Degree project in application mode

A discrete event simulation of students’ flow through an undergraduate academic program.

Cristian Ferney Hortúa Ortiz a,c, Diego Alejandro Pardo Suárez a,c, Laura Vanessa Prieto Galindo a,c, David Uribe Mora a,c.

Ricardo Fernando Otero Caicedo b,c

a Student of Industrial Engineering
b Professor, Director of the Degree Project, Department of Industrial Engineering
c Pontificia Javeriana University, Bogotá, Colombia

Abstract

For educational institutions, students’ flow is the phenomenon which represents all movements of students throughout the subjects of a specific curriculum. Such flow is a complex phenomenon in which variables such as: choice of subjects, failure of subjects and desertion affect such flow; in addition, all these variables depend on each student specific characteristics and their progress in the career.

Due to the need of Universities to measure and understand indicators related to: curriculum completion time, courses demand and bottlenecks, in this project we propose to build a simulation model that represents this phenomenon in the Industrial Engineering career at the Pontificia Universidad Javeriana. There will be a simulation model of discrete events, where students will have characteristics of agents to both influence and be influenced by the system. It is expected that the simulation model promotes a more informed decision-making process.

Key words: Students flow; discrete simulation; Markov chains.

1. Introduction

The role played by universities positively affects social development because it allows improving educational level of individuals, making them better trained to face current problems. Therefore, it is essential that decisions of educational establishments are oriented to the adequate planning, organization and structuring of academic programs, as well as the physical and human resources. Management of these resources impacts on the quality of education, because it guarantees the correct functioning of all the processes and tools that are necessary for the correct learning of students, thus improving their life project and opportunities of growth.

Within management of resources, decisions must be made about the curriculum or mesh. The curricular grid describes an academic program, its structure, its content, activities and results (Schellekens, Paas, Verbraeck, & Van Merriënboer, 2010). Currently, several universities manage academic programs based on credits, where students register the available subjects for each academic period, considering course requirement and the maximum enrollment capacity of student. This curricular flexibility transfers to students’ the responsibility of enrollment, adapting its academic intensity according to personal criteria and decisions throughout the academic program.

In general, academic faculties are responsible for the correct planning of the quotas of each subject for students to enroll in their academic programs. In this sense, although there is control over the courses available in each academic period, there is no control over the various movements between subjects of students from one period to another. Understanding how students flow through the curriculum would allow faculties to improve decision
making. In the first place, it allows a better planning of the resources associated with the number of courses available for each subject. Also, it allows identifying the patterns that students follow when they move through the curriculum (McFarland, 2006).

This flow is conditioned by the profiles of students, which contemplate their academic performance and progress in the curriculum. This performance is affected by different social, family and personal variables. Therefore, although not all students have the same academic performance, they share similarities that allow them to be classified in different profiles and to model their progress over time (Reamer, Ivy, Vila-Parrish, & Young, 2015).

Furthermore, the categorization of students has been addressed by different authors. For example, (Barger, Wormington, Huettel, & Linnenbrink-Garcia, 2016) mentions that students have different levels of motivation, different attitudes of learning and very different responses to certain academic environments. Thus, student profiles are a set of variables associated with academic, social, personal and behavioral criteria, among others, that allow to know more closely the way of thinking, acting and interacting based on the common or individual characteristics of students (Felder & Brent, 2005). In turn, these different situations and motivations can cause students to change from one category to another. According to (Gillet, Morin, & Reeve, 2017), the behavior, the way of thinking and the decisions of students are modified according to the objectives that are set in order to achieve specific activities and achievements.

Besides that, desertion is a situation that a student faces when he or she must or wants to leave the academic program. According to (Suárez-Montes & Díaz-Subieta, 2015), this is affected by factors associated with socioeconomic, individual, institutional and academic conditions. The understanding of dropout, as a parameter, is important to determine future movements of students through the academic program with more veracity. This allows knowing those students who continue and those who retire in each period. In addition, it facilitates decision-making to improve the quality of the program, as part of the improvement plans for the institutions and its operational plans (López, Posada, Cardozo, & Cuartas, 2010).

Finally, the selection of subjects by students in the different semesters is a fundamental part for studying students’ flow. Understanding the logic and criteria that students use to select their courses each semester helps pinpoint the number of students who are interested in each of the program courses. However, this phenomenon is susceptible to great variability because decisions not only vary among students but also can change over time. Additionally, each student prioritizes different factors when enrolling subjects seeking to complete their curriculum. The dynamics associated with decision-making as opposed to the choice of subjects, represent a matter of transcendental interest for universities since they must be capable of providing an efficient service (Fiallos & Ochoa, 2017).

For universities, providing quality education and having processes of continuous improvement for their policies and guidelines is essential to enable students’ academic programs more suited to their needs and to meet the requirement and rigor necessary, projected by faculties. Thereby, it is necessary to continuously measure and improve indicators such as the subjects’ demand in the different students’ categories, the time spent in the program, the number of graduates and the number of people who fail each subject. However, these indicators depend on the guidelines established for each study plan, such as the precedence of the subjects, the number of subjects or credits of the program, the probability of approving the subjects in the different categories of students, the number of people entering the academic program and the dropout rate of the program. These affects, both individually and collectively behavior and dynamics of student flow. Therefore, it is necessary to understand and analyze how the variation of these parameters could impact students’ flow to have accurate and grounded indicators that help in the decision-making of each faculty.

The modeling of the flow can be done through a simulation model. This has the advantage of allowing to carry out experiments that could not be done in real life, it is also possible to have control over experimentally complex factors (Fiallos & Ochoa, 2017). The simulation to be performed is discrete with an agent-based component. That is because the model design is centered on the curricular process. Hence, this system has pre-established rules however, students adopt characteristics of agents because they make decisions and have an objective related to their curriculum, they are affected, they affect, and they adapt to the system.

As mentioned in the previous paragraph, simulation allows taking control of factors that experimentally could not be manipulated. Being able to take control over the parameters of the simulation makes it possible to create
several scenarios that allow evaluating the effect of the curriculum structure on the variables of interest related to students’ flow. In this case, the simulation could give indications to questions such as: When modifying the structure of requirements, how does the average time of permanence in the program change? If the rates of loss of subjects are modified, how is the demand for the courses affected? Among other questions that could contribute to the decision-making process of the faculties.

Given the above, the purpose of this project is to simulate the students’ flow and their behavior within the curriculum. The data is collected from the Industrial Engineering career of the Pontificia Universidad Javeriana, considering the various variables that are linked to desertion, students’ categorization and students’ criteria for decision making.

2. Literature Review

In literature several authors have addressed relevant aspects of students’ flow, using different methodologies and tools to represent them.

For example, (Valle et al., 2015) establishes that academic profiles are focused on motivational aspects based on the combination of academic goals and expectations of self-efficiency. This interaction and predominance of certain characteristics predict student’s behavior. While (Reamer et al., 2015) categorize students through academic profiles according to their results. For both studies, Markov chains were used as a quantitative tool to represent this factor and its impact.

