



Università degli Studi di Firenze

Dipartimento di Statistica "G.Parenti"

**DOTTORATO DI RICERCA IN STATISTICA APPLICATA**

**XVIII ciclo**

**FOLLOWING STUDENTS:  
A COMPACT COHORT-BASED MODEL  
FOR STUDENT CAREERS AND  
EDUCATIONAL SYSTEM ANALYSES**

Coordinatore del corso

Prof.ssa **FABRIZIA MEALLI**

Tutore:

Prof. **ACHILLE LEMMI**

Co-Tutore:

Prof. **GIULIO GHELLINI**

Tesi di Dottorato di

**FRANCESCO MACCARI**

# INDEX

<b>INTRODUCTION</b>	<b>1</b>
<b>CHAPTER ONE</b>	<b>3</b>
<i>“Review of models for educational system analysis in the framework of social account matrices”</i>	
1.1. The measurement of social systems	3
1.2. Markov models for social processes	5
1.3. The role of data	9
1.4. Indicators of school performance	11
1.5. Beyond the past experiences: looking for systems’ dynamics	14
<b>CHAPTER TWO</b>	<b>18</b>
<i>“Following students: a cohort-based model for the educational system analysis”</i>	
2.1. Overview of existing models	18
2.2. Definitions: school system and phenomena of interest	20
2.3. Data needed for model building	21
2.4. The proposed model	23
2.5. Measurement of school paths	28
2.6. Definition of indicators	30
2.7. An overview to the model: what has still to be done?	34
2.8. Conclusions and proposals for further research	37
<b>CHAPTER THREE</b>	<b>39</b>
<i>“Following students: a cohort analysis of Pisa Province students’ experience in secondary education”</i>	
3.1. Introduction	39
3.2. The case of study	40
3.3. The estimation of the model	42
3.4. The comparison of cohort performance	43
3.5. Cohort analysis by entry variables	46
3.6. Accounting for schools in the analysis	47
3.7. Conclusions and further applications	49
<b>APPENDIX</b>	<b>51</b>
<b>REFERENCES</b>	<b>53</b>

## **TABLES**

### **CHAPTER TWO**

Table 2.1. – Structure of “cohort history” data sets _____	23
Table 2.2. – An overview to the model _____	35

### **CHAPTER THREE**

Table 3.1. – Distribution of entry variables per cohort _____	41
Table 3.2. - School titles provided in the Italian Educational System _____	43
Table 3.3. - Comparison between cohorts 1994/95 and 1997/98 performance Indicators _____	45
Table 3.4. - Cohort 1994/95 performance indicators by entry variables _____	46
Table 3.5. - Distribution of students by school of enrollment in cohort 1994/95 _____	47
Table 3.6. - Cohort 1994/95 performance indicators by school of enrollment _____	48
Table 3.7. - Distribution of entry variables by school of enrollment _____	48

## **FIGURES**

### **CHAPTER TWO**

Figure 2.1. - Evolution of a school cohort through time _____	22
Figure 2.2. - A graphical representation of a school path _____	30

### **CHAPTER THREE**

Figure 3.1. – Regular School Life Expectancy (rsle) on 5 cohorts _____	44
Figure 3.2. - Regular Exit with Diploma (%) on 5 cohorts _____	44

## INTRODUCTION

Educational system has always been one of the key processes of human society. In particular, the quality of the overall educational process, which is ultimately measurable through students' success both in school life and in subsequent working life, has a strong influence on the destiny of society.

After a rather long period of substantial stability in their organization, school systems in every country, and moreover in Italy, have gone through a phase of radical changes that involved all their major organizational aspects. Reforms have been applied very frequently in the last decade, aiming to an increase in efficiency, as regards both financial aspects and school capability to form students for their future transition to active life.

Under a global point of view, every national school system has nowadays to compete with other countries, since school is going to become, as well as economy, a global market, where more efficient systems will be likely to attract the best students and also more investments from school stakeholders. Indeed, national school systems' performance is constantly evaluated by international organizations (i.e. OECD) that produce statistics on various aspect of their outcome.

For these reasons, the set-up of statistical models describing and measuring school system's activities is of high value for policy-makers, who can benefit from their applications in planning structural reforms and in subsequently evaluating their effects.

In traditional statistical approaches, the school system is usually seen as a social process formed of a sequential state space. In simplest models, states are represented by school grades, and the unit of this process is the student, who moves along states during his whole school career. Under this point of view, the formal representation of the system is provided by a set of matrices that contain the state-to-state transition probabilities and, hence, synthesize the behavior rules of the system.

As often happens in statistics, the major threatens to the effectiveness of models are mostly due to the lack of proper data to feed them. Especially at national level, only gross data of the stock type are usually provided. However, they are not sufficient to detect students' real flows through the system.

The primary impulse to this work has come from the availability of unusually detailed longitudinal information about an Italian local school sub-system. This data was provided by the *Osservatorio Scolastico Provinciale* of Pisa Province, which has been leading a project of school career monitoring since the beginning of last decade.

The data provided are also individual, thus giving a unique opportunity to look for an innovating model inspired by two main objectives:

- i) preserving as much as possible the desirable properties of traditional stochastic models, with particular regard to their compactness in the system's representation and to the availability of simple but meaningful school performance indicators;
- ii) introducing a cohort oriented data structure, aimed to catch the real dynamic component of the school system.

With regard to this second point, the work finds an important background in Légaré's proposal (1972). This approach though dating back to three decades ago still represents a very convincing way to interpret the cohort as the key instrument of explication of the main phenomena involved in school's evolution.

However, the available set of data permits to go beyond the limits that usually characterized traditional models, releasing several unrealistic assumptions. In particular, we tried to take the time dimension into account in the creation of transition matrices. That is to say that transition probabilities are not homogeneous in time, unlike it used to be in Légaré's approach. Under a cohort-oriented point of view, this implies that the student school record (at least in terms of number of past repetitions) influences the current probabilities to progress to another school grade, to repeat the current grade and to quit school.

A central objective of the work has been to create a compact software package that is manageable at micro-level (by local administrations, for instance) to obtain simple indicators of school system performance. This package, though still at an experimental level, is already operating in its basic version, allowing the estimate of all the transition matrices and, most of all, to calculate the "school paths" matrix, which represents our major proposal to compactly describe school dynamic behavior. Finally, the realized software provides for a first set of indicators of student performance.

In detail, chapter one presents an overview of existing models that we took as references for the approach proposed. This review is aimed in particular to point out the strongest constraints to models' effectiveness.

Chapter two explains the analytical aspects of our model, with particular regard to those that, in our opinion, bring some substantial innovations to the existing ones.

In the end, an example of analysis to the available data is shown in chapter three, where the performance of various cohorts, relating to different student profiles, is evaluated by means of the indicators that are available at the moment. This part is only limited to some opportunities of analyses. A more complete application, especially with regard to the comparison of results with traditional models and to the predictive capability of the proposed model, will hopefully be carried out in further research.

## Chapter One

# REVIEW OF MODELS FOR EDUCATIONAL SYSTEM ANALYSIS IN THE FRAMEWORK OF SOCIAL ACCOUNT MATRICES

### **Abstract:**

In this paper we take into account the methods of representation and measurement of social processes, with particular regard to educational system. The paper starts with a brief review of existing models, focusing on their most valuable objectives and features. From the one hand, we refer to stochastic processes, which are particularly suitable to represent the process structure; from the other hand, we resume a cohort-based approach in order to better capture the dynamic behavior of school system. The resulting model manages to synthesize student careers' evolution into a "school paths" matrix and provides for the definition of a rather complete set of student performance indicators.

### **1.1. The measurement of social systems**

From a theoretical point of view, the problem of measuring and representing social phenomena with statistical tools dates to nearly two centuries ago, when Quetelet (1849) sustained that the study of social sciences could well be approached in an analogous way as physics. Such a conviction bases on the idea that, even if human behavior is naturally unpredictable, social patterns, being the results of the interaction of a large set of human beings, can be interpreted through the application of probability laws.

However, the first practical applications of this substantially innovating vision, which has indeed reduced the broad distance that once separated social sciences from physics, came only at around half of last century, when a sufficiently complete theory on probabilistic models was actually available (Bush e Mosteller 1955, Steindhl 1965).

From that time on, the greater part of research on this field has mostly concentrated on the use of probability models, since the centrality of random component has always been recognized in social systems. More specifically,

stochastic processes, such as Markov chain models, have been the most used instrument in these applications. Bartholomew (1982) clearly points out the key factors of time and chance, which govern the evolution of social systems and make stochastic processes a very suitable instrument for their analyses. The basic idea is to represent the system of interest as a set of possible status in which an individual can be, and to postulate the fundamental laws that rule status changes along time.

Stochastic processes permit to represent social systems' behavior in a compact way and, in many cases, to do that at different levels of detail, induced by the definition of status.

Referring to the educational system, which will be the case of study in this research, we can simply define status as the school grade attended by the student in a certain year, but we can also think of more analytical models, defining the status as the combination of, say, school grade and type of school attended at a certain time. Multiplying in this way the set of possible status, we obtain a more complex representation of the phenomenon, taking into account a consequently wider set of possible transitions in time.

A model providing for a compact representation of a social system can be used to achieve many purposes, which can be linked to two general needs:

- a) The analysis of specific aspects regarding the structure and evolution in time of the considered social process, including the measurement of flows linking it to other social systems (entry and exit of units);
- b) The exigency, on a broader scale, of a social accounting system, regarding the set of all social processes in which a population is involved.

The first aim is with no doubt more familiar to the social researcher and comprehends the following objectives:

- provide for a synthetic description of the system considered;
- make projections and simulations of the system's evolution.

We will return shortly on these aspects later.

As regards the second point, we basically refer to the work by Stone (1966, 1971 and 1973), who recalling a firmly established input-output scheme in economic analysis (Leontieff 1951), has first underlined the great importance assumed by the set-up of a complex system of overall social accounting, gathering information about the composition and evolution of society under different aspects (education, working activities, etc.).

Going slightly beyond the pure accounting matter, the benefits arising from the disposal of integrated models for different social sub-systems would be mostly relevant for policy making. Of course, this presumes all social process to be

considered as components of an overall system, in which every person takes part in the whole of his life.

In the next sections we shall concentrate mainly on the first of the two points of view introduced above, trying to put in evidence the main features of the most common models and paying a particular care in underlying the problems related to the commonly available data and to the constraints that they usually induce in the modeling phase.

## 1.2. Markov models for social processes

In this section, we will focus on discrete time models applied to social processes, with particular reference to the educational system. The advantages of choosing a discrete time scale for this field of application are widely discussed in literature (Stone, 1965 and 1972; Thonstad, 1969; Bartholomew, 1982).

The usual representation of school system that we are going to show is subject to the following set of assumptions:

- i) the school system is formed of a finite set of status, that can be either transient ( $n$  school activities) or absorbent ( $m$  levels of education);
- ii) the system considered is closed, and each unit can be in only one status at a certain time;
- iii) units being in a transient status can either remain in the same status, or move on to the following one, or leave the system only in the discrete times of a given time set (which corresponds to the succession of school years);
- iv) the probability of units to be in a certain school activity or final education level at a given time  $t+1$  depends only on the activity they were in at time  $t$ ;
- v) transition probabilities are constant in time.

Under this set of assumptions, the model suitable to represent the considered system is a finite absorbent Markov chain (Kemeny and Snell, 1976).

The transition probabilities are then indicated as follows:

- I)  $p_{i,j}$  is the probability to pass from school activity  $i$  to school activity  $j$  ( $i, j = 1, \dots, n$ );
- II)  $l_{i,k}$  is the probability to pass from school activity  $i$  to final level  $k$  ( $i = 1, \dots, n$ ;  $k = 1, \dots, m$ ).

Omitting instead assumption v), the resulting stochastic model comes out to be a finite Markov process. In this case, we assume that the two above seen



probabilities are time-dependent and the notation has to be consequently changed to  $p_{i,j}(t)$  and  $l_{i,k}(t)$ .

Referring for simplicity to a finite absorbent Markov chain, the two sets of probabilities can be arranged in the following matrices, representing the whole probabilistic structure of the system considered.

$$\mathbf{P} = [p_{i,j}], \quad \mathbf{L} = [l_{i,k}].$$

Hence, the canonical form of this Markov chain is given by:

$$\mathbf{S} = \begin{bmatrix} \mathbf{I} & \mathbf{O} \\ \mathbf{L} & \mathbf{P} \end{bmatrix},$$

where  $\mathbf{I}$  is the  $m$ -order identity matrix and  $\mathbf{O}$  is the  $m \times n$  null matrix.

As a consequence of the hypotheses made, in particular of hypothesis ii), given a certain school activity  $i$ , the following result is immediate:

$$\sum_{j=1}^n p_{i,j} + \sum_{k=1}^m l_{i,k} = 1 \quad (i = 1, \dots, n).$$

The above listed assumptions induce the following reflections.

- a) Assumption i) is well respondent to reality, and only implies the need of a precise definition of the set of transient and absorbent status.
- b) Assumption ii) represents instead a big simplification, since school system is not closed in reality, but involves output and input flows from and towards every status and at any time, especially when referring to a small geographical scale.
- c) Assumption iii) is less complex: actually, a student in a certain school activity can progress on to the next one only at the end of the year, but can instead leave school at any time. This assumption, on the contrary, imposes to consider together all the exits from school happening during the same school period. However, this does not cause any significant distortion, since all students leaving school from a given grade actually exit with the same education level (or, in other words, make a transition from the same activity to the same absorbent status). Another implication of this assumption is that students can neither skip grades nor pass to a lower one at any time.
- d) Assumption iv) is fundamental, since it constitutes the key property of a Markov chain. At the same time, it entails the most severe and unrealistic simplifications to the model: first, no differences in transition and leaving probabilities are assumed between students with different past career; in addition to that, no limitation is imposed to the number of possible repetitions.

e) Assumption v), as we have seen, is omitted in case of a finite Markov process.

