Data Pre-processing and Data Generation in the Student Flow Case Study

Luís Cavique[1,2(✉)], Paulo Pombinho[2], Antonio J. Tallón-Ballesteros[3], and Luís Correia[2]

[1] Universidade Aberta, Lisbon, Portugal
luis.cavique@uab.pt
[2] MAS-BioISI, FCUL, Lisbon, Portugal
pmimatos@fc.ul.pt, luis.correia@ciencias.ulisboa.pt
[3] University of Huelva, Huelva, Spain
antonio.tallon.diesia@zimbra.uhu.es

Abstract. Education covers a range of sectors from kindergarten to higher education. In the education system, each grade has three possible outcomes: dropout, retention and pass to the next grade. In this work, we study the data from the Department of Statistics of Education and Science (DGEEC) of the Education Ministry. DGEEC maintains those outcomes for each school year, therefore, this study seeks a longitudinal view based on student flow. The document reports the data pre-processing, a stochastic model based on the pre-processed data and a data generation process that uses the previous model.

Keywords: Data pre-processing · Data generation · Student flow · Stochastic model

1 Introduction

Models of the student flow throughout the educational system can be very important for planning infrastructure resources, providing the distribution of human resources in the labor market, and identifying problems across the educational system, as well as to propose specific actions to overcome them.

In a broader definition the student flow is characterized by a number of features that changes over time. Two time-variables should be distinguished: curricular year (or grade), and the school (or academic) year. The school year is usually made up of a couple of consecutive years, for instance 2019–2020. To simplify the nomenclature, sometimes in this work we use the year of enrollment, so for the given example the value of 2019.

In the grade-by-grade student flow models, each grade has three possible outcomes: dropout, retention and pass to the next grade. The cohort definition is a group of individuals who share specific characteristics, usually in the same period, such as year of birth. Some studies use the concept of cohort and calculate the survival rate of the cohort.

The goal of this paper is to present the student flow approach in an educational system, by detailing the model and proposing techniques. This work is part of a wider project, the student flow modeling in the Portuguese educational system, with the acronym ModEst.

C. Analide et al. (Eds.): IDEAL 2020, LNCS 12490, pp. 35–43, 2020.
https://doi.org/10.1007/978-3-030-62365-4_4

In our study, the Public Administration organism is the Department of Statistics of Education and Science (DGEEC), of the Portuguese Education Ministry. DGEEC has a vast amount of data regarding 2 million students per year on the Portuguese school system, from pre-scholar to doctoral programs, with annual in/out flow of more than 0.5 million students. The data provided by DGEEC have been anonymized in order to meet the requirements of the General Data Protection Regulation (GDPR).

In this work, we only used the data from the 1st to the 12th grade, leaving the study of higher education for future work. A stochastic model, based on the Markov chain process, is applied in order to generate new data volumes.

The procedure used in this work is developed in R language and can be summarized in three steps: data pre-processing, modeling and data generation. In step (i) the data is pre-processed and the structure of the problem is formalized, in Step (ii) a stochastic model is generated, based on the pre-processed data, that explains the past, and in step (iii), data generation in the prediction phase, we forecast the future number of students in the system. The proposed procedure can be presented in the following data pipeline: data pre-processing → stochastic model → data generation.

The paper is organized in five additional sections. In Sect. 2, background information is presented. Section 3 details the data pre-processing for the student flow case study. The stochastic model is introduced in Sect. 4. Section 5 reports the data generation and the prediction of the number of students. Finally, in Sect. 6 conclusions are drawn.

2 Background Information

2.1 Student Flow

Education covers a range of sectors from kindergarten, basic (1st to 9th grade) and secondary school (10th to 12th grade, referred as High School in some countries), to higher education, from level 0 to level 8 of the International Standard Classification of Education, ISCED 2011 classification [4].

The education sector can be seen as a series of components where each student follows a pathway, which meets his/her own aspirations.

Many examples of the application of mathematical models to education planning exist since the late 1960s [7]. In this state-of-the-art in student flow problem, we identified different approaches such as Key Performance Indicators (KPI), Visualization, Markov models and What-if simulation. These approaches are ordered by the usual sequence of data analysis.

In Science Education two KPI are usually established in student flow: dropout and failure or retention [5]. The knowledge and prediction of dropout and retention in the student flow is highly relevant since these KPI directly influence the performance of the education system.

