THE PERFORMANCE OF AFFIRMATIVE ACTION STUDENTS AND ANALYSIS OF SUBJECTS GRADES USING LATENT CLASS AND LATENT BUDGET MODELS

Enio Galinkin JELIHOVSCHI[1]

■ ABSTRACT: Latent class analysis (LCA) is used to analyse data about performance of students. The original performance variables in the data set are course grade and approval status. However, those variables were not used directly, instead four new variables were calculated from those previous two, variables which are much more informative about the student performance. Coupled with another variable, say affirmative action, the results give light to an understanding about the performance of the divided by affirmative action, yes or no. Besides of showing the results it is also shown how the changing of the original variables by some suitable transformation of the original ones gives more reliable results. The main result is that affirmative action students have a lower performance than those coming from private schools. The paper also analyses the subjects grades using latent budget analysis (LBA), and it is found that the variables cited above have a real effect in characterizing the subjects. It is also shown that those results can be used in a process of evaluation on how the subject is being taught.

■ KEYWORDS: Latent class analysis; latent budget analysis; students performance; affirmative action.

1 Introduction

Brazilian universities are about to complete a decade since the introduction of the affirmative action programs, so that the time is ripe to study how all the students that manage to get into brazilian public university system, either state or federal, by means of affirmative action programs, performed. This will indeed help

[1]Universidade Estadual de Santa Cruz - UESC, Departamento de Ciências Exatas e Tecnológicas, CEP 45662-900, Ilhéus, BA, Brazil. E-mail: *eniojelihovs@gmail.com*

on the improving the program itself and how the universities are going to deal with the students who entered through those programs.

The paper refers to the analysis of the performance of students based on their grade of all subjects they have studied. The study covers the students who entered the university from the year of 2008, when the affirmative action program was instituted, with the exception of engineering courses which started at the second semester of 2011, up to 2014. Besides, the data also provides whether the student entered the university through the affirmative action program or through the regular examination, which in Brazil means that the student studied in a private school.

All variables studied were transformed into categorical variables so that the structure of the data set became a multivariate categorical data which were analyzed using latent class analysis (LCA) and latent budget analysis (LBA) which are a powerful and useful statistical technique for the analysis of this kind of data. The results of those methods may be interpreted from matrices generated by algorithms which solve maximum likelihood estimation method, however using correspondence analysis (CA) graphs generated from those matrices it becomes much easier to interpret the results. When the number of manifest variables are large those graphs are indeed very helpful. Graphical results are much nicer and easier to interpret than tables and the CA results are, in this regard, very representative and go further in the understanding of the many relations among the variables in the data set.

Our original data set consists of the results on two main variables from the students, measuring their academic performance at a state university located in the state of Bahia, Brazil. Those original variables were then used to create four new variables which proved to be much more informative.

The results clearly show that there is a gap between students coming from public schools i.e. affirmative action students and those coming from private schools. The former consistently rank lower as long as the academic performance is concerned.

In section 2 we describe the models used in this paper. In section 3 we show all the results using the dataset cited above.

2 Methods

Latent class analysis is a powerful and useful statistical technique for the analysis of multivariate categorical data. When observed data take the form of a series of categorical responses as, for example, in opinion surveys, longitudinal surveys over time, or consumer behaviour it is often of interest to investigate sources of relationship among the observed variables, identify and characterise clusters of similar cases, and estimate the distribution of observations across the variables being studied. The latent class model seeks to stratify the cross-classification table of observed (or manifest) variables by an unobserved (latent) categorical variable that eliminates all dependences among the manifest variables. This indeed is the crucial and brilliant idea of the LCA, that conditional upon values of this latent variable, responses to all of the manifest variables are assumed to be statistically

independent; an assumption referred to as conditional or local independence, see (GOODMAN, 2002; McCUTCHEON, 1987).

Mathematically it goes as follows. The Latent Class Model (LCM) assumes multinomial distribution over all cells of the cross-classification table .

Let us assume that the number of manifest variables are 3; A, B and C. Also let X be the latent variable with T categories. Following (GOODMAN, 1974a,b) notation the basic equations maybe stated as follows.

$$\pi_{ijk}^{ABC} = \sum_{t=1}^{T} \pi_{ijkt}^{ABCX} \tag{1}$$

where

$$\pi_{ijkt}^{ABCX} = \pi_t^X \pi_{ijkt}^{\overline{ABC}X} = \pi_t^X \pi_{it}^{\overline{A}X} \pi_{jt}^{\overline{B}X} \pi_{kt}^{\overline{C}X} \tag{2}$$

π_t^X is the probability that a subject belongs to $X = t$ and so it is a measure of the size of latent class t. $\pi_{ijkt}^{\overline{ABC}X}$ denotes the conditional probability that a subject belongs to category (i, j, k) of the joint manifest variable ABC, given $X = t$. $\pi_{it}^{\overline{A}X}$ is the conditional probability that a subject obtains score $A = i$, given that this subject belongs to latent class t of X. The same for the other parameters. Equation (2) shows, in mathematical language, the already mentioned property called conditional independence.