Within this approach, we specifically refer to model SFINGE (*Stima dei Flussi, Indicatori e Generazione di Effettivi scolastici*), proposed for the application to the Italian educational system during the 1980s (Trivellato 1980, Bernardi and Trivellato 1980, Trivellato and Bernardi 1994).

SFINGE, in his original version, entirely recalls the structure of a finite Markov process, which we have described above.

A refined version has then been formalized with the proposal of the so called “Repeating – Non repeating students” model. It introduces the assumption that the behavior of a student is different with respect to whether he is enrolled in a certain school grade for the first time or as a repeating. Consequently, this specification provides for a duplication of every school activity and adopts a different set of transition probabilities from each grade for the two groups. Notwithstanding this, the desirable properties of a Markov chain (or process) are preserved.

A further refinement is represented by the “Triplicate” model, which, according to the same principle, provides for three different status for each school grade: non repeater, repeater and multi-repeater (Bernardi et al. 1986).

These proposal, though solving much of the bias problem in estimating transition and leaving probabilities, do not face the most unrealistic assumption, that is the independence of school patterns from the overall past career. In addition to that, other threatens to the model efficacy are often caused by the data used in the estimation phase. We shall discuss this aspect more clearly in next section.

The recognition of the above seen big lacks in these kind of models, very recurrent in literature, brings us to take into account a slightly different view to the problem, the cohort-based approach. In this case, the social process is analyzed under a cohort point of view, in an analogous way as is commonly found in demography.

Models inspired to this point of view of social processes are specifically thought for the definition of indicators aimed to measure the outcome and evolution of the considered system. The first specific examples in literature date back from 1960s on (Stockwell and Nam 1963, Blot 1965, Légaré 1972). We mainly refer here to the work from Légaré in order to show the nature of the approach.

Légaré is very clear in stating that school events, such as repeating a grade or passing to the next one, have sense only when considered within a cohort approach. Transitions must be interpreted along time, the natural dimension in which a cohort evolves. On regard to this point, a cohort in education should be analyzed with respect to both age and grade attended by the students. The combination of these two variables contains all the information needed to detect the pattern followed by a student.

Unfortunately, age and grade of students are hardly ever available at the same time. Therefore, the following discussion will be based only on the second one, which is necessary to measure the repeat enrolment phenomenon.

The model proposed provides for three possible events occurring to students enrolled in grade  $i$  at time  $t$ : either they pass to grade  $i+1$ , or they repeat grade  $i$ , or they leave the school.

Subsequently, the following probabilities are defined for every time  $t$ :

$p_i$ , the transition probability from grade  $i$  to grade  $i+1$ ;

$r_i$ , the probability to repeat grade  $i$ ;

$a_i$ , the leaving probability.

A set of equations is then proposed in order to estimate the given probabilities. We report here only the “transition” equation as an example, where  $E_i$  is the number of students enrolled in grade  $i$  at time  $t$ .

$${}_{t+1}E_i = {}_t p_{i-1} \cdot {}_t E_{i-1} + {}_t r_i \cdot {}_t E_i$$

However, in the estimation phase, a crucial assumption is done, in order to simplify calculations, of homogeneity of the cohort in time. In other words, the transition probabilities are assumed to be independent of the students’ previous school career, and, hence, to be constant in time. As we have already discussed in the review of the previous models, this simplification is hard to be accepted, being clearly unrealistic if one thinks to the real school process characteristics and behavior. Performance indicators are finally defined on the basis of this strong assumption.

In addition to that, Légaré underlines that no limitations are assumed for the possible number of total repetitions, even if this can be quite acceptable most of the times, on condition that the repeat rate is low enough.

Despite the problems encountered in the application phase, it looks clear how this kind of approach to the study of social systems is more effective than a traditional one for many aspects. In particular, we want to underline the following:

- a) It is more suitable to catch the dynamic component of the system, represented as a succession of events occurring to an initial cohort along time.
- b) Models are more likely to be linked to the study of other social systems, since they detect well the flows into and out from the system considered.

In his work, Légaré recognizes how much of the potential of this approach is dramatically vanished by the quality of available data, which are very often of the stock type and, therefore, do not fully adapt to a cohort approach. This would preferably need, instead, data of the flow type, in order to link events occurring to the same population in different times.

In the next section, we shall specifically deal with the aspects relating to the usually available data for social analyses, with particular regard to the Italian educational system.

### **1.3. The role of data**

It is well recognized that an effective analysis of social systems should involve, at least in part, the collection of longitudinal data in order to investigate the real flows along time.

On the contrary, most of the existing applications of the models presented in the previous section are carried out on the basis of cross-sectional data, which can usually be gathered from “official” sources. The main problem with this kind of data lies in the impossibility to measure real flows, which have actually to be estimated.

Another point to take into consideration regards the detail of data: individual information would be the ideal scenery, though utopian at this time, since it would permit, according to the available variables, to produce statistics referred to groups formed in any desired way.

Again, the real situation is very far from the ideal one: public statistics for social systems are usually of the stock type, and, moreover, they hardly ever go beyond a classification per gender and geographical variables.

This is particularly evident if we refer to the Italian educational system. Both ISTAT and the Ministry of Public Education produce very limited stock data suitable to be used in the estimation of models seen before (Martelli, 1995, Trivellato et al., 1995). Nowadays, SFINGE, is still the most valuable example of application of statistical models aimed to give measures of school system performance.

Two desirable perspectives of the Italian situations are to be underlined:

- i) a more intensive exploitation of public statistics that are available at the moment, through the application of models based on stock data;
- ii) the introduction and development of a project of longitudinal data collection, possibly providing for a national student data-set recording the individual events involved in the school system.

Of course, this second objective in particular cannot leave a full involvement of local school institutions out of consideration, since proper data can only be collected at micro-level. However, the design of such project should be carried out by central institutions, defining common rules to follow in the collection phase and providing for the assignment of identification codes to all the subjects involved, from schools to single students. The collection phase, in fact, is crucial and subject to many practical difficulties. In particular, an effective method must

be implemented in order to follow students' migration flows. This is particularly tasking when these migrations occur among schools located in considerably far areas and, therefore, belonging to different local public institutions.

Although some European countries have already undertaken similar experiences, Italy is still behind the times, apart from some sporadic and lucky local realities, even if a projecting phase of a student central record has recently been started, in strong connection with the incoming reform project of the Italian educational system.

Going back to the modeling aspects, it is worth to concentrate on the consequences of scarcity of "good" data on the quality of statistical analysis. In particular, we will refer first to SFINGE model and, then, to Légaré's cohort approach.

As regards SFINGE, it has always been applied to the Italian school system on the basis of the stock data available from ISTAT sources. For this reason, it provides for an estimation of transitions by means of differences between stock data collected at different times. This already serious bias in the detection of real flows is made even worse by the assumption of closed system: new entrants into the school are here confused with internal transitions.

Another consequence is that the estimation of drop-outs has to be drawn from successive stock data differences, in coherence with the overall yearly data constraints.

In addition to that, much of the potential improvement in estimates brought by the introduction of the two refined versions of the model (repetents-non repetents and the "triplicate" model) is in practice undone by the unavailability of exact information on students' enrolment status. The number of repeaters (or multi-repeaters) used to be actually estimated by means of a sample survey (which is no more led nowadays) giving the distribution of repeaters by age and school grade.

The cohort-oriented approach obtains a more correct measure of flows, on condition that longitudinal information are actually available. At least, a repeated collection of cross-sectional data would be needed for a certain number of school years, in order to define the set of transition probabilities for the cohort. In practice, instead, such flows are again estimates using contemporary stock data. This, together with the assumption of homogeneity of the cohort in time, induces a serious threaten to the real capability of the model to represent cohort evolution.

## 1.4. Indicators of school performance

Indicators are normally used to measure a wide range of aspects of educational system, relating to the actors involved (i.e. students), to the status of financial investments, to the overall quality of the system, and many more.

A compact framework through which the most common indicators can be seen is provided by OECD (Organisation for Economic Co-operation and Development), which regularly publishes a great amount of statistics in education regarding its member Countries (for recent editions, see OECD, 2000 and 2005).

OECD classifies educational indicators with respect to the following three dimensions:

- i) the actors involved in the system (i.e. students, educational institutes, learning environment, up to the educational system as a whole);
- ii) the object of measurement (education outcomes, policy levers influencing outcomes and constraints or antecedent to policy);
- iii) the policy issue of reference (education quality, education equity and opportunities, resource management).

Within this framework, we refer only to indicators measuring student educational outcomes. As regard this aspect, OECD statistics are limited essentially to graduation and educational attainment rates relating to different levels of education.

The probabilistic models seen in previous sections can provide for more analytical measures of the system's outcome, taking into account aspects such as regular school life, duration of permanence into school activities, average educational attainment, and so on.

SFINGE model in particular proposes the following three sets of indicators:

### I) Permanence indicators

- i) Regularity tables
- ii) Permanence tables
- iii) Expected length of regular schooling
- iv) Expected permanence at school
- v) Expected number of school grades attended
- vi) Expected permanence per school grade attended

### II) Indicators of education levels attainment

- vii) Tables of regular achievement of school titles
- viii) Tables of achievement of school titles

### III) Inefficiency indicators

- ix) Expected additional time spent per school grade
- x) Expected additional time spent per school grade

Referring to the probabilistic representation of the school system as a finite absorbent Markov chain, the definition of all indicators is rather compact and is founded upon two important theorems, which can be summarized as follows (Kemeny and Snell 1976):

- $\mathbf{P}^\psi = \overbrace{\mathbf{P} \times \mathbf{P} \times \dots \times \mathbf{P}}^{\psi \text{ times}}$  is the matrix of transition probabilities between two transient status in  $\psi$  years ( $\psi = 0, 1, 2, \dots; P^0 = I$ );
- $\mathbf{L}^\psi = \mathbf{P}^{\psi-1} \times \mathbf{L}$  is the matrix of transition probabilities from a transient status to an absorbent one in  $\psi$  years ( $\psi = 1, 2, \dots$ ).

Another theorem is very useful to simplify the formalization of indicators, stating that, for every finite absorbent Markov chain, matrix  $(\mathbf{I} - \mathbf{P})$  is invertible, and its inverse matrix is given by

$$\mathbf{X} = (\mathbf{I} - \mathbf{P})^{-1} = \mathbf{I} + \mathbf{P} + \mathbf{P}^2 + \dots = \sum_{\psi=0}^{\infty} \mathbf{P}^\psi .$$

The general element of  $\mathbf{X}$ ,  $x_{i,j}$ , represents the expected permanence in the activity  $j$  starting from activity  $i$  (Bartholomew 1982).

From the theorem, it is also possible to show that the general element  $l_{i,k}^{(tot)}$  of matrix

$$\mathbf{L}^{(tot)} = \mathbf{X} \times \mathbf{L}$$

is equal to the transition probability from the transient status  $i$  to the absorbent one  $k$  in whatever number of years.

Given these useful results, we shall now show the analytical definition of four of the above mentioned indicators, belonging to the first group, which are of great importance in our research.

#### *Regularity tables*

They indicate the students' transition probability of progressing from school activity  $i$  ( $i = 1, \dots, n-1$ ) to each of the following ones  $i+\psi$  ( $\psi = 0, 1, \dots, n-i$ ) in  $\psi$  years (that is, following a regular path).

$$r_i = \left[ p_{i,i+\psi}^{(\psi)} \right] \quad (\psi = 0, 1, \dots, n-i).$$

#### *Permanence tables*

They indicate the students' transition probability of progressing from school activity  $i$  ( $i = 1, \dots, n$ ) to school activity  $j$  ( $j = i, \dots, n$ ) in  $\psi$  years.

$$perm_i = \left[ \sum_{j=i}^n p_{i,j}^{(\psi)} \right] \quad (\psi = 0, 1, 2, \dots).$$

#### *Expected length of regular schooling*

It indicates the expected number of school years to be completed without repetitions, starting from school activity  $i$  ( $i = 1, \dots, n-1$ ).

$$er_i = \sum_{\psi=0}^{n-i} p_{i,i+\psi}^{(\psi)}.$$

#### *Expected permanence at school*

It indicates the expected number of school years to be spent into the system, starting from school activity  $i$  ( $i = 1, \dots, n$ ).

$$ep_i = \sum_{j=1}^n x_{i,j}.$$

The two indicators of education levels attainment are of great importance to us too; they can easily be obtained through a proper composition of matrices  $\mathbf{P}^\psi$  and  $\mathbf{L}$  (as regards indicator vii) and through the values of matrix  $\mathbf{L}^{(tot)}$  (as regards indicator viii)). We do not report here their exact formalization for simplicity (Trivellato 1980 for details).

The major problem with these indicators, giving indeed a wide range of useful measures of educational system's quality, is that they are not sensible to the past career of students, due to the set of constraints to the structure of matrices  $\mathbf{P}$  and  $\mathbf{L}$ . In the so-called "triplicate" model, we can obtain different measures of performance, starting from a certain school grade, for non-repeaters, for repeaters and for multi-repeaters, but no more subdivision of students is possible.