The visual representation of student's march to graduation can be analogous to Napoleon's march to Moscow created by Charles J. Minard in late 1800 s [3]. The visual-analytics capabilities are relevant for two reasons. First, these tools allow a rapid exploration of large data sets, providing the ability to easily detect trends and anomalies in data related to student progress. Second, the ability to supply visual proof of student progress grounded in facts, rather than speculation or supposition.

In business intelligence, what-if analysis fills this gap between data mining and decision making, by enabling users to simulate and inspect the behavior of a complex system under some given hypotheses. What-if dynamic simulation model allows to analyze the impact of potential changes in core curriculum policy, prerequisite structure, and staffing capacity to be tested prior to implementation [8]. Discrete Event Simulation is more adaptable to real-world applications than pure a Markov model. It accommodates more easily the complexities and interdependencies of the many components involved in the system [6].

2.2 Markov Chains

Stochastic processes and, in particular, Markov chains are very useful tools to deal with uncertainty [1, 2].

A stochastic process consists of a set of indexed random variables X(t), for example the number of students at the beginning of the school year *t*. The time parameter *t* can take discrete values, $t = \{0, 1, 2,\ldots, n\}$, or take continuous values, i.e. $t \geq 0$. Variable X can assume a set of states, $S = \{1, 2,\ldots, m\}$. A stochastic process is said to have a finite state space, if |S| is finite, and the stochastic process has a space of continuous states, otherwise.

The Markov chain is a special case of stochastic process, with the following propriety: if the process is currently in state *i*, then it will occupy state *j* in the next period with probability P(i, j), i.e. $\text{Prob}(s_{t+1} = j \mid s_t = i) = P(i, j)$, where P(i, j) is a parameter fixed for each pair of states $(i, j) \in S \times S$.

An important property in Markov chains is the stationarity of the process. The stochastic process X is said to be stationary when its behavior is independent of time and the system is in a steady state. In a stationary process, as *t* becomes very large, X^t converges to the probability vector X such that $X = X.P$.

3 Data Pre-processing

The available dataset contains the enrollment information of pre-primary, primary, secondary and post-secondary education levels, from level 0 to 5 of ISCED 2011 [4], of a period of 10 school years, from 2008–2009 to 2017–2018. The dataset deals with around 17,000,000 instances, which corresponds to an average of 1,700,000 students per school year, which supports several enrolments of around 5,500,000 individual students during 10 school years.

In the data pre-processing the ETL (extract, transform and loading) process was applied to the dataset. During the cleaning phase the classifications were uniformed, the duplicate student registrations were removed, and missing data were also removed.

In this section, firstly, we detail possible transitions from a generic grade in order to find the outcomes of the student flow. And secondly, we identify the most relevant dimensions of DGEEC dataset.

To better explain the developed work, we introduce the concept in three steps: the data collection, where the data is prepared and the attributes selected, the selection of dropout and loyalty rules with a decision tree algorithm, and the creation of loyalty actions to prevent dropout.

3.1 Grade Transitions

The sequence of grades can be seen as a sequence of states in a state transition system with specific transitions, as shown in Fig. 1. The three possible outcomes, or output transitions, are dropout, retention or pass to the next grade. On the other hand, the input transitions are retention in the same grade, pass from the previous grade and ingress of new students coming from outside the system.

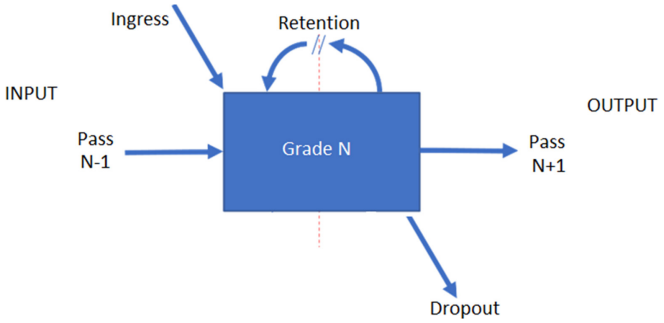

Fig. 1. A state (concerning a grade) with input and output state transitions

In the ModEst project, given the anonymous identification of the student, the school year, and his current grade, we can identify his transition state:

- Dropout: a student drops out if there is no information about him/her in the next school year;
- Retention: in the following school year, a student is retained if he/she is in the same grade;
- Pass: in the following school year, a student passes if he/she is enrolled in a higher grade;
- Ingress: a student ingresses if there is no information about him/her in the previous school year.

