Analysis of the salary trajectories in Luxembourg : a finite mixture model approach

Jean-Daniel Guigou[1], Bruno Lovat[2], Jang Schiltz[1]

[1] Luxembourg School of Finance
University of Luxembourg, Luxembourg
(e-mail: jang.schiltz@uni.lu)
[2] Université de Nancy II
Nancy, France
(e-mail: bruno.lovat@univ-nancy2.fr)

Abstract. We analyse the salaries of about 700.000 employees who worked in Luxembourg between 1940 and 2006 with the aim of detecting groups of typical salary trajectories with respect to some covariants like sex, workstatus, residentship and nationality. We use the proc traj SAS procedure from Bobby L. Jones to classify the workers and descriptive statistical methods like the CHAID procedure or multinomial logistic regression to get a caracterisation with respect to the covariants of the different groups.

Keywords: Finite mixture models, Salary trajectories, CHAID, Multinomial logistic regression, Proc-Traj procedure, Economic modeling.

1 Introduction

Knowing the salary structure in a country is of great importance for a host of economic applications, for instance for an analysis of its pension system. We highlight the evolution of salaries in Luxembourg. To this end, we use the recent statistical group based trajectory model of D. Nagin [8]. We estimate model parameters from a single database, provided by the general social security inspection office (IGSS) and containing annual salaries of all wage earners in the Luxembourg private sector. As a result we divide up the population into nine groups, each with its own mean salary trajectory in time and its relative weight in society.

In a second part of the paper we give a socioeconomic description of the nine groups and adress the question of the prediction of group membership for a given individual. These kind of results are of great interest for insurance companies and banks who like to know the evolution of the career of their customers to be better able to advice them on the possibilities of money investments in a pension fund for instance.

2 A statistical method based on clustering

Longitudinal data are the empirical basis of research on various subjects in the social sciences and in medicine. The common statistical aim of these various application fields is the modelisation of the evolution of an age or time based phenomenon. In the 1990s, the generalized mixed model assuming a normal distribution of unobserved heterogeneity [1](Bryk and Raudenbush 1992), latent growth curves modeling (Muthén 1989 [7]) and the nonparametric mixture model, based on a discrete distribution of heterogeneity (Jones, Nagin and Roeder 2001 [5]) have emerged. We choose this variation of the generalized mixed model because of the growing interest in this approach to answer questions about atypical subpopulations (see Eggleston, Laub and Sampson 2004 [3]).

The SAS procedure Proc Traj, programmed by Daniel Nagin and Bobby Jones[5], allows to estimate the parameters of a semiparametric mixture model for longitudinal data that are either normal (censored) distributed or follow a Poisson or Bernoulli distribution. The subgroup trajectories can be modeled by polynomials up to the fourth degree. The procedure enables to calculate the posterior probability of group membership in terms of risk factors that are stable in time. Moreover, time-dependent covariates can influence the trajectories and cause different effects in different subgroups.

Nagin's nonparametric mixed model [8] starts from a set of individual trajectories and tries to divide the population into a number of homogeneous sub-populations and to estimate a mean trajectory for each of these subpopulations.

Consider a statistical variable Y defined on a population of size N. Let $Y_i = y_{i1}, y_{i2}, ..., y_{iT}$ denote a longitudinal sequence of measurements on individual i over T periods.

Let $P(Y_i)$ dennote the probability of Y_i. The purpose of the analysis is to find r trajectories of a given type, in general polynomials of degree 4, $P(t) = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \beta_4 t^4$. Let $P^j(Y_i)$ denote the probability of obtaining the observed data for individual i given membership in group j and π_j the probability of an individual chosen at random to belong to the group number j.

We try to estimate a set of parameters $\Omega = \{\beta_0^j, \beta_1^j, \beta_2^j, \beta_3^j, \beta_4^j, \pi_j; j = 1, ..., r\}$ which maximises the probability of Y_i. The ideal number of groups r is also an outcome of the analysis. For a given group, conditional independence is assumed for the sequential realisations of the elements of Y_i, y_{it}, over the T periods of measurment. The likelihood L of the sample is then given by

$$L = \frac{1}{\sigma} \prod_{i=1}^{N} \sum_{j=1}^{r} \pi_j \prod_{i=1}^{T} \phi \left(\frac{y_{it} - \beta^j x_{it}}{\sigma} \right),$$

where ϕ denotes the density function of the standard normal distribution. These equations are too complicated to hope to obtain an algebraic solution.