Notwithstanding this, the set of indicators available do constitute a very effective instrument of knowledge about the main aspects of a school system. We also wish to underline that they possess the very desirable property of compactness, making them rather easy to apply and manage.

In Légaré's cohort approach, three indicators are proposed: graduation rate, average length of schooling and average educational attainment. While the first two can be related to indicators provided also by SFINGE model (that is, respectively, to indicators viii) and iv)), the last one is rather original and arises from the fact that Légaré establishes a correspondence between every school grade and a different educational attainment (for formal aspects, see again Légaré 1972).

The definition of indicators is not the only, though maybe the major, purposes achieved by the class of models here recalled. They are aimed to make projections and previsions of future evolution as well. This research, however, is



substantially voted to the descriptive and evaluative aspect of social processes. Therefore, this last field of application will not be examined here.

### **1.5. Beyond the past experiences: looking for systems' dynamics**

After this synthetic review of models for the description of social processes, and in particular of the school system, the following general, but central, considerations have arisen which deserve to be clearly focused.

In the first place, we have pointed out the centrality of dynamics when we are trying to describe any social process. The peculiar events have sense mostly if considered under a time evolution optic. This is typically the case of educational system, where transitions and drop-outs occur at the end of every time unit and their temporal sequence determines the overall destiny of student generations. This reflection has resulted more clearly when approaching the system through a cohort point of view.

In the second place, we realized how existing models largely applied to the study of school system, although formally well built, suffer from strong constraints due to the assumption of too unrealistic hypotheses. These constraints are often inevitable, since assumptions are usually made in order to render the model simple enough to be applicable or, even more, they are due to the type of data available to feed the model.

We subsequently come to the third, crucial, point, which we have thoroughly discussed in the chapter the role covered by the collected data. We have seen how these data, especially when referring to a large scale, do very seldom possess the desirable properties that would be needed to fully exploit the above seen models' potential. Information of the stock type, providing for a very small number of variables (as gender), is usually the only one available from official sources, particularly at Country level. This scenery represents a big limit for any development perspective of nowadays used models, since the relax of heavy assumptions can only be pursued if suitable analytical data are provided about the process of interest.

As for this last point, with regard to Légaré's proposal of a cohort approach, we have remarked that some of the author's aims remained unachieved for the lack of proper data to feed the model, which had indeed to be run on the basis of subsequent sets of stock observations.

It is mainly from these considerations about the existing methods that we have drawn the impulse to look for a substantially new method for the representation and analysis of a social sequential system, specifically referred to the educational system.

The opportunity to lead this research was given by the availability of an unusually detailed data collection, regarding students' careers of Pisa province public high school system for a long period of time (ten school years). This data-warehouse, that we will accurately describe in chapter three, consists of individual information that could be managed in order to obtain a longitudinal data structure recording the real school events occurred to each student in the population considered. In addition to that, it has been possible to reconstruct cohort histories on the basis of the students' year of entry into the system. Some very interesting personal variables (such as gender, lower level school career and familiar background) were also available for the greater part of units involved.

The model and the subsequent proposal of school performance indicators will be explained in chapter two. We wish to underline here some important points that, in our opinion, make this approach innovative with respect to the above described ones.

Our model is strictly cohort-oriented. Its basic idea is to consider cohorts as the natural way to group students. A cohort represents the concept of evolution of a system much better than a set of cross-sectional data collection does. In his article, Légaré (1972) is very clear about this point and we will not discuss it any further.

Therefore, the main purpose of our approach is to look for a coherent representation of school system's dynamics, under a cohort optic, based on student career events (transitions and exits from school) along a clear time dimension.

On the other hand, we wanted to maintain some of the good features of existing models, particularly as regards the set of indicators provided by SFINGE. At the same time, we focused on the need to solve some of the most serious constraints. With respect to this point, our model is subject to the following assumptions, that we will first list and then discuss, paying a particular care in their comparison to the case of finite absorbent Markov chain.

- i) The school system is represented through a finite number of states, either transient or absorbent. Transient states are represented by the combination of school grade attended and number of past repetitions suffered in the same system. Absorbent states are the possible exit levels of education.
- ii) The system is closed in a cohort optic (no students belonging to other cohorts enter the system at any time) and each student can be only in one status at a certain time.
- iii) Students change status at the end (and only at the end) of every time unit.
- iv) A student's probability to be in any status at time  $t+1$  depends only on the status he was at time  $t$ .

Assumption i) is certainly the most important one, since it provides for a different way to define school transient status. In fact, they are no more

represented only by the grade attended by the student, but also by the number of past repetition occurred to him till now. This additional variable is a measure of the past career, that we strongly believe to dramatically influence students' future behavior (this evidence will result more clearly from the analyses reported in chapter three).

Assumption ii) is only apparently equivalent to the corresponding one of Markov models. The model is in fact closed with respect to students belonging to other cohorts, but it provides for the possibility of monitoring re-entries of students who had entered the system as units of the cohort considered. We will return to this feature later in this section.

Assumption iii) regards the choice of a discrete time dimension in representing the system. It is to notice that, due to assumption i), a student's status always changes at the end of each period, since one of the two variables involved (school grade and number of repetitions) increases as a result of, respectively, a promotion or a repeat enrolment into the current grade.

Assumption iv) is the same as for Markov processes, but, again, the way of defining school status implies that the status a student is into at a certain time depends on his whole past career, at least with regard to the number of repetitions previously occurred.

We want now to underline two major consequences of the described set of assumptions:

- a) students enrolled in the same school grade, but with different careers (in terms of number of past repetitions), have, in general, different probabilities to pass to any school grade or absorbent status;
- b) no homogeneity in time is assumed for the cohort analyzed.

As we can see, two serious simplifications threatening traditional models' efficacy have been released in our proposal.

At this step of research, we looked for a compact instrument able to represent all the dynamic process involved in cohort evolution. More precisely, we defined a structure describing the probabilities of all possible school paths that can be identified by means of the available data: we called it "school paths matrix" and indicated it as **W**.

Keeping technical details aside (they will be thoroughly treated in next chapter), **W** contains the probability to pass from an initial status to a final one in a given number of years. We remark again that initial and final states are identified by both school grade attended and number of past repetitions. It is clear that this matrix only refers to paths that are "inside" the school system; in other words, it does not include leaving events. For this reason, **W** can be estimated by means of the probabilities contained in the transition matrix **P**.

School paths matrix represent a useful instrument to synthesize the complex dynamics involved in the cohort history. Of course, it is still a simplification of that history, since it distinguishes paths only on the basis of their starting and ending status, regardless of the exact succession of transition that have generated such paths.

However, we believe that this can be a valuable proposal of an alternative approach to the study of social processes, on condition that suitable longitudinal data are actually available.

The school paths matrix allows, in our representation, to define a set of school system performance indicators, regarding the same aspects as the ones seen for SFINGE model, as will be shown in next chapter.

Great impulse to the approach we presented was given by the possibility to run more complex computer procedures than there was some decades ago, when the presented models have been implemented.

First of all, the programming phase has been definitely crucial for the creation of cohort data sets by integrating individual information on several years.

Secondly, in the modeling phases, we exploited the big opportunities given by informatics instruments in the formalization of the following features of the model:

- i) the multiplication of the set of status, which has forced to built multidimensional matrices in order to effectively describe the probability patterns of the cohort;
- ii) the creation of “school paths” matrix, which needed to be defined in five dimensions;
- iii) the management of students’ re-entries into the system after the initial time, which has been led through an additional matrix, whose structure and properties are explained in next chapter.

A further development would be needed, consisting in the introduction of a school dimension in the set of status. This would add complexity to the overall structure of the model, but would permit to include in it all the transitions of students between different types of school. This, in our opinion, is a very important component of the total dynamics, which deserves to be taken into account.

## Chapter Two

# FOLLOWING STUDENTS: A COHORT-BASED MODEL FOR THE EDUCATIONAL SYSTEM ANALYSIS

### Abstract:

In this paper we propose a model for the analysis of school system performance under a cohort-based approach. The model recalls the structure of Markov processes, but, using a longitudinal set of individual data, introduces the concept of “school path”. The definition of paths allows to take into account the time dimension of school careers, represented by the number of school years spent by the student into the system until a certain school grade.

A set of indicators measuring student performance is defined as well, in coherence with those provided by traditional Markov models. The paper ends with a thorough discussion on the possibility of future development of the approach, as regards in particular the theoretical aspects of the model.

### 2.1. Overview of existing models

The application of stochastic models to the study of social processes, with particular regard to their evolution, dates to several decades ago (Bartholomew, 1982), but it bases upon principles that are general enough to preserve an absolute validity through time. Indeed, most social systems are easily represented by means of a succession of states that a single unit (a person in the system) can visit along time. This is also the case for educational system, where the identification of possible states comes natural (grades, school titles, and so on).

Referring here to this last specific application, it is useful to recall the main formal aspects of the model (for a complete review, see: Stone, 1965 and 1972, Bartholomew, 1982, Thonstad, 1969).

The main assumptions underlying the following representation are:

- a) the school system can be represented through a finite number of states, that can be either transient ( $n$  school activities) or absorbent ( $m$  levels of education);
- b) each unit can have only one state at a certain time;

- c) units that have in a transient state can either move to another state or leave the system only in the discrete times of a given time set (which corresponds to the succession of school years);
- d) the probability of units being in a certain school activity (state) at a given time  $t+1$  depends only on the activity they were in at time  $t$  (that is, no limitation is fixed for the number of repetitions);
- e) transition probabilities are constant in time.

Under these assumptions, it is possible to define the following transition probabilities:

- i)  $p_{i,j}$  is the probability to pass from school activity  $i$  to school activity  $j$  ( $i, j = 1, \dots, n$ );
- ii)  $l_{i,k}$  is the probability to pass from school activity  $i$  to final level  $k$  ( $i = 1, \dots, n; k = 1, \dots, m$ ).

Two matrices can be built with the above seen probabilities, as follows:

$$\mathbf{P} = [ p_{i,j} ], \quad \mathbf{L} = [ l_{i,k} ].$$

Given the set of hypotheses, we have:

$$\sum_{j=1}^n p_{i,j} + \sum_{k=1}^m l_{i,k} = 1, i = 1, \dots, n.$$

Hence, we have obtained a finite absorbent Markov chain (see Kemeny, Snell, 1960, for a theoretical background), whose canonical form is given by:

$$\mathbf{S} = \begin{bmatrix} \mathbf{I} & \mathbf{O} \\ \mathbf{L} & \mathbf{P} \end{bmatrix},$$

where  $\mathbf{I}$  is the  $m$ -order identity matrix and  $\mathbf{O}$  is the  $m \times n$  null matrix.

In Italy, in particular, a typical application of the above illustrated model is represented by SFINGE (*Stima dei Flussi, Indicatori e Generazione di previsioni di Effettivi scolastici – Flow estimation, indicators and previsions of future enrolments*), which has been developed during the 1980s with the purpose of achieving a compact but effective set of indicators of the public school system performances (Bernardi et al. 1986, Trivellato 1980, for theoretical background). SFINGE takes into account not only the aspects related to the education system in general, but also the results of student careers. The model concentrates mainly on the following aspects, and, consequently, on the related groups of indicators: i) probability of permanence at school; ii) probability of attaining different levels of education; iii) expected duration of school life; iv) efficiency/inefficiency of the school system as a whole.

A great part of the existing research work on measurement of school performances, of which SFINGE is a good example, is based on stock data that

refers to a certain point in time. It is to underline that this is mainly due to the opportunity to use “official statistics”, which are generally of the stock type, to feed the models (Martelli, 1995, Trivellato et al., 1995).

Nevertheless, the advantages of alternative approaches based on flow-type data, firmly established in demographic analysis, are well recognized (Légaré, 1972), especially when such models are applied to making projections. This kind of approach considers the student as the single unit of the school system and belonging to a certain cohort that is defined in the same way as in demography. A cohort of students, in fact, can be defined as the set of all those students who have enrolled to the first grade of the considered school system for the first time in the same instant  $t$  (where  $t$  usually indicates the beginning of the school year – e.g.  $t = 2005$  for the students who have entered the school at the beginning of 2005/06).

The advantages, together with some disadvantages, of a cohort-based approach with respect to a cross-section one, will be detailed in section 2.7.

## **2.2. Definitions: school system and phenomena of interest**

In this chapter we refer to the structure of the Italian high school (*scuola secondaria superiore*) system in the recent past (from 1990s to early 2000s). Before formalizing the proposed model, we will briefly describe the peculiarities of the analyzed system and, subsequently, focus on the phenomena our model mainly takes into account and represents.

Italian students can enter high school after successfully attending eight years of studies (that is, five years in *scuola elementare* plus three in *scuola secondaria inferiore*), at the end of which they get a *licenza*. In the period, which this application refers to, these two cycles of school used to represent the minimum compulsory education.

A standard high school course is composed of five classes, after which students get a degree that gives them the right to matriculate at the university. Besides, Professional schools give the opportunity to receive a professional instruction diploma after the third class; Normal Schools release a diploma for teaching in low level schools at the end of fourth class (this kind of degree has indeed been abolished in late 1990s); Art Schools too provide for a professional art diploma at the end of grade four. The existence of different rules in different kinds of school underlines the need for a model that is general enough to allow analyzing all of them at once.