All quantities in equations 1 and 2 are probabilities and as such, are subject to standard restrictions, they cannot exceed 1 and are larger than 0 and their sum is 1 after summation over the appropriate subscripts, for example,

$$\sum_{t=1} \pi_t^X = \sum_{i=1} \pi_{it}^{\overline{A}X} = \sum_{j=1} \pi_{jt}^{\overline{B}X} = \sum_{k=1} \pi_{kt}^{\overline{C}X} = 1.$$

The estimation procedure of the above probabilities using the method of maximum likelihood estimation was first outlined by (GOODMAN, 1974a,b) and has been implemented in many computer programs, see also (AGRESTI, 2002). In practice, the estimation of the latent class model is carried out by maximizing the log-likelihood function

$$lnL = \sum F_{ijk} ln(\pi_{ijk}^{ABC}) = \sum F_{ijk} ln(\sum_{t=1}^{T} \pi_{ijkt}^{ABCX})$$

with respect to π_{ijkt}^{ABCX} using the expectation-maximization (EM) algorithm

(DEMPSTER et al., 1977), where F_{ijk} is the observed cell counts. As with any finite mixture model, the EM algorithm is applicable because each individual's class membership is unknown and may be treated as missing data (McLAHAN; KRISHNAN, 1997, 2000).

The conditional probabilities represent a measure of the degree of association between each manifest variable and each latent class, as higher that probability greater the representation of that variable in that specific latent class. This is used to interpret the latent classes. Those probabilities can be organised as a matrix having the observed variables levels as its rows and the latent classes its columns. In this paper we use this matrix to run a CA and get its graphical result.

The latent variable is assumed to explain away the hidden relationships among the observed variables, therefore, if we know what the latent variable represents, conditioning on that variable there should be no more hidden relations among them. The model groups each observation into a latent class, which shows how that observation will respond on each manifest variable. Although the model does not automatically determine the number of latent classes in a given data set, it does offer a variety of parsimony and goodness of fit statistics that the researcher may use in order to make a theoretically and empirically sound assessment. In other words, the model offer tools for the interpretation of the latent classes, which is one of the most important parts of the analysis. In short,the solution of LCA has 2 parts; one confirmatory, in the form of likelihood ratio goodness of fit test used to find the number of latent classes and the other interpretive in the form of a matrix containing the estimated class-conditional outcome probabilities. LCA can also be modelled as a Loglinear model, see (HAGENAARS, 1993). For a more detailed description see (JELIHOVSCHI; SANTANA, 2013)

Latent budget analysis (LBA) is a method for the analysis of contingency tables, and it is used to understand the relation between rows and columns of the table whenever the rows are explanatory and the columns are response variable. LBA uses the matrix of conditional probability of the response given the explanatory variable, or compositional data. The LBA allows us to find which categories of the response are related to different groups of the explanatory categories. If the table has a product multinomial distribution we can understand the latent budget model (LBM) as explaining the relationship between the explanatory and the response variables assuming that conditioned on the latent variable they are independent. In that sense, the latent budgets, which are categories of a latent variable, are hidden values which explain the relationship between the explanatory and response variables, exactly like the LCA.

The table used in LBM is called compositional data matrix P which is defined as follows. Consider a two way table with observed frequencies F_{ij} with I rows and J columns. Let F_{i+} and F_{+j} be the row sums and column sums respectively, the table having the the rows defined as

$$p_i = \frac{F_{ij}}{F_{i+}} \quad \text{for} \quad 1 \leqq j \leqq J$$

is called a compositional data matrix. Each row vector p_i of $\mathbf{P}$ is called observed

budget and is approximated by the expected budget π_i which is a mixture of $K(K \leqq min(I, J))$ latent budgets. The row vectors π_i $(i = 1, \ldots, I)$ form the expected matrix Π which has a lower rank and, in LBM, approximates $\mathbf{P}$.

The latent budgets are represented by β_k $(k = 1, \ldots, K)$ and the model is written as

$$\pi_i = \alpha_{1|i}\beta_1 + \ldots + \alpha_{k|i}\beta_k + \ldots + \alpha_{K|i}\beta_K,$$

where $\alpha_{k|i}$ are the mixing parameters.

The elements of Π are $\pi_{j|i}$ and are called *expected components*.

The elements of β_k are $\beta_{j|k}$ called *latent components*. In scalar notation,

$$\pi_{j|i} = \sum_{k=1}^{K} \alpha_{k|i}\beta_{j|k},$$

and in matrix notation $\Pi = AB^t$ where Π is an IxJ matrix whose rows are the expected budgets. A is an IxK matrix of mixing parameters and B is a JxK matrix whose columns are the latent budgets. LBM(K) is then the latent budget model with K latent budgets. Similar to the observed components, the parameters of LBM are subject to the sum constraints

$$\sum_{j=1}^{J} \pi_{j|i} = \sum_{k=1}^{K} \alpha_{k|i} = \sum_{j=1}^{J} \beta_{j|k} = 1$$

and the non-negativity constraints $0 \leqq \pi_{j|i}, \alpha_{k|i}, \beta_{j|k} \leqq 1$.

In this way, all parameters are proportions what further facilitates the the interpretation of the model. See (JELIHOVSCHI et al., 2011) for a more detailed description of LBM.

Both solutions of LCA and LBM are in the form of matrices. In LCA the matrix corresponds to the conditional probabilities of the occurrence of the manifest variables categories given the latent variable which has many values as the number of classes. In LBA the results comprise two matrices; one for the latent components which shows the conditional probabilities of the occurrence of the column (response) variables categories given the latent variable, the second corresponding to the mixing parameters which shows the conditional probabilities of the occurrence of the latent variable categories given the row (explanatory) variables in other words, which values of the row variables correspond to every value of the latent variable. Nevertheless, instead of studying those matrices straightforwardly, a two dimensional graph is plotted from them by making use of the correspondence analysis methodology. Those graphs are then used as a tool to the interpretation of the model. When the number of categories of either explanatory and response variables are large those graphs are indeed very helpful. Graphical results are much nicer and easier to interpret than tables and the CA results are, in this regard, very

representative and go further in the understanding of the many relations among the variables in the data set.

