European Journal of Economic and Social Systems **14** N° 1 (2000) 27-36

The sequence of steps in the analysis of youth trajectories

JEAN-PIERRE FÉNELON, YVETTE GRELET AND YVETTE HOUZEL

Abstract. – The study of the school-to-work transition is based on longitudinal data, in the form of personal chronologies which require specific handling. Drawing on French research based on Céreq's longitudinal surveys, this article presents the different options that emerge in the course of statistical processing and explores their respective implications and challenges.

Over the past fifteen years, public-sector statisticians involved with the analysis of the training-employment relationship have been amassing not only the usual socio-demographic data on students and the jobs they hold but also a body of longitudinal information about their itineraries. The resulting treatment has now given rise to a yearly workshop where results and approaches can be compared [1]. Notwithstanding the large number of studies and samples alike, it seems to us that little profit has been drawn from this research in terms of accumulation attainments. While the shifting economic or institutional contexts in which these itineraries are defined make the comparison of results difficult, the harmonisation of the different approaches and the kind of conclusions they lead to is not lacking in interest (Fénelon *et al.*, 1997).

In the analysis that follows, we begin by indicating the specificities of these data and the main approaches used to process them and then propose a detailed indication of what the factorial and typological methods have to offer.

Complex data for analysing the school-to-work transition

The data to which we are referring, and which has been the subject of our research and analysis, comes from the so-called 'itinerary' surveys conducted by Céreq in order to

[1] These meetings, the "Journées du Longitudinal", are organised by Céreq and the Institut du Longitudinal-Iresco (CNRS). The papers presented are published in Céreq's Documents series [1994, 1995, 1996, 1997].

Keywords: coding, trajectory, entry into working life, longitudinal, data analysis, transition.

analyse the school-to-work transition of young people exiting the school system, or in other words, the beginnings of working life [2].

TABLE 1. Status chronology: example from Céreq's survey of young people exiting general or technical secondary education in 1986.

Year	**1985**			**1986**	**1987**	**1988**	**1989**		
Month	10	11	12	...	...	...	...	10	11
In employment	01	02	03					54	55
In military (or alternative) service	01	02	03					54	55
In training	01	02	03					54	55
In full-time studies	01	02	03					54	55
Unemployed	01	02	03					54	55
Absent from labour market	01	02	03					54	55
On parental leave	01	02	03					54	55

We shall not enter into the problems and pitfalls that economists encounter in the concrete translation of the 'transition' concept, in terms of defining the successive steps and parameters of the process as well as its chronology. Suffice it to say that in recent years the employment crisis and transformations of the employment system have complicated youth labour-market entry. The forms of entry are more numerous and more diversified; the amount of time required has increased and government policies on employment and training have come to play a growing role. Numerous questions have thus emerged in relation to the evaluation of these policies (*e.g.*, the impact of government employment schemes on the unemployment rate or the comparative performance of the different streams within the educational system in terms of employment), and they pose new problems for the handling of the data that has been collected over the past fifteen years.

These are individual data that offer highly detailed information on an individual's immediate situation in relation to the employment system at every moment of the observation period. They are generally obtained through retrospective inquiry (one or more times) of samples of school leavers with the help of questionnaires including a detailed chronology of their work status. There are several broad reasons for the complexity of the usable data:

a) First of all, the individual's position within the employment system may be described by different kinds of variables. In most of the analyses carried out to date, it has

[2] Céreq (Centre d'études et de recherches sur les qualifications) is the French centre for research on education, training and employment. It operates under the joint supervision of the Ministry of Education and the Ministry of Labour.

