise necessary to consider whether obtaining a degree entails completing the observation of this individual. We have opted for this solution, considering that a graduated student has reached the position where no longer has to obtain credits. If we end thus the observation of graduates, they remain in this status and are missing from that moment on, regardless of whether they go on to do a Master's degree or another degree.

1.1.2. Installation and syntax

1.1.2.1 Installation and Syntax

Traj can be installed by issuing the following commands within Stata.

```
. net from https://www.andrew.cmu.edu/user/bjones/traj

. net install traj, replace
```

The syntax to perform the GBTM is as follows

```
traj [ if exp ] , var(varlist) indep(varlist) model(string) order(numlist)

[min(real) max(real) iorder(numlist) risk(varlist) tcov(varlist) plottcov(matrix) start(matrix) weight(varname) exposure(varlist) refgroup(integer) dropout(numlist) dcov(varlist) obsmar(varname) outcome(varname) omodel(string) detail ]
```

According to this structure, an example of a command to perform the GBTM analysis for university trajectories with our data could be the following:

```
> traj, var(sup1-sup7) indep(c1-c7) model(cnorm) min(0) max(240) order(1 1 1 1)
```


Co-funded by the Erasmus+ Programme of the European Union

From this syntax we can identify different elements:

- Trajectory Variables
  - var(varlist) dependent variables, measured at different times or ages (required). In our example, var (sup1-sup7) are the credits passed (sup) in the different moments in time that we have observations (sup1 = credits passed in 2012, ..., sup7 = credits passed in 2019.
  - indep(varlist) independent variables (i.e. when the dependant variables were measured) (required). In the syntax of the example, "indep(c1-c7)" refers to the variables containing information about the different observed times (c1 = 2012, c2=2013, ..., c7 = 2019).
- Model
  - Model (string) probability distribution for the dependent variables (required).
  - order(numlist) polynomial type (0=intercept, 1=linear, 2=quadratic, 3=cubic) for each group trajectory (required).
  - min(real) minimum value for the censored normal model (required for cnorm). In our example, min = 0 ECTS passed.
  - max(real) maximum value for the censored normal model (required for cnorm). In our example, max = 240 ECTS passed.

Two of the most important elements that can generate more confusion in this syntax are the command "model" and "order". The first element ("model") refers to probability distribution for the dependent variables. The second (order) refers to the development of the trajectories over the time analyzed (or the polynomial function) and the number of groups of the typology.

Before estimating the number of groups, we will need to define from the distribution of the dependent variable, one of the three alternative ways to calculate the probability. STATA supports these models:

- Censored Normal (CNORM) Model
- Zero Inflated Poisson (ZIP) Model
- Logistic (LOGIT) Model

Based on this distribution, we could modify the "model" part of the command as follows:

```
> traj, var(sup1-sup7) indep(c1-c7) model(cnorm) min(0) max(240) order(1 1 1 1)
> traj, var(sup1-sup7) indep(c1-c7) model(zip) min(0) max(240) order(1 1 1 1)
> traj, var(sup1-sup7) indep(c1-c7) model(logit) min(0) max(1) order(1 1 1 1)
```


Co-funded by the Erasmus+ Programme of the European Union

Since in the example we are analysing the number of credits passed by students each academic year, we will use the censored normal model (CNORM):

```
> traj, var(sup1-sup7) indep(c1-c7) model(cnorm) min(0) max(240) order(1 1 1 1)
```

1.1.3. Estimation of optimal number of groups

Based on the literature review and previous empirical results, we hypothesize a plausible number of groups of students' university trajectories. We must take into account the variable on which we define these university trajectories: performance, enrolment, success rate, change of studies, graduation, etc.

The specific procedure for the estimation of the groups consists, in a first stage, of successively estimating different models with a different number of groups. The model with the highest (least negative) value of BIC is the preferred one, therefore, we will test different models until the model fit does not improve according to the values of the BIC statistic. Normally, models with k=1-7 groups are tested. Thus, the number of groups will correspond to the model where the BIC values stop decreasing.