CA is a dimension reduction method for data analysis of multivariate categorical data. CA graphical results show what are the relation among the rows and columns of a contingency table, namely which rows and columns relate to each other and also which rows (columns) can be grouped together. Indeed, CA graphics capture most of the information contained in the numbers of a contingency table and lay it down on a 2 dimensional graphic, it is one of the best visualisation tools of categorical data analysis, see (JELIHOVSCHI; FERRAZ, 2010) for a detailed description and (GREENACRE, 2007) for the complete theory.

3 Results

3.1 The data set

The original data variables are the grades on every subject the students have taken since they start their course at the university and also the results of the students on every subject, that is whether they passed or not. Besides, the data also give the year the student entered the university plus whether or not he or she entered through the affirmative action program. The first two variables were then used to create four other measurements. Those four variables make a very informative data set which, applied together in the methods used, give a very reliable result.

The following variables were measured:

1. Grade point average (gpa), it is the average of the grades on all subjects taken by a student. It is a quantitative measure of performance and has 3 levels.
   1 - ($gpa < 5.0$), 2 - ($5.0 \leqq gpa < 7.5$), 3 - ($7.5 \leqq gpa \leqq 10$).

2. Approval (Pass) rate (pr), which consists of the proportion of the approved subjects, i.e., the number of all subjects taken by a student in which he/she passed divided by the number of subjects he/she took. It has 3 levels."
   1 - ($pr < 50\%$), 2 - ($50\% \leqq pr < 75\%$), 3 - ($75\% \leqq pr \leqq 100\%$).

3. Final exam (fe), it is the ratio between the number of passing subjects on which the student had to take the final exam (approved with final) and the number of subjects he or she passed without having to take final exam (direct approval). If the average of the grades of the tests a student do within a semester in a given subject is greater or equal seven this student pass with grade equals to that average without need to take a final exam. This variable is an odds, not a rate. This measure is a refinement of the pr and serves to highlight the best students and also help in the separation of affirmative action students from non-affirmative action students. It has 6 levels.
   1 - ($fe \leqq 0.33$), 2 - ($0.33 < fe \leqq 0.66$), 3 - ($0.66 < fe \leqq 1$), 4 - ($1 < fe \leqq 3$), ($3 < fe \leqq 6$), ($6 < fe$).

4. Total (to), it is the sum of the numerator plus the denominator of the fe, that is the total amount of passing subjects. This measure serves to "keep the fe on the line", that is, when the fe is close to zero, which means only "direct pass" and besides that, the total is small, it means that the student is just starting his studies and may not really be a good student. It has 3 levels.
   1 - ($1 < to \leqq 8$), 2 -($8 < to \leqq 16$), 3 - ($16 < to$).

5. Affirmative action (afa) has 2 levels.
   N - Students from private schools, Y - Students from public schools.

3.2 Students' performance analysis

All the analysis has been performed using R (R CORE TEAM, 2016), the graphical interface Tinn-R (Tinn-R TEAM, 2016), the Rpackage poLCA (LINZER; LEWIS, 2011) for latent class analysis, the Rpackage lba (JELIHOVSCHI; ALLAMAN, 2016) for latent budget analysis.

The analysis of four data sets will be shown: the first contains all engineering courses together, the second is the nursing school the third is the law school and the fourth is civil engineering.

3.2.1 Engineering courses

The results of five engineering courses were put together. The courses are civil, electrical, mechanical, chemical and industrial.

The LCA method first requires the evaluation criteria by the means of a goodness of fit test using the likelihood ratio statistic so that we may test the adequacy of the model, that is, the amount of classes which best explain the variation of the observed variables.

In table 1, the evaluation criteria likelihood ratio chi-square statistic G^2 is presented for the LCM with three latent classes for engineering courses.

Table 1 - LCM evaluation criteria; five engineering courses

Model	G	df	pvalue
two-class LCM	332.22	194	0.00
three-class LCM	189.19	183	0.36

The results in that table indicate that we do not accept the two class model, but can accept the hypothesis of three-class LCM.

The conditional probabilities of the observed variables given the latent variable are used to interpret the model by checking one by one and comparing the probabilities for every line, what becomes a cumbersome task as the number of variables and variable levels increase. Instead of doing that, we will use a correspondent analysis graph of the conditional probabilities table and the results become much easier to interpret. The graph has three latent classes positioned in

the map. The variable levels surrounding are used to interpret and name each latent class.

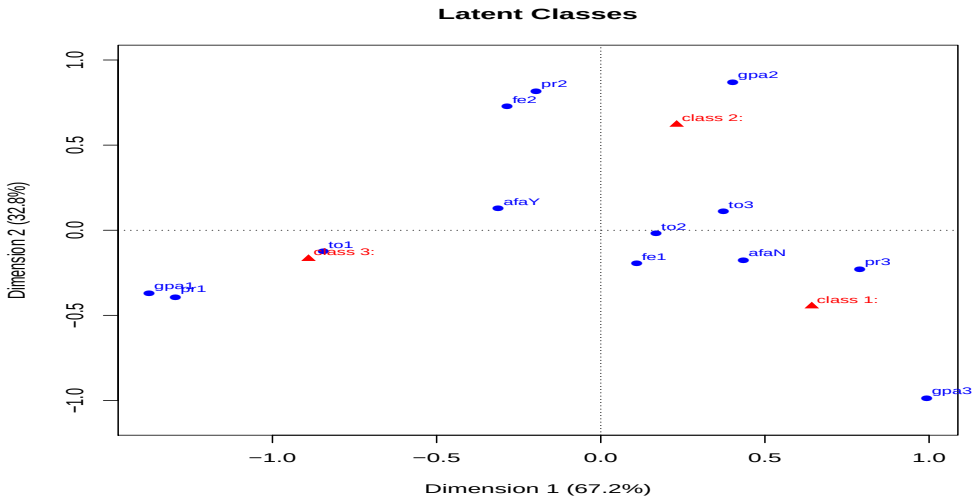

Figure 1 - All engineering graphics, students performance, 3 classes.

