Application of Non–Supervised Learning Tools and Visualization Techniques to Understand the Segmentation Dynamics of First–Year Engineering Students

Patricio Salas[a], Rodrigo De la Fuente[a,*], Andres Riquelme[b]

[a] *Department of Industrial Engineering, Universidad de Concepción, Concepción, Chile*
[b] *Facultad de Economía y Negocios, Universidad de Talca*

Abstract

Multivariate data visualization techniques are powerful methods for extracting information from large databases. Among these techniques, grouping and projecting data from high–dimensional to low–dimensional spaces play an important role, allowing to discover hidden structures in the data. This work presents a methodology for the segmentation and visualization of academic information of first–year students from the college of engineering at the University of Concepción, Chile. First, we use Kohonen's Self–Organizing Map (SOM) to generate a non–linear projection to a two–dimensional space of the academic records. Then, students' are clustered applying the k-means algorithm to the SOM results. Next, we use the Local Interpretable Model-agnostic Explanation (LIME) algorithm to determine the relative contribution of each feature to the centroid of each cluster. After that the clusters are geo–spatially projected to observe patterns among schools and network visualizations are generated to understand the dynamics of how high–schools students flow to the programs offered by the College of Engineering. Finally, using the squarified treemap algorithm, we visualize the participation that different schools have had on enrollment over time. The proposed methodology clearly showed that students coming from the clusters with better academic performance, which correlates with privates schools, prefer degrees in either Chemical, Industrial, Civil, or Mechanical Engineering, while those coming from the worst performing groups composed mainly of public schools, enrolled in Telecommunications, and Materials Science Engineering. Additionally, we observe that subsidized private schools have gained participation in all clusters, becoming a natural replacement for public schools and a serious competitor to private schools.

1. Introduction

Since the emergence of a global economy, education has played the most crucial role and contributed the greatest to the development of a country. In particular, post-secondary education has become vital for all developed countries to maintain domestic prosperity and promote international competitiveness (Ma and Frempong, 2013). With increased enrollment, dropouts become a concern (Ma and Frempong, 2013). Dropout is a problem that affects both developed and developing countries. For example, in the USA, several studies have shown that on average, 60% of students do not complete their undergraduate studies (Shapiro et al., 2015; McFarland et al., 2019). In Europe, the situation is similar; as a consequence, they have defined strategies to decrease this educational indicator (Vossensteyn et al., 2015). While in Latin America and the Caribbean, according to Marta Fer-

* Corresponding author
Email addresses: patricioasalas@udec.cl (Patricio Salas), rodelafuente@udec.cl (Rodrigo De la Fuente), juriquelme@utalca.cl (Andres Riquelme)

reyra et al. (2017) on average, about half of the population aged 25-29 who began higher education at some point did not complete their studies, either because they are still studying or because they dropped out. In Chile, according to the information provided by the *Servicio de Información de Educación Superior* (SIES), on average, the dropout rate in freshmen is 24%.

A large body of literature (Sanchez et al., 2005; Donoso and Schiefelbein, 2007; Chacon et al., 2012; Miranda and Guzmán, 2017) indicate that there is high variability in the factors that motivate the entrance and permanence of students in universities, ranging from economic to academic issues. Consequently, desertion is been addressed more broadly as a problem that generates negative impacts not only at the individual level, but also at institutional, regional, and national levels. Additionally, studies show that desertions occur more significantly during the first year of studies (Donoso and Schiefelbein, 2007; González and Uribe, 2018). Thereby, educational institutions have started to generate policies and strategies leaning to increase retention of freshmen.

Additionally, Braxton and Hirschy (2005) indicated that academic managers should know their students' characteristics to adopt the most adequate actions for their formative development. Given that the structure and nature of these characteristics have an impact on the students' graduation rate (and/or dropout rate), their profiles can be configured based on three categories of data (Gianoutsos, 2011): Demographic profile (gender, socio–economic level, parents' educational level, place of residence, etc.), Academic profile (high school grades, entrance examinations scores, etc.) and Enrollment profile (number of taken or passed credits, undergraduate degree, etc.)