In the following example, we will analyse the trajectories of students in relation to their performance rate in the period 2012-2018. The values of the dependent variable are a ratio from 0 to 1 that refers to the credits passed over credits enrolled. Thus, the dependent variable follows a censored normal distribution (i.e. "model(cnorm)"). Accordingly, we will test the different models with groups of k=1 to 7 following the syntax below:

```
> traj, var(tr1-tr7) indep(c1-c7) model(cnorm) min(0) max(1) order(1)
> traj, var(tr1-tr7) indep(c1-c7) model(cnorm) min(0) max(1) order(1 1)
> traj, var(tr1-tr7) indep(c1-c7) model(cnorm) min(0) max(1) order(1 1 1)
> traj, var(tr1-tr7) indep(c1-c7) model(cnorm) min(0) max(1) order(1 1 1 1)
> traj, var(tr1-tr7) indep(c1-c7) model(cnorm) min(0) max(1) order(1 1 1 1 1)
> traj, var(tr1-tr7) indep(c1-c7) model(cnorm) min(0) max(1) order(1 1 1 1 1 1)
> traj, var(tr1-tr7) indep(c1-c7) model(cnorm) min(0) max(1) order(1 1 1 1 1 1 1)
```

Following this approach, we analyse how the BIC and the log Bayes Index (2 log (B10)) evolve to make a selection of the number of groups. The log Bayes Index 2 log e (B10), as explained by Nagin et al. (Jones & Nagin, 2013), is a fit index used to compare competitive models that include different numbers of trajectories. The nested models allow comparing the model fits according to the log Bayes Index as follows:

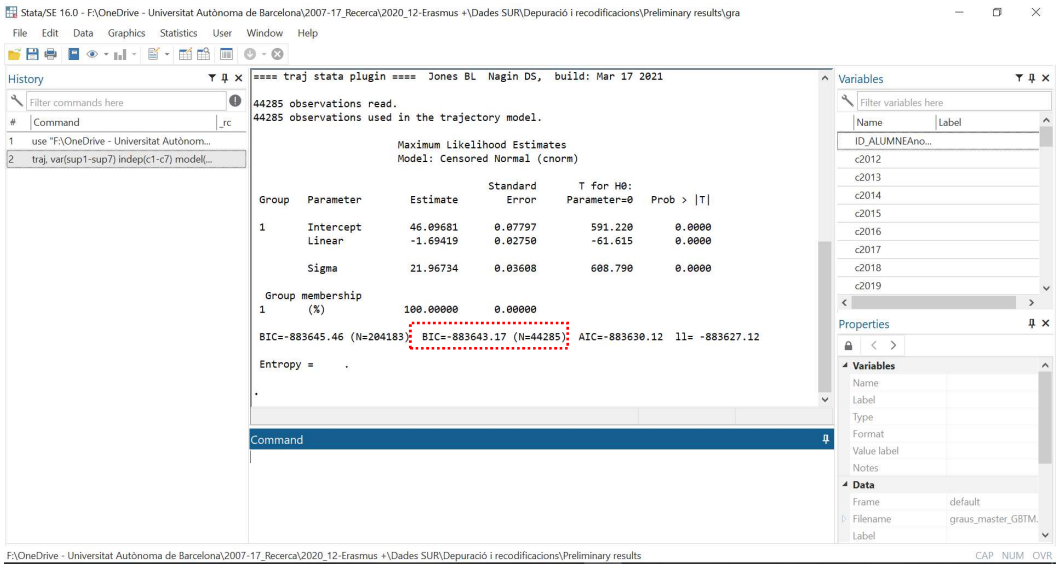
Co-funded by the Erasmus+ Programme of the European Union

Table. Interpretation of 2 log $_e$ (B_{10})

2 log$_e$ (B_{10})	**(B_{10})**	**Types of evidence against H0**
0 to 2	1 to 3	Weak evidence
2 to 6	3 to 20	Moderate evidence
6 to 10	20 to 150	Strong evidence
> 10	> 150	Very strong evidence

> traj, var(sup1-sup7) indep(c1-c7) model(cnorm) min(0) max(240) order(1)

Here below we see the STATA screenshot where we can find the value of the BIC:

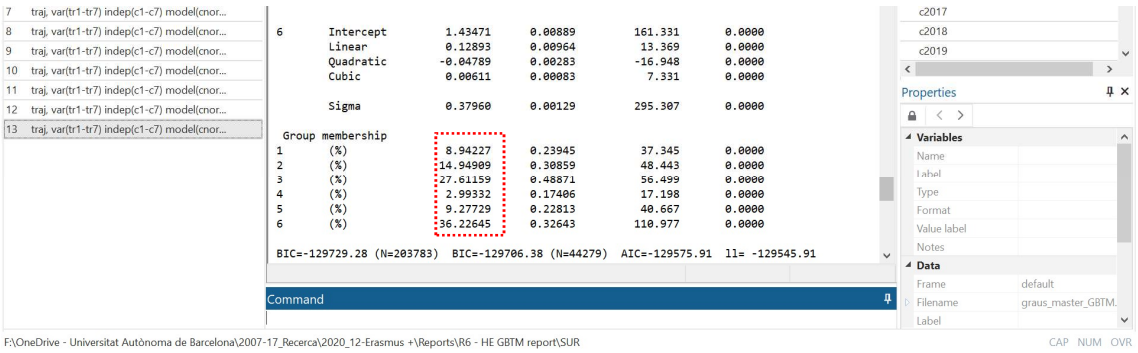


The following table shows the results when we analyse some data from the example of the Catalan university system as a whole. When analysing the number of groups that best fits the data, it is observed that the BIC decreases every time we introduce a new model until k = 7. Although the BIC reduces up to k = 7, it is observed that the reduction is very small and perhaps does not provide enough information. In this sense, another criterion for selecting the number of groups is that none of the resulting groups should include less than 5% of the total cases (students in our case).

Co-funded by the Erasmus+ Programme of the European Union

Groups	BIC	2 log e (B10)[2]	G < 5%
1	-160152,45		
2	-138369,11	43566,68	0
3	-133725,77	9286,68	0
4	-132082,90	3285,74	0
5	-130704,34	2757,12	0
6	-130003,80	1401,08	1
7	-129856,68	294,24	1
8	-129402,31	908,74	1

From the results of the previous table, we will select 5 groups, since after entering the model with 6 groups, one of the resulting groups is less than 5% of the total. The percentage of each group can be found in the first column of the "Group membership" section of the results table. As we can see in the image, in the model of 6 groups, one of the resulting groups has a percentage lower than 5% of the total cases.



1.1.4. Optimal shape of the trajectories

Once a number of groups has been estimated, the models of the trajectories are estimated based on the polynomial functions (the shapes of the trajectories). These trajectories can be modelled as linear, quadratic or cubic. In general, we can identify:

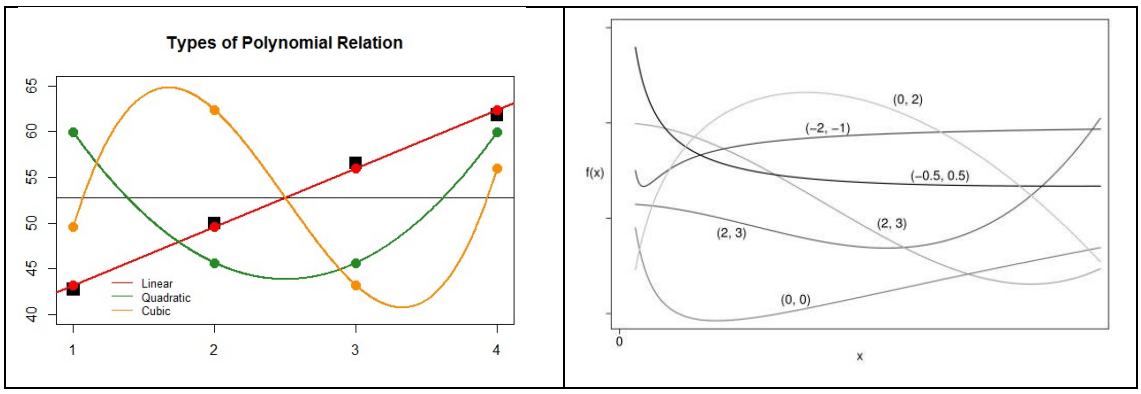
- A linear change trajectory is defined by a linear trend that can increase or decrease steadily or remain stable (red line).

[2] Excel will be very useful here, as this column can be easily calculated introducing the formula where, for instance, BIC of group 2 minus BIC of group 1 then multiplied by 2 results in the 2 log e (B10) corresponding to group 2.

Co-funded by the Erasmus+ Programme of the European Union

- The quadratic trajectory is defined by an increase, decrease or stability until a certain point in time when it changes in magnitude or direction (green line).
- Cubic trajectory is defined by various changes in magnitude or direction over time and will not remain stable (yellow line).

In the first graph, we can see the main trajectories according to 4 moments in time while on the right we see some of the multiple forms of trajectories when we have many observations:



In our case, we have 7 observations or academic years, therefore, the trajectories can be linear, cubic or quadratic. We will have to analyse which combination of polynomial forms is the most optimal for each of the groups we have selected. There are different articles that explain how to make the selection of these trajectories and explain the procedure to follow. The document "Latent Class Growth Modelling: A Tutorial" explains how to try different models and make this selection of groups taking into account, at the same time, the shape of the polynomial function (Åhlin, Westerlund, Griep, & Magnusson Hanson, 2018; Andruff et al., 2009).