Looking at figure 1 it becomes very easy to figure out the latent classes:

- Class1 = gpa3, pr3, fe1, afaN: The highest grade point average and pass rate, also the highest number of direct past in relation to final exam. This class have a higher proportion of non-affirmative action students than affirmative action ones. This class may be called "best students".

- Class2 = gpa2, pr2, fe2: The middle gpa, pr2, and fe. This class may be called "intermediate students".

- Class3 = gpa1, pr1, to1: The lowest gpa and pr, the smallest number passing subjects taken, that is, mostly beginners at the university. This class may be called "lowest performance students"

The interpretation of the CA plot shows that the most important variable to influence the latent classes is the gpa followed by pr. The variable afaN clearly belongs to class one whereas afaY is midway between class1 and class2. This means that students coming from private schools are better than the ones coming from public schools.

The variable fe levels 3 and 4 are represented by a very low number of students so they were taken out of CA analysis because otherwise they could either mislead results by over influence in the graphical results, or just have no influence at all.

3.2.2 Nursing school

In table 2, the evaluation criteria likelihood ratio chi-square statistic G^2 is presented for the LCM with two latent classes for nursing school.

Table 2 - LCM evaluation criteria; nursing school

Model	G	df	pvalue
two-class LCM	202.96	194	0.31
three-class LCM	75.6	102	0.98

The results in table 2 indicate that we should accept both class models, in fact, whenever we accept a k latent classes we will accept the model for any number greater than k. The reason that the smallest k should be chosen is called parsimony criterion.

Parsimony criterion seeks to strike a balance between over and under-fitting the model to the data. Usually, the researcher must take into account that parsimony is the best help in order to achieve a good interpretation of the model, that means a close resemblance between the observed and expected data, with as few parameters as possible, see (DeLEEUW et al., 1990) and (VAN der ARK, 1999). Therefore, in this case, the preferred model is the two latent classes model.

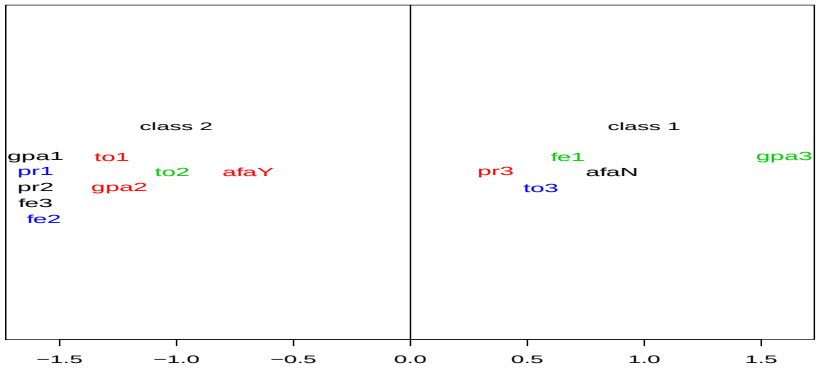

Figure 2 - Nursing school graphics, students performance, 2 classes.

The figure 2 shows the following classes:

- Class1 = gpa3, pr3, fe1, to3, afaN: The highest grade point average and pass rate, also the highest number of direct past in relation to final exam and the greatest amount of passing subjects, the senior students. This class have a higher proportion of non-affirmative action students than affirmative action ones. This class may be called “best students”.

- Class2 = gpa1, gpa2, pr1, pr2, fe2, fe3, to1, to2: The middle gpa, pr2, and fe. This class may be called "intermediate students and lowest performance students".

3.2.3 Law school

In table 3, the evaluation criteria likelihood ratio chi-square statistic G^2 is presented for the LCM with three latent classes for engineering courses.

Table 3 - LCM evaluation criteria; law school

Model	G	df	pvalue
two-class LCM	193.9	142	0.0
three-class LCM	104.8	102	0.96

The results in table 3 indicate that we do not accept the two class model, but do accept the hypothesis of three-class LCM. The data set of law school students have 664 rows of data. Among those rows, 637 have the variable fe = 1 and only 27 to other values of fe. Therefore this variable was taken out of the analysis since it would not add any information to the model.

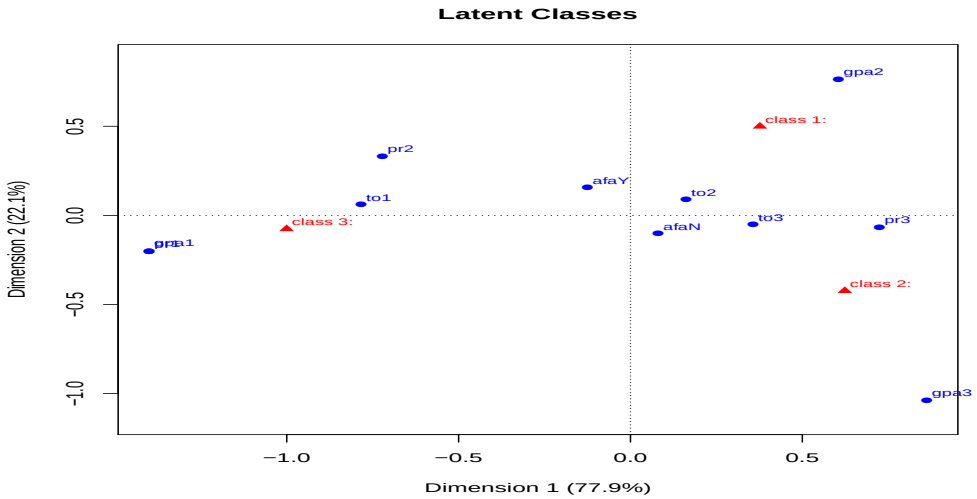

Figure 3 - Law school graphics, students performance, 3 classes.