We have concentrated on the following aspects of school career, which we believe to be the most relevant:

- a) the evolution of cohorts of student along time, i.e. the transition probabilities from state to state;

- b) the exits from school, with particular regard to the state (class or level of education) at which they occur;
- c) the “memory” of the past school record of the students (only limited to what occurred in the considered school system), measured in terms of years students have spent in order to reach the current state (class or level of education);
- d) the re-entry phenomenon, i.e. the case when a student enrolls again after having previously quitted school.

The objective of the analysis is to define a set of general indicators that help evaluate the performance of the school system under various aspects. We come back to this point later on in this chapter.

### **2.3. Data needed for model building**

It is quite clear that in order to build a model that represents the complete evolution of a school cohort we need to have data about the complete careers of all students belonging to that cohort. This is obviously possible only in case the data has been gathered year by year from the generation of the cohort until its definitive extinction, i.e. that is until the last student has left the school system.

In addition to that, data of the stock-type, usually available in “official” statistics, is not sufficient for our purpose: the suitable data should be made of a set of longitudinal records that report the school career of any single student in the system.

The history of a cohort of students is complex, since it is made of the effects of the interaction of several different events taking place in different times and maybe in different places (for a review, see Légaré, 1972).

The graph 2.1. can help focus on the typical pattern of cohort evolution through time.



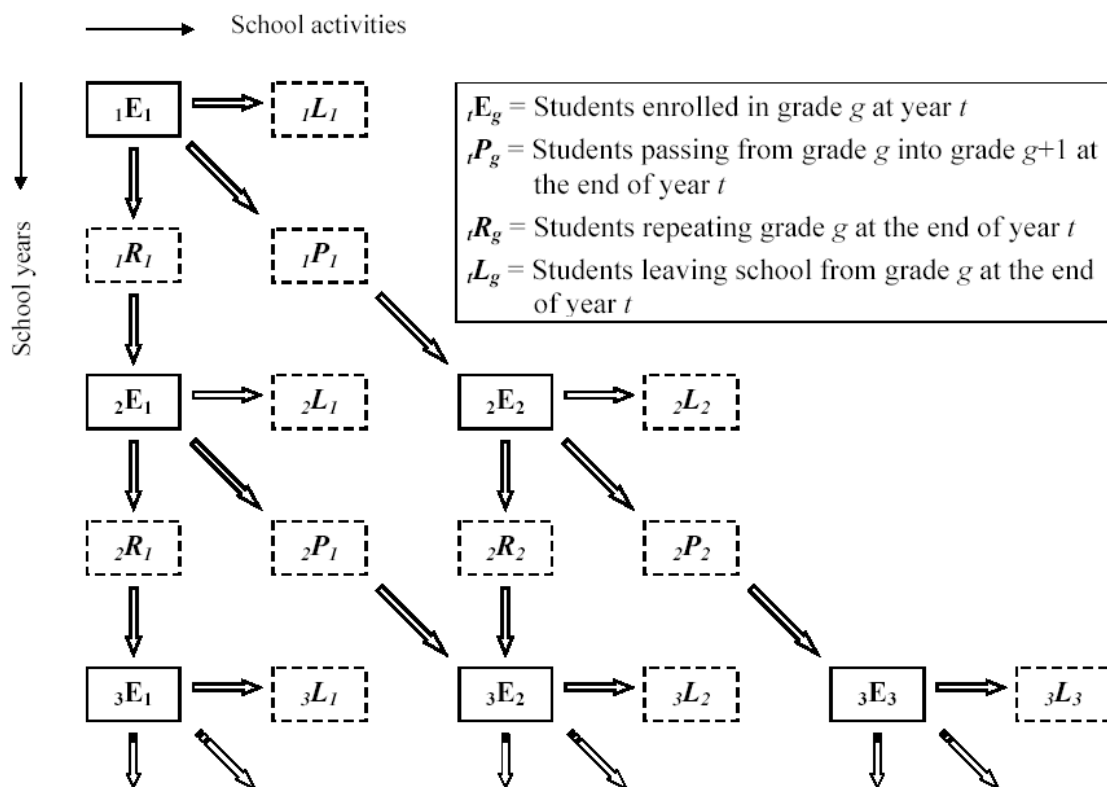


Figure 2.1. - Evolution of a school cohort through time

In order to effectively represent the cohort history, a collection of the following pieces of information should initially be available for every student:

- a) the year of entrance into the selected school system, to be used to relate every student to the cohort he belongs to;
- b) the type of school attended, possibly for every school period;
- c) the grade that every student attends in every school period of his career;
- d) the final result of every school period, which is particularly useful when leaving the school, also in terms of level of education achieved;
- e) the year and grade of possible re-entries in the school after a previous exit;
- f) information on some personal characteristics, such as gender, age of first school enrolment, family background (parents' jobs and levels of education), place of residence, and so on.

Such information can then be arranged, for a given cohort, into some data sets that record the history of every student in the cohort.

- **G**, containing the sequence of grades attended by every student of the cohort;
- **S**, containing the sequence of schools at which every student is enrolled;

- **R**, containing, for each student in the cohort, the final result of every school year;
- **A**, a dataset collecting, for every student in the cohort, all the variables that can be considered time-invariant with respect to the school system (entry variables).

Apart from this last dataset **A**, all the others have a similar structure: every row relates to a single student, while every column is a year of the cohort, from the year of generation to the one of extinction (or, anyway, the last year of observation).

This structure is reported in table 2.1., where  $N$  is the number of students in the cohort,  $t_1$  is the year of generation of the cohort and  $t_T$  is the year of extinction or related to the last observation available.

In the next section we show how this kind of information can be arranged in order to formalize a cohort-based model.

Table 2.1. – Structure of “Cohort history” data sets

		Cohort years			
		$t_1$	$t_2$	...	$t_T$
Students	1				
	2				
	⋮				
	$N$				

## 2.4. The proposed model

The model we show in this section is in fact a way to organize the data set that was described above into a compact structure that can represent the dynamics of the school careers of a given cohort, giving particular emphasis to the grade-to-grade transitions and to the drop-out of students at the various educational levels.

For a good reference about the scheme of school cohorts’ dynamics, see Légaré, 1972. We essentially refer to the same basic idea to represent cohort life, but also introduce some quite important modifications that we will underline in the following sections.

The input data sets that we showed in the previous section refer to a time dimension that is exactly the same for all the students involved, since it is represented by the age, expressed in school periods, of the analyzed cohort. In other words, referring for instance to **G** matrix, let  $T_1$  be the index for the initial school period of the cohort, then, for a given student  $i$  (row), the  $g_{i,t}$  variable indicates the grade that student is taking in the  $t$ -th year of life of the cohort (or rather in school period  $T_t$ ), in case he is still at school, and 0 otherwise.

Before passing to the explanation of the model, we shall list the major underlying assumptions:

- a) a student belongs to the cohort  $T_l$  only if he enters the school system, in first grade, at the beginning of school period  $T_l$ , so that pupils who enter the following stages from other systems (i.e. private schools or other geographical areas) are never included, even when they belong to the same demographic cohort;
- b) at the end of each school period, a student can either i) progress on to next grade, or ii) repeat the current grade, or iii) quit school;
- c) school is represented as a discrete-time system, so that students can change their status only at the end of the corresponding school period;
- d) no limitation is assumed for the number of grade repetitions that can occur in the whole career of a student.

As regards assumption b), it comes out that, in case a student either progresses on to the next grade or repeats the current one, the same student cannot quit school, because the three possible events (promotion, repetition and quitting) are clearly incompatible. For this reason, in case a student quits school at the end of a certain school year, but, at the same time, gets his final result (promotion or repetition), we consider two different events: first, a transition from the current school grade either to the following one (in case of promotion) or to the same one (in case of repetition), and, then, a leaving event from the last reached school grade.

Therefore, we introduced a fictitious school grade  $n+1$  in order to represent the event of final school title achievement ( $k = K$ ) in coherence with all other school leaving events, that is a transition from grade  $n$  to  $n+1$  and a subsequent exit from grade  $n+1$ .

On the basis of the shown assumptions, we are able to define the matrix  $\mathbf{T}$  of cohort events, which measures the number of times every possible event (transitions and exits) has occurred during the cohort observed life.  $\mathbf{T}$  is a three-dimensional matrix (  $(n+1) * (n+1+K) * rmax$  ), whose general element  $t_{g,j,r}$  represents the number of students with  $r-1$  past repetitions who have passed from grade  $g$  to status  $j$  in the cohort. This representation comprehends both transition and leaving events, since  $j$  can indicate either  $j$ -th school grade (when  $j = 1, \dots, n+1$ ) or  $k_j$ -th level of education ( $j = n+2, \dots, n+1+K; k_j = j-n-1$ ).

Given the number of past repetitions  $r$ , matrix  $\mathbf{T}_r$  can be graphically represented as follows:

$$\mathbf{T}_r = \begin{bmatrix} t_{1,1,r} & t_{1,2,r} & 0 & \dots & \dots & 0 & t_{1,n+2,r} & \dots & t_{1,n+K,r} & 0 & \left| \begin{array}{l} N_{1,,r} \\ N_{2,,r} \\ \dots \\ N_{n,,r} \\ N_{n+1,,r} \end{array} \right. \\ 0 & t_{2,2,r} & t_{2,3,r} & 0 & \dots & 0 & t_{2,n+2,r} & \dots & t_{2,n+K,r} & 0 & \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \\ 0 & \dots & \dots & 0 & t_{n,n,r} & t_{n,n+1,r} & t_{n,n+2,r} & \dots & t_{n,n+K,r} & 0 & \\ 0 & \dots & \dots & \dots & \dots & 0 & 0 & \dots & 0 & t_{n+1,n+1+K,r} & \end{bmatrix}$$

where  $r = 1, \dots, rmax$ ,  $N_{g,,r} = \sum_{j=1}^{n+1+K} t_{g,j,r}$  and, consequently,  $N_{1,,1} = N$ .

Our purpose is to synthesize the cohort careers information contained in  $\mathbf{T}$  into the following set of three matrices:

- **P**: a matrix representing the cohort grade-to-grade observed transition probabilities;
- **L**: a matrix representing the cohort probabilities of school leaving, from every grade to every level of education;
- **E**: a matrix representing the re-entries into the school system by those students who have previously quitted it.

In further detail, **P** is a three-dimensions matrix  $((n+1) * (n+1) * rmax)$ , whose general element  $p_{g_1, g_2, r}$  represents the observed transition probability from grade  $g_1$  to grade  $g_2$  for a student with  $r-1$  previous repetitions ( $r = 1, \dots, rmax$ ). The probability is given by:

$$p_{g_1, g_2, r} = \frac{t_{g_1, g_2, r}}{N_{g_1,,r}} \quad (g_1 = 1, \dots, n; g_2 = 1, \dots, n+1; r = 1, \dots, rmax).$$

**L** is a three-dimensions matrix  $((n+1) * K * rmax)$ , whose general element  $l_{g,k,r}$  is the observed probability of a student with  $r-1$  past repetitions to leave school from grade  $g$  with level  $k$  of education, where  $K$  is the number of different levels of education provided in the considered school system, included the entry level ( $k = 1$ ). The probability is given by:

$$l_{g,k,r} = \frac{t_{g,n+k+1,r}}{N_{g,,r}} \quad (g = 1, \dots, n+1; k = 1, \dots, K; r = 1, \dots, rmax).$$

The following properties come out easily from the ones we showed for matrix **T**:

- i)  $\sum_{g_2=1}^{n+1} p_{g_1, g_2, r} + \sum_{k=1}^K l_{g_1, k, r} = 1 \quad (g_1 = 1, \dots, n+1; r = 1, \dots, rmax);$
- ii)  $l_{n+1, K, r} = 1 \quad (r = 1, \dots, rmax).$

Given the number of past repetitions  $r$ , matrices  $\mathbf{P}_r$  and  $\mathbf{L}_r$  can be graphically represented as follows:

$$\mathbf{P}_r = \begin{bmatrix} p_{1,1,r} & p_{1,2,r} & 0 & \dots & \dots & 0 \\ 0 & p_{2,2,r} & p_{2,3,r} & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & \dots & p_{n,n,r} & p_{n,n+1,r} \\ 0 & \dots & \dots & \dots & \dots & 0 \end{bmatrix}$$

$$\mathbf{L}_r = \begin{bmatrix} l_{1,1,r} & \dots & l_{1,k,r} & \dots & l_{1,K,r} \\ \dots & \dots & \dots & \dots & \dots \\ l_{g,1,r} & \dots & l_{g,k,r} & \dots & l_{g,K,r} \\ \dots & \dots & \dots & \dots & \dots \\ l_{n,1,r} & \dots & l_{n,k,r} & \dots & l_{n,K,r} \\ 0 & \dots & \dots & 0 & 1 \end{bmatrix}$$

The structure of matrix  $\mathbf{E}$  is more complex, since the re-entry phenomenon involves several variables. Therefore, we shall not give here a graphical representation.

When taking re-entries into account, we assume that the school system is open, and this is actually what happens in real life. It is to underline that we analyze re-entries still in a cohort approach. In fact, we consider only the input of students who were part of the initial cohort.

We essentially focus on two aspects of this phenomenon:

- i) the status that the student had when leaving the system;
- ii) his actual status at the moment of re-entry.