Another relevant aspect mentioned by some authors is university dropouts. (Spady, 1971) states that the predominant reason that leads to academic wear and ultimately to desertion is centered on academic performance. Although dropout has been addressed with sociodemographic factors, (Otero, Bolívar, & Palacios, 2016) they emphasize that the degree relevance of the subject’s complexity has not been studied yet. Thus, these authors focus on measuring academic desertion according to the number of times a student enrolls in a subject and fails it, providing a quantitative approach through Markov chains to measure the impact of this factor on dropping out of university. On the other hand, (Rojas Naranjo, Vanegas Ardilla, Parada Pinzón, & Otero Caicedo, 2017) used the supervised learning algorithm Support vector machine (SVM) in order to classify students into two types of categories, deserters and not deserters for each academic period.

Another factor that affects the students’ flow is subject selection through the academic program. Some authors have modeled this selection from two points of view: a probabilistic approach where (Fiallos & Ochoa, 2017) made a model to determine the enrollment decision using historical information and factors such as the current academic period of the student, as well as the period where the subject should be enrolled and if this subject has precedence. On the other hand (A. G. Parameswaran, Koutrika, Bercovitz, & Garcia-Molina, 2010) propose a data mining model with precedence which estimates the probability that a user must consume a product in the future given its past behavior. However, it is not natural that subject selection has a random component on the basis that students select their courses thinking of a specific objective, where the fastest completion of the academic program is sought.

Another approach is optimization, where the individual scores for each subject, obtained by the existing recommendation algorithms, add restrictions and requirements to search for suggestions. (A. Parameswaran, Venetis, & Garcia-Molina, 2011). Similarly, (Robayo, Cote, & Otero, 2016), conducted a study to determine the decision criteria of students at the Pontificia Universidad Javeriana. They found that they make rational decisions under certain conditions, such as avoiding assigning certain subjects simultaneously in the same academic period. Therefore, a heuristic was developed based on the Bin Packing Problem with precedence’s to represent the decision-making process of students. This algorithm aims to minimize the number of periods needed to complete the academic program.

Finally, the main issue to be addressed is students’ flow, this problem has been previously worked in literature by some authors.

For example, (Shah & Burke, 1999), used Markov chains to model students’ flow through Australian higher education in order to obtain the average time students spend in the system. (McFarland, 2006) also used Markov chains to determine the critical path in the academic trajectory, as well as the periods that comprise it,
considering curricular movements such as abandonment, academic interruption and subjects’ failure. In addition, (Nicholls, 2007) focused on determining through Markov chains the different probabilities of finishing the doctorate and master’s degree, in a business university in Australia, according to academic load.

On the other hand, (Saltzman & Roeder, 2012), addressed the problem of students’ flow with a different methodology. They used a discrete simulation to calculate the students’ flow in the business program of a public university in San Francisco, considering factors such as blockages, withdrawals and utilization rates. This simulation includes limitations of resources, pre-requisites and entry parameters such as the total number of students in the program, the number of courses included in the syllabus, among others. While the most recent study was conducted (Fiallos & Ochoa, 2017) through a simulation, in order to predict the average time a student finishes the degree.

It is important to notice that certain authors have recognized the superiority of simulation over other methodologies to address the problem of student flow. For example, (Johnes, 2015) mentions different procedures, highlighting simulation as a tool that has proven to be useful for evaluating the effects of different factors such as changes in the curricular grid, changes in the allocation of resources due to budget cuts and control of how many students can enter the system per academic period.

Beneath, there is a summary table that allows identifying the main characteristics and methodologies used in articles related to students’ flow, as well as which indicators they used to evaluate their results.

<table>
<thead>
<tr>
<th>Article</th>
<th>Technique</th>
<th>Indicators</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Shah &amp; Burke, 1999)</td>
<td>Markov chains</td>
<td>Transition probability between states (semesters), average time in the system.</td>
</tr>
<tr>
<td>(Nicholls, 2007)</td>
<td>Markov chains</td>
<td>Total probability of reaching the different absorbing states: desertion and completion.</td>
</tr>
<tr>
<td>(Saltzman &amp; Roeder, 2012)</td>
<td>Discrete simulation</td>
<td>Real time spent in years studying the academic program compared to simulated time. Real graduation rate of students compared to the simulated rate.</td>
</tr>
<tr>
<td>(Fiallos &amp; Ochoa, 2017)</td>
<td>Discrete simulation</td>
<td>Percentage variation between the actual dropout rate and the simulated rate. Percentage variation between the real rate of records in the courses per semester and the simulated rate.</td>
</tr>
</tbody>
</table>

Table 1: Flow selection. Own authorship

In conclusion, between the different studies, those factors that proved to have a high impact on students’ flow were chosen. Among the most relevant factors are: desertion with the study of (Rojas Naranjo et al., 2017) and the criteria for subject selection, being that of (Robayo et al., 2016) the most pertinent model for the process of student’s decision making when selecting subjects. In the same way, the categorization of the students is considered as support to determine their behavior in the flow. For this, the study carried out by (Reamer et al., 2015) is considered, which not only classifies the students according to academic aspects, but also determines the probabilities of jumping between categories over time. Finally, a simulation will be developed, which contains the aforementioned factors, which allows to represent students flow over time.
3. Objectives

3.1. Main Objective.

Design an application that simulates students’ flow through the curriculum subjects of the Industrial Engineering program at Pontificia Universidad Javeriana.

3.2. Specific Objectives.

- Profile students in different categories depending on their academic characteristics.
- Use and adapt a heuristic that represents the process of subject selection by the students in the academic periods.
- Design a simulation application that integrates desertion and categorization factors to represent students’ flow in the curriculum.
- Conduct a sensitivity analysis using model parameters to measure their impact on indicators such as average time to finish the academic program, subjects demand and jumps between categories.
- Perform a descriptive analysis by comparing real life data against simulation results in order to estimate model accuracy.

3.3. Design statement.

Design an application which, supported by Industrial Engineering databases of the Pontificia Universidad Javeriana, processes a simulation model in such a way that it represents students’ flow throughout the curriculum.

3.4. Expected design requirements.

The quality of the design will be evaluated by comparing the results obtained by the simulation model with the historical data of the faculty. For this, the application must be able to:

- Use student categories developed for the model.
- Include a model with desertion as a possible exit of the system within the model.
- Represent the flow of students by using real life data to train the model.
- Calculate indicators in real time such as subjects demand, graduation time by categories and drop-out rates per semester.
- Present the indicators clearly and precisely.
- Estimate, according to academic categories how they affect students’ grades.

The validation methodology will be explained below, and it consists of using only a couple semesters of information from the databases to train all the algorithms and compare their results with the information of the semesters that were not used. Markovian assumptions and distributions will be presented below as well.

3.5. Design constraints.

There are two constraints that can prevent further depth and accuracy of the findings in the project to be carried out:

- Quality in the database of the faculty that will be used as input data of the simulation model.
- Computational capacity of the equipment available to execute the model, since this could prevent recreating very robust scenarios with all the information that is handled.

3.6. Standards.

The project will base its quality standards on the ISO 9126 standard: (Sugawara & Nikaido, 2014). This standard will be used since it shows characteristics and guidelines for the quality use of software. For this, the standard
defines six quality characteristics: quality for the users’ needs, internal and external quality requirements, internal and external quality of the product and the quality of the project in use. ISO requirement fulfillment can be seen in annex 1

4. Case of study.

The project will be carried out in the career of Industrial Engineering at the Pontificia Universidad Javeriana, using information from the students of the databases in past academic periods.

5. Methodology

5.1. Simulation summary.

Simulation is divided in the following parts: simulation preparation, run of the simulation, simulation results, validation and test of scenarios.