In order to select which functions best fit the data, we have several criteria for good model fit:

- The Average Posterior Probability Assignment (APPA) of each group to a trajectory examines whether individuals are assigned with high probability and the overall average probability of assignment to each group. The criterion for a good model is that the APPA > 70% for each class.
- Odds of correct classification (OCC) is the proportion of the odds of a correct classification in each group based on the maximum likelihood classification rule and the estimated proportions of class members. The criterion for a good model is that OCC > 5.0 for each class.
- Entropy is a global measure of classification uncertainty, which takes into account all subsequent probabilities. Entropy takes values from [0, ∞), with higher values indicating greater uncertainty. The criterion of a good model is when the entropy values are closer to 0 (since they correspond to models with lower classification uncertainty).

Co-funded by the Erasmus+ Programme of the European Union

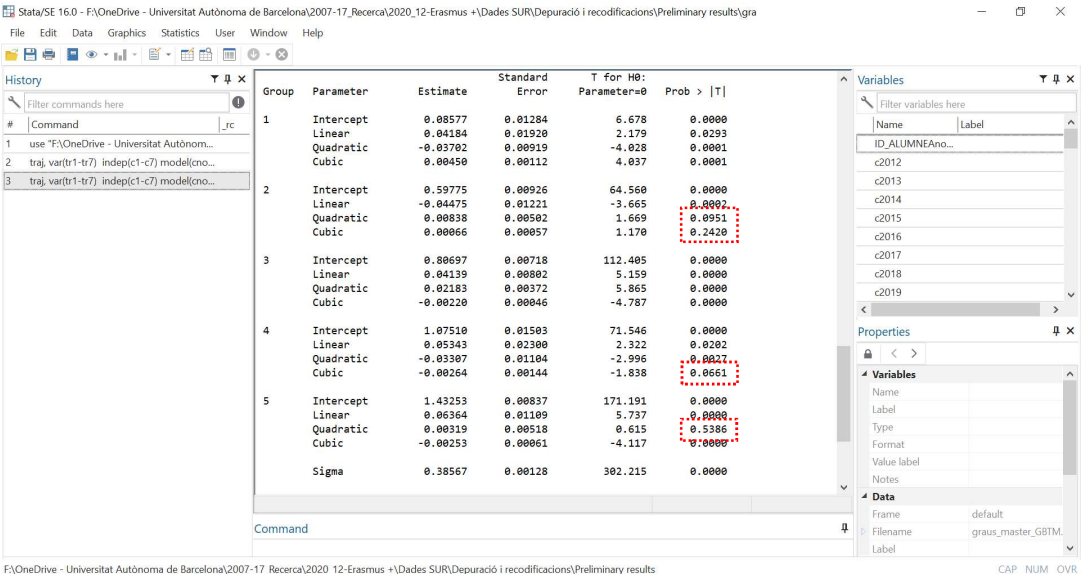
Therefore, to analyse which model best fits the data, we will test different models combining different polynomial forms. Here we can see the STATA command for the analysis of the 6-group model of the previous example. The different polynomial options we can test are 0=intercept, 1=linear, 2=quadratic, 3=cubic. In our case, we start with a model with the 5 cubic trajectories (3) and we will vary the function if the model does not improve or if there is a non-significant trajectory.

> traj, var(tr1-tr7) indep(c1-c7) model(cnorm) min(0) max(1) order(3 3 3 3 3)
> traj, var(tr1-tr7) indep(c1-c7) model(cnorm) min(0) max(1) order(3 2 2 3 1)
> traj, var(tr1-tr7) indep(c1-c7) model(cnorm) min(0) max(1) order(3 2 2 2 1)
> traj, var(tr1-tr7) indep(c1-c7) model(cnorm) min(0) max(1) order(3 2 2 1 1)

...