The information of the outcome of each student can be easily implement in SQL using sub-queries, as show in Fig. 2.

3.2 Data Dimensions

For each enrolled student, there is a set of information associated that can be classified into dimensions. The fact table of ModEst Project has a set of six dimensions which includes:

- School year of enrollment (e.g. 2008–2009);
- Curricular year (or grade) in which the student enrolls (e.g. 10^{th} grade);
- Modality of the course (e.g. regular or professional education);
- Nature of course (e.g. public or private education);

```
-- Update Dropout
UPDATE factTable AS F1
SET outcome = "dropout"
WHERE F1.student NOT IN (SELECT F2.student
                         FROM factTable AS F2
                         WHERE F1.schoolYear+1 = F2.schoolYear);
-- Update Retention
UPDATE factTable AS F1
SET outcome = "retention"
WHERE EXISTS (SELECT *
              FROM  factTable AS F2
              WHERE F1.student = F2.student
              AND F1.schoolYear+1 = F2.schoolYear
              AND F1.grade = F2.grade);
-- Update Pass
UPDATE factTable AS F1
SET outcome = "pass"
WHERE EXISTS  (SELECT *
               FROM factTable AS F2
               WHERE F1.student = F2.student
               AND F1.grade+1 = F2.grade
               AND F1.schoolYear+1 = F2.schoolYear);
```

Fig. 2. SQL code to update outcome information of each student

- Geographic information at the level of 'nomenclature of territorial units for statistics' NUTS-II and NUTS-III (e.g. Lisbon Metropolitan Area);
- Outcome of the enrollment: dropout, retention or passing grade.

The student data was anonymized with a specific identifier. Each line of the fact table corresponds to a student enrolled in only one school year, with the additional information of the grade, the course Modality, the course Nature, the NUTS and the outcome. The education level (primary education and secondary education) aggregates the dimensions of grade and course modality. Fact table and respective dimensions are shown in Fig. 3.

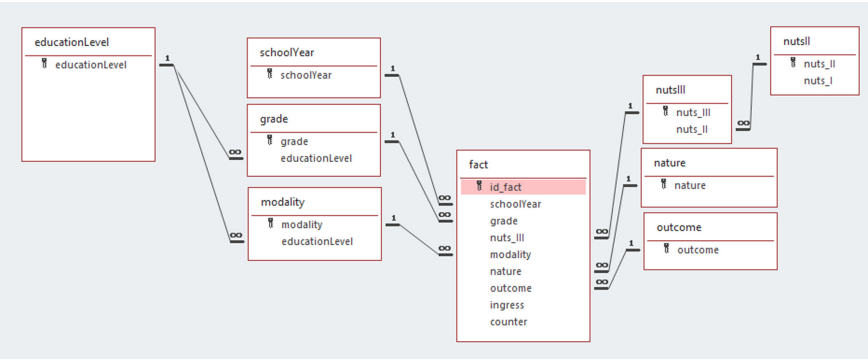

Fig. 3. Fact table and student dimensions

In the fact table the number of ingress can also be found. For each combination of the dimensions we have a counter with the number of students in a fact table with 85,000 instances.

As it is usual in projects with real data, data cleaning procedures were performed. In particular, the dimension grade was normalized from 0 (kindergarten) to 13th (post-secondary) for the different course modalities.

4 Stochastic Model

As previously defined, in a stochastic model with *m* states and *n* time periods, the variable $X = [x_1, x_2, \ldots, x_m]$ in the time period $n + 1$ is given by $X^{n+1} = X^n.P$. In this section, we define the structure of the transition matrix with the respective *m* states.

Given the number of students, for each school year, and for each grade, we obtain the number of ingresses, dropouts, failures and transitions, as the sample shown in Table 1. With this we aim to define the structure of the transition matrix.

Table 1. Sample of the input data of the transition matrix

Year	In/outcome	Grade1	Grade2	Grade3	Grade4
20XX	ingress	98,627	10,382	10,382	10,382
20XX	dropout	7,135	10,112	9,823	11,466
20XX	retention	1,019	6,741	3,274	4,586
20XX	transit	93,774	95,499	96,043	98,605

In our study the student flow presents an open-loop system, with an input flow, corresponding to the ingresses of new students, and two output flows, which correspond to dropout students and completion of studies.