5.1.1. Simulation preparation.

This part describes received databases filtering, debugged databases, algorithm training, data loading for real life students, Markovian assumptions and variable selections for students’ categorization.

Data loading and filtering: Data is taken from csv files into python data frames, in addition, it is also filtered and debugged. This process allows to create new databases and variables in order to feed the simulation running process, upload simulated information of real-life students to represent them and Markovian verification assumptions.

Different algorithms and methodologies were used throughout the simulation, each one of them have a specific process for variable selection and assumptions verification. Therefore, they will be explained on the section in which the role of the algorithm and methodology in the simulation is addressed.

5.1.2. Run of the simulation.

In this part, freshmen students are created, both real life and simulated students enroll subjects, their grades are estimated, and their status is updated. Python entire code is annex number 2

Students initialization: Real life student data is uploaded to the programed agents to represent real life students and their category jump, as well as new agents are created and categorized.

Subjects enrollment: At this stage students decide the number of credits and subjects to enroll on the following semester.

Subjects grades: Simulation uses students’ categories to assign a grade to each enrolled subject.

Update of student data: Student category, current career status, pending subjects, GPA (grade point average), among other variables are updated when a semester is finished.

5.1.3. Simulation results, validation and test of scenarios.

In this final section, simulation results are analyzed and contrasted with real life data. In addition, possible scenarios are created and tested.

Indicators estimation: Relevant indicators such as subjects’ demand, average time of graduation, dropout rates and mean absolute difference for the grades of students, are estimated.

Validation: During the validation process, indicators obtained throughout all replicates are contrasted with real life data to estimate how accurate the simulation is.
Simulation of possible scenarios: In the last stage, different scenarios that could happen in real life are simulated in order to obtain predictions of possible outcomes or values for the previously mentioned indicators.

Figure 1: Simulation pseudo-code. Own authorship.

5.2. Data loading and filtering

5.2.1. Databases description:

Three databases were selected, provided by the Faculty of Industrial Engineering, where academic, institutional and socio-demographic information of each student is detailed until the first semester of 2018. In order to facilitate the development of the project, a new database named “Data Simulation” was created with new variables, using the given data. In this new database, information was synthesized and adjusted. In the following table information of each database is detailed:

<table>
<thead>
<tr>
<th>Databases</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estados (Status)</td>
<td>It contains information regarding students’ status, both academic and sociodemographic.</td>
</tr>
<tr>
<td>Neos (Freshmen)</td>
<td>It contains freshmen’s information such as their current academic status, sociodemographic situation, scholarships, Icfes results, among others.</td>
</tr>
<tr>
<td>Notas (Grades)</td>
<td>It contains students’ grades throughout semesters.</td>
</tr>
<tr>
<td>Énfasis (Emphasis)</td>
<td>Contains information regarding emphasis selection from students who are close to graduate.</td>
</tr>
</tbody>
</table>

Table 2: Databases provided by faculty. Own authorship.

Furthermore, these databases were used both as initial parameters for the simulation and to obtain and create relevant data. A table with the variables created is attached

5.2.2. Data filtering

Databases had the following inconsistencies: missing data, mistaken courses codes and contradictory information between databases. Missing information was handled by either dropping student’s information or by estimating its value using mean function. When contradictory information among databases appeared, it was solved by using the database which showed more information regarding the searched value. For last, mistaken
courses codes were fixed manually because some subjects codes were changed across time and therefore, they must be compatible with the current ones.

5.3. Students initialization

5.3.1. Students creation

For this simulation two types of students were created. First, real life students which have initial parameters and their further flow through the curriculum is estimated. Second, non-existent students that enter the system as freshmen and which entire parameters and flow are simulated.

Real life students have a specific set of parameters that must be uploaded to the simulation. These parameters are approved subjects, grade point average, grades for each enrolled subject and the estimation of a specific choice regarding their emphasis and internship preferences. After uploading this data, students are classified, and their flow is simulated.

Simulated students are those who are created by the program and appear as freshmen. They are programmed with a specific initial category and chosen emphasis. The number of students created per semester depends on which it is first semester of the year or second one. After plotting data using a histogram, it was determined that the number of students who enter the program as freshmen follows a uniform distribution but with different parameters for first and second semester of the year. This probability function determines how many freshmen will enter the simulation.

5.3.2. Students Categorization

Categorization implies that there is at least one variable that a group of students share that affects their grade results in a way that differentiates them from other groups. Therefore, categorization allows to obtain more accurate estimations of students’ results and by extension, better graduated students indicators per semester, subjects’ demand estimations, and a better simulation in general terms.

In order to have a proper classification, three requirements had to be fulfilled in order to consider a variable as a factor. First, since the grades obtained by the students vary during their careers, both categorization and the variable responsible for it, must vary as well. Second, the variable must have a direct impact on grades results. Finally, it must be either implicit or explicit on databases. Despite other combination of variables that fitted previous conditions, the chosen variable for classification was grade point average per semester due to its correlation with grades.

Dispersion diagrams shows there is a linear correlation between grade point average per semester and grades. As an example, the subject’s differential calculus, algorithm thinking, and science of materials are shown below.

![Figure 2: Grade Point Average dispersion diagram for three core subjects. Own authorship.](image)

Three options for categorizing students were:

1. 5 categories using a hierarchical algorithm.
2. 6 categories using the following ranges: [0,1), [1,2), [2,3), [3,3.5), [3.5,4) and [4,5].
Initial categorization of real-life students and simulated ones had the following process. For real life students, grades are on databases and consequently their own grades decide their initial categories. For simulated students, a frequency table was built with data from real life students, which was used as a discrete uniform probability distribution for their initial categorization.

Since the grades of students depend on the category in which they are, categorization changes must not be determined by the grade point average on the simulation, otherwise an infinite cycle is created. Therefore, these changes were modeled using Markov chains, as it is shown further in the text.

5.4. Subjects enrollment

In the process of understanding and representing students’ flow, subject selection is core, because it represents student’s decision of when and what courses to enroll. Consequently, an accurate representation of subject selection allows the simulation to have a more accurate estimation of subjects’ demand. The goal is to find which and when to enroll subjects through the algorithm. In this simulation two assumptions were made regarding subject selection. First, students act rationally when enrolling subjects. Second, students’ only goal is to finish their career as soon as possible.

As it was previously mentioned, (Robayo et al., 2016) developed an heuristic that shows which subjects a student must enroll per semester in order to minimize the number of semesters to finish his or her career, giving specific conditions regarding students’ preferences both on subjects and number of credits to enroll.

This algorithm consists in splitting the career in the different decision semesters and determine for each one a set of possible subjects to enroll. This is determined with a precedencies binary matrix $P_{ij}$ which indicates whether the subject $i$ is precedence of subject $j$. If the sum of all $i$ subjects for each $j$ of precedencies matrix is equal to cero, the subject no longer has requisites, it means the subject can be added to this subset of possible subjects to enroll.

After that, the bin packing algorithm is applied to the semester which is currently being evaluated. This algorithm consists in maximize the relevance factor of the selected objects. In this case, those objects are subjects and the relevance factor indicates how important is a subject in order to minimize the number of semesters to graduate. This relevance factor is calculated with an algorithm that reviews the verticality and horizontality of the subjects that become available when approving the subject in question. This means that the algorithm checks how many subjects this subject opens and how many semesters are left to end the whole subjects set that come after this subject. At the same time, the binary variable $X_j$ indicates whether the subject

---

Figure 3: Subjects grades in category jumps sequence. Own authorship.
j is enrolled in the semester. This mathematical model is subject to an established capacity, which in this case is the number of credits to enroll obtained in the NCA algorithm later mentioned in this paper.