Bobby L. Jones (Carnegie Mellon University) has programmed a SAS procedure based on a quasi-Newtonian maximum search method (Dennis, Gay & Welsch 1981[2]). The estimated standard deviations are obtained by inverting the observed information matrix.
Nagin's model also allows to determine to wich group a given individual belongs. The posterior probability $P(j/Y_i)$ for an individual i to belong to group number j is indeed given by the Bayes theorem:

$$P(j/Y_i) = \frac{P(Y_i/j)\hat{\pi}_j}{\sum_{j=1}^{r} P(Y_i/j)\hat{\pi}_j}$$

A large posterior probability estimate for a small group requires that Y_i be so strongly consistent with the small group that $P(Y_i/j)$ for that group is very large in comparison to its companion probabilities for the big groups (Nagin 2005).

3 The IGSS database

The analysis relies on a file containing the salaries of all employees of the private sector in Luxembourg. The data cover the period from 1940 to 2006. Since the file contains the careers of those started to work from the beginning of the forties onwards, it is not complete during the first years, but becomes so gradually. In particular it includes all the employees of the private sector in Luxembourg from the beginning of the seventies till 2006.

This file originates from the General Inspectorate of Social Security (IGSS). The main variables are the net annual taxable salary, measured in constant euros (2006 euros), sex, age at first employment, residence and nationality (Luxembourg national living in Luxembourg, foreigner living in Luxembourg and commuters) and the type of employment contract (blue or white collar worker). Initially, the file consisted of about 7 000 000 lines showing the salaries of some 718 054 workers. Many careers are incomplete for many reasons. Moreover, for immigrant workers, we know only the part of their careers made in Luxembourg and know nothing about what they have done in their country of origin. Finally, the percentage of employees who quit prematurely with a disability pension or take pre-retirement or quit early for family reasons (women stopping work or interrupting their work to look after their children for example) is around 50 per cent. The domestic employment (which includes the commuters working in Luxembourg) has experienced strong growth since the mid-eighties, with an average increase of 3.5% annually and an increase of more than 110 000 jobs between 1986 and 2001 (compared to 20 000 jobs in the period 1975-1985) (Source: STATEC). The development of the financial and the growing needs of the public sector are key drivers of this evolution. Today, the services sector represents more than three-quarters of total employment. These changes are not without consequences in terms of professional status, so that changes in careers before the 80s are necessarily significantly different from those of twenty-five years. We have therefore decided to pay interest careers of individuals who began working in Luxembourg between 1982 and 1986 and who have worked for at least 20 years. The final file used for our analysis includes data from 22 203 employees and private workers. Note that in Luxembourg, the maximum contribution ceiling on pension insurance is 5 times the minimum wage, or 7 577 (Euro 2006) per month. Wages in our data are also limited by that number.

4 The mean salary trajectories in Luxembourg

We used the SAS procedure Proc Traj, programmed by Daniel Nagin and Bobby Jones, to determine the mean salary trajectories of 22 203 people who began working between 1982 and 1987 in the private sector in Luxembourg and who worked for at least twenty years.

We established the trajectories for models with between 4 and 20 groups. As the salary trajectories form more or less a continuum in the continuous functions from [1000, 4000] with values in the interval [1200, 7577], the BIC adjustment criterion for determining the ideal number of groups is not well suited. Indeed, BIC increases with the number of groups. This is quite normal, since it just shows that if one assumes more groups, one can necessarily represent reality with more details. On the other hand, one creates smaller groups and an explanatory model more complicated to use. After discussion with the IGSS, we decided to retain a 9 groups solution, for it gives a good representation of the career development in Luxembourg. Solutions with more groups add essentially parallel paths to those present in our model.