According to our model, the following variables are included in the definition of matrix  $\mathbf{E}$ :

- $g_l$ : the school grade attended when leaving the system;
- $r_l$ : the number of past repetitions in the moment of leaving;
- $y$ : the number of school years spent out of the system;
- $g_e$ : the school grade attended after re-entering the system;
- $s_e$ : the status (either grade or leaving) achieved at the end of re-entry school year.

Given this set of variables, the number of student's actual repetitions at the moment of re-entry comes out to be:

$$r_e = r_l + y - (g_e - g_l).$$

Hence, the leaving status is given by  $(g_l, r_l)$ , while the re-entry status is given by  $(g_e, r_e)$ .

The inclusion of variable  $s_e$  is useful if one wants to separate the outcome of re-entered students from all the others.

In conclusion,  $\mathbf{E}$  is a five-dimension matrix  $(n * rmax * T * n * (n+1+K))$  and its general element  $e_{g_l, r_l, y, g_e, s_e}$  indicates the number of students (of the considered cohort) who have left school at grade  $g_l$  with  $r_l$  past repetitions and have then re-entered the system after  $y$  years, attending grade  $g_e$  and with a final result  $s_e$ .

This representation allows to relate all re-entries to the corresponding school drop-outs.

It looks clear that the past career of the students, in terms of number of repetitions occurred until the current grade, is key information in this way of representing the cohort evolution. In fact, for any possible number of past repetitions, we obtain a different set of transition probabilities and of drop-off probabilities as well. This is to say that, unlike traditional models, transition and leaving coefficients are not independent of the school record or, in other words, are not constant in time.

The third dimension of matrices  $\mathbf{P}$  and  $\mathbf{L}$  cannot be fixed in advance, since, as we stated above, we choose not to assume any limitation to the number of repetitions. Therefore, the procedure of construction of the matrices, which is led through a year by year updating of transitions for every student's career, provides for an augmentation of that dimension when a further level of repetition occurs for the first time in the cohort. As a result, the final value for the third dimension comes out to be equivalent to the maximum number of repetitions that has actually occurred in the cohort during the observation period.

One may object that, in practice, complete data relating to the entire cohort life cannot always be available, and therefore an inevitable constraint to the number of repetitions, as well as a bias of transition and leaving probabilities, is due to the presence of censored records. This is actually true, but we can well assume that these negative effects are negligible on condition that the observation period of the cohort is long enough, which implies that a very small number of students is still attending school after the last observed school period.

The above presented data structure has been built by means of an R language program.

It is to underline that this way of proceeding can be followed both for the whole cohort and, with no loss of generality, when subdividing the cohort itself in groups formed by values of a variable (or of a combination of variables) contained in the data set  $\mathbf{A}$ .

## 2.5. Measurement of school paths

In traditional Markov chain models for the analysis of school system outcome, which we have shortly recalled in chapter one and in section 2.1., the use of the matrix indicated as  $\mathbf{X}$  by Bartholomew is fundamental for the description of student careers dynamics. Indeed, through the estimation of  $\mathbf{X}$ , we can easily define the main indicators of regularity, staying at school, achievement of titles, and so on.

With reference to this matter, it is important to notice that, in the model we propose, matrices  $\mathbf{P}$  and  $\mathbf{L}$ , though possessing good analytical properties in terms of identification of the possible status of a student, are quite complex in structure, due to the specification of repetitions, and cannot be used simply enough in order to measure the probability of school paths and to define coherent indicators.

Our basic idea in this step of research was, therefore, to look for a data structure corresponding to matrix  $\mathbf{X}$ , in order to formalize an equivalent set of indicators to be used in case of a cohort-based analysis of the school system.

First of all, we introduce a definition of school path, which can be useful to subsequently define some indicators. It must be noted that, given the following definition, we only consider paths that are fully “inside” the school system, hence not including drop-off events; therefore, it will be possible to measure paths probability by means of  $\mathbf{P}$  matrix only.

After these preliminary statements, a path can be identified through the following parameters:

- $g_b$ : the path’s initial school grade ( $1 \leq g_b \leq n$ );
- $r_b$ : the number of repetitions that the student underwent before passing to grade  $g_b$  ( $0 \leq r_b \leq rmax$ );
- $g_e$ : the path’s final school grade ( $1 \leq g_e \leq (n+1)$ ); it is clear that, if  $g_e = n+1$ , the corresponding path actually ends with the achievement of final degree.
- $y$ : the length of the path, that is the number of school years that are needed to pass from grade  $g_b$  to  $g_e$  ( $y > 0$ ).

Given these parameters, the number of repetitions corresponding to the path’s final grade comes out to be:  $r_e = r_b + y + g_b - g_e$ .

We are aware that a path identified in this way can actually be formed of a set of several real paths; in other words there is generally more than one single way to progress from  $g_b$  to  $g_e$  in  $y$  years, starting from  $r_b$  past repetitions. Notwithstanding this, such a path is suitable for the definition of the main indicators, as we will show later.

In order to calculate the matrix of school paths probabilities, that we indicate here as **W**, we have implemented an R language program, shown in appendix, based on the scheme explained below.

**W** has to be a four-dimension matrix ( $n * rmax * (n+1) * Y$ ), where  $Y$  represents the number of observed school years for the cohort and, consequently, the maximum length of observed school paths. We want the general element  $w_{g_b, r_b, g_e, y}$  to be equal to the observed probability to progress from grade  $g_b$  to grade  $g_e$  in  $y$  years starting with  $r_b$  past repetitions.

In order to calculate that probability, we have to locate all possible distinct single paths corresponding to the specified set of parameters and sum up their individual probabilities.

To get to this point, let us first consider a path identified by the set of parameters ( $g_b, r_b, g_e, 1$ ): given the definition of matrix **P**, the probability of such a path is clearly equivalent to the simple transition probability  $p_{g_b, g_e, r_b}$ .

Consider then a general path with parameters ( $g_b, r_b, g_e, y$ ) such that  $y > 1$ . In this case, any possible path of that kind must be formed of  $y$  steps, the first starting from grade  $g_b$  and the last ending in grade  $g_e$ , and its probability will be equal to the product of the transition probabilities connected to the sequence of status ( $g_t, g_{t+1}, r_t$ ) that form the considered pattern.

Focusing on the first step, we can divide the general path, which, as we have shown, is a group of paths, into two sub-groups of paths, the former corresponding to an initial progress from grade  $g_b$  to grade  $g_{b+1}$  and the latter to an initial repetition of grade  $g_b$ . The probability  $w_{g_b, r_b, g_e, y}$  will then be equal to the sum of these two sub-path probabilities, once identified. Since the probability of the two possible initial steps are well known, being respectively equal to  $p_{g_b, g_{b+1}, r_b}$  and  $p_{g_b, g_b, r_b}$ , it is possible to run a recursive procedure expressing the probability of any kind of path as a composition of single transition probabilities in the following way:

$$w_{g_b, r_b, g_e, y} = p_{g_b, g_{b+1}, r_b} * w_{g_{b+1}, r_b, g_e, y-1} + p_{g_b, g_b, r_b} * w_{g_b, r_b+1, g_e, y-1}$$

The graph 2.2. represents, as an example, the composition of a path starting from school grade 1 at time 1 and ending in school grade 2 after two school years (hence, at time 3). The four factors that compose the probability of the path are evident.



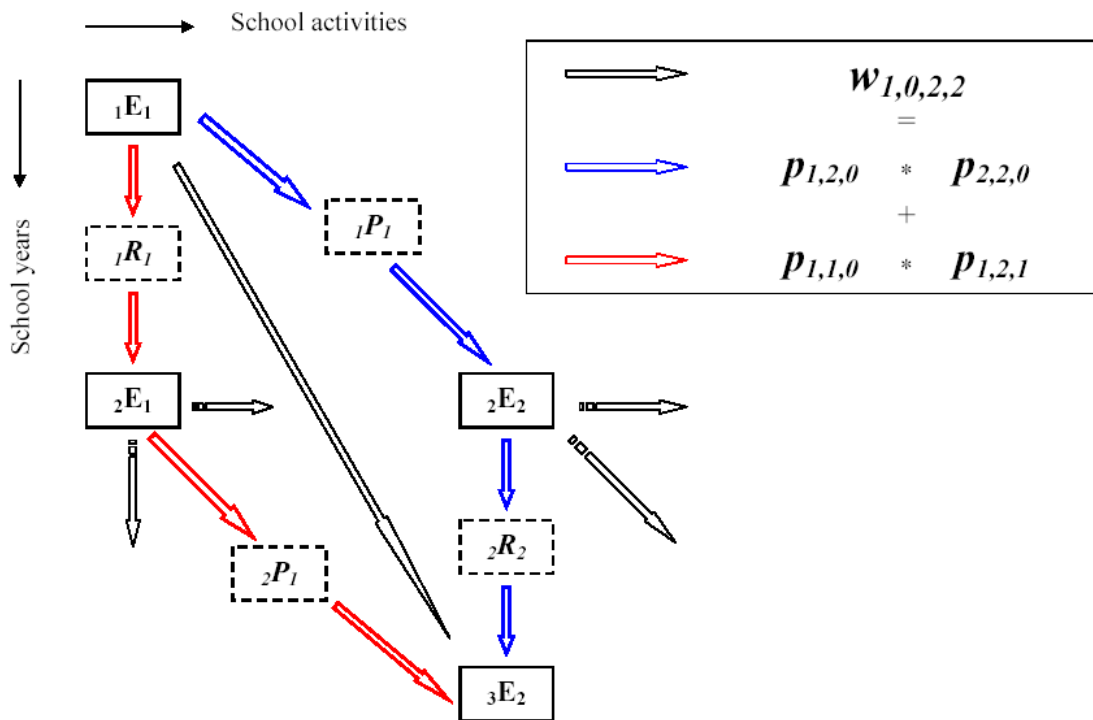


Figure 2.2. - A graphical representation of a school path

The introduced concept of path appears to effectively synthesize the dynamics of the analyzed cohort. This will result more clearly in the next section, where we will define some performance indicators for the school system.

## 2.6. Definition of indicators

The indicators we are now going to introduce aim to provide some simple measures of students performance, relating to the following phenomena:

- a) staying at school;
- b) attainment of different education levels;
- c) school life expectancy.

This is made in coherence with an equivalent set of indicators proposed for the Markovian contemporary-based SFINGE model (Trivellato, 1980). Though focusing here only on these three aspects, we do not foresee any particular problem in the definition of several other indicators based on this model.

However, this would probably go beyond the aim of this research, which is more oriented to the various formal aspects of the model.

Before going into definitions, we remark that all the indicators we are going to propose contain, together with parameter  $g$  (the starting school grade) also parameter  $r$ , representing the number of repetitions occurred until the student reaches grade  $g$ . This aspect confirms the importance given in our approach to the student past career: we base this choice on the conviction that measures of permanence and probabilities to achieve school titles are influenced much more by the overall past career (number of repetitions occurred) than by the last experienced event (repetition / non repetition of current grade), upon which, on the contrary, traditional models like SFINGE are based.

### ***Staying at school indicators***

The first two indicators we are showing relate to the probability of staying at school, and can be easily obtained using **W** matrix:

#### a) *Regularity tables*

They describe students probability of progressing from a given school grade  $g$  ( $g = 1, \dots, n-1$ ), with  $r-1$  past repetitions ( $r = 1, \dots, rmax$ ), to each of the following grades  $g+y$  ( $y = 1, \dots, n-g+1$ ) in  $y$  years (that is, with no further repetitions).

$$\mathbf{reg}_{g,r} = [w_{g,g+y,r,y}] \quad (y = 1, \dots, n-g+1)$$

#### b) *Permanence tables*

They describe the probability to stay into the considered school system, starting from grade  $g$  and  $r-1$  past repetitions ( $g = 1, \dots, n$ ;  $r = 1, \dots, rmax$ ), regardless of any further repetition.

$$\mathbf{perm}_{g,r} = \left[ \sum_{j=g}^n w_{g,j,r,y} \right] \quad (y = 1, \dots, Y)$$

### ***Indicators of attainment of different education levels***

In the formalization we propose,  $k$  indicates the different levels of education ( $k = 1, \dots, K$ ;  $K$  being the number of education levels at which students can leave school in the considered system). We indicate the entry education level with  $k = 1$  and the final one with  $k = K$ . Referring, as an example, to the Italian high secondary school system in late 1990s, we can recognize four levels of education (hence,  $K = 4$ ), that are in detail:

- $k = 1$ : low secondary school license (entry education level)
- $k = 2$ : professional schools intermediate degree (granted by professional institutes after three years)

- $k = 3$ : low level teaching degree (granted by teaching institutes after four years)
- $k = 4$ : final degree (*diploma*, granted by all high schools after five years)

As regards this phenomenon, the indicators we propose are:

c) *Tables of regular school leaving with title  $k$*

They describe students probability of regular school leaving after achieving school title  $k$  (where  $k = 1$  represents the entry education level of the considered school system) starting from school grade  $g$  and  $r-1$  past repetitions ( $g = 1, \dots, n$ ;  $r = 1, \dots, rmax$ ).

In order to calculate this indicator, we have to know the minimum number  $d_k$  ( $k = 1, \dots, K$ ) of school years needed to achieve education level  $k$  (where  $d_1 = 0$  and  $d_K = n$ ).