A similar algorithm is internally used by the simulation, however the parameters are determined by resampling databases and not by a user. In addition, (Robayo et al., 2016) algorithm calculates all the subjects to enroll until the end of the career meanwhile the algorithm used in this simulation only calculates a semester at a time. However it was proven that (Robayo et al., 2016) algorithm performance is the same as the algorithm adjusted in the simulation.

5.4.1. Subjects career enrollment

Students’ subject’s enrollment can be divided in two subproblems, how many credits should a student inscription and what subjects should be enrolled.

Regarding students’ preferences on how many credits to subscribe, the following steps are considered: First, the algorithm validates how many subjects a student can enroll in each semester, regardless the number of credits. To do this, three requirements are verified for each subject First, all subjects that are prerequisites must be approved. Second, fulfill any additional requirement, for instance, a minimum number of credits to be able to course the subject. Third, verify if the subject is necessary for its emphasis. The sum of credits for all the subjects that meet these requirements is the number of credits available.

Once the number of available credits is obtained, the preferences of students regarding the number of credits to effectively enroll is estimated, which is the second step of career subject’s enrollment. This estimation uses conditional probability, by estimating the number of credits a student enrolls, given that he or she has certain amount of available credits to enroll. After these probabilities are calculated for each possible number of available credits, a resampling process is used, having them as the filtered parameter for the resampling process.

Finally, once the amount of credits to be enrolled is obtained, the algorithm decides what subjects should be enrolled. Since (Robayo et al., 2016) algorithm considers a set of subjects that according to the preferences of students should not be enrolled on the same semester, the simulation must reflect these preferences as well. This was made by analyzing for each pair of subjects, of the previously mentioned set, the number of students that enroll both, against the total number of students that were able to do so. This frequency was applied to the subject selection algorithm to estimate if a student would have those preferences or not. Subjects obtained by this algorithm are the core subjects that the student enrolls.

5.4.2. Elective subjects’ enrollment

For elective subjects a similar process is performed, but unlike career ones, the number of credits enrolled is the only variable to be estimated, since the grades for elective subjects are not calculated.

In order to estimate the number of credits a conditional probability was also used. In this case, is the probability to enroll certain amount of credits for elective subjects given that the student enrolled certain amount of career subjects. Data was filtered using this probability and a resampling methodology is used as well to estimate the number of elective credits to be enroll.

5.5. Subjects grades

Once all agents have their subjects enrolled, simulation begins to estimate grades for every student and for each subject currently enrolled by it. The methodology used for this estimation is resampling selection. In it, grades of the specific subject being evaluated are filtered, considering only those on which students share category with the current agent. In case there are no students that have shared both subject and category, grades are obtained using the closest category as the resampling one. From this resampling process, a grade is chosen randomly. In addition, for each subject, only data grades of first enrollment of each student were used, in order to improve the accuracy of the failure rates.

5.6. Update student data
Student data is divided in two, student’s status and category. Student status refers to the university position towards the student, they are: active, expelled, graduated and deserted. On the other hand, a student’s category means the relation between the student and his or her grades.

5.6.1. Status Update

5.6.1.1. Academic Expulsion.

After obtaining all agents grades, simulation will determine which students get expelled for an insufficient score at their grade point average. Along with the university policies, there are two scenarios in which a student gets expelled for academic reasons. The first one is if a student is in probationary status, which means that its grade point average is equal or less than 3.4, for three semesters in a row. The second one is if a student has a grade point average below 2.5 unless the student is freshmen. Both conditions are considered in the simulation to expelled agents from the student flow.

5.6.1.2. Dropouts

According to the regulations of the Pontificia Universidad Javeriana, drop out is the situation in which a student does not renew his or her enrollment for one year or more and therefore, loses his or her student status on the respective program. Desertion is register on the databases, assigning specific status labels for drop out students, those labels are described furthermore. Dropout directly affects simulation on the variation of the desertion indicators in the curriculum. In addition, it indirectly affects subjects’ demand.

As it was mentioned before, dropout is a complex phenomenon in which a student decides to leave his or her career without finishing it. This decision is not random, because it involves both social and academic variables and consequently, it can be predicted as (Rojas Naranjo et al., 2017) mentioned. The methodology used is support vector machine algorithm, which can be trained to predict if a student will drop out its career due to academic reasons.

(Rojas Naranjo et al., 2017) support vector machine model was trained to predict if a student is going to drop out at the end of the next semester. For this model, a genetic algorithm was used to select which variables should be considered for training it, as well as three additional variables given the problems nature. Final variables were: age, number of failed subjects, linear regression slack for grade point average, number of enrolled core credits, career total number of credits enrolled, binary variable for student status regarding probationary status, binary variable for scholarship, grade point average for core subjects, maximum number of times for failing a subject without approving it and percentage of failed credits. For last, cross validation was used for tuning the algorithm parameters.

(Rojas Naranjo et al., 2017) support vector machine was adapted to the simulation using the same variables. Since the simulations needs to predict if a student will drop out at the end of the current semester and new data is used, a parameter tuning was performed through a grid search algorithm. It was found that every test had a better accuracy when using a Radial Basis Function kernel instead of a Linear one. On the other hand, to avoid overfitting a cross validation with several splits was used, in order to verify which was the model that better fitted data.

In order to train the support vector machine, it is necessary to do an initial validation on databases to determine the status of each student that is registered. The original databases involve the variable "student's status", associated with the current situation of each person enrolled in the curriculum. The values that the variable can take are: active in the program, canceled, expelled, interrupted, permission, finished program and temporary suspension. However, it is necessary to validate through the semesters if the previous condition, not enrolled for more than one year in the program, is met. SVM was trained with the dropouts found on the database.

If the SVM shows that a student is going to drop out, the simulation will remove all its academic information from the database of active students in the program and its information will be added to the database of program deserters. Once the information of the defector student has been withdrawn, the flow of that person will not be simulated, since the simulation only considers the students active in the program. The above represents the exit of the system, of the deserting students.
5.6.1.3. Graduate students

A student is considered graduated when it has enrolled and approved all core subjects, at least one emphasis and the total number of complementary and elective subjects assigned. After this, the agent is removed from the simulation and its data is stored separately.

5.6.1.4. Active students

If a student is not expelled from the university nor has finished all his subjects or deserted the academic program, it is considered an active student. Therefore, this agent will continue the simulation until it fulfills the conditions of one of the previously mentioned status.

5.6.2. Category update

The category on which a student is clustered could change over time. Therefore, categorization modeling must consider not only specific time periods but also how can these discrete jumps between semesters affect students clustering. By using discrete Markov chains (Ross, 2014), it was possible to model these category changes over time.

Let \( \{X_n, n = 0, 1, 2, \ldots, \} \) be a stochastic process that takes on a finite or countable number of possible values. If \( X_n = i \), then the process is said to be in state \( i \) at time \( n \).