The figure 3 shows that the 3 classes are:

- Class1 = gpa2 : The middle gpa is the only one variable-level that affects this class. This class may be called "middle students".

- Class2 = gpa3, pr3, to3: The highest grade point average and pass rate, the highest total, the senior students is loosely related. This class may be called "best students".
- Class3 = gpa1, pr1, pr2, to1: The lowest gpa and pr, the smallest and the middle number passing subjects taken, that is, mostly beginners at the university. This class may be called "lowest performance students".

The affirmative action, afa variable is close to the origin. This means that it is not strongly connected to the classes, notwithstanding afaN is closer to class2, the best students and afaY is divided between class1 and class3.

It is interesting to note that in nursing school, the variable afa is strongly marked in each one of each class, that is, the students not in affirmative action strongly belong to best students class and those in affirmative action do belong to the other class. On the other hand in engineering courses the division is similar to nursing school, nevertheless not so strongly and in law school the connection of afa and classes is the most loosely one. The reason behind that is that the law school has a tradition of being a very competitive course, so that the law students are very well prepared, even the afaY students. Engineering requires a strong mathematical background and the afaY students have a weaker background than the others, however the background of afaN students in this university are not strong, nevertheless stronger than afaY students.

Pedrosa *et al.* (2007) shows a study made at the University of Campinas (Unicamp) The main result is that students coming from disavantaged backgrounds, in both educational and socioeconomics aspects, have a higher relative performance than than their complementary group. They show a quality which they called *educational resilience in higher education* which is somewhat comparable to the comments made in the last paragraph concerning the law school students. The main difference between the affirmative action acceptance at the university is that in Campinas an ammount of points were added to the affirmative action students result and then they were classified among all the students whereas the university of the dataset of this paper reserved 50% of all the vacancies of the year to the affirmative action students. Their model of acceptance is designed to capture better and more qualified affirmative action students whereas the other one was designed to have half of all the students at the university coming from affirmative action background.

3.3 Civil engineering subjects analysis based on LBA

The dataset used in this section is part of the subjects belonging to the syllabus of the course of civil engineering. The variables are similar to those used in the prior sections but, their definition are somewhat different. This time all the calculations were made per subject. They are:

1. Grade (gr), the grade of each student who took a specific subject for all the subjects used in the analysis. It is a quantitative measure of performance and

has 3 levels.
1 - ($gr < 5.0$), 2 - ($5.0 \leqq gr < 7.5$), 3 - ($7.5 \leqq gr \leqq 10$).

2. Result (res),
"p" - passed; "fe" - passed in the final exam, "f" - failed, in the subject.

3. Pass rate (pr). It consists of the ratio of the number of students who passed in a subject and the total number of students who took that subject. It has 3 levels.
1 - ($pr < 50\%$), 2 - ($50\% \leqq pr < 75\%$), 3 - ($75\% \leqq pr \leqq 100\%$).

4. Number of students who failed per subject (nf). 1 - ($0 \leqq nf < 10$); 2 - ($10 \leqq nf < 40$); 3 - ($40 \leqq nf$).

5. Final exam (fe), it is the ratio between the number of passing students on which the student had to take the final exam (pass with final) and the number of subjects he or she passed without having to take final exam (direct pass), per subject.
1 - ($0 \leqq fe < 0.34$); 2 -($0.34 \leqq fe < 0.66$); 3 - ($0.66 \leqq fe < 1.0$).

This study will be divided in two parts; the first one are basic core subjects. The following subjects were used in the analysis.

- Calculus I, Calculus II, Calculus III, Physics I, Physics II, Physics III
- Chemistry I, Vectorial Mechanics, Strength of Materials I.

The following graphs are results from *latent budget analysis*.

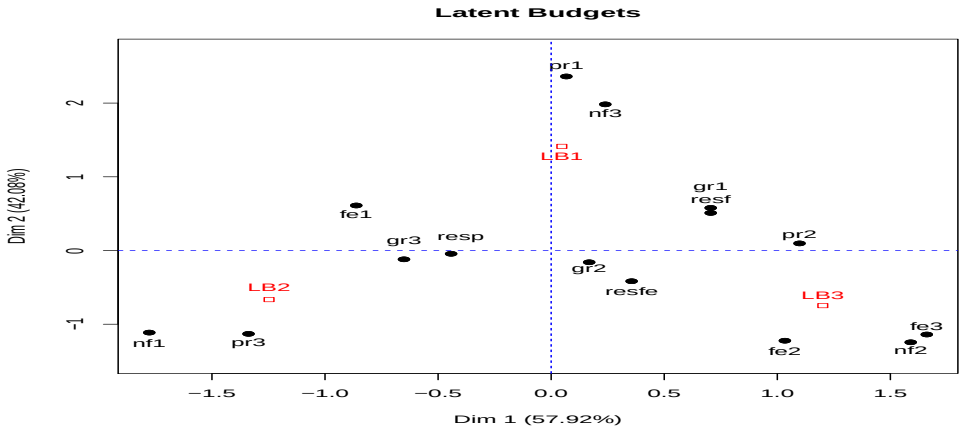

Figure 4 - Variables graphics, subject analysis, 3 Latent Budgets, civil engineering.