Model	Polynomic function	BIC	2 log e (B10)	G < 5%
0	11111	-130704,34		
1	33333	-130253,82	901,0	1
1.2	32231	-130261,69	885,3	1
1.3	32221	-130394,42	619,8	0
1.4	32211	-130389,14	630,4	0

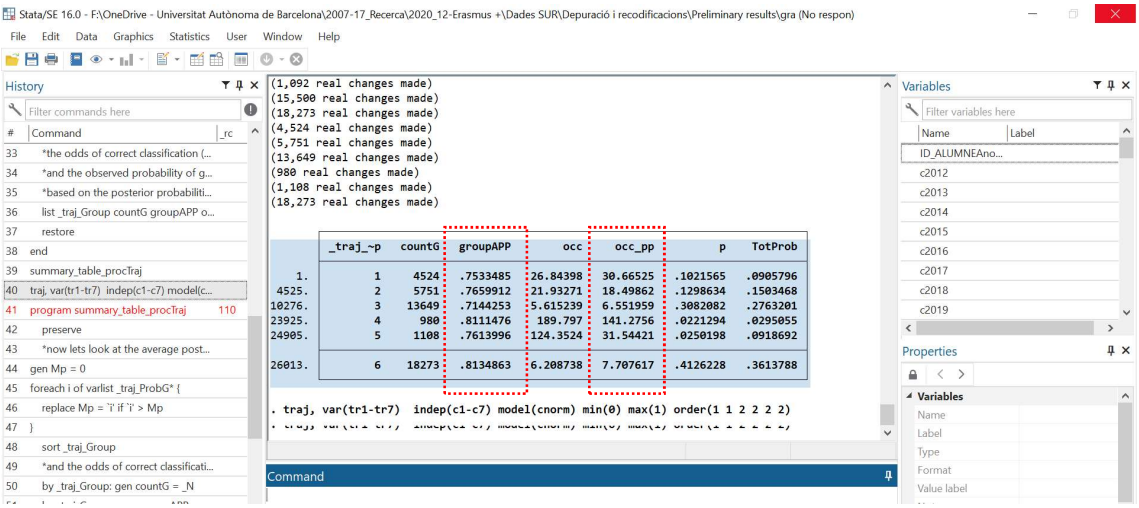
To make the selection of the type of trajectories to choose for each group we will also analyse the significance for each parameter displayed in the Prob > |T| column. The T for H0: Parameter = 0 column provides a value for the test of the null hypothesis that determines whether the parameter is significant or not.



Co-funded by the Erasmus+ Programme of the European Union

As we can see in the STATA screenshot, different trajectories are not significant (e.g. cubic and quadratic of the second group). So, we will try different models combining different polynomial functions. The **optimal group** will be the one that has **the lowest BIC value**, **does not contain groups smaller than 5%** and, at the same time, the **trajectories are significant** according to their polynomial function.

Other criteria that will allow us to select the model that best fits the data will be the APPA and OCC for each model. We can use a small macro of the annex from which we will obtain a table with several indicators as in the following screenshot:



The following table is constructed from the STATA outputs for the different combinations analysed. As can be seen, the model that best fits the data would be model 1.4 of the table since it is the model where the BIC stops decreasing, where the APPA > 70% for each class, where the OCC > 5.0 for each class and there is no group lower than 5% of the total.

		APPA					OCC				
model	Polynomic function	G1	G2	G3	G4	G5	G1	G2	G3	G4	G5
0	11111	0,71	0,71	0,78	0,77	0,88	23,6	13,9	63,4	7,3	11,4
1	33333	0,74	0,75	0,76	0,81	0,89	27,2	17,1	7,2	83,7	11,7
1.2	32231	0,75	0,81	0,75	0,88	0,76	28,8	81,1	17,5	11,1	7,1
1.3	32221	0,75	0,81	0,75	0,76	0,88	28,8	80,5	17,6	7,1	11,5
1.4	32211	0,75	0,81	0,75	0,76	0,88	28,8	80,5	17,6	7,1	11,5

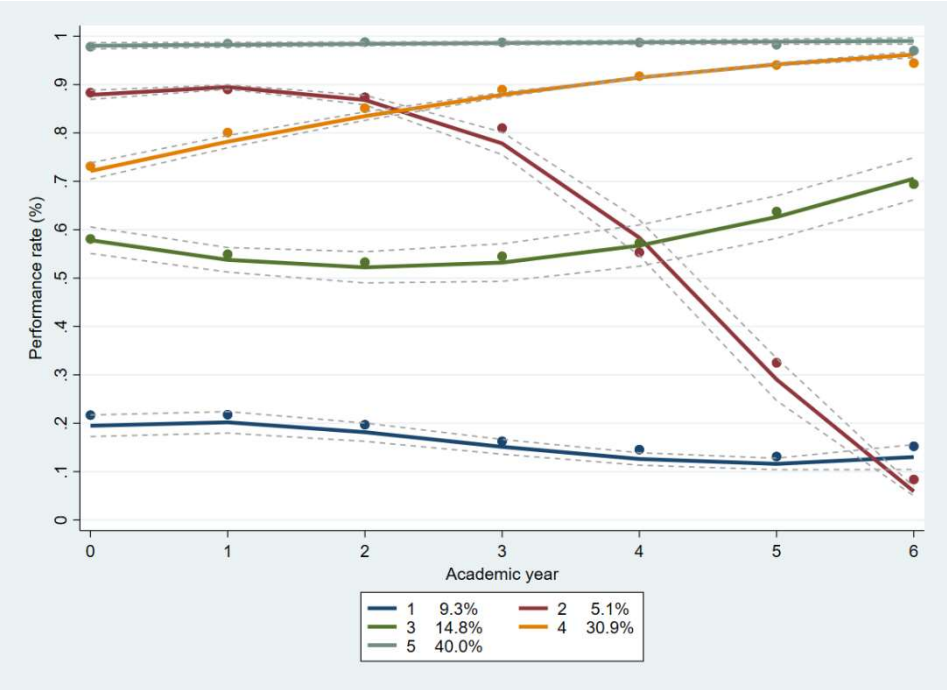
Finally, the model that best fits the data is a 5-group model with one cubic trajectory (3), two quadratic (2) and two linear (1).