To test the stability of trajectories in time, we also established the trajectories for the first 15 years of the careers (careers starting between 1985 and 1992) and for full career of 40 years (careers starting between 1960 and 1967). The trajectories of the first 15 years are very close to the first 15 years of trajectories of 20-year careers. The sizes of the groups vary between 2 and 6% compared to the ones we found and the changes are due to gains or losses to groups with similar salaries. Since moreover the macroeconomic situation has not changed dramatically during the last twenty years, we are fairly confident that the trajectories remain valid, except in cases of severe economic shock that could certainly change the situation completely. The only thing that might change over the coming years is the percentage of commuters in the different groups. Since the total number of employees increased faster than the population can do this percentage will continue to grow in all groups. The trajectories of the complete careers are also quite similar to ours, except that they show a clear decline of the wages during the steel crisis for the trajectories representing the high wages.
Figure 1 below shows the average salary trajectories in our 9 groups.

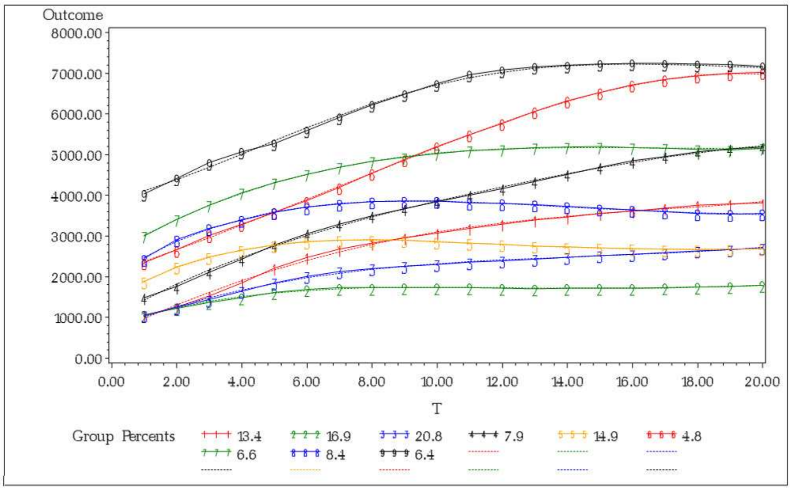


Fig. 1. Salary trajectories of the 9 groups

5 Partition of the labour force

We have investigated the nine groups obtained here above by means of numerous data analysis techniques in order to get a description of each of them. The details of the undertaken tasks can be found in [4]. Summerising our results, we can conclude that the different group trajectories are differentiated mainly by two factors: the starting salary, strongly related to the age at first employment (and therefore to the academic degree) of the employee and the dynamics of his career. There are actually three different types of dynamics: Groups 2, 5 and 8 show "flat" careers, meaning that people from these groups have almost no increase in their salary after the first five years of their career; groups 3 and 7 show a "normal" salary increase of about one per cent per year and groups 1, 4 and 6 show "dynamic" careers, in which wages increase much over time. The ninth group of trajectories is somewhat apart, since it contains the high salaries that exceed the ceiling of 7 577 contained in our data set. Taking this limitations into account, their path resembles that of the normal dynamics. Another interesting discovery is that in most of the groups workers of Luxembourg nationality have, on average, a more dynamic career than foreign workers and commuters. A difference between men and women can also be shown, but only in the groups with lower salaries. Due to lack of information, it is unfortunately impossible to give a more detailed socioeconomic descriptions of the nine groups. We hope to get more variables about the population in our dataset in the future to be able to obtain a better characterisation of the nine groups.

6 Group membership predictions

Bayes' formula allows to compute the probability of individual i to belong to group number j. It gives the possibility to classify correctly almost all persons, without any ambiguity about group membership. In analysing our full sample, we find indeed an average probability of belonging to one of the groups varying between 92.41 % (group 3) and 99.23 % (group 9). The median probability varies even between 99.12 % (group 3) and 100 % (groups 6 and 9). The practical disadvantage of the Bayes' formula is that we need the salaries of the first twenty years of the career. Hence, it is just useful for people who have already completed more than half of it. Analysing the distribution of salaries in the different groups, we find that in many groups the first three to five years show a relatively high dispersion usually paired with a bi- or even trimodal distribution. After the first years however the salary distribution becomes a normal law with a fairly low dispersion. Figure 2 shows the distribution of the salaries in the first twenty years of their career for the people belonging to group 1. It should hence be possible to correctly predict the group membership if we know the salaries of the first years of the career plus some socioeconomic information. That's what we will try in the sequel.