According to M. Vernières, entry into working life is the process through which individuals who have never belonged to the labour force (in this case, young people) attain a stabilised position in the employment system. *Cf.* Vernières (1997), p. 11.

been roughly summed up by the status of the main occupation, in other words, a variable with mutually exclusive items defining, with more or less precision depending on the surveys, situations of precarious or stable employment, job seeking, training, military service or absence from the labour market. The limitations of such a definition, which have been brought out in numerous studies, call for the integration of other descriptive elements. In particular, we know that such a variable should include aspects such as the activity sector or the size of the workplace in addition to the kind of occupations involved. Likewise, job features such as wages or working time are not irrelevant to the quality of the entry into working life as the evaluation procedures seek to measure it.

b) In addition, the individual's status changes over time, and it is necessary to take into account both the moments where one or another of the preceding parameters changes as well as the duration of a given situation. These temporal variables can lead to problems of definition as well as distortions in measuring. Thus, the status chronology, as it is defined at the outset, is limited in time, while the end of the transition is not conceptualized and its observation remains vague. As a result, the observed data are, for the most part at least, censored on the right [3]. It thus becomes necessary to distinguish between the number of months in a status at a given moment and the total time spent in that status, which is only partially known.

The transition (passage from status S1 to status S2) takes place on the date *t* of the chronology, which may be absolute or relative (*e.g.*, to the date of the survey or the time of the exit from the educational system). This transition can be handled independently of its dating, by using it as a significant variable in itself. But if we want to date it, the problem becomes more complicated. There are various methods for including a dating element – using the order in which the transition occurs (the 1st, 2nd, 3rd etc. change since the beginning), modelling in the form of process (Markovian or not) or use of chance functions and time-based models.

In our view, none of these means is sufficient for grasping the subject of analysis that is needed for the evaluation. This is a complex subject that we call 'trajectory' and which, in a synthetic form, traces not only the succession of relevant statuses but also the length of time that they last. The trajectory is the result of a specific structuring of the individual chronology [4]. As a result, when two authors speak of a young person's 'trajectory', it is unlikely that this term applies to the same content because they are referring to subjects that have been structured differently.

c) Last of all, the identification of potential explanatory variables is also complicated. Some of these variables, such as entry into a public employment scheme or the formation of a couple, can be dated and may play a causal role, but they may also simply be included in the trajectory. Still other variables may be considered external to the transition process per se (*e.g.*, gender, geographical location of studies).

For the sake of explanation, these complex data may be seen as corresponding to a cube. Indeed, although the concrete plotting of trajectories takes very different forms, the basic information can always be reduced to a cube $I \times S \times T$, where I is the totality of the

[3] This is, among other things, what justifies the use of time-based models.

[4] See Box for an example of the calendar used in the Céreq surveys.

individuals, *S* the totality of the statuses and *T* the totality of the discrete time elements. The space containing the complex data is "nearly empty" – with a chronology of seven statuses over forty months, for example, there are potentially 7^{40} different trajectories for a sample of several thousand individuals. Furthermore, the space is very structured, and many trajectories are similar. Thus, the classic methods are difficult to apply, and two paths of simplification are generally chosen:

– reducing the potentially multi-dimensional model to its one-dimensional projection, as Courgeau and Lelièvre do in their analysis of biographies because the present state of research does not permit the treatment of the multi-dimensional version (*cf.* Courgeau and Lelièvre, 1989);

– drastically limiting the number of parameters to be estimated for the model considered as theoretically relevant. This is exemplified by the work of Bonnal, Fougère and Sérandon (1994).

A general framework for analysis

A great number of studies on these data have been carried out in France by research units related to Céreq. We have already shown that these studies mainly involve a procedure of evaluation – identifying 'good paths' and attempting to define their main determinants (*cf.* Fénelon *et al.*, 1997). The analysis of this approach thus follows the same basic outline:

– identification of the evaluation criterion (result variable);

– analysis of its links with the 'explanatory' features;

– interpretation of results and formulation of conclusions.

At the same time, however, in methodological and statistical terms, these studies take different paths that are often considered as opposed: econometric modelling (in the form of transition analysis or models over time) and the use of the multidimensional data analysis (simple or multiple factorial analyses, neuronal or other clusterings). The positions reflected in these studies may be summarised schematically by showing how the different steps are reached in each case.

Result variable

In the econometric approach, it is identified, whether it is constructed or available in the data. It involves, for example, the length of time spent in unemployment or employment, or the probability of being unemployed – in other words, variables approaching the dimensions of labour-market entry used in the theoretical construction. The identification of this result variable is a prelude to the calculations.