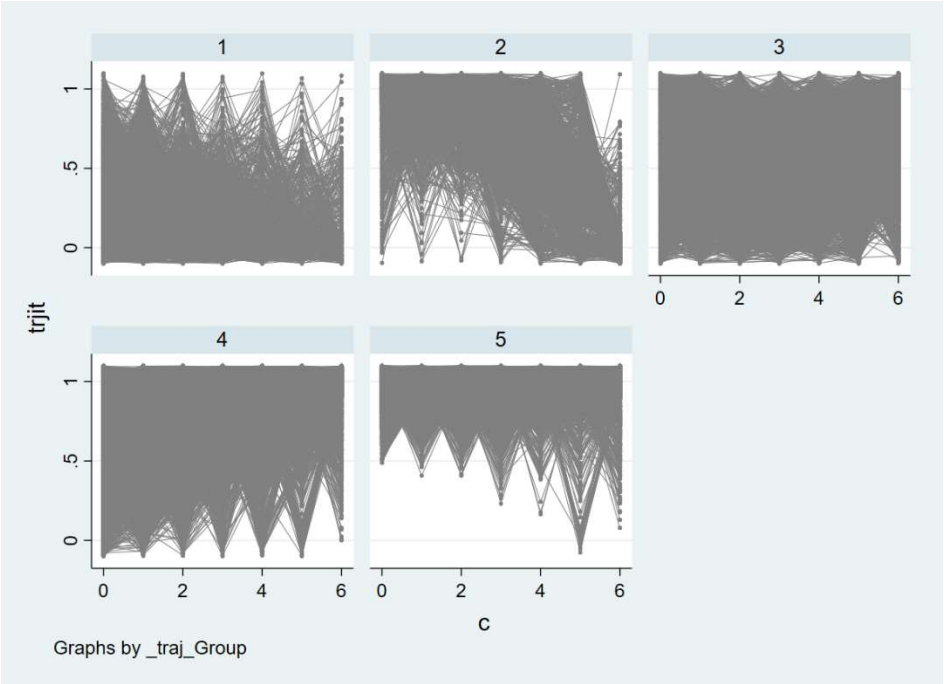
Co-funded by the Erasmus+ Programme of the European Union

1.1.5. Visualization

Finally, a third criterion for selecting the best model refers to the visualization of the graphical representations and the evaluation and interpretation according to theoretical framework of the graphs.

trajplot, xtitle(Academic year) ytitle(Performance rate (%)) xlabel(0(1)6) ylabel(0(0.1)1) ci





Co-funded by the
Erasmus+ Programme
of the European Union

2. Group-Based Multi-Trajectory Modelling (GBMTM)

Another interesting option that can be useful for analysing students' trajectories is to introduce more than one variable to create the types of trajectories. The procedure is very similar to the one we have seen so far but introducing different variables at the same time. However, it is necessary to define the number of groups beforehand and it has to be the same number of groups for the different variables.