The latent budgets LB, explain the variance of the data variables, they are defined in the same way as LCA.

- LB1: gr1 - subjects with the lowest grades, pr1- subjects with the lowest passing rate, nf3 - subjects with more than 40 failing students, resf - subject with high number of failing students.

- LB2: gr3 - subjects with the highest grades, pr3- subjects with the highest passing rate, nf1 - subjects with less than 11 failing students, resp - subject with high number of direct passing students, fe1 - subjects with the highest number of direct passing (three to one) in comparison to final exam passing.

- LB3: fe2, fe3 - subjects with the middle number of direct pass (three to two and one to one) in comparison to final exam passing, nf2 - subjects with 11 to 40 failing students, pr2 - subjects with the middle passing rate.

The three latent budgets are thereof: LB1 represents the most difficult subjects. There are many reasons why a subject is consistently difficult. If the teachers change from semester to semester, the students may be ill prepared for the level of the subject. LB2 represents the easiest subjects, the reason may be either that the teachers make them a low demanding ones or that the students are very well prepared. LB3 represents the middle way subjects. They have the greatest variation among students and are the ones more in line with the level of the students.

The mixing parameters of LBA are the subjects themselves i.e. the graph will show to which latent budget each subject mentioned above belongs. They are:

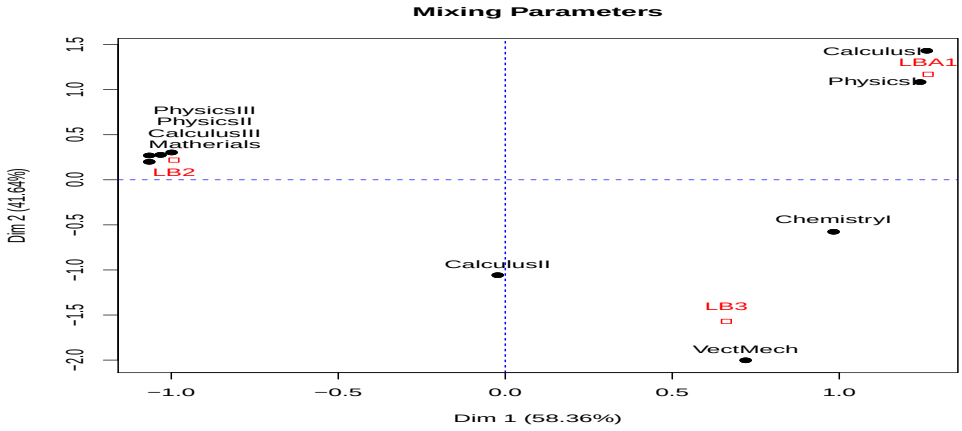

Figure 5 - Subject graphics, 3 Latent Budgets, civil engineering.

- LB1: Calculus I, Physics I, the most difficult subjects.

- LB2: Physics II, Physics III, Calculus III, Strength of Materials I. the easiest subjects.

- LB3: Vectorial Mechanics, the middle way difficulty subject.
- Chemistry is in between LB1 and LB3. Calculus II is in between LB3 and LB2.

Table 4 shows the number of students (n.s.)who took each subject.

Table 4 - Number of students per subject; civil engineering

Subject	n.s.	Subject	n.s.	Subject	n.s.
Strength of Materials I	12	Physics I	200	Physics III	79
Calculus I	166	Calculus III	73	Chemistry I	197
Calculus II	71	Physics II	59	Vectorial Mechanics	93

The subject Strength of Materials I had only 12 students during period when the data was collected and all of them passed direct. Therefore it is still too early do classify as a very easy subject. On the other hand, the difficult subjects had around 200 students, this means a large number of failing students and so, many them took it more than once. The easier subjects had about 70 to 80 students meaning very low repetition.

At this point, we could ask: why the difficult subjects are constantly difficult and easiest ones constantly easy? By observing that the most difficult subjects are those given in the first semester, part of the answer could be just that the beginners are ill prepared for the level required in those subjects.

The second group of variables are the professional subjects.

The following subjects were used in the analysis.

- Architecture and Urbanism, Structural Analysis I, Building Materials I, Soil Mechanics I,
- Strength of Materials II, Structural Analysis II, Building Technology II, Reinforced Concrete Structure I

Three of the variables are the same as before and three have somewhat different levels. The later are:

1. Pass rate (pr),
   1 - ($pr < 70\%$), 2 - ($70\% \leqq pr < 100\%$).
2. Number of students who failed per subject (nf),
   1 - ($0 \leqq nf < 4$); 2 - ($4 \leqq nf < 8$); 3 - ($8 \leqq nf$).
3. Final exam (fe),
   1 - ($0 \leqq fe < 0.2$); 2 -($0.2 \leqq fe < 0.66$); 4 - ($1.0 \leqq fe < 4.1$).

The resulting LBA graphics, the latent budgets (variables) and mixing parameters (subjects) are shown below.