Chile is no stranger to the phenomenon of desertion of freshmen, it is so that in recent decades. Moreover, rapid economic development increased the demand, provoking a sudden rise of institutions offering higher education. This new scenario generated changes that are currently difficult to observe in the students' selection and admission processes. Besides, the increase in coverage and enrollment in higher education has brought significant changes in the socio-economic profile of incoming students. According to data presented in Gallegos et al. (2018), in the early 1990s, there were 245,000 students enrolled in post-secondary education, while by 2017 this number reached a total of 1,162,306. The study also shows that the higher education system has switched from being for an elite to be for a broader target (Gallegos et al., 2018). This new reality makes more relevant the study of the factors that explain entry to higher education and the determinants of school dropout (Acuña et al., 2010). For example, Rolando et al. (2012) showed that sex, high school education quality, and standardized admission test results have a strong influence on the risk of dropping out during the first year. Similarly, (Manzi et al., 2008) showed that the results obtained in the admission test are good predictors of university success. Regarding engineering students, (Díaz, 2009) found that the most relevant factor for dropout is the quality of the preparation obtained during high school.

Summarizing, is imperative to study the phenomenon of desertion in higher education because leaving university without a degree has significant consequences for individuals, institutions, and society (Bernardo et al., 2017; Sarra et al., 2018). To successfully reduce school dropout, it is essential to understand what the underlying determinants of dropout are and which students are at risk of dropping out of school (Berens et al., 2018).

To better understand this new scenario we propose the use of ML algorithms which are a competitive alternative to statistical approaches. This article presents a methodological procedure that loosely couples proposes unsupervised learning to organize and then clusterize student data, and visualization techniques to understand the clusters' dynamics from four different view points: 1) Covariate contribution (LIME), 2) Geographic configuration (GIS), 3) Student–major dynamics (Social Networks), and 4) High-schools dynamics (TreeMaps). Figure 1 illustrates the algorithmic structure of this article.

We believe our methodology provides a complete radiography of first-year students dynamics,

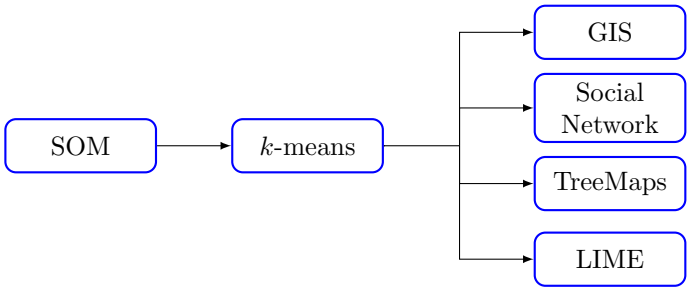

Figure 1: Algorithmic structure of the proposed methodology

and these result could be used to develop internal marketing strategies focused on both capturing better students and reducing dropout rate.

The rest of this article is structured as follows: Section 2 presents the literature review. Section 3 describes the methods and data used in the application, followed by the results in Section 4. Next, Section 5 contains a brief discussion of our analysis. Finally, Section 6 closes the paper with concluding remarks.

2. Literature Review

Since the seminal paper by Becker (1962), investment in human capital is modeled as a financial decision. In one hand, high school graduates can enroll in a higher education institution and forgo part of the wage they can make in the job market while studying. On the other hand, they can go full time to the job market. The rationale behind the decision is that after graduation, the increase in earnings can offset the present value of the foregone income. Dillon and Smith (2017) find that college quality increases earnings ten years after graduation. This setup has been broadly extended. Strayer (2002) study a two–staged decision process, in which the school quality affects the college choice, which in turn affect earnings. Each stage has different determinants and motivations.