We suppose that whenever the process is in state \( i \), there is a fixed probability \( P_{ij} \) that it will next be in state \( j \). That is, we suppose that

\[
P(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \ldots, X_1 = i_1, X_0 = i_0) = P_{ij}
\]

for all states \( i_0, i_1, \ldots, i_{n-1}, i \), \( j \) and all \( n \geq 0 \). Such a stochastic process is known as a Markov chain. Equation (4.1) may be interpreted as stating that, for a Markov chain, the conditional distribution of any future state \( X_{n+1} \), given the past states \( X_0, X_1, \ldots, X_{n-1} \), and the present state \( X_n \), is independent of the past states and depends only on the present state. The value \( P_{ij} \) represents the probability that the process will, when in state \( i \), next make a transition into state \( j \). Since probabilities are nonnegative and since the process must make a transition into some state, we have

\[
P_{ij} \geq 0 \quad i, j \geq 0 \quad \sum_{j=0}^{\infty} P_{ij} = 1 \quad i = 0, 1, \ldots
\]

In order to establish that changes in categorization can be modeled by using a discrete Markov chain two assumptions must be proven. First, probabilities of changing from state \( i \) to state \( j \) will remain the same through time. Second, as it was seen before, the probability of being in state \( j \) this semester will only depend of being in state \( i \) in the previous semester.

To prove independency between categories jumps and time periods, six contingency tables were created using a level of significance of 5%, one for each state. Each table considers one of the states as the initial one, comparing the independence between the changes regarding that initial state to each possible final state, between all time periods. For these tests, all the inter-semester periods were not considered. Since initial results did not showed independency between all initial states and time periods, the oldest semesters were eliminated one by one until test results were consistent with a markovian process for all of them. Codes and databases that were used to prove markovian assumptions is annex 3. Final p-values were the following:

<table>
<thead>
<tr>
<th>Status/Criteria</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>P value</td>
<td>0.9999</td>
<td>0.9999</td>
<td>0.0641</td>
<td>0.0519</td>
<td>0.1963</td>
<td>0.4811</td>
</tr>
</tbody>
</table>

Table 3: Final independency results. Own authorship.
5.6.2.1. First Order Markov Chain.

To test whether the student category breaks are a first-order Markov chain, \( n - 1 \) contingency tables were created were \( n \) are the total number of semesters that are proven to be markov. States for each transition matrix are clusters 0, 1, 2, 3 and 4 and each one considers transitions from semester \( t \) to \( t + 1 \), for \( t < n \). Results showed that for all semesters, the current cluster depends on the previous one.

Results for order one markov chain appear in the following table:

<table>
<thead>
<tr>
<th>Status/Criteria</th>
<th>1630 to 1710</th>
<th>1710 to 1730</th>
<th>1730 to 1810</th>
</tr>
</thead>
<tbody>
<tr>
<td>P value</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

*Table 4: First order Markov chain test results. Own authorship.*

On the other hand, it is necessary to consider whether the category jumps follow a second order Markov chain. However, in the project these contingency tables were not estimated because there was not enough data for all of the possible transitions, given that some of them were practically impossible because of the rules regarding expulsion mentioned above.

5.6.2.2. Transition Matrix.

The first-order transition matrix is shown below:

<table>
<thead>
<tr>
<th>Cluster</th>
<th>0 [0,1)</th>
<th>1 [1,2)</th>
<th>2 [2,3)</th>
<th>3 [3, 3.5)</th>
<th>4 [3.5, 4)</th>
<th>5 [4, 5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 [0,1)</td>
<td>0.4286</td>
<td>0.1429</td>
<td>0.1429</td>
<td>0.2857</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>1 [1,2)</td>
<td>0.0930</td>
<td>0.2093</td>
<td>0.2326</td>
<td>0.2326</td>
<td>0.1395</td>
<td>0.0930</td>
</tr>
<tr>
<td>2 [2,3)</td>
<td>0.0079</td>
<td>0.0491</td>
<td>0.2936</td>
<td>0.3777</td>
<td>0.2252</td>
<td>0.0465</td>
</tr>
<tr>
<td>3 [3, 3.5)</td>
<td>0.0021</td>
<td>0.0072</td>
<td>0.1554</td>
<td>0.4240</td>
<td>0.3394</td>
<td>0.0719</td>
</tr>
<tr>
<td>4 [3.5, 4)</td>
<td>0.0008</td>
<td>0.0023</td>
<td>0.0544</td>
<td>0.2732</td>
<td>0.4607</td>
<td>0.2087</td>
</tr>
<tr>
<td>5 [4, 5)</td>
<td>0.0000</td>
<td>0.0020</td>
<td>0.0109</td>
<td>0.0631</td>
<td>0.2984</td>
<td>0.6256</td>
</tr>
</tbody>
</table>

*Table 5: First-order transition matrix. Own authorship.*

Where values in parentheses correspond to the range of the grades to which that state is equivalent. This probability matrix was used to simulate students change of category. Final category jump is chosen randomly.

Another approach for students’ grades behavior was to consider multiple transition matrices for each enrolled semester since it could reflect psychological aspects for students such as academic pressure and maturity. For instance, a matrix for freshmen, a different matrix for second semester students and so on. However, this methodology was discarded for two main reasons. First, there was insufficient data for certain semesters. Second, since a student can only be freshmen once, or been on a specific semester once, this behavior is more likely a conditional probability than a Markov chain.

5.7. Indicators estimation

The most relevant information of the simulation will be a descriptive analysis of the indicators and results, which are shown below.

1. Subjects demand: shows for each subject in the curriculum, how many students need to enroll it. In this case, it was calculated by adding up the number of students who enrolled each subject for each semester.

2. Time of graduation: shows the career mean time for students that are graduated. In this case, this was calculated by filtering databases in order to find all graduated students and computing their mean time in the university.
3. Number of dropouts: shows the number of students who abandoned the career. In this case, this was calculated by filtering databases in order to count all the students who dropped out. When students abandon their career due to academic reasons, the faculty interprets it as opportunities to improve the program.

4. MAD: shows the mean absolute deviation between the grades obtained in the simulation against the grades of real-life students that enroll each subject of the curriculum in a given academic period. The purpose of this indicator is to measure how accurate is the simulation in terms of estimating grades for each subject.

5. Average grades per subjects per categories: shows the grade means per subject obtained for each category created. In this case, this was calculated by filtering databases by category and subject in order to estimate the grade mean for each combination. This indicator is important since it shows the impact of categories, for each course, in the grades mean.

5.8. Validation

Validation is the stage on which the program indicators are faced against reality, in order to test simulation accuracy. Simulation accuracy was tested by contrasting real life data against the indicators previously found. In order to compare the indicators with the real-life data, databases semesters were divided into training semesters and test semesters, that way, simulation results were not fed with test data. Validation data is presented in annex 4

5.8.1. Subjects demand.

In order to validate subjects, demand the following formula was used as a measurement for the error between simulated results and databases information:

$$\text{Error} = \frac{|\text{Mean real life demand} - \text{Mean simulated demand}|}{\text{Mean real life demand}}$$

The three sources of error were found when validating this indicator. The most significant errors are:

1. Databases errors: There are some cases in which databases establish a subject demand for less than twenty students, which would be almost impossible. Other cases of this type of error are detected based on experience

2. Faculty choices: Faculty can affect the demand of certain subjects by either changing courses credits, prerequisites or by enrolling it or not to freshmen. Such cases either modifies students’ choices or even the number of active students on the program now to enroll it. These changes were not considered when filtering databases because if this had been the case, some subjects would not have had enough observations.