In the data analysis approach, the search for the result variable is central. It synthesises the individual trajectory, which is then expressed by its membership in a class or its position on a factorial axis.

Analysis of links with 'explanatory' variables

This step constitutes the core of econometric modelling. The goal of the model is to quantify the effects of the explanatory variables or at least to test their existence and meaning. The model is presumed to be complete and exhaustive, and the amplitude of the effect of one variable is conditional on the presence of the other variables.

In the factorial or typological approach, these links are usually established by analysing the cross-tabulations of the result variable with the explanatory variables. An additional step consists of quantifying the effects through regressions (simple, logistic, by proximities, etc.).

Interpretation

Econometric modelling often leads to an explicit determination of causality, especially when there is a temporal effect. The results are directly expressed in terms of probabilities. In the second approach, the interpretation is most often based on complex links expressed in terms of the proportion of variance explained. In either case, the results are applied to the parent population by extrapolation from the sample, either on the basis of hypotheses concerning notably the residuals and the form of the distributions, or on the basis of empirical estimates derived from sampling simulations.

It is clear that these two approaches differ radically in their objectives: in the econometric approach, the result variable is identified at the outset, the evaluation criterion is clearly defined and the modelling bears on its more or less sophisticated links with the individual's other characteristics. By contrast, the data analysis approach is mainly concerned with seeking regular features in the diversity of the trajectories in order to propose a synthesis that will serve as a result variable. It is thus an intermediate step between the structuring of the data and the estimating of a model – the structuring of the trajectory which in fact leads, as we shall see below, to proposing an elaborate synthetic result variable. What is modelled is the trajectory, rather than its links with other individual variables.

In this approach, the individuals constitute a set whose form is studied 'per se', contrary to econometrics, which formalises a unique behaviour pattern that applies to each individual. However, the statistical results, whether principal components or clusters, are not known beforehand but emerge from the statistical treatment. The properties of these operations must be examined in terms analogous to those of the econometric models. These operations lead in turn to new variables, either explanatory or result variables, which combine the properties of interpretability and optimality in very specific forms.

With the preceding vocabulary, the common situations may be represented in the following Table 2.

Steps of statistical treatment

Structuring of trajectories: different uses

This step occurs at the first level of the chain, where the raw data is transformed in order to arrive at the result variable. In modelling, the structuring step is relatively simplified insofar as it involves constructing an indicator that will be taken as a summary, describing a particular feature of the trajectory. Structuring involves the definition of a function

TABLE 2. Successive steps of analysis.

Raw data cube $I \times S \times T$		
	econometric modelisation	data analysis
Step 1: Constructing of result variable	direct	indirect by data analysis
Step 2: Links with explanatory variables	regression transitions duration models logistic regressions	cross-tabulations regression on principal components logistic regressions regressions by proximity

(often a simple summation) from I to $\Re$: each individual is associated with a scalar (total period of unemployment, period of access to first unlimited-term contract, number of transitions, etc.), which we shall call the primary statistical result.

Through the use of data analysis, the operation aims, at this level of the chain, to produce one or several synthetic variables summarising as much as possible all the information on the trajectory. The structuring of the basic data is broken down into two operations:

1a) Prior to the analysis of the data, the information deemed relevant is extracted from the raw data (the cube $I \times S \times T$), and if the variables envisioned for the future treatment are not already directly accessible in the individual chronologies, other elements bringing together this information are constructed. For each individual i, a group of elements is constituted, either simple scalars or more complex items such as 'words', counting vectors or transition matrices (secondary statistical elements).

1b) Afterwards, only the data thus transformed are submitted to analysis. The results, in the form of classes or factors, will then constitute second-degree constructions elaborated from the basic data (tertiary statistical elements).

Structuring in preparation for the data analysis

We shall return below to the necessary validation of these instrumental synthetic variables that serve to summarise the trajectory. For the moment, let us consider the preparatory structuring of the data (step 1a). Since the choice of classifications and categories has already been made, it is up to the expert to decide whether they are relevant, and we shall not discuss them here. Rather, we shall simply indicate the different kinds of coding found in the sizeable body of literature we have reviewed (see Box 1).