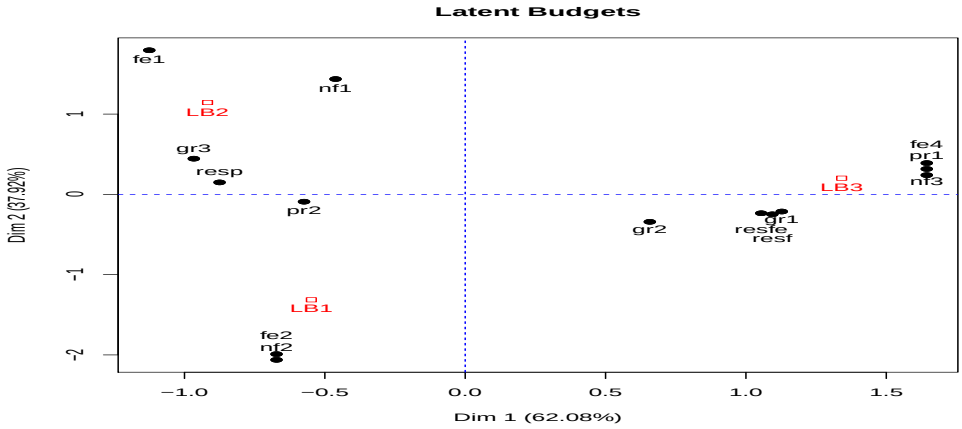

Figure 6 - Variables, 3 Latent Budgets, civil engineering.

The latent budgets are defined as follows.

- LB1: fe2 - subjects with the middle number of direct pass (10 to 2 up to 3 to 2) in comparison to final exam passing, nf2 - subjects with 4 to 7 failing students.

- LB2: gr3 - subjects with the highest grades, nf1 - subjects with the lowest number of failing students, less than 4, fe1 - the highest, 10 to 2 or more, resp - passing result

- LB3: gr1 - subjects with the lowest grades, pr1- subjects with the lowest passing rate, nf3 - subjects with the highest failing students, more than 8, resfe and resf - results final exam and fail.

- pr2 is in the midway between LB1 and LB2.

LB3 characterizes the most difficult subjects, LB2 characterizes the easiest subjects and LB1 the subjects in between LB3 and LB2 subjects.

As before, we will look at the plot of the mixing parameters of LBA which are the professional subjects themselves.

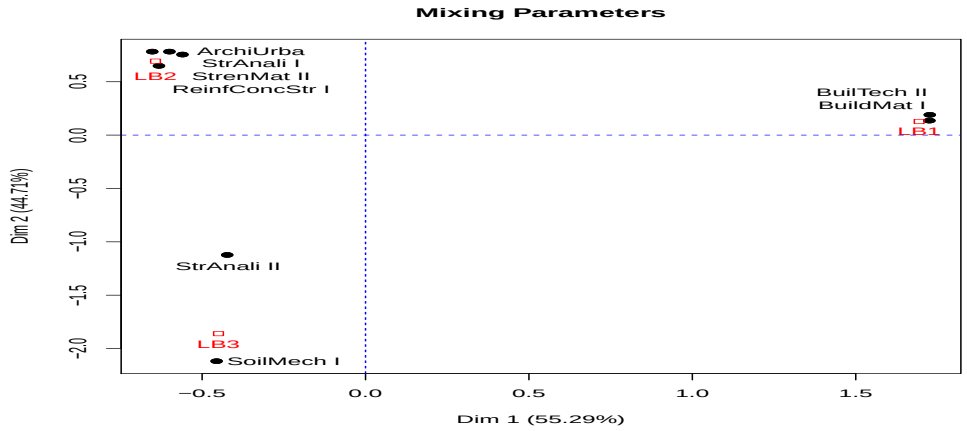

Figure 7 - Subjects graphics, 3 Latent Budgets, civil engineering.

- LB1: Building Technology II, Building Materials I. *Subjects with middle degree of difficulty.*
- LB2: Architecture and Urbanism, Structural Analysis I, Strength of Materials II, CET1066 - Reinforced Concrete Structure I. *Easiest subjects.*
- LB3: Soil Mechanics I, Structural Analysis II. *Most difficult subjects.*

Table 5 shows the number of students who took the above subjects. Structural Analysis II and Reinforced Concrete Structure I have been offered only once. The first belongs to LB3, the most difficult subjects, and the second belongs to LB2, the easiest subjects. All students who took those subjects passed, however, in Structural Analysis II all passed with resfe, while in Reinforced Concrete Structure I all passed direct resp.

Table 5 - Number of students per subject; civil engineering

Subject	n.s.	Subject	n.s.	Subject	n.s.
ArchiUrba I	68	SoilMech I	24	BuilTech II	42
StrAnali I	43	StrenMat II	39	ReinfConcStr I	13
BuildMat I	83	StrAnali II	5		

Table 6 shows the passing rate per subject (pr.s.). Excepting soil mechanics I, practically all the students passed and, in their great majority, with resp. That is why most subjects are considered to be easy.

Table 6 - Number of students per subject; civil engineering

Subject	pr.s.	Subject	pr.s.	Subject	pr.s.
ArchiUrba I	1.0	SoilMech I	0.63	BuilTech II	0.90
StrAnali I	0.98	StrenMat II	0.95	ReinfConcStr I	1.0
BuildMat I	0.93	StrAnali II	1.0		

All subjects have a small number of failing students. The significance of that could be either that the students are very academically good, or that the teachers make it easy for them, or that the subjects are academically easy.

Conclusions

What are the possible ways regarding the affirmative action policy to be followed after ten years of it´s implementation? That is the question that should be asked by policy makers at the brazilian universities.

Although quantitative results do not completely respond to the first paragraph question, they offer a basis over which a thought process may begin. It is clear that students from private schools perform academically better than those from public schools in this particular university. Other studies must follow from different universities and regions of the country. Latent class analysis proved to be a powerful tool to analyse this kind of data, though other methods might also prove useful.