3. Students enrollment choices: One of the assumptions on the simulation program is that students minimize the number of semesters in the university. However, certain subjects that are part of the critical path are not always enrolled on the same semester by students due to personal reasons and therefore, affects simulation precision.

On average, the error of all subjects is approximately of 0.26. However, excluding subjects with the errors mentioned before, the mean error is around 0.0557. Therefore, subjects demand estimation can be considered accurate for the critical path.

After performing an analysis, all subjects that have an error equal or higher than 1 has a type one error. There are some cases of subjects affected by an error higher than 0.75 are affected by type errors one or two. Subjects with errors higher than 0.5 are usually affected by errors type two or three.

As a conclusion of the last four tables, even though subjects’ estimation for any given subject are relatively accurate, those subjects that did not had any of the previously mentioned problems, have a very high accuracy. Therefore, the accuracy of the simulation increases with the stability of the system and the quality of databases
5.8.2. **Time of graduation**

For this indicator the only period considered is 1810. The following table shows results for average time of graduation for each cluster:

<table>
<thead>
<tr>
<th>Cluster</th>
<th>5%</th>
<th>Median</th>
<th>95%</th>
<th>Min</th>
<th>Mean</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>11</td>
<td>13.7</td>
<td>10</td>
<td>11.66</td>
<td>14</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>11</td>
<td>14</td>
<td>9</td>
<td>11.53</td>
<td>17</td>
</tr>
<tr>
<td>4</td>
<td>9</td>
<td>11</td>
<td>14</td>
<td>9</td>
<td>11.17</td>
<td>19</td>
</tr>
<tr>
<td>5</td>
<td>9</td>
<td>11</td>
<td>13</td>
<td>9</td>
<td>10.77</td>
<td>19</td>
</tr>
</tbody>
</table>

*Table 6: Average time of graduation per cluster. Own authorship.*

As it was expected, lower clusters have a longer average time of graduation. Therefore, clustering methodology was accurate. Clusters zero and one never get graduated because they get expelled before graduating.

In real life, there were many changes in the curriculum and the complexity of subjects after 2015. Since there are no students that started after 2015 and finished their career, it was necessary to change some parameters of the simulation such as grades, subjects' complexity and precedencies in order to obtain a more accurate result. For example, the resampling methodology only used data available before 2015 and Optimization was no longer precedence of Processes Engineering. Then, general results of graduation time are shown below:

<table>
<thead>
<tr>
<th>Source</th>
<th>5%</th>
<th>Median</th>
<th>95%</th>
<th>Min</th>
<th>Mean</th>
<th>Max</th>
<th>Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real-life</td>
<td>10</td>
<td>11</td>
<td>14</td>
<td>9</td>
<td>11.23</td>
<td>17</td>
<td>1.23</td>
</tr>
<tr>
<td>Simulation</td>
<td>9</td>
<td>11</td>
<td>14</td>
<td>9</td>
<td>11.04</td>
<td>19</td>
<td>1.44</td>
</tr>
</tbody>
</table>

*Table 7: Descriptive statistic for real-life vs simulation time of graduation results. Own authorship.*

As shown in the table above, real life 5th percentile is ten semesters while simulation is nine semesters. Whereas real life 95th percentile are 14 semesters both in simulation and real life. However, mean time of graduation differs between real life and simulation with 11.23 and 11.04 semesters respectively. This means that simulation has a better performance in the subject’ enrollment decisions. At the same time, simulation does not take into account the complexity of subjects when enrolled in the same semester, so there is no reason for simulation to avoid enrolling subjects due to the difficulty of them.

5.8.3. **Number of dropouts**

Validation of dropout indicator follows two phases: first, the value of real drop-out per semester is calculated. To do so, database that records the student information is filtered. Faculty databases register students who voluntarily withdraw from the career with one of the following states: “Interrupción”, “Permiso” y “Suspensión temporal”; this was the filter applied. Therefore, the number of people who reflect this situation is counted and the total value is the real dropout of the academic period that is being analyzed. Second, the strategy used to contrast the real value against the simulation was to compute values of the 5th and the 95th percentile. The reason was to know the number of students who, at these percentiles, have deflected. In turn, the minimum, maximum, median and mean values were used to analyze simulation results.

In order to validate the accuracy of this indicator over time, the position and central tendency statisticians were created for the periods 2017 first semester, 2017 second semester and 2018 first semester. The results of this indicator show that, for all validated semesters, the actual dropout value is within the established percentiles. The results are shown below:

<table>
<thead>
<tr>
<th>Semester</th>
<th>Real</th>
<th>5%</th>
<th>Median</th>
<th>95%</th>
<th>Min</th>
<th>Mean</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>2017 - 1</td>
<td>31</td>
<td>28.90</td>
<td>38</td>
<td>49.55</td>
<td>27</td>
<td>38.20</td>
<td>50</td>
</tr>
</tbody>
</table>
Table 8: Results of desertion indicator per semester. Own authorship.

The real value of semester 2018-01 is close to the average value and the median value of the simulation; the real value of 2017-01 is below the average and the median of the simulation; and the real value of the semester 2017-03 is outside the percentile interval created, being below the position and central tendency statisticians.

5.8.4. Average grades per subjects per categories

Validation of student’s categorization is intrinsic and does not require a comparison with real-life data, since categorization is a strategy developed in the project to obtain students' grades more accurately. To validate the categorization, it must be considered that the lower cluster must have a lower grade point average than the upper cluster.

The following table shows the percentage in which this assumption is fulfilled among clusters for all the curricular subjects. Additionally, they also show the percentage of the number of subjects that meet the assumption for the five clusters through the periods 1710, 1730 and 1810.

<table>
<thead>
<tr>
<th>Semester</th>
<th>Cluster Grades Validation</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0 ≤ 1</td>
<td>1 ≤ 2</td>
</tr>
<tr>
<td>1710</td>
<td>79.31%</td>
<td>84.48%</td>
</tr>
<tr>
<td>1730</td>
<td>84.48%</td>
<td>82.76%</td>
</tr>
<tr>
<td>1810</td>
<td>77.59%</td>
<td>81.03%</td>
</tr>
</tbody>
</table>

Table 9: Percentage of categorization compliance for career subjects. Own authorship.

For the set of subjects belonging to the Industrial Engineering curriculum, only from cluster 4 to cluster 5 the established assumption is respected. Almost all clusters, and for the three validated semesters, are over 80% the compliance assumption between the lower cluster and the upper cluster. However, since there are more cases where the lower clusters have better grades than the higher ones, only 62.07% to 65.52% of the subjects fulfill the assumption perfectly.

The main reason of why it is more likely that the assumptions fail in lower clusters is that, is more probable for a student to improve their grades when been in a low cluster and it is also due to motivational factors.

5.8.5. Accuracy for students’ grades

As it was mentioned before, mean absolute deviation was used to determine how accurate grades are. The mean result of MAD for the critical path subjects is 0.84, considering a scale for grades from 0 to 5. The grades median of the prediction errors is between 0.5 and 0.7, which indicates that half of the predictions are at maximum 0.7 units up or below the real notes. All subjects minimum error is 0.1 or 0, which means a 100% accuracy for at least one grade. The following table shows the average of the indicators considering all the subjects.
As the table above shows, for every semester the mean absolute deviation is less than 0.63 which means that the model is accurate regarding grades estimation. Although, qualifications may be affected by non-academic variables, mean absolute deviation shows that the simulation is accurate, even though changes in these variables generate some atypical data that affects percentile 95 and max values indicators.