Many authors focus on the first stage (e.g, Zvoch, 2006; Venegas-Muggli, 2019), in which students have to decide which college to apply and enroll, and colleges have to choose the best students using limited information of the students quality, usually from standardized tests. There is a broad consensus that the Standardized Admission Test (SAT) are valid predictors of many aspects of student success across academic and applied fields (Manski and Wise, 1983; Camara and Echternacht, 2000; Kuncel and Hezlett, 2007) and that the correct sorting maximizes the efficiency on a human capital production function (Sallee et al., 2008). In Chile, the admission process to universities is governed by the use of SAT, and assume that this mecanism guarantees the permanence of the best students at the university (Donoso and Schiefelbein, 2007; SIES, 2014). Rubio (2011) shows that access to funding reduces the likelihood of students dropping out of college. Also, it presents evidence that students from public schools are more likely to drop out than those from subsidized or private schools.

The matching process between students and colleges is not an easy task, and it differs among different students and college profiles. From the students perspective, Zemsky and Oedel (1983) examine SAT distributions to find market profiles defined by socioeconomic characteristics of the student's families and geographic aspirations of the applicants. Manski and Wise (1983) find evidence that students weight costs, academic quality, and distance when choosing a college. Hoxby and Avery (2013) suggest that neighborhood characteristics and high school characteristics correlate with the college choice. From the college perspective, (Robert, 2010) discusses the effect of market approaches to the education system under several setups, such as public–private schools and admission systems. Another literature focuses on simultaneous analysis: Fu (2014) proposes a structural equilibrium model where heterogeneous students apply to colleges that choose among them using noisy measurements of the student's ability. Dillon and Smith (2017) find that the mismatch between college and applicants is driven by student's decisions (enrollment and application) rather than college admission decisions, where the mismatch is defined as the difference between student's cognitive ability and the relative college quality. Using the same definition, in a later study, (Dillon and Smith, 2017) find that most informed students undematch less and overmatch

more. The consequences of mismatch are multiple, and the economic literature has focused the dropout rates. Dillon and Smith (2018) find that college quality reduces the dropout rates, whereas Stinebrickner and Stinebrickner (2014) use simulations to find that the 45% of the dropout in the first two years of college can be explained by what students learn from their academic performance.

In universities, marketing approaches allow market segments to be identified among future students, building a conceptual model that goes beyond demographics (Angulo et al., 2010). By understanding the people served by the university, it is possible to develop offers that satisfy the target market (Lewison and Hawes, 2007) and the segments that conform it (Bloom, 2005). Grouping study subjects into homogeneous clusters is one of the most commonly used techniques for identifying consumer segments (Saarenvirta, 1998; Davari et al., 2019) because this technique allows study a problem from multidimensional perspective.

A suitable tool to visualize complex multidimensional data is the *Self–Organizing Map* (SOM) of Kohonen (Kohonen and Maps, 1995; Kohonen, 1982). SOM is a very popular unsupervised neural network model for the analysis of high-dimensional patterns in data mining applications (Shieh and Liao, 2012). It can project high-dimensional patterns onto a low-dimensional topology map (Shieh and Liao, 2012). SOM has been used in several fields, including environmental sustainability, hidrology, innovation, computer vision, credit risk assessment, quality control, and financial situation of companies (Mostafa, 2010; Haselbeck et al., 2019; Segev and Kantola, 2012; Hajek et al., 2014; Ortega-Zamorano et al., 2016; Bao et al., 2019; Alkahtani et al., 2019; Chen et al., 2013). In education context, Alias et al. (2006) identify significant patterns of student behavior, Saadatdoost et al. (2011) make knowledge discovery from data of an institute of higher education and Nogales et al. (2019) analyze the characteristics of undergraduate students in Psychology.