Figure 4 shows the transition diagram of the open-loop Markov chain, of a sequence of school years in the years 2000, with four grades. The state ingress, I, is larger in the 1st grade and reduced in the remaining grades. Grades 1 to 4 have three possible outcomes: going to the next grade (pass), remaining in the same grade (retention) or leaving the system (dropout). The students that pass in the 4th grade go to the absorbing state named 5th grade. The dropout state, D, is also an absorbing state, since once entered on that state, it cannot be left. In all the lines, *i*, of the transition matrix, $\sum P(i, j) = 1$ should be verified.

Generalizing for any number of grades, the number of states, *m*, is equal to the number of grades plus 3 (ingress, conclusion and dropout). The transition matrix allows to identify a set of transitions (dropout, retention, transit) for each school year.

To find the P matrices of the last two years, an iterative process is used, which varies the fitting methods that capture the trend of variables dropout, retention, transition and ingress, of the school years from 2008 to 2016. Then by combining the information of the two P matrices, the transition matrix that best predicts the last two years is found.

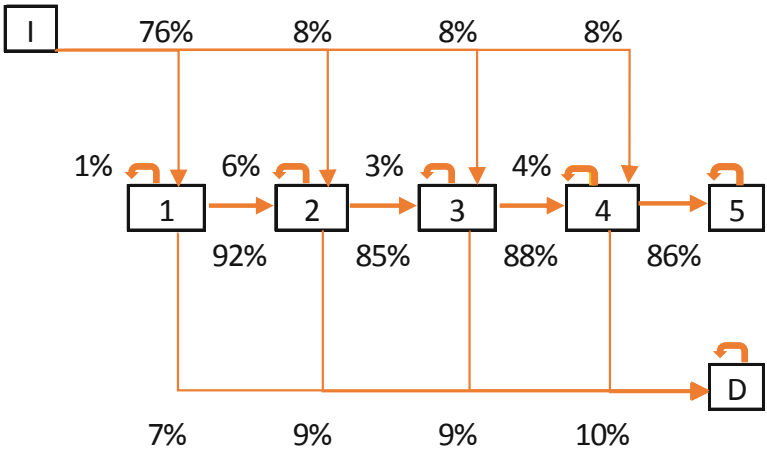

Fig. 4. Transition diagram

Using Y to express the prediction and X the real data, the prediction for the year n + 1 is given by $Y^{n+1} = X^n.P$. The error of the prediction, for a given year *n*, is calculated by the expression $error^n = \sum(abs(X^n - Y^n))/\sum X^n$, where Y^n corresponds to the prediction and X^n to the real data. In this study, the following error percentages were found:

- to predict 2015, error2015 = 4.93%,
- to predict 2016, error2016 = 6.22%.

The error of the predictions of the school years 2015 and 2016 is about 5%, i.e., ensuring an average accuracy of 95%.

5 Data Generation and Prediction

As already mentioned, the prediction of the next time period is given by $X^{n+1} = X^n.P$ where P is the probability matrix described in the previous section. Based on the previous stochastic model, data can be generated in order to predict the number of students. In this section we develop a what-if approach, with four scenarios, to predict X in the year 2022, i.e., X^{2022}.

The stochastic variable X must be compatible with P, having the following structure X = [ingress, grade1, grade2,…, grade12, conclusion, dropout]. The number of ingresses is a new variable that changes each year, and should be inserted into the stochastic model. The other variables of X are given by the previous year.

The reduction of the moving average of the number of ingresses in the last 6 school years varies from 10% to almost 30%. In this work, we create four conservative scenarios of the number of ingresses with the reduction of 0%, 5%, 10% and 15%.

In the procedure to predict the number of students, the input is data from school year 2016, X^{2016}, the transition matrix P, and a new vector Ingress with the information of

the number of students that ingress in the system. The output of the procedure is a set of vectors $\{X^i,\ldots, X^{i+k}\}$ with the information about each grade, conclusion and dropout.

The procedure should run for each scenario with the respective reduction of the number of ingresses. This what-if system allows us to find the impact of the scenarios in the number of students in 2022, as shown in Fig. 5. The reduction of 0%, 5%, 10% and 15% in the ingress, corresponds in 2022, of 1.02, 0.97, 0.93, 0.89 million students, respectively.