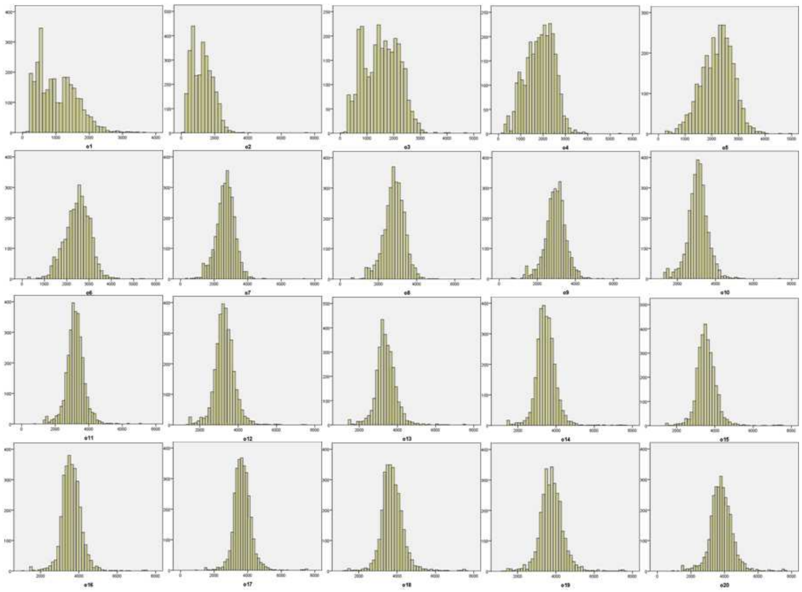
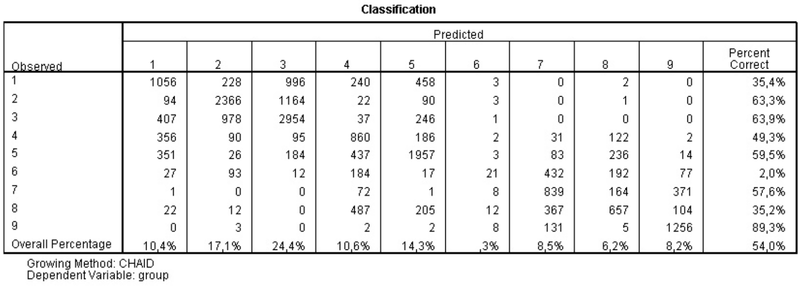


Fig. 2. Salary evolution during the 20 first year of the career for group 1

We tried to predict membership of a given person to one of the nine groups by means two statistical classification methods, the CHAID algorithm (Chi-

squared Automatic Interaction Detector) and multinomial logistic regression. The CHAID algorithm is a technique of decision tree type, published in 1980 by Gordon V. Kass citeKass. It can be used to detect interactions between variables or for prediction purposes. Its a mainly visual and easily interpretable method.
Trying to predict group membership of the entire sample using only the variable "sex", "status", "nationality and residence" and "age at first job in Luxembourg" gives a very bad result. Only 38.3 % of the individuals are correctly classified. The only group that is correctly predicted is group 9 (72.1 %). Moreover, we observe that classification faults are done to the benefit of nearly all groups.

If we consider the socioeconomic variables and the first 6 years of salary, the CHAID algorithm can correctly classify 54 % of the individuals. Groups 2 (63.3 %), 3 (63.9 %), 5 (59.5 %) and 7 (57.6 %) are fairly well predicted, the allocation to the group 9 starts to be good (89.3 %). The second method we used for classification is the multinomial logistic regression. The results are quite similar to those discussed above. Trying to predict group membership of the entire sample using only the variable "sex", "status", "nationality and residence" and "age at first job in Luxembourg" gives a very bad result. Only 35.0 % of people are correctly classified and group 9 is the only correctly predicted group (61.8 %). The pseudo R-square of Cox and Snell, which gives the percentage of the total variability explained by the model is 0.563.
If we consider the socioeconomic variables and the first 6 years of salary, the multinomial logistic regression correctly classifies 53.9% of individuals and the pseudo R-square of Cox and Snell is 0.851, which is a fairly good result. Groups 2 (61.1%), 3 (59.7%), 5 (68.9%) are fairly well predicted, the allocation to the group 9 begins to be good (85.9%).