When the current state of the individual chronologies is considered adequate for the rest of the analysis, the raw data is taken as is, without transformation, and we can speak of neutral coding [5].

[5] *Cf.* Box 1.

Box 1

Different coding of individual chronologies
(example for one year : T = training, E = Employment, U = Unemployment)

Basic data:	TTTEEEEEUUEE
Neutral coding	TTTEEEEEUUEE
Example of prior coding	total time in Employment = 7 time in Unemployment = 2 time in Training = 3
Other example of prior coding	T3 E5 U2 E2
Example of adjusted coding	T short E long U short E short
Other example of adjusted coding	unemployment<3 mos. employment>6 mos.

Another type of coding, which we shall call prior coding, consists of carrying out blind operations on the data without taking their form into account [6]. This would apply, for example, to the calculation of the number of instances of a given situation (number of periods of unemployment or employment) or the length of time in this situation (time spent in unemployment, time of access to employment, time in training). However, such indicators do not describe the transition process insofar as the order of the different events is not taken into account. This order is partly restored in certain typological analyses which divide the status chronology into semesters – or any other time periods – and calculate the time spent in each situation during each semester [7]. These operations, which allow the time factor to be reintroduced, are fairly automatic and systematic.

The other types of recoding, by contrast, do not have this feature of complete reproducibility. Indeed, they depend on an empirical treatment that remains closer to the data and is aimed at expressing their specific features. These could be called *a posteriori* or data-adapted coding [8]. They translate the chronologies in terms of sequences of different statuses (employment, non-work, etc.) and time periods ('short' or 'long') while once again retaining all the episodes recorded in the chronologies. Their validity is determined solely by an exogenous, content-based evaluation of their relevance. In this sense, even if the procedures are well described and allow another experimenter to arrive at an identical result with the sample involved, the transferability of the protocol to another cohort depends on the stability of the external conditions (economic situation, state of the youth labour market, changes in the educational system, etc.).

[6] *Cf.* Box 1.

[7] It may be noted that the number of periods used and their length have only a marginal effect on the results of the factorial and typological analyses, probably because the data on entry into working life is highly structured.

[8] *Cf.* Box 2.

After the analysis: reliability of the construction of the result variable

The question of the stability of the result, clustering or factors, depends basically on two aspects – its robustness and its reproducibility. Robustness is confirmed if the results remain stable in random simulation tests such as elimination of lines or disruption within the columns (*cf.* Box 2).

The problem of reproducibility, meanwhile, is common to all inferential statistics. Indeed, in the data analysis approach, what is sought is a structure that would be present in the population and conveyed by the sample as an empirical observation of this reality. This constitutes a concrete case of the law of large numbers, whose hypotheses are sufficiently non-restrictive to allow its use here. This situation is equivalent to the fit of a model to reality. It is no more trivial to say that the model is discovered at the same time that it is tested than to say that the most probable value of the mean is the average found in the sample. Since the sample is assumed to be representative, there is no reason for the result variable to be biased, and analogous results – the same forms – should be found with another sample from the same population. The result variable is a statistic whose sample yields a random variable, but this must be validated for an immediate re-use in the last step of the processing chain (search for links with individual characteristics)[9].

In any case, clustering is only used as a result variable when the separability of the types is guaranteed. This separability is estimated on the sample. The reality is marked by two extreme cases – that of absolute separability, where the intra-class variance is only a sampling variance, and that where the borders between the types are blurred and the class centres are only points of reference in a space of reduced dimensions (the space of the factors). In the one case, we have a discrete variable, and in the other, a small number of continuous variables. If we have been able to identify a factor as a factor of transition quality, then we have a quantitative result variable. We may note that in the case where the result variable is discrete, it is not ordinal – even if we can read a hierarchy among the classes, this order depends on an interpretation rather than a construction.