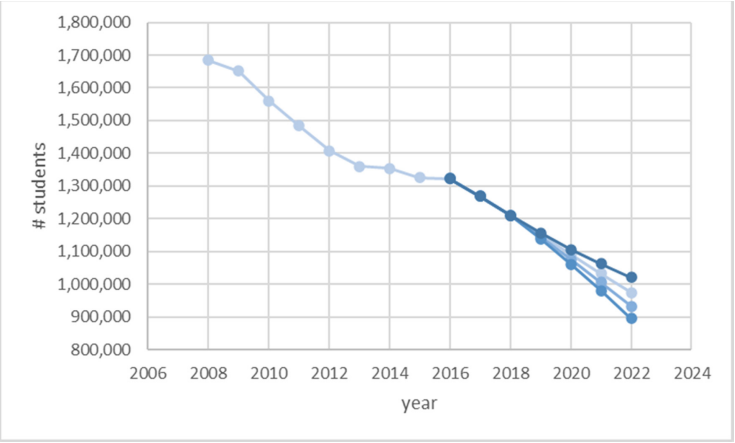

Fig. 5. Impact of the scenarios in the number of students in 2022

6 Conclusions and Future Work

In this work, we focused on the data from the 1st to the 12th grade, provided by the DGEEC dataset. The document reports the data pre-processing and the stochastic model that allows the data generation.

Data pre-processing included the ETL, the definition of the grade/state transition and the data warehouse with one fact table and six dimensions.

The stochastic model includes the formulation of the problem and the model tuning: (i) formulation: define *m* and the transition matrix P for each school year; (ii) stochastic modelling: given the data of time periods n from 2008 to 2016, find the transition matrix P that best predicts the last two years. In the model tuning we obtained an accuracy around 95% for the years 2015 and 2016.

Data was generated in order to predict the number of students, from 2017 to 2022. In the prediction phase, three of four scenarios indicate that the number of students from grade 1st to 12th, in the enrolment year of 2022, would be less than one million.

Two major contributions can be highlighted related to the student flow case study. The first one corresponds to the dichotomy between the annual surveys and the view of the data flow, i.e., DGEEC validates the annual data collected from the schools and ModEst considers a longitudinal view of the student flow obtained in the pre-processing phase.

Given the stochastic model, the second contribution concerns the variety of possibilities of data generation, in order to overcome the limitations imposed by just 10 years of real data.

Since this work is part of a wider project, for future work we point out the following topics:

– to apply the tools developed to predict the number of students using filters on the dimensions (modality, nature, NUTS);
– to visualize the student flow with cohorts using Sankey diagrams, among others.

Acknowledgments. The authors would like to thank the FCT Projects of Scientific Research and Technological Development in Data Science and Artificial Intelligence in Public Administration, 2018–2022 (DSAIPA/DS/0039/2018), for its support. LCav, PP and LCor also acknowledge support by UID/MULTI/04046/2103 center grant from FCT, Portugal (to BioISI).

References

1. Bronson, R., Naadimuthu, G.: Schaum's Outline of Operations Research, 2nd edn. McGraw-Hill, New York (1997)
2. Grinstead, C.M., Snell, J.L.: Introduction to probability. American Mathematical Society, Providence (1997)
3. Heileman, G.L., Babbitt, T.H., Abdallah, C.T.: Visualizing student flows: busting myths about student movement and success, change. Mag. High. Learn. **47**(3), 30–39 (2015)
4. ISCED: International Standard Classification of Education. UNESCO Institute for Statistics, Montreal, Quebec, Canada (2011)
5. Junior, P.L., Silveira, F.L., Ostermann, F.: Survival analysis applied to student ow in undergraduate physics courses: an example from a Brazilian university. Revista Brasileira de Ensino de Física **34**(1), 1403 (2012)
6. Fiallos A.X.: Ochoa Discrete event simulation for student flow in academic study periods, Twelfth Latin American Conference on Learning Technologies (LACLO). IEEE, Argentina (2017)
7. Lovell, C.C.: Student Flow Models: A Review and Conceptualization. National Center for Education Research and Development, Washington, D.C. (1971)
8. Saltzman, R.M., Roeder, T.M.: Simulating student flow through a college of business for policy and structural change analysis. J. Oper. Res. Soc. **63**(4), 511–523 (2012)