In the field of market segmentation, SOM can help market managers easily recognize market segments accurately, as well as monitor market responses for each segment (Wei et al., 2012). For example, Vellido et al. (1999) realized a exploratory segmentation of on-line shopping market, Hung and Tsai (2008) realized market segmentation of multimedia on-demand in Taiwan, Wei et al. (2012) developed a market segmentation of a children's dental clinic. Moreover, cluster analysis can be used as a complementary tool to determine the functional relationship of all the attributes that characterize individuals and to obtain groupings of different sets of them, based mainly on their spatial proximity. Several studies used cluster analysis to perform segmentation, for example, Park and Lee (2014) who segmented consumers into four distinct groups based on their beliefs and motives regarding pro–environmental consumer behavior, Rundle-Thiele et al. (2015) use cluster analysis to identify segments in the context of physical activity. In the educational field, Angulo et al. (2010) propose a market segmentation approach for higher education based on rational and emotional factors Casidy and Wymer (2018) develop a taxonomy of university students based on their orientation to achievement and sensitivity to prestige, Davari et al. (2019) realized a combination of quantitative and qualitative approaches to identify different market segments in the education industry. Finally, good segmentation contributes towards a better understanding of the market and customer demands (Zhou et al.), in this sense, the combined use of SOM and cluster method allows a better understanding of market segmentation (e.g., Bao et al., 2019).

Another well studied application in higher education through the use of ML methods is the prediction of dropout risk, under the a classification problem context (Siri, 2015; Aulck et al., 2016; Beaulac and Rosenthal, 2019). The emphasis of these applications is maximize the prediction accuracy, while interpretability is studied simply by performing a graphical analysis, thus establishing a certain degree of importance of each explanatory factor within the prediction. However, Ribeiro et al. (2016) propose LIME, a method capable of explaining locally the predictions of a ML model through simpler and more accessible models. The

predictive methods offer an interesting framework for brands trying to understand their customers (Zhu et al., 2015).

Lastly, evidence indicates that market knowledge obtained from segmentation derived from the application of ML methods is improved through the use of information visualization techniques and algorithms, such as: First, the GIS provides an even more solid basis for consumer segmentation, as well as for the selection and deselection of entire geographical areas (Pridmore and Hämäläinen, 2017). In addition, Quesada-Pineda et al. (2017) proposes that GIS techniques can be employed in different marketing areas of a company. Second, the analysis of social networks makes it possible to understand the dynamics that link individuals within the market, which is a vital support for marketing (Webster and Morrison, 2004). In education context, Shields (2013) analyses changes in the international network of student mobility in higher education over ten years (1999-2008), using data provided by the UNESCO Institute for Statistics. Sirer et al. (2015) analyze student enrollment flows in public schools in Chicago, USA. Finally, hierarchical information it is possible to visualize using treemaps, developed by Shneiderman (1990), this representation has been widely used in different applications such as business, the stock market and manufacturing (see, Keivanpour, 2019), in educational Keivanpour (2019) developed an interface that allows online student performance monitoring based on the use of treemaps.

3. Methods and data

In this section we describe the methods used in this article to understand the segmentation dynamics of first-year engineering studen ts. Also, the empirical records and variables used in this work are described.

3.1. Data

The University of Concepción, located in the Bío–Bío Region, Chile, is the largest university in the southern part of the country having an enrollment of 20.000 students. This study evaluated data of first–year students who entered the college on engineering between 2005 and 2017 and attended high–schools in the Bío–Bío Region (7866 students).

The variables that characterize the students are those that define a student's profile, namely: gender, socio–economic level[1], distance from his/her high school to the university, scores obtained in the university selection test, number of credits passed in the first year and the undergraduate degree pursued. Categorical variables were coded using *one hot encoding* to facilitate their inclusion in ML models.