The search for links

The different procedures evoked in Table 2 are generally used to estimate relationships between result variables and explanatory variables and do not fall within the same field of statistical methodology. In their implementation, however, they all rely on a group of hypotheses that we cannot treat in detail here. Suffice it to say that these involve at once the laws of residuals, the kinds of interactions, the forms of sampling distributions, the play of explanatory variables and so on.

[9] It must be stated, as with any statistic, that it also depends on the way it is calculated – in this case, the choice of distance and the aggregation criterion for classification.

Box 2

Interpretative properties of the data-analysis methods

Optimality is ensured by the two-by-two non-correlation of the main axes of inertia, the selection of a limited number of the latter and the concomitant ignorance of the others leading to an economical result in terms of decision parameters. As a weighted combination of initial variables, each of these axes constitutes a synthetic variable. The redundancies between initial variables as well as the biases owing to effects of structure or an inappropriate selection of the variables of the model can safely be eliminated. At the same time, these axes guarantee the possibility of reconstituting the original table at each step and returning to the initial data; thus, the interpretative decision can, in all cases, be understood in geometric terms and referred to the initial system. The individual data thus remain within the space of the explanatory variables and provide the basis for recognising the main forms of their cluster.

The data is thus reconfigured by a procedure as automatic as the calculation of means can be, and such a "re-presentation" constitutes a more comprehensible reflection of a complex reality. This could be termed an operation of geometrical statistics. But the scope can be enlarged by considering individuals as technical intermediaries for the elaboration of secondary or tertiary statistical objects [10]. The latter represent the world of the parent population in a new form. In this case, probability hypotheses should be added, and we go beyond the framework of geometric statistics.

We would simply add that estimates based on multiple or logistic regressions correspond to "all things being equal" models – those in which the quantification of effects depends on the selection and form of the explanatory variables. We shall indicate a few of the less commonly used methods which are based on developments of factorial techniques. This is the case for the use of so-called supplementary variables. If Y is the previously constructed result variable, we can study the link between Y and X group of explanatory variables (socio-economic factors, training curriculum, etc.) by including this variable in a supplementary column of a factorial analysis in the table (I,X) crossing individuals and variables. To validate this information, it is possible to carry out the analysis on a fraction of the individuals (80%) and repeat the operation on the remaining fraction (20%). The comparison of squared cosines between Y and X obtained in repeated simulations gives an estimate of the variance of the link between Y and X.

When the relationship between Y and X is not linear, this estimate can be made through the intermediary of Burt table analysis in which the result variable, or variables, that we were seeking to explain is/are added to the whole of the explanatory variables analysed. The same kind of validation used earlier permits an estimate of the variance.

In conclusion, it seems to us that satisfactory methods of analysing trajectories for the evaluation of public policies have not yet been perfected. The approach that we have

[10] *Cf.* below.

attempted to describe here, which we might term 'Adaptative Data Mining', seems to us to be an efficient, rigourous means of dealing with these complex data. As a succession of constructions and validations of new statistical objects, it can be extended to many other fields.

References

Acts of the first longitudinal workshop (1994) Céreq Document no. 99, edited by M. Ourtau and P. Werquin.

Acts of the second longitudinal workshop (1995) Céreq Document no. 112, edited by A. Degenne, M. Mansuy and P. Werquin.

Acts of the third longitudinal workshop (1996) Céreq Document no. 115, edited by A. Degenne, M. Mansuy, G. Podevin and P. Werquin.

Acts of the fourth longitudinal workshop (1997) Céreq Document no. 128, edited by A. Degenne, M. Mansuy, Y. Grelet, J.F. Lochet and P. Werquin.

Bonnal L., Fougère D., Sérandon A. (1994) L'impact des politiques d'emploi sur le devenir des jeunes chômeurs : une évaluation économétrique sur données longitudinales, *Economie et Prévision* no. 115, pp. 1-29.

Courgeau D., Lelièvre E. (1989) *Analyse démographique des biographies*. Paris: INED.

Fénelon J.P., Grelet Y., Houzel Y. (1997) Modéliser l'insertion, *Formation Emploi* no. 60, pp. 37-48.

Vernières, M. (ed.) (1997) *L'insertion professionnelle, Analyse et débats.* Paris: Economica.