5.9. Simulation of possible scenarios

Testing possible scenarios allows to estimate how student flow gets affected by changing parameters along the simulation. These scenarios are divided in five types, for all of them, indicators average time of graduation and subjects demand are evaluated. The following scenarios are being evaluated. First, stress test for entering freshmen. Second, change the precedence of certain subjects. Third, change the approval rates of the bottlenecks. Fourth, allow freshmen to enroll their subjects in the first semester. Fifth, Include Stochastic Process and Operation Optimization into core career’ subjects. All results for scenario testing as well as additional graphs, analysis, hypothesis tests are in annex 5.

5.9.1. Stress test for freshmen students.

In this scenario, a stress test was performed where values were incrementing or decrementing, approximate, until 50% of the mean of students’ income per semester. These levels are -75, -50, -25, +25, +50, +75, meaning that these values are added up to the mean of students’ income per semester.

![Figure 4: Subjects' demand for stress test. Own authorship.](image)

As the graph below shows, subject’s demand does increment and decrement proportionally in accordance to the new number of students entering the career. That is because the rates of failure are the same but the number of students increment, so the subjects demand get amplified. However, it has variation in consequence of the students’ entrance distribution.

In addition, the table below shows the mean time graduation for every level inside this scenario.
For every pair of levels, T student tests were performed, were it was found that the majority of them do not reject the null hypothesis of mean equality and in only two cases this null hypothesis was rejected. Therefore, it can be concluded that there is no relationship between the number of students that enter the career and the mean graduation time.

5.9.2. Changes subject’s precedence.

5.9.2.1. Integral Calculus is prerequisite of Algorithm Thinking.

For every pair of levels, T student tests were performed, were it was found that the majority of them do not reject the null hypothesis of mean equality and in only two cases this null hypothesis was rejected. Therefore, it can be concluded that there is no relationship between the number of students that enter the career and the mean graduation time.

5.9.2. Changes subject’s precedence.

5.9.2.1. Integral Calculus is prerequisite of Algorithm Thinking.

Table 11: Mean time of graduation for stress test scenarios. Own authorship.

<table>
<thead>
<tr>
<th>Current</th>
<th>-75</th>
<th>-50</th>
<th>-25</th>
<th>25</th>
<th>50</th>
<th>75</th>
</tr>
</thead>
<tbody>
<tr>
<td>11.95</td>
<td>11.96</td>
<td>11.95</td>
<td>11.93</td>
<td>11.96</td>
<td>11.96</td>
<td>11.94</td>
</tr>
</tbody>
</table>

5.9.2. Changes subject’s precedence.

5.9.2.1. Integral Calculus is prerequisite of Algorithm Thinking.

For every pair of levels, T student tests were performed, were it was found that the majority of them do not reject the null hypothesis of mean equality and in only two cases this null hypothesis was rejected. Therefore, it can be concluded that there is no relationship between the number of students that enter the career and the mean graduation time.

5.9.2. Changes subject’s precedence.

5.9.2.1. Integral Calculus is prerequisite of Algorithm Thinking.

Table 11: Mean time of graduation for stress test scenarios. Own authorship.

<table>
<thead>
<tr>
<th>Most Relevant Subjects</th>
<th>Percentage Variation</th>
<th>People Variation</th>
<th>Graduation Time</th>
<th>Mean Test P-Value for Graduation time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algorithmic thinking</td>
<td>-8.12%</td>
<td>-12.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Operations optimization</td>
<td>-2.52%</td>
<td>-1.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Optimization</td>
<td>-1.07%</td>
<td>-1.8</td>
<td>11.97</td>
<td>0.16</td>
</tr>
<tr>
<td>Processes Engineering</td>
<td>-0.35%</td>
<td>-0.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Integral calculus</td>
<td>2.00%</td>
<td>2.8</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 12: Integral Calculus most relevant subjects’ results. Own authorship.

Scenario results, shows that subject demand increases more significantly than subjects which demand decreases. In addition, the p-value fails to reject the hypothesis that establish equality between both means for the graduation times between the scenario and without its precedence changes.

5.9.2.2. Markets Logistics is not prerequisite of Logistics.

For every pair of levels, T student tests were performed, were it was found that the majority of them do not reject the null hypothesis of mean equality and in only two cases this null hypothesis was rejected. Therefore, it can be concluded that there is no relationship between the number of students that enter the career and the mean graduation time.

5.9.2.2. Markets Logistics is not prerequisite of Logistics.

Table 13: Markets Logistics most relevant subjects’ results. Own authorship.

<table>
<thead>
<tr>
<th>Most Relevant Subjects</th>
<th>Percentage Variation</th>
<th>People Variation</th>
<th>Graduation Time</th>
<th>Mean Test P-Value for Graduation time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimization</td>
<td>-2.08%</td>
<td>-3.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Industrial processes</td>
<td>-1.78%</td>
<td>-1.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Production</td>
<td>-0.96%</td>
<td>-1.1</td>
<td>11.89</td>
<td>0.0000</td>
</tr>
<tr>
<td>Logistics</td>
<td>-0.90%</td>
<td>-1.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Markets logistics</td>
<td>-0.96%</td>
<td>-1.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Projects preparation and evaluation</td>
<td>-0.83%</td>
<td>-0.8</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 13: Markets Logistics most relevant subjects’ results. Own authorship.

Scenario results, shows that subject demand increases more significantly than subjects which demand decreases. In addition, the p-value fails to reject the hypothesis that establish equality between both means for the graduation times between the scenario and without its precedence changes.

5.9.2.2. Markets Logistics is not prerequisite of Logistics.

Table 13: Markets Logistics most relevant subjects’ results. Own authorship.

<table>
<thead>
<tr>
<th>Most Relevant Subjects</th>
<th>Percentage Variation</th>
<th>People Variation</th>
<th>Graduation Time</th>
<th>Mean Test P-Value for Graduation time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimization</td>
<td>-3.54%</td>
<td>-6.0</td>
<td>11.88</td>
<td>0.0000</td>
</tr>
<tr>
<td>Operations Optimization</td>
<td>-2.92%</td>
<td>-2.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logistics</td>
<td>-1.97%</td>
<td>-2.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Projects preparation and evaluation</td>
<td>-1.31%</td>
<td>-1.3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Subjects demand for this scenario showed that most affected subjects are “Optimization”, “Operations optimization” and “Processes engineering”. For the first two subjects, demand decreases and for the last one increased. Concerning average time of graduation, p-value analysis failed to reject null hypothesis, in consequence this scenario shortens the time of graduation.

5.9.2.4. University Social Project is prerequisite of Degree Project.

As equal as previous scenarios results, demand variations are more significant negatively than positively. According with results, if this scenario were to be applied, demand of the latest career’ subjects would be affected. The average graduation time is affected when this scenario is preceded. P-value for average mean time establishes that there is a significant change in the time of graduation. Change goes from 11.95 semesters to 12.44 semesters.