For example, in relation to the Italian high secondary school system we have introduced above, the corresponding vector  $\mathbf{d}$  is given by:

$$\mathbf{d} = [0 \quad 3 \quad 4 \quad 5]$$

Then, the  $k$ -th element of the table of regular school leaving ( $\mathbf{rsl}_{g,r}$ ), starting from grade  $g$  and  $r-1$  past repetitions, is given by:

$$rsl_{g,r,k} = \begin{cases} 0 & k = 1 \\ w_{g,d_k+1,r,d_k+1-g} * l_{d_k+1,k,r} & k = 2, \dots, K \end{cases}$$

When  $k = 1$ , the probability of regular leaving is obviously zero, since a student cannot be regularly leaving with entry education level once he has enrolled to the considered school system.

For  $k > 1$ , the first factor of  $k$ -th element of the table corresponds to the  $(d_k+1-g)$ -th element of  $\mathbf{reg}_{g,r}$ . The second factor, instead, represents the leaving probability, with  $k$ -th education level, from grade  $d_k+1$ , and not from grade  $d_k$  since, as we have already explained in section 2.4., when a student quits school after a promotion, the model reports quitting at the beginning of the following school year.

d) *Tables of school leaving with title  $k$*

They describe the probability of leaving the school at various levels of education, regardless of the number of school years needed to achieve them, starting from school grade  $g$  and  $r-1$  past repetitions ( $g = 1, \dots, n$ ;  $r = 1, \dots, rmax$ ).

$$\mathbf{sl}_{g,r} = \left[ l_{g,k,r} + \sum_{y=1}^Y \sum_{j=g}^{n+1} w_{g,j,r,y} * l_{j,k,r+y+g-j} \right] (k = 1, \dots, K)$$

The first member of the sum is the probability to leave the school during the current year and, consequently, at the current number of repetitions. The first factor in the sum represents the probability of following a school path getting to each possible school grade  $j$  in every possible number of years  $y > 0$ . In the second factor, we correctly consider the leaving probability at the number of repetition corresponding to path's end, that is indeed  $r+y+g-j$ .

The two above shown indicators do not measure, in general, the probability of school titles achievement, but the probability of leaving school with the various school titles (of course, this makes no difference as far as the final title is concerned, but it is relevant in case of intermediate ones).

A workaround for this constraint, apart from running the model separately on each school type, can be adding another dimension, representing the kind of school attended by the student, to the matrix system; that further specification may then be ignored for the major part of the indicators, and considered, instead, in those analyses, and this would be the case, where the influence of school type on the measured aspect is thought to be relevant.

### ***School life expectancy indicators***

The last two indicators we are going to propose regard the expected duration of school life and can be easily formalized by means of regularity and permanence tables.

#### e) *Expected regular school life*

This indicator measures the expected length, expressed in school years, of regular school life (without repetitions), starting from school grade  $g$  and  $r-1$  past repetitions ( $g = 1, \dots, n-1$ ;  $r = 1, \dots, rmax$ ).

$$ersl_{g,r} = 1 + \sum_{y=1}^{n-g} w_{g,g+y,r,y}$$

where the element in the sum is clearly the  $y$ -th element of  $\mathbf{reg}_{g,r}$ .

#### f) *Expected permanence at school*

It is a measure of expected length, expressed in school years, of school life, regardless of the number of repetitions, starting from school grade  $g$  and  $r-1$  past repetitions ( $g = 1, \dots, n$ ;  $r = 1, \dots, rmax$ ):

$$eps_{g,r} = 1 + \sum_{y=1}^Y \sum_{j=g}^n w_{g,j,r,y}$$

where the interior sum is equal to the  $y$ -th element of  $\mathbf{perm}_{g,r}$ .

As we assumed that students can change their status only at the end of every school year, then  $ersl_{g,r} \geq 1$  and  $eps_{g,r} \geq 1$ , since permanence is guaranteed at least for the whole current year.

We want to point out once again that several other indicators that refer to further aspects of the school system may be proposed, but this would go beyond the scope of this work. Indeed, the indicators that we have seen above investigate some phenomena of major interest and can also represent a good example of how synthetic measures can be easily drawn with the helpful use of  $\mathbf{W}$  matrix, as we will discuss further in the next section.

## **2.7. An overview to the model: what has still to be done?**

We have so far gone through the formalization of a cohort-based model that has been specifically thought for the study of dynamics implied in the school system. The definition of transition and leaving probabilities and the formal construction of what we called school paths matrix have led to proposing a set of indicators that aim to evaluate school cohort performance from different, though connected, points of view.

We now enumerate the main advantages of our proposed modeling approach, compared to a traditional contemporary-based Markov Chain model, together with some limits that may reduce its analysis potentialities in some cases.

A quick overview is reported in table 2.2. Each aspect is then explained more in detail.

As for the information needed, this model investigates cohort dynamics with the use of coherent longitudinal data. Models based on stock data, on the contrary, need to assume that the behavior of contemporary students, actually belonging to different cohorts, can be used to analyze the school system evolution. Instead, we are firmly convinced that the analysis of dynamics in a system must imply the use of longitudinal data, preferably collected on individuals.

In relation with this aspect, the original data needed is quite simple to collect and store, since it consists of very few variables that are usually readily available in all schools. Also their processing, done in order to obtain the suitable structure for model feeding, is easy enough to perform, as we showed in section 2.4.

As regards the model specification, we have made two significant innovations with respect to existing models, even to cohort-based ones (Légaré, 1972), as we are going to explain further on.

First of all, the model takes the number of past repetitions into account for all the presented analyses and indicators, meeting the fundamental assumption that future career of students strictly depends on what students experienced along the overall past career, more than from the most recent events only. This assumption implies that the transition and leaving probabilities from a given school grade are not constant in time.

Table 2.2. – An overview to the model

<b>Phases of modeling</b>	<b>Points of strength</b>	<b>Points of weakness</b>
<b><i>Data collection and arrangement</i></b>	<p>Use of longitudinal data (more suitable to catch time flows)</p> <p>Row data are simple to collect and maintain along time</p> <p>The data sets used to feed the model are easy to set up</p>	<p>Cohorts need to be followed for a long time (until their extinguishment)</p> <p>Row data are usually not provided by “official” sources</p>
<b><i>Model specification</i></b>	<p>The re-entry phenomenon can be taken into account</p> <p>The transition and leaving probabilities depend on the past school career (they are variable in time)</p>	<p>The transition and leaving probabilities do not depend on the condition of enrolment in the actual school grade</p>
<b><i>Application aspects</i></b>	<p>The model is applicable to the whole educational system, with an appropriate definition of the set of possible status</p> <p>All indicators can be obtained by means of the school paths matrix</p> <p>Indicators can be calculated starting from any possible school status (good for comparing behaviors with respect to past career)</p> <p>The model looks suitable for simulations of the impact of changes that may occur into the school system</p>	<p>The model does not manage to represent well short period tendency and behaviors</p>

In addition to that, this approach permits to choose whether to take into consideration possible re-entries in the school system; moreover, the behavior of students after such re-entries can be analyzed separately from all other school events. Considering the phenomenon of re-entries into the analysis is particularly

desirable in case the model is applied to high-level education systems and on a small scale, where student migrations are more likely to occur.

It is also important to underline how, despite these two substantial refinements, the final model is still characterized by an acceptably low complexity in structure, and, therefore, keeps on being very manageable and easily interpretable.

Finally, as far as the model application opportunities are concerned, we stress the following points of strength.

Firstly, it is possible to extend the analysis to the whole school system, from primary education onwards, using the same model, with no need to modify its theoretical structure at all, by only appropriately defining the set of status and of possible exit levels.

Secondly, since this model is based on cohort follow-up with longitudinal data, it appears to be particularly suitable to catch the effects of the introduction of structural changes in the school system that are likely to influence the evolution of cohorts. This makes the model more suitable for simulation purposes than those based on cross-sectional data.

Another valuable feature regards the set of indicators proposed: being all indicators derived from a composition of school path probabilities, each of them can be generally calculated for every possible initial status and not only for the generation of the cohort. This is very helpful if one wants to compare student performance with respect to past career. This opportunity will be exploited in the analysis shown in next chapter, where various indicators for the whole cohort will be compared with the ones relating only to students repeating the first grade.

Together with these positive aspects, we believe that the proposed approach presents also some limitations. We are going to discuss these in detail following the same sequence we have adopted so far in this section.

As regards the data structure, the main issue is the lack of suitable information to be recovered from “official” sources, which, at least in Italy, do not usually consist of longitudinal data. Therefore, it makes necessary to collect them from “ad hoc” sources, though this can become problematic in case the model needs to be applied on a vast scale.

In addition to that, since cohorts of students should be followed at least until they are nearly extinguished, data needs to be collected for a long period of time. Therefore, assuming that data collection starts in a certain year, it is impossible to gather the necessary information to feed the model for a very long time.

A partial remedy for this problem would be to estimate SFINGE model for school year  $t$  on the basis of real transition events from year  $t$  to year  $t+1$ , instead of simply estimating flows on the basis of stock data per year. This may reduce some of its important bias, but would not represent a cohort-based approach anymore.

As for the model specification, the most evident limitation is due to the fact that transition and leaving probabilities, though dependent from the overall past career, do not depend strictly from the fact of being at present in a repeat enrolment. As a consequence, two scholars that enroll to the same grade and have the same number of past repetitions are assumed to have the same leaving and transition probabilities, even if one of them is repeating that grade, while the other is not.

This assumption is actually quite unrealistic and directly leads to a threat to the model application capabilities. Indeed, it is clear that this model does not effectively apply to the estimation of short-term student behavior.

Nevertheless, this is partially due to the choice that we made at the beginning of the research to look for a model that is able to analyze school system dynamics in the medium and long period. We believe this to be more relevant as far as the objective of study is represented by a cohort.

## **2.8. Conclusions and proposals for further research**

In this final section we focus on the aspects of the proposed approach that we believe deserve a more accurate formal development. We also suggest improvements to testing and application of the presented model.

As for the first aspect, the proposed indicators should be integrated with new ones, in order to make the model really useful to analyze all major school system phenomena. In particular, we point out the great usefulness of measures of expected educational attainment, as well as the need to evaluate the school system's overall grade of efficiency.

With regard to this point, we are convinced that the school paths probability matrix, defined as explained in this chapter, can well represent the basis for the definition of any indicator connected with students' careers.

In addition to the introduction of new indicators, another improvement of the model would be to add further variables in its specification. These time-variant variables imply an increase in the number of possible states. In general, the model can provide for a subdivision of matrices based on the values of all the variables that concur to the definition of school paths.

A typical time-variant variable in the considered system is the type of school attended by students every year of their career. The subdivision of matrices per school is likely to help considerably in the detection of different behaviors among groups of students. It would also make possible to detect migrations among schools with subsequent advantages in the estimation of transition coefficients. Moreover, in every application it would be possible to choose which dimensions to take into account. This aspect will be treated more deeply in chapter three.



As regards the empirical aspect, next chapter shows the application of the model to high secondary school students' careers in the Province of Pisa, referring to the data collected from school year 1993/94 to 2002/03.

## Chapter Three

# FOLLOWING STUDENTS: A COHORT ANALYSIS OF PISA PROVINCE STUDENTS' EXPERIENCE IN SECONDARY EDUCATION

### **Abstract:**

The paper shows the results of the application of an innovative cohort-based model to a longitudinal data collection about Italian school system at local level. After synthetically describing the case of study, we show some interesting evidence deriving from the calculation of a set of school performance indicators under a cohort point of view. Analyses are led aimed to look for different school outcomes among cohorts, taking into account the influence of student background and school of enrolment towards school performance. In this first application, indicators seem to have a good sensitivity in catching some major determinants of school careers.

### **3.1. Introduction**

In recent times, a growing interest has regarded the analysis of school system in Italy. This interest arises from the need to evaluate the quality of educational process in a historical phase characterized by frequent changes in the organization of the whole national education system.

The effect of radical reforms are very complex to detect and represent by means of synthetic measures like performance indicators, since the variables involved in such changes are numerous and often very correlated to each other.

However, the correct approach to these kinds of analysis should be cohort-based, for the reasons explained in previous chapters. Some recent applications to this matter in the Italian school system, often referring also to the study of school-to-work transitions, can be found in Bernardi and Ghellini (1997), Ghellini et. al. (1999 and 2000), Ghellini (1996 and 1997) and Zaccarin (1994).

In this paper we will show some possible results that can be drawn through the application of the model presented in chapter two on a data set with very desirable properties to give rise to a cohort analysis.

Although this application is quite partial, it should put in evidence some valuable features of the defined cohort approach model.

### **3.2. The case of study**

The analyzed data relates to the high secondary school system of Pisa province. They have been gathered by the *Osservatorio Scolastico Provinciale*, within a project that consists of the creation and management of an informatics system recording student school careers. This project has recently been extended to the whole local school system (that is from primary school onwards) and is intended to involve all the other provinces of Tuscany in the near future.

For the application that we will show in this paper, we have used data referring to students enrolled in public high schools from school year 1993/94 to 2002/03. That makes a total of ten years of observation.

The original data we collected from the above mentioned source consists of the three following data sets.

- a) An “Entry” data set that contains those variables that can be considered time-invariant with regard to the period spent by students into high school; these variables refer essentially to personal characteristics, familiar background and past school career and are believed to be in quite strong relationship with the future school career. We will refer to these variables as entry variables.
- b) A “School Years” data set that reports, for every school year spent by each student in the considered school system, the name and the type of school attended, the school grade of enrolment and the final result achieved.
- c) A “Schools” archive that contains the main characteristics of public high schools located in the province of Pisa.

The first two data sets are linked by means of the personal identification number, assigned univocally to every student. In addition to that, information reported in the “Schools” data set can be linked to records in “School Years” data set by means of the identification code relating to every school in the system considered.