3.2. Segmentation

3.2.1. Self Organizing Map (SOM)

The SOM is an unsupervised ML method for multivariate data analysis. It is considered the non–linear version of Principal Components, and it has been applied to group and explore data in various areas (Kohonen, 2013). SOM models are associated with regular grid nodes and use competitive learning, in which a set of neurons compete among themselves to activate. Once the *winning neuron* is found, it is set as an anchor to update the networks' weights. The fundamental idea behind SOM is to convert a signal from a multidimensional vector to a lower–dimensional space, generally to a one– or two–dimensional map. In this paper, we considered the particular type of SOM known as the Kohonen network (Kohonen, 1982, 2013), which does not require prior assumptions of the data, more than the proper encoding of categorical variables. This kind of SOM has a *feed–forward* structure with a single computational layer arranged in rows and columns.

To adjust a SOM, first, a set of small random values must be generated and assigned to the connection weights of each input with the nodes of the regular grid. Then, for each input, the neurons calculate a discriminant function that provides the basis for competition. Next, the neuron with the lowest value in the discriminant function is declared the winner. This neuron determines

[1] Deduced from the type of high school the student attended.

the spatial location of a neighborhood of excited neurons, thus generating the basis for cooperation between neighboring neurons. Finally, new neurons decrease their values of discriminant function through appropriate adjustment of connection weights; therefore, improving the response of the winning neuron to input with similar characteristics.

3.2.2. SOM Algorithm

The basic SOM algorithm includes the following steps (see, Kohonen and Maps, 1995; Shieh and Liao, 2012):

1. Select the parameters that define the map topology. Then, randomly generate the weights of the initialization vector corresponding to each neuron.

2. The network is fed with the data under analysis to find the best *matching unit* for each input vector. Each X record includes quantitative values of n attributes:

$$X = [X_1, \ldots, X_n] \quad \in \mathbb{R}^n \tag{1}$$

The initial weight vector of the i-th neuron is defined as:

$$m_i = [m_{i1}, \ldots, m_{in}] \quad \in \mathbb{R}^n \tag{2}$$

Then, for each input record, the *winning neuron* is the one that satisfied Eq. (3)

$$c = \arg\min_i d(X, m_i) \in \mathbb{R}^n, \tag{3}$$

where $d(X, m_i)$ is the Euclidean distance between a register and the weight vector of the i-th neuron.

3. Update the initial weight vector for each neuron using the Eq. (4)

$$m_i(t+1) = m_i(t) + \alpha(t) \\ h_{ci}(t)[X(t) - m_i(t)], \tag{4}$$

where $0 < \alpha < 1$ is the learning rate and $h_{ci}(t)$ indicate the neighborhood rate of the i-th neuron with the *winning neuron*. This is obtained from the Gaussian function

$$h_{ct} = \exp\left(\frac{||r_c - r_i||^2}{2\sigma^2(t)}\right). \tag{5}$$

In the Eq. (5), σ is the controller of the function domain and decreases as the training process progresses. In addition, r_i and r_c are the position of the i-th neuron and the *winning neuron* (c) in the grid defined by SOM, respectively.

3.2.3. SOM Visualization

The visualization technique used in this research was introduced by Ultsch (1990). This technique uses heatmaps to relate the average values of each of the attributes, in the weight vector, to a given color, to display the values in a color spectrum. Another outstanding aspect of heatmaps is that they provide the possibility to evaluate the correlation between numerical attributes. Similar color shades in two different maps indicate the direct correlation of the corresponding attributes. Likewise, the intensity of the difference or similarity of color between the maps can show the correlation rate between two variables in different parts of the space.

In general, to visualize groups, the distance between adjacent neurons is calculated, and the result presented in what is known as U–*matrix*[2]. If the attributes of two parts of the space under analysis are similar, then the distance between the weight vectors of the related neurons is small, indicating that both neurons are in the same group of the space under analysis. However, the use of *clustering algorithms* have proven to be more efficient in the generation of clusters (Kohonen, 2013).

Finally, the number of nodes that make up the grid is related to the number of observations or inputs available. Vesanto and Alhoniemi (2000)

[2]Unified distance matrix.

showed empirically that a good rule of thumb for the optimal number of neurons in the SOM is:

$$S = 5 \cdot \sqrt{N}, \tag{6}$$

where N is the number of inputs.