Is it the university responsibility to try to remedy that situation or the responsibility should be placed only upon the students?

Those are important questions whose answers will point the direction of future studies and actions.

The second part of the paper deals with the way subjects are taught at universities and how is the learning process in different subjects. What does it mean when practically every student pass directly in a subject, not once, but whenever the subject is offered? And the reverse of that situation, when most of students fail. The first might make it easy for the students to graduate, the second might create a backlog in the graduation process. Nonetheless all those questions are important issues if university policy is a high academically level of the students. Latent budget analysis is a powerful statistical method to help dealing with those methods quantitatively.

JELIHOVSCHI, E. G.; FERRAZ, M. I. F. Análise do conjunto dos candidatos ao vestibular da UESC no ano de 2008 usando análise de correspondência. *Rev. Bras. Biom.*, Lavras, v.36, n.4, p.827-845, 2018.

- *RESUMO: Dados sobre desempenho de estudantes foram analisados usando latent class analysis (LCA). As variáveis de desempenho originais no banco de dados são nota de cada disciplina por aluno e o resultado, se ele passou ou não. No entanto, estas variáveis não foram usadas diretamente, outras quatro variáveis, muito mais informativas, foram calculadas a partir daquelas duas. Junto com a variável affirmative action (ação afirmativa), elas geram um resultado que dão um entendimento mais profundo sobre o desempenho quando dividido pela variável affirmative action, sim ou não. Além disso também é mostrado que as novas variáveis criadas por transformações adequadas dão resultados mais confiáveis. Os dois principais resultados são que os estudantes que entraram na universidade por meio das ações afirmativas tem um desempenho mais baixo em relação aos estudantes que estudaram em escolas particulares. No artigo também são analisados as notas das disciplinas usando latent budget analysis (LBA), mostrando que todas as variáveis usadas acima são importantes para caracterizar as disciplinas. Além disso os resultados podem ser usadas nos processos de avaliação das disciplinas.*

- *PALAVRAS-CHAVE: Análise de classes latentes; análise de budgets latentes; desempenho de estudantes; ações afirmativas.*

References

AGRESTI, A. *Categorical Data Analysis*. 2.ed. New Jersey: Wiley-Interscience, 2002.

DE LEEUW, J.; VAN DER HEIJDEN, P. J. M; VERBOON, P. A latent time budget model. *Statistica Neerlandica*, n.44, p.1-21, 1990.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via EM algorithm. *Journal of the Royal Statistical Society B*, v.39, p.1-38, 1977.

GOODMAN, L. A. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, v.61, p.215-231, 1974a.

GOODMAN, L. A. The analysis of systems of qualitative variables when some of the variables are unobservable : A modified latent structure approach. *The American Journal of Sciology*, v.79, p.1179-1259, 1974b.

GOODMAN, L. A.; HAGENAARS, J. A.; Latent class analysis: The empirical study of latent types, latent variables, and latent structures *Applied Latent Class Analysis*. Cambridge: Cambridge University Press, 2002.

GREENACRE, M. *Correspondence Analysis in Practice*. Boca Raton: Chapman & Hall / CRC, 2007.

HAGENAARS, J. A. *Loglinear Models with Latent Variables*. London: Sage University Paper series on Quantitative Applications in the social Sciences, 1993.

JELIHOVSCHI, E. G.; FERRAZ, M. I. F. Analysis of the set of candidates to the State University of Santa Cruz (UESC) entrance exam in 2008, using correspondence analysis. *Revista Brasileira de Biometria*, v.28, n.4 , 117-136, 2010.

JELIHOVSCHI, E. G.; ALVES, R. R.; CORREA, F. M. Interacting latent budget analysis and correspondence analysis to analyze beauty salon management data. *Revista Brasileira de Biometria*, v.29, p.657-673, 2011.

JELIHOVSCHI, E. G.; SANTANA, C. R. University students performance: an interaction between latent class analysis and correspondence analysis. *Revista Brasileira de Biometria*, v.31, p.310-326, 2013.

JELIHOVSCHI, E. G.; ALLAMAN I. B. *lba: Latent Budget Analysis for Compositional Data* R package version 2.3, URL: https://github.com/ivanalaman/lba, 2016.

LINZER, D. A.; LEWIS, J. B. poLCA: An R Package for Polytomous Variable Latent Class Analysis. *Journal of Statistical Software*, v.42, p.1-29, 2011.

McCUTCHEON, A. L. *Latent Class Analysis*. London: Sage University Paper series on Quantitative Applications in the social Sciences, 1987.

McLACHLAN, G. J.; KRISHNAN, T. *The EM Algorithm and Extensions*. New York: John Wiley & Sons, 1997.

McLACHLAN, G. J.; KRISHNAN, T. *Finite Mixture Models*. New York: John Wiley & Sons, 2000.

PEDROSA, R. H. L.; DACHS, J. N. W.; MAIA, R. P., AANDRADE, C. Y. Academic Performance, Students' Background and Affirmative Action at a Brazilian University. *Higher Education Management and Policy*, v.19, n.13, OECD 2007.

R CORE TEAM. *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/, 2016.

Tinn-R TEAM. *Tinn-R Editor - GUI for R Language and Environment*. URL http://nbcgib.uesc.br/lec/software/editores/tinn-r/en, 2016.

VAN der ARK, L. A. *Contributions to Latent Budget Analysis*. Phd Thesis,1999. 217p. Utrecht: University of Utrecht, 1999.

Received on 19.05.2017.

Approved after revised on 27.12.2017.