Therefore, we have been able to reconstruct the school career of every student, including his enrolment profile (time-invariant entry variables) and his year-by-year path through the observed school system.

On the basis of the year of enrolment into high school, it has been possible to assign every student to his cohort, taking into account only the cohorts that

starting during the observed period. That makes a total of ten cohorts<sup>1</sup>.

The variables contained in the “Entry” data set are in detail:

- *Gender*, male (M) or female (F);
- *Age of entry*, which indicates whether a student has enrolled into the school system in “regular” age ( $\leq 14$  years of age) or later ( $> 14$  years of age);
- *Low secondary school result*, which is *low* for students who have achieved low secondary school *licenza* with *sufficiente* or *buono* marks, while it is *high* for those who have achieved it with *distinto* or *ottimo* marks;
- *Parents’ education level* is *low* if both student’s parents have a low level of education (no title, primary school *licenza*, low secondary school *licenza* or professional school diploma), while it is *high* if at least one of the student’s parents has a high level of education (high secondary school diploma or university degree).

Table 3.1. shows the distribution of entry variables for all cohorts.

Table 3.1. – Distribution of entry variables per cohort

Cohort	93/94	94/95	95/96	96/97	97/98	98/99	99/00	00/01	01/02	02/03
<b>Gender</b>										
M	50.4	49.4	50.7	48.4	48.0	50.5	50.1	49.9	52.8	51.5
F	49.6	50.6	49.3	51.6	52.0	49.5	49.9	50.1	47.2	48.5
<b>Age of entry</b>										
$\leq 14$ years	81.8	85.7	86.9	88.6	89.2	88.8	88.6	87.8	88.7	90.7
$> 14$ years	18.2	14.3	13.1	11.4	10.8	11.2	11.4	12.2	11.3	9.3
<b>Low secondary school result</b>										
Low	68.9	65.5	64.0	63.9	64.9	63.8	64.8	63.7	66.3	60.3
High	30.4	33.1	33.3	34.7	32.7	31.8	32.3	34.0	32.4	33.4
Not Available	0.8	1.4	2.6	1.4	2.4	4.4	2.9	2.3	1.3	6.3
<b>Parents’ educ. level</b>										
Low	52.6	50.6	48.7	44.8	44.1	41.4	37.7	39.5	35.2	31.6
High	33.8	37.3	35.2	37.0	38.8	39.5	37.6	41.1	40.5	37.2
Not Available	13.6	12.1	16.1	18.2	17.1	19.1	24.7	19.4	24.3	31.2
	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>
<i>No. of students</i>	3539	3226	3171	3008	2989	2856	2978	2837	2769	2983

<sup>1</sup> In this work, we define cohort *t* as the set of all those students who have enrolled, for the first time, into the first grade of the considered school system in school year *t*.

We can underline the following evidence:

- a) the gender distribution is quite stable in the period considered and the proportions of males and females in all cohorts are very close to each other;
- b) the proportion of students entering the school system in “regular” age considerably increases in time (growing from 81,8% in cohort 93/94 till over 90% in cohort 02/03);
- c) the distribution of the low secondary school final result shows a quite constant pattern in most of the cohorts, with the exception of the first and last one (the latter is also characterized by a rather high rate of missing data);
- d) the parents’ educational level is with no doubt the least stable variable in this group for the following reasons: firstly, the frequency of low educational levels has dropped in ten years time to nearly half of the first cohort value (from 52.6% to 31.6%), while the frequency of high educational levels has slightly grown; secondly, the percentage of missing data, especially in the most recent cohorts, is very relevant.

Looking at the entry profile of cohorts before passing to any phase of modeling is very important in order to point out, when possible, the presence of selection effects existing in performance among the various groups of student. The above underlined aspects will be therefore taken into great account in the following analysis.

### **3.3. The estimation of the model**

In order to calculate the set of matrices involved in the proposed model, we actually need to refer to cohorts being sufficiently covered by the available observations. For this reason, we mainly focused on the first six cohorts. Among these, we have decided to concentrate on two in particular. As the first one, we have chosen 1994/95 cohort, since it gave more guarantees of data quality, with respect to the previous one. As the second one, we have referred to 1997/98 cohort, which, though being separated enough in time from the other one, still provides for six observed years, which give us the possibility to examine the behavior of regular students as well as of those who finish school after one repeat enrolment.

The described raw data has been processed in order to set up the four data structures that permit to run the model chosen and that, in chapter two, we have indicated with **G**, **S**, **R** and **A**. In this phase, we have used only some of the available information.

Referring to the formalization seen in chapter two, in our case of study the central parameters of the model are:

- $n = 5$  (number of grades into high secondary school)
- $K = 4$  (number of possible education levels of exit)
- $\mathbf{d} = [0 \ 3 \ 4 \ 5]$  (number of school years needed to achieve every education level as a regular)
- $t_1 =$  starting school year of the cohort (i.e.  $t_1 = 1994$  for 1994/95 cohort)

Table 3.2. clarifies the types of education levels provided in the Italian school system in the period considered.

*Table 3.2. – School titles provided in the Italian Educational System*

<b>K</b>	<b>d<sub>k</sub></b>	<b>Title description</b>
1	0	No title
2	3	Intermediate Professional Degree
3	4	Normal School “Low Level Teaching” Degree + Professional Art Degree
4	5	Diploma

In the following sections we will show some results of the application of the model to the observed cohorts, taking into account the following indicators:

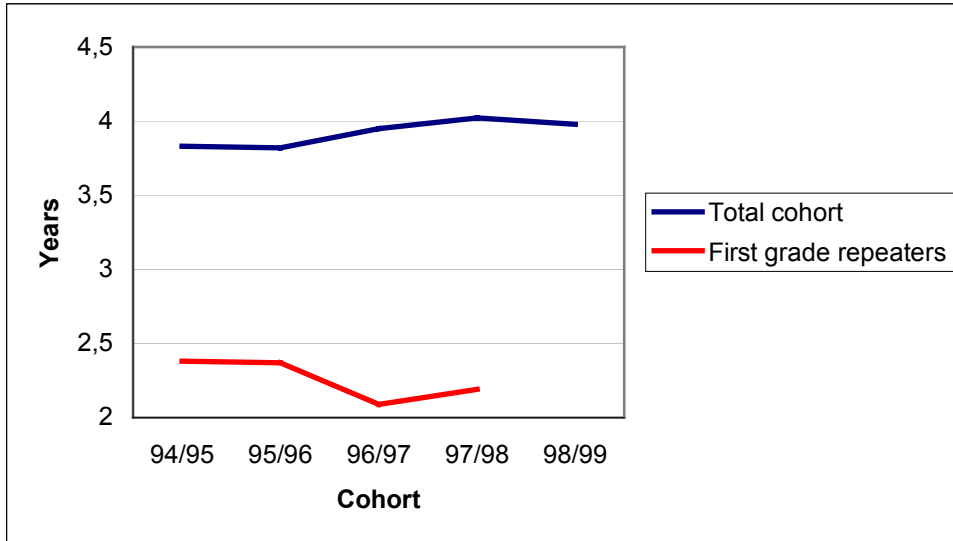
- *Regularity table*. It is referred to a cohort (usually fixed at 10.000 units) starting at time  $t_1$  and reports the number of students who have had a regular school career after  $y$  years ( $y = 1, \dots, n-1$ )
- *red* (Regular Exit with Diploma). This indicator is drawn from the table of regular school leaving with  $k = K$ .
- *ed* (Exit rate with Diploma in any number of years). This indicator is drawn from the table of school leaving with  $k = K$ .
- *ewt* (Exit rate Without any Title). This indicator is drawn from the table of school leaving with  $k = 1$ .
- *ersl* (Expected Regular School Life).
- *eps* (Expected Permanence at School).

### 3.4. The comparison of cohort performance

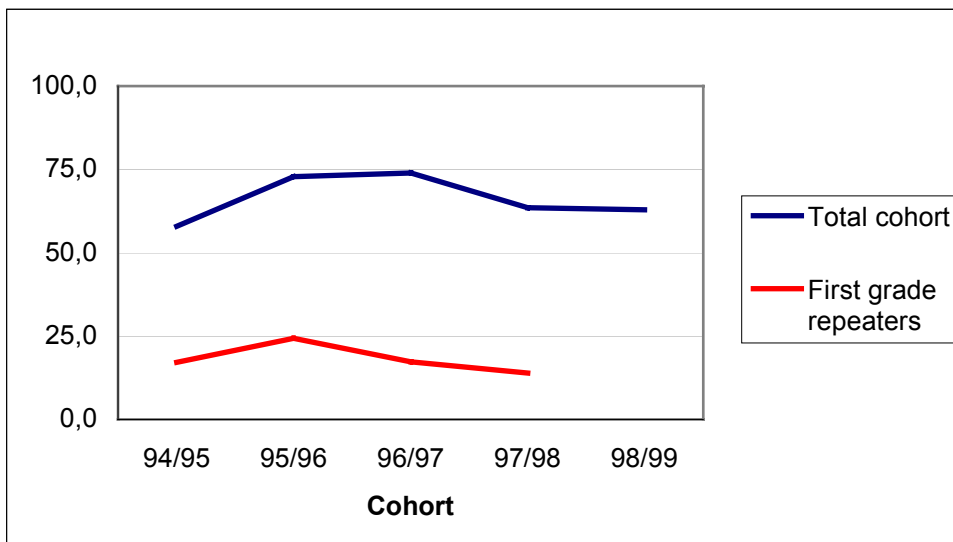
The first results that we are going to show regard the comparison of some performance indicators among the cohorts. Cohort 1993/94 has been excluded from the following applications, due to a lack of data quality detected in

preliminary analysis. Hence, we will refer here to cohorts going from 1994/95 to 1998/99.

Figure 3.1. shows the values of *ersl* and *red* for the five cohorts. For each cohort, indicators have been calculated i) for the total initial enrolment ( $g = 1, r = 1$ ) and ii) for students repeating first grade ( $g = 1, r = 2$ ). Due to the too short time of observation, only the former values are provided for cohort 1998/99.



*Figure 3.1 – Regular School Life Expectancy (rsle) on 5 cohorts*



*Figure 3.2. – Regular Exit Diploma (%) on 5 cohorts*

As for regularity, the overall performance of cohorts has increased in the period considered: regular school life expectancy, indeed, has grown from 3,83 years for 1994/95 cohort to about 4 years for 1998/99 cohort. On the contrary, such expectancy has decreased for students repeating first school grade. This suggests that the impact of an initial repeat enrolment has progressively become more dramatic.

As for the achievement of diploma, the situation is slightly different for several reasons. Firstly, the value of *red* for both groups, after an initial increase in time, tends to decrease again for most recent cohorts. Secondly, the impact of repeating first grade is much more dramatic for the probability to achieve the diploma as a regular: in the cohort 1997/98, for instance, *red* is around 64% for the total cohort, but only 14% for students repeating first school grade. This suggests that most of the students who, after repeating first school grade, manage to achieve the final school title experience further repetitions during their subsequent career.

It is interesting to notice how, in this case, the two indicators tend to give information that, at first glance, look rather in contrast with each other. A high rate of diploma achievement as a regular does not emerge as a simple consequence of a proportionally high regular school life expectancy.

Table 3.3 shows other opportunities of comparison among school cohort performance. In particular, we refer here to two cohorts, 1994/95 and 1997/98, and introduce the analysis of regularity patterns.

*Table 3.3. – Comparison between cohorts 1994/95 and 1997/98 performance indicators*

	Total students enrolled (T)		Students repeating first grade (T+1)	
	94/95	97/98	94/95	97/98
<i>Regularity table</i>				
Initial cohort	10.000	10.000	10.000	10.000
After 1 year	8.342	8.692	5.545	5.028
After 2 years	7.306	7.845	3.725	3.086
After 3 years	6.550	7.066	2.627	2.232
After 4 years	6.100	6.641	1.895	1.543
<b>ERSL (years)</b>	3,83	4,02	2,38	2,19
<i>RED (%)</i>	57,9	63,6	17,1	14,0
<b>ED (%)</b>	74,0	-	25,6	-
<b>EWT (%)</b>	22,1	-	70,0	-

The values in regularity tables confirm the evidence arising from the previous analysis. Cohort 1997/98 shows a better overall regularity pattern for every time lag considered: for instance, 6641 students out of a total of 10.000 are still regular after 4 years, with respect to 6100 of the other cohort. On the other hand, regularity tables show a contrary state for students repeating first grade.

The values of *ersl* and *red* have already been commented.



The two last indicators, *ed* and *ewt*, have been calculated only for 1994/95 cohort, since the other one is not fully observed (too many students were still at school after six years of observations). The diploma rate, which is equal to 74% for the whole cohort, drops to 25,6% for students repeating first grade, underlining again the negative effect of an initial repeat enrolment on the possibility of achieving the final school title. The same conclusion can be drawn by looking at the values of *ewt*.

### 3.5. Cohort analysis by entry variables

School outcome should not be seen solely as a consequence of the school system's "quality", but, with no doubt, it is partly due to some student individual characteristics. In our case of study, as we have explained in first section, we dispose of some very interesting entry variables that represent the profile of students at the moment of their enrolment into the high school system.

In this section, therefore, we will show some results of an analysis of school performance indicators that take into account the differences in the students' entry profiles. All results are referred to cohort 1994/95.