2.1. GBMTM with 3 variables: success rate, ECTS passed and graduation

An example syntax for Group-Based Multi-Trajectory Modelling with three variables would be:

traj, multgroups(5) var1(tr1-tr7) indep1(c1-c7) model1(cnorm) min1(0) max1(1) order1(3 2 2 1 1) var2(sup1-sup7) indep2(c1-c7) model2(cnorm) min2(0) max2(240) order2(1 1 1 1 1) var3(t1-t7) indep3(c1-c7) model3(logit) min3(0) max3(1) order3(1 1 1 1 1)

As we can see, we define the name of groups with the command "multgroups(5)" and then we concatenate the syntax for each one of the variables:

traj, multgroups(5) var1(tr1-tr7) indep1(c1-c7) model1(cnorm) min1(0) max1(1) order1(3 2 2 1 1) var2(sup1-sup7) indep2(c1-c7) model2(cnorm) min2(0) max2(240) order2(1 1 1 1 1) var3(t1-t7) indep3(c1-c7) model3(logit) min3(0) max3(1) order3(1 1 1 1 1)

multtrajplot, xtitle(curs) ytitle1(Success rate (%)) ytitle2(ECTS passed) ytitle3(graduation) xlabel(0(1)6) ylabel1(0(10)60) ylabel2(0(0.2)1) ylabel3(0(10)70)

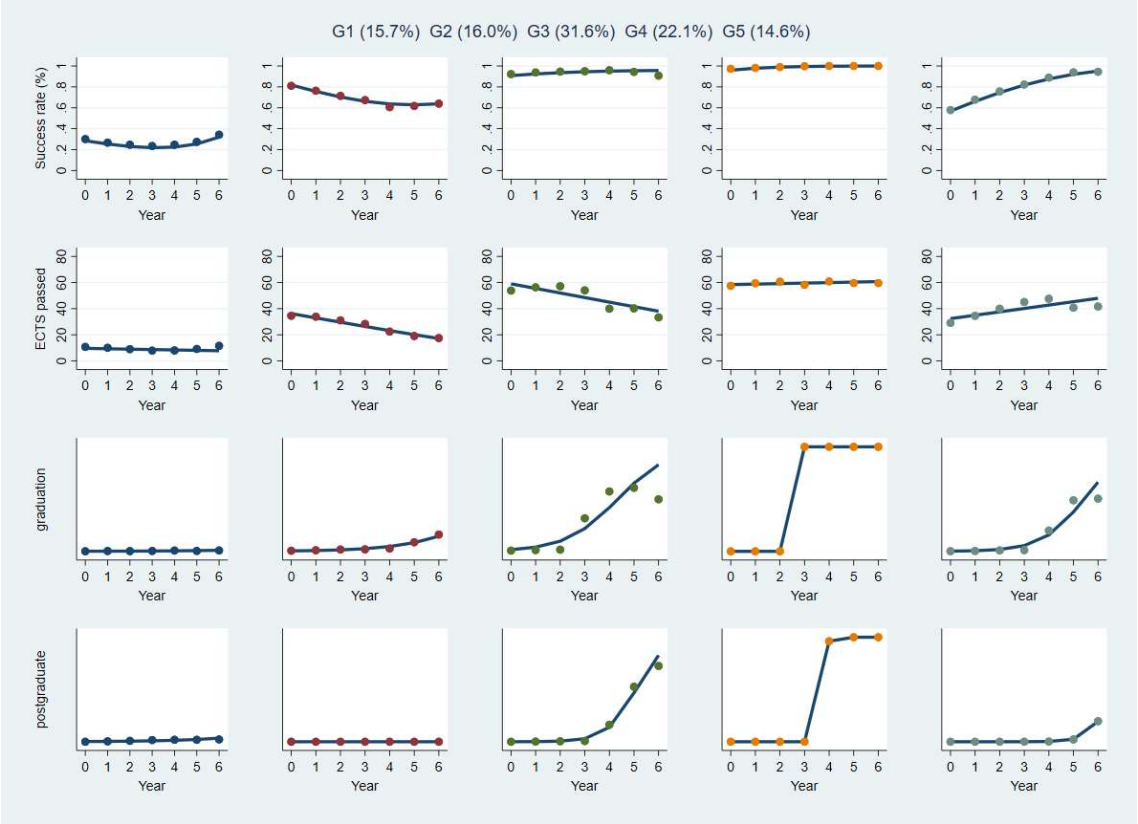
Co-funded by the Erasmus+ Programme of the European Union

2.2. GBMTM with 4 variables: success rate, ECTS passed, graduation and Postgraduate Enrolment

traj, multgroups(5) var1(tr1-tr7) indep1(c1-c7) model1(cnorm) min1(0) max1(1) order1(3 2 2 1 1) var2(sup1-sup7) indep2(c1-c7) model2(cnorm) min2(0) max2(240) order2(1 1 1 1 1) var3(t1-t7) indep3(c1-c7) model3(logit) min3(0) max3(1) order3(1 1 1 1 1) var4(grau_master1-grau_master7) indep4(c1-c7) model4(logit) min4(0) max4(1) order4(1 1 1 1 1)

multtrajplot, xtitle(Year) ytitle1(Success rate (%)) ytitle2(ECTS passed) ytitle3(graduation) ytitle4(postgraduate) xlabel(0(1)6) ylabel1(0(0.2)1) ylabel2(0(20)80) ylabel3(0(0)1) ylabel4(0(0)1)