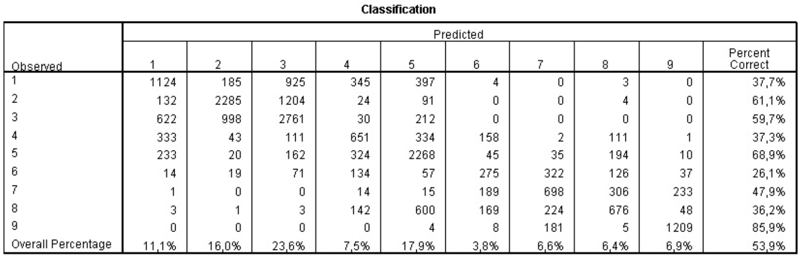


Classification

	Predicted									
Observed	1	2	3	4	5	6	7	8	9	Percent Correct
1	1056	228	996	240	458	3	0	2	0	35,4%
2	94	2366	1164	22	90	3	0	1	0	63,3%
3	407	978	2954	37	246	1	0	0	0	63,9%
4	356	90	95	860	186	2	31	122	2	49,3%
5	351	26	184	437	1957	3	83	236	14	59,5%
6	27	93	12	184	17	21	432	192	77	2,0%
7	1	0	0	72	1	8	839	164	371	57,6%
8	22	12	0	487	205	12	367	657	104	35,2%
9	0	3	0	2	2	8	131	5	1256	89,3%
Overall Percentage	10,4%	17,1%	24,4%	10,6%	14,3%	,3%	8,5%	6,2%	8,2%	54,0%

Growing Method: CHAID
Dependent Variable: group

Fig. 3. Results of the CHAID procedure with 6 years of salary



Classification

Observed	Predicted									
	1	2	3	4	5	6	7	8	9	Percent Correct
1	1124	185	925	345	397	4	0	3	0	37,7%
2	132	2285	1204	24	91	0	0	4	0	61,1%
3	622	998	2761	30	212	0	0	0	0	59,7%
4	333	43	111	651	334	158	2	111	1	37,3%
5	233	20	162	324	2268	45	35	194	10	68,9%
6	14	19	71	134	57	275	322	126	37	26,1%
7	1	0	0	14	15	189	698	306	233	47,9%
8	3	1	3	142	600	169	224	676	48	36,2%
9	0	0	0	0	4	8	181	5	1209	85,9%
Overall Percentage	11,1%	16,0%	23,6%	7,5%	17,9%	3,8%	6,6%	6,4%	6,9%	53,9%

Fig. 4. Results of the multinomial logistic regression with 6 years of salary

7 Conclusion

We have established a classification of the careers in the private sector in Luxembourg into nine groups by means of Nagin's semiparametric mixture model and given a socioeconomic description of the groups. He have seen that the problem of the correct classification of a person in one of nine groups of salary trajectories is a rather complex problem. Considering only the socioeconomic variables, the results are very bad. Add a few years of salary greatly improves the situation and can give a correct result. For the future we try to get some additional socioeconomic variables and to program a classification software that will achieve a good result by combining these variables with the first years of salary.

References

1. Bryk, A.S., *Hierarchical linear models*, Sage, Newbury Park, CA (1992).
2. Dennis, J.E., Gay, D.M., and Welsch R.E., "An Adaptive Nonlinear Lesat-Squares Algorithm",*ACM Transactions on Mathematical Software* 7,348-383 (1981).
3. Eggleston, E.P., Laub, J.H., and Sampson, R.J., "On the Robustness and Validity of Groups", *Journal of Quantitative Criminology* 20-1, 37-42 (2004).
4. Guigou, J.D., Lovat, B., and Schiltz, J., *Les retraites au Luxembourg: modélisation et évaluation d'un système diversifié avec répartition et capitalisation*, Technical report, 112 pages (2010).
5. Jones, B.L., Nagin, D.S., and Roder, K.,"A SAS procedure based on mixture models for estimating developmental trajectories", *Sociological Methods and Research* 29, 374-393 (2001).
6. Kass, G.V.,"An Exploratory Technique for Investigating Large Quantities of Categorical Data", *Journal of Applied Statistics* 29-2, 119-127 (1980).
7. Muthén, B.O., "Latent Variable Modeling in Heterogeneous Populations", *Psychometrika* 54-4, 557-585 (1989).
8. Nagin, D.S., *Group-Based Modeling od Development*, Harvard University Press, Cambridge, Massachusets (2005).