*Table 3.4. – Cohort 1994/95 performance indicators by entry variables*

	Gender		Age of entry		Parents education level		
	M	F	Regular	Later	High	Low	Missing
<b>ERSL (years)</b>	3,52	4,13	4,04	2,57	4,22	3,63	3,46
<b>ESP (years)</b>	4,81	4,92	5,05	3,82	5,00	4,81	4,71
<b>RED</b>	47,4	68,1	63,5	24,2	70,7	50,6	48,6
<b>ED</b>	65,8	82,0	80,7	33,7	83,8	68,5	66,3
<b>EWT</b>	29,4	15,0	16,0	58,1	14,4	25,9	29,8

As for students' gender, all indicators show a better performance for females than for males. The difference is particularly relevant in the rates of exit with diploma: 68,1% of female students achieve diploma with no repetitions, while only 47,4% of males do the same. In addition to that, males' exit rate with no title nearly doubles females' one.

The age of entry into the system is even more relevant with respect to school performance.

First of all, students entering school after "regular" age have a very low school permanence expectancy (3,82 years), which suggests a very high proportion of drop-outs. Moreover, *ersl* and *red* indicate a much lower probability to have a regular school life with respect to students who have entered school in "regular" age.

Maybe, the worse consequence of a non-regular initial enrolment is represented by the dramatically low graduation rate (*ed*), which tells us that only about a third of the students entering school after “regular” age manages to get the final diploma. Hence, high schools do not seem to be able to give reasonably good school career perspectives to those who have already experienced some delays during their previous school life.

Familiar background too, synthesized here by the parents’ level of instruction, clearly influences student school performance, as is shown by all the indicators in the table. Gaps between the two-formed groups (high and low parents’ education level) are substantial under every considered aspect. This can be the mirror of the school’s difficulty in achieving good results apart from the student family cultural context.

In conclusion to this section, we have put in evidence the strong influence of entry variables on the future school career of students. This enforces the need to perform this kind of analyses taking into account as more as possible all those external variables that can help explain much part of the dynamics involved in the school system.

### 3.6 Accounting for schools in the analysis

Another important aspect to be taken into account when analyzing the school system’s outcome is the presence of different kinds of schools. In our case of study, schools have been classified into four categories: “Liceo”, Technical Institutes, Normal and Art Schools, Professional Institutes. Their rather different internal organization induced us to look for separate analyses to lead on each of them, in order to detect possible differences in student career results among them. This part of the analysis relates only to 1994/95 cohort.

Students have been assigned to the school type of first enrolment in the system. Hence, we assumed the absence of school change during the career. The distribution of enrolments by school type is shown in table 3.5.

*Table 3.5. – Distribution of students by school of enrollment in cohort 1994/95*

	No. of Enrolments	Percentage
“Liceo”	840	26.0
Technical Schools	1360	42.2
Normal and Art Schools	633	19.6
Professional Schools	393	12.2

For every group, we have calculated the same set of performance indicators as we did in previous section for entry variables. The results are shown in table 3.6.

*Table 3.6. – Cohort 1994/95 performance indicators by school of enrollment*

	<b>School</b>			
	<b>“Liceo”</b>	<b>Technical Schools</b>	<b>Normal and Art Schools</b>	<b>Professional Schools</b>
<b>ERSL (years)</b>	4,45	3,69	3,84	2,97
<b>ESP (years)</b>	5,07	4,98	4,79	4,19
<b>RED</b>	79,8	53,8	56,3	27,7
<b>ED</b>	92,2	74,5	70,5	38,2
<b>EWT</b>	7,7	24,7	22,4	43,4

It is to notice that the last two groups (Normal or Art Schools and Professional Institutes) are not fully comparable with the others, since, as we have already mentioned in chapter two, they provide for intermediate diplomas that students can achieve in a shorter time, hence reducing the overall permanence at school.

Looking then in particular at the first two columns of the table, we observe a much better performance for “Liceo” students than for Technical Institute ones. Nearly 80% of students enrolled in “Liceo” achieve the final title as regular (against less than 54% of those enrolled in Technical Institutes); the total graduation rate is even more noticeable, being higher than 92% (against 74,5% of Technical Institutes) and, consequently, the exit rate without any title is negligible with respect to the other school type.

As regards this last indicator, we underline the dramatically high drop-out rate related to Professional Institute students (43,4%).

It is to say, however, that the shown gaps are partly due to substantial differences existing in the background of students who initially choose the various types of school. This is clearly confirmed by table 3.7., which shows the joint distribution of the type of school attended with respect to each entry variable.

*Table 3.7. – Distribution of entry variables by school of enrollment*

	<b>School</b>			
	<b>“Liceo”</b>	<b>Technical Schools</b>	<b>Normal and Art Schools</b>	<b>Professional Schools</b>
<b>Gender</b>				
M	43.3	59.5	25.0	66.9
F	56.7	40.5	75.0	33.1
<b>Age of entry</b>				
Regular	98.2	90.0	79.5	53.7
Later	1.8	10.0	20.5	46.3
<b>Parents’ education level</b>				
High	60.6	32.1	33.0	12.2
Low	23.1	58.5	57.4	71.5
Missing	16.3	9.5	9.6	16.3

The considerations made before would recommend to analyze the school system performance separately for both entry variables and type of school attended, even though this part is lacking in this paper.

Another way, even more radical, to account for the type of school attended in the used model is to introduce this variable in the specification of the model. The school of enrolment, in fact, being a time-variant variable, cannot be considered as entry variables are: indeed, school-to-school transitions should be taken into the same account as grade-to-grade ones in order to capture most of the dynamics involved in the system. This proceeding can be lead following the same general method of construction of the model explained in chapter two, with a consequent augmentation in the number of states and in matrices dimensions.

### **3.7. Conclusions and further applications**

In this paper we have put in evidence, even if through a very limited set of analyses, some interesting aspects regarding school performance referred to cohorts of students. Some behaviors emerged have also given rise to the perspective of further research and application. In addition to the ones mentioned in this chapter, we point out the following applicative aspects that would deserve to be seriously considered.

As we have already introduced in chapter two, the model proposed is applicable to the whole education system, with no need to change its theoretical structure. It would therefore be desirable to extend it to other school levels, even if the definition of status, transitions and education levels would probably be much more complicated.

In this optic, an application to the University system is with no doubt a purpose to achieve, given the importance that evaluation processes have assumed in the academic world in the recent past. However, such application would probably need a deep revision of the model, since University system is much more complex in structure than school one and even a proper definition of the state space is likely to be very tasking.

We also wish to perform an accurate comparison with the results of the application of models based on contemporary data: this may eventually lead to the definition of relative efficiency indicators between different models and, hopefully, to the proposal of theoretical rules aimed to correct biases produced in the estimation of flows by models that make use of stock and cross-sectional data.

Finally, the model's predictive capability should be verified, since the prediction of school future performance must certainly be one of the application opportunities to mostly take into account. It is to say, indeed, that the data of our

case of study would be suitable for this purpose, since they give information on a considerably high number of cohorts.

## Appendix – R program for the creation of “school paths” matrix (W)

```
school_paths <- function (P, N_years)

# Parameter P is the probability matrix referred to the cohort analyzed
# Parameter N_years is the maximum length of paths that we want to calculate

{

##
Recursive function prob_t: it calculates the probability to go from grade g1 with
rep past repetitions to grade g2 in y years.
##

prob_t <- function (g1, g2, rep, y)

{

if (y == 1)
{
prob <- p [g1,g2,rep]
}
else
{
if (rep < max_rep)
{
rep1 <- rep + 1
w1 <- p [g1,g1,rep] * prob_t (g1,g2,rep1,y-1)
}
else

## No further repetitions have been observed, so the probability is approximated by
p [g1,g1,rep]

{
w1 <- p [g1,g1,rep]
}
if (g1 < N_grades)
{
w2 <- p [g1,g1+1,rep] * prob_t (g1+1,g2,rep,y-1)
}
else

## No further promotions are possible, so the promotion probability is zero

{
w2 <- 0
```

```

}

# The result is the sum of the two sub-paths probabilities

result <- w1 + w2
}

result
}

## MAIN FLOW

P <- P [,-ncol(P),]
N_grades <- nrow (P)
max_rep <- dim (P)[[3]]
W <- rep (0, N_grades*N_grades*max_rep*N_years)
dim (W) <- c (N_grades, N_grades, max_rep, N_years)
dimnames (W) <- list (dimnames (P)[[1]],dimnames (P)[[2]],dimnames (P)[[3]],NULL)
for (y in 1 : N_years)
{
  for (rep in 1 : max_rep)
  {
    for (g1 in 1 : (N_grades-1))
    {
      for (g2 in g1 : N_grades)
      {
        W [g1,g2,rep,y] <- prob_t (g1,g2,rep,y)
      }
    }
  }
}

W
}

```

## References

- Bartholomew D. (1982), *Stochastic models for social processes*, Wiley, Chichester
- Bernardi L., U. Trivellato (1980), “Un modello markoviano del processo scolastico: (II) suo impiego per l'analisi della selettività del sistema preuniversitario italiano”, *Rivista di Statistica Applicata*, 13 (2), 55-90
- Bernardi L., G. Lovison, U. Trivellato (1986), “Markovianità ed allargamento dello spazio degli stati in un modello stocastico del processo scolastico”, *Atti della XXXIII Riunione Scientifica S. I. S.*, Bari, I, 341-349
- Bernardi L., G. Ghellini (1997), “Sistema formativo e imprese: problemi e metodi per lo studio della transizione dei giovani al lavoro”, *Proceedings of the Intermediate SIS Conference “La Statistica per le Imprese”*, Torino, 513-528
- Bush R. R., C.F. Mosteller (1955), *Stochastic models for learning*, Wiley, New York
- Ghellini G. (1996), “Studio pilota sui flussi scolastici e stima dell'offerta di lavoro per titolo di studio e sesso”, *Flash Lavoro Quaderni*, 39, Regione Toscana, ORML
- Ghellini G. (1997), “Indagine sperimentale per lo studio dei percorsi scolastici e degli sbocchi professionali dei diplomati”, *Flash Lavoro Quaderni*, 50, Regione Toscana, ORML
- Ghellini G., A. D'Agostino, A. Mulas (1999), “Indagine longitudinale per lo studio dei percorsi scolastici e degli sbocchi professionali dei diplomati delle scuole medie superiori in Toscana”, *Primo Rapporto Intermedio*, ORML, Firenze
- Ghellini G., A. D'Agostino, A. Mulas (2000), “I percorsi scolastici lavorativi dei diplomati toscani: prime analisi sulla coorte 1995/96”, *Flash Lavoro Quaderni*, 78, Regione Toscana, ORML
- Kemeny J. G., J. L. Snell (1976), *Finite Markov chains*, Springer, New York
- Légaré J. (1972), “Methods for measuring school performance through cohort analysis”, *Demography*, 4, 617-624
- Leontieff W. (1951), *The Structure of the American Economy 1919-1939. An Empirical Application of Equilibrium Analysis*, Oxford University Press, New York



- Martelli C (1995), “Disponibilità e limiti dei dati ufficiali per l’analisi dei fenomeni demografici attraverso lo studio delle biografie”, *Proceedings of the SIS Conference “Continuità e Discontinuità nei Processi Demografici”*, Rubbettino, Soveria Mannelli (CZ), 379-386
- Nam C. B., E. Stockwell (1963), “Illustrative Tables of School Life”, *Journal of the American Statistical Association*, 58, 1113-1124
- Organisation for Economic Co-Operation and Development (2005), *Education at a Glance*, Centre for Educational Research and Innovation, OECD Indicators, Paris
- Organisation for Economic Co-Operation and Development (2000), *Investing in Education : analysis of the 1999 World Education Indicators*, OECD Publishing, Paris
- Quetelet A. (1849), *Letters on the theory of probabilities as applied to the moral and political sciences* (translated by O.G. Downes), C. & Layton, London
- Steindhl J. (1965), *Random processes and the growth of firms*, Griffin, London
- Stone R. (1965), “A model of educational system”, *Minerva*, 3, 177-186
- Stone R. (1966), “Input-output and demographic accounting: a tool for educational planning”, *Minerva*, 4, 365-380
- Stone R. (1971), *Demographic accounting and model-building*, OECD, Paris
- Stone R. (1972), “A markovian education model and other examples linking social behaviour to the economy”, *Journal of the Royal Statistical Society*, series A, 135, 511-543 (with discussion)
- Stone R. (1973), “A system of social matrices”, *Review of Income and Wealth*, 19, 143-169
- Thonstad T. (1969), *Education and Manpower: theoretical models and empirical applications*, Oliver & Boyd, Ltd., Edimburg
- Trivellato U. (1980), “Un modello markoviano del processo scolastico: (I) specificazione e procedimento di stima”, *Rivista di Statistica Applicata*, 1, 3-20
- Trivellato U., L. Bernardi (1994), “Scolarità e formazione professionale nel Mezzogiorno: nuove evidenze da un’analisi dei flussi”, *Economia & Lavoro*, 3-4
- Trivellato U., G. Ghellini, C. Martelli, A. Regoli (1995), “Prospettive per possibili analisi longitudinali nella statistica ufficiale italiana”, *Rapporto di ricerca per la Commissione di Garanzia sull’Informazione Statistica*, Presidenza del Consiglio dei Ministri, Roma
- Zaccarin S. (1994), “Indagini longitudinali sulla transizione scuola-lavoro e sull’entrata nella vita attiva”, *Economia & Lavoro*, 2, 27-46