Co-funded by the Erasmus+ Programme of the European Union



Co-funded by the
Erasmus+ Programme
of the European Union

3. References

Åhlin, J. K., Westerlund, H., Griep, Y., & Magnusson Hanson, L. L. (2018). Trajectories of job demands and control: risk for subsequent symptoms of major depression in the nationally representative Swedish Longitudinal Occupational Survey of Health (SLOSH). *International Archives of Occupational and Environmental Health*, *91*(3), 263–272. https://doi.org/10.1007/s00420-017-1277-0

Andruff, H., Carraro, N., Thompson, A., Gaudreau, P., & Louvet, B. (2009). Latent Class Growth Modelling: A Tutorial. *Tutorials in Quantitative Methods for Psychology*, *5*(1), 11–24. https://doi.org/10.20982/tqmp.05.1.p011

Jones, B. L., & Nagin, D. S. (2012). *A Stata Plugin for Estimating Group-Based Trajectory Models*.

Jones, B. L., & Nagin, D. S. (2013). A Note on a Stata Plugin for Estimating Group-based Trajectory Models. *Sociological Methods and Research*, *42*(4), 608–613. https://doi.org/10.1177/0049124113503141

Jongbloed, J. (2021). Group-based trajectory modeling. *Complex Trajectories Methodological Group (MGroup)*. Bourgogne: IREDU, Universitéde Bourgogne.

Lennon, H., Kelly, S., Sperrin, M., Buchan, I., Cross, A. J., Leitzmann, M., … Renehan, A. G. (2018). Framework to construct and interpret latent class trajectory modelling. *BMJ Open*, *8*(7). https://doi.org/10.1136/bmjopen-2017-020683

Nagin, D. S. (2005). *Group-Based Modeling of Development*. Cambridge, MA.: Harvard University Press.

Nagin, D. S. (2014). Group-based trajectory modeling: An overview. *Annals of Nutrition and Metabolism*, *65*(2–3), 205–210. Karger Publishers. https://doi.org/10.1159/000360229

van der Nest, G., Lima Passos, V., Candel, M. J. J. M., & van Breukelen, G. J. P. (2020). An overview of mixture modelling for latent evolutions in longitudinal data: Modelling approaches, fit statistics and software. *Advances in Life Course Research*, *43*(January), 100323. https://doi.org/10.1016/j.alcr.2019.100323

Co-funded by the Erasmus+ Programme of the European Union

4. Appendix

4.1. Function to obtain the APPA and the OCC

The following function to obtain the APPA and OCC indicators has been created by Andrew Wheeler and extracted from the website https://andrewpwheeler.com/tag/group-based-trajectory/:

```
program summary_table_procTraj
   preserve
   *now lets look at the average posterior probability
   gen Mp = 0
   foreach i of varlist _traj_ProbG* {
      replace Mp = `i' if `i' > Mp
   }
   sort _traj_Group
   *and the odds of correct classification
   by _traj_Group: gen countG = _N
   by _traj_Group: egen groupAPP = mean(Mp)
   by _traj_Group: gen counter = _n
   gen n = groupAPP/(1 - groupAPP)
   gen p = countG/ _N
   gen d = p/(1-p)
   gen occ = n/d
   *Estimated proportion for each group
   scalar c = 0
   gen TotProb = 0
   foreach i of varlist _traj_ProbG* {
     scalar c = c + 1
     quietly summarize `i'
     replace TotProb = r(sum)/ _N if _traj_Group == c
   }
   gen d_pp = TotProb/(1 - TotProb)
   gen occ_pp = n/d_pp
   *This displays the group number [_traj_~p],
   *the count per group (based on the max post prob), [countG]
   *the average posterior probability for each group, [groupAPP]
   *the odds of correct classification (based on the max post prob group assignment), [occ]
   *the odds of correct classification (based on the weighted post. prob), [occ_pp]
   *and the observed probability of groups versus the probability [p]
   *based on the posterior probabilities [TotProb]
   list _traj_Group countG groupAPP occ occ_pp p TotProb if counter == 1
   restore
end
summary_table_procTraj
```


Co-funded by the Erasmus+ Programme of the European Union