5.9.2.5. Simulation is prerequisite of Production.

Demand behavior, when applying this change in precedence, is similar to the results of Markets Logistics. In terms of people, the demand does not vary in more than 4 people for any of the subjects. A general average indicates that the variation in demand would be one more person or one less person for all the subjects taught in the program. In addition, according with the p-value, average time of graduation does not affect the time of graduation of students.

5.9.3. Changes on approval rates of the bottlenecks.

The following tables shows how average time of graduation and relevant subjects demand changes throughout all scenarios. As well as p-values for independence test.
5.9.3.1. **Optimization changes on approval rates.**

<table>
<thead>
<tr>
<th>Scenario</th>
<th>P-Values Range for mean different tests</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Time of Graduation</td>
<td>0.0001</td>
</tr>
<tr>
<td>Theory of probability</td>
<td></td>
</tr>
<tr>
<td>Statistical inference</td>
<td>0.0001</td>
</tr>
<tr>
<td>Operations optimization</td>
<td></td>
</tr>
<tr>
<td>Stochastic processes</td>
<td></td>
</tr>
</tbody>
</table>

Table 17: Optimization changes on approval rates scenarios summary. Own authorship.

The main reason for the average graduation time and “Optimization” grades correlation is that this subject is both part of the critical route, and also is most frequently fail by students, making it a bottleneck subject as well.

5.9.3.2. **Algorithm Thinking changes on approval rates.**

<table>
<thead>
<tr>
<th>Scenario</th>
<th>P-Values Range for mean different tests</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Time of Graduation</td>
<td></td>
</tr>
<tr>
<td>Algorithmic thinking</td>
<td></td>
</tr>
<tr>
<td>Optimization</td>
<td></td>
</tr>
<tr>
<td>Process engineering</td>
<td></td>
</tr>
<tr>
<td>Degree work</td>
<td></td>
</tr>
</tbody>
</table>

Table 18: Algorithm Thinking changes on approval rates scenarios summary. Own authorship.

Even though “Algorithmic thinking” is prerequisite for “Optimization” and a bottleneck subject. However, this subject is not part of the critical path, since students have some slack to approve it, in case of losing it, their time of graduation would not be affected. This slack is the main reason of why the null hypothesis on the ANOVA did not failed to be rejected.

5.9.3.3. **Probability Theory changes on approval rates.**

<table>
<thead>
<tr>
<th>Scenario</th>
<th>P-Values Range for mean different tests</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Time of Graduation</td>
<td></td>
</tr>
<tr>
<td>Theory of probability</td>
<td></td>
</tr>
<tr>
<td>Statistical inference</td>
<td></td>
</tr>
<tr>
<td>Operations optimization</td>
<td></td>
</tr>
<tr>
<td>Stochastic processes</td>
<td></td>
</tr>
</tbody>
</table>

Table 19: Probability Theory changes on approval rates scenarios summary. Own authorship.

Grades from “Probability Theory” have a direct influence on students’ average graduation time. As well as “Optimization”, “Probability Theory” is also part of the critical path and a bottleneck subject and therefore, these results are reasonable as well.

5.9.4. **Freshmen enroll their own subjects.**

Actually, the Pontificia Universidad Javeriana in the faculty of industrial engineering assigns all subjects to freshmen students in order to enroll them in a balanced semester. However, this policy is not the most common
among universities hence, this scenario seeks to simulate the impact on flow indicators when letting freshmen students to enroll any subject they want in their first semester. In this case, assignment of subjects occurred under the assumption of optimality looking for the minimal time of graduation.

![First semester subjects' demand](image)

**Figure 5: First semester subjects’ demand for scenario 4. Own authorship.**

According to the graph above, is evident that there are 3 subjects that increase their mean demand because they are part of the critical path but are not usually enrolled for the University in the first semester. On the other hand, there is a first semester core subject that decrease its demand. This can be explained since the maximum number of credits available to enroll per semester is 20, the optimization algorithm chooses the subjects with higher weight in the critical path.

Regarding the mean graduation time, it was proven by a T-student test that there is a significant difference between both current and scenario 4, having 11.95 and 11.51 semesters respectively. This makes sense because in scenario 4, students are choosing to enroll subjects that reduce the number of semesters to graduate at the same time that they enroll all the 20 credits in core subjects, while current assignation of subjects by the faculty to freshmen students only includes 13 core credits.

5.9.5. **Replace two core subjects for two emphasis subjects.**

The current industrial engineering study plan at the Pontificia Universidad Javeriana does not include the subjects “Stochastic processes” and “Operations Optimization” in their core subjects but in the emphasis ones. In spite of that, there is an urgent need to include them because they complement certain topics of further subjects that are already included in the core. With this new methodology, every industrial engineer student will have to enroll “Stochastic processes” and “Operations Optimization” in order to graduate.

Therefore, 3 scenarios were run in order to see how flow indicators would be affected. The first scenario was to only include “Stochastic processes” in the core subjects, second scenario was to only include “Operations Optimization” in the core subjects and finally the third scenario was to include both “Stochastic processes” and “Operations Optimization” as core subjects.

Mean time graduation results are shown below:
Table 20: Mean time of graduation results for scenario 5. Own authorship.

According to the table, the mean time graduation of “Stochastic processes” only raises very little as a result of the increment of students failing this subject. Despite the inclusion of “Stochastic processes” in the core subjects, “Simulation” which is the following subject, is still enrolled by majority of students in 7º semester, hence the mean time does not substantially get affected. Moreover, when “Operations Optimization” is included in the core subjects, the mean time graduation does increment significantly due to the need of students to approve both “Optimization” and “Operations optimization”, which have very high failure rates, in order to enroll “Degree project”. Finally, the bigger increment is caused when “Stochastic processes” + “Operations Optimization” is included in the core subjects in cause of the previously mentioned reasons.

Additionally, a T-student test for mean difference between the graduation times was performed, giving as a result that there is no significant difference between the current mean graduation time and the one in the “Stochastic processes” scenario. However, there is significant difference between the current mean graduation time and the graduation times of “Operations Optimization” and “Stochastic processes” + “Operations Optimization” scenarios. It means if these two scenarios became applied in the curriculum the mean graduation time would increment.

Finally, both subjects increment their failure cases in comparison to current scenario due to the need of students to enroll these subjects obligatory. However, “Stochastic Processes” increments the number of students that fail the subject in a bigger way than “Optimization” because in the current curriculum “Optimization” is included in three emphases so even though this subject in not currently included in the curriculum, is very likely that many students will have to enroll it, meanwhile “Stochastic Processes” is only included in one emphasis. Hence, when they become core subjects, “Stochastic Processes” suffers a considerable increment in its demand and failure cases.

6. Conclusion

- As it was seen previously, validation for subject grades and dropout rates both were consistent with real life data. Therefore, it can be concluded that grade point average per semester is useful for students’ categorization.
- Concluding Markovian jumps for categorization, there is evidence that students’ grades not only depend on their current semester situation but on the previous one too. In addition, it also shows that students’ relation with their grades change.
- Simulation results and estimations improve as long as databases information is properly debugged as well as subjects remain stable in their program both in their prerequisites and inscription time.
- Simulation program is able to estimate and predict outcomes for changes in stress tests, precedencies changes, changes in subjects means, changes of subjects that freshmen students enroll and the current situation.
- It is concluded that subjects and credits enrollment choices are in fact very similar to results of Robayo-Cote’s algorithm.
- Validation results proved that simulation estimations are accurate in front of reality data.