3.2.4. K–means Algorithm

The $k-$means algorithm (MacQueen et al., 1967) is a clustering method that generates a partition of the data set into k heterogeneous classes between them and homogeneous classes within them. Here, homogeneity represents the spatial proximity of the vectors of observations based on the Euclidean distance. The process begins with k vectors, randomly selected from the dataset, and used as centroids of the temporary clusters. Then, the algorithm calculates the distances between the centroids and all vectors in the dataset, matching each vector to its nearest centroid. After assigning all data, the computation of the new centroids of each cluster is calculated according to the following criteria

$$c_i = \frac{1}{m_i} \sum_{j=1}^{m} x_{ji}, \tag{7}$$

where c_i is the centroid of the C_i conglomerate and m_i is the number of instances collected in the i-th cluster.

The performance of the algorithm depends heavily on the number of groups to search, a number that is unknown *a priori*. In this work, we applied the *elbow method* criterion, which is a method that analyzes the percentage of variance explained as a function of the number of clusters (Bholowalia and Kumar, 2014). However, other criteria, such as the *silhouette* method, could also be used.

3.3. Visualization

3.3.1. Hierarchical Data Visualization

An efficient method to organize and visualize elements with hierarchical structure is the *treemap* (Johnson and Shneiderman, 1991). Initially designed to display files on a computer disk, this type of graph subdivides a rectangular region into small sub–regions that represent the importance of each part in the hierarchy (Marson and Musse, 2010). The algorithm starts with an outer rectangle, defined as the root. Then, its inner space is filled recursively with rectangles in a hierarchical father/son relationship. Each parent is assigned an area of the root rectangle according to its relative weight. When, the problem recreates itself, and the same procedure is applied to each parent, i.e., subdivided into the necessary number of rectangles to accommodate the children within the area assigned to the parent in the hierarchy.

The algorithm proposed by Johnson and Shneiderman (1991) is known as the standard method and suffers from generating very elongated and thin rectangles that make it difficult to distinguish the importance of the information contained. Several types of trees have been proposed (Shneiderman and Wattenberg, 2001; Bederson et al., 2002; Cesarano et al., 2016). We use the grid treemaps proposed by Bruls et al. (2000), a method that allows comparisons between different groups. It recursively inserts, from one of the vertices of the parent rectangle, rectangles that try to keep the aspect ratio wide/height as close to one as possible. Using this procedure, the area represented by the data approximates a square. Consequently, the area of the square is relative to the importance of the data it represents, which simplifies interpretation and evaluation.

Treemaps are simple to use because they only require three graphical parameters: the display area, the position within the display area, and the color–coding (Tu and Shen, 2007). Also, the use of different colors helps to distinguish between different groups and allows to show the relationship between the hierarchical levels of parent/child (Tu and Shen, 2007).

3.3.2. Local Interpretable Model–agnostic Explanations (LIME)

Many of the ML algorithms are considered black-boxes with limited interpretability. To over-

come this shortcoming, Ribeiro et al. (2016) proposed Local Interpretable Model-agnostic Explanations (LIME), which is a implementation of local surrogate models. Local surrogate models are interpretable models that are used to explain individual predictions of black-box ML models. Following the definition given by Molnar (2019), the main idea of LIME is quite intuitive and consists in understanding why the machine learning model made a specific prediction. LIME shows what happens when predictions disturbances are added to the data in the ML model. The algorithm generates a new dataset consisting of exchanged samples and the corresponding predictions from the black-box model. In this new dataset, LIME trains an interpretable model, which is weighted by the proximity of the sampled instances to the instance of interest. The interpretable model can be, for example, linear regression or a decision tree. The model learned should be a good approximation of the predictions of the machine learning model locally, but it does not have to be a good approximation globally. This type of precision is also called local fidelity.

Mathematically, local surrogate models with interpretability constraint can be expressed as follows (Molnar, 2019):

$$explanation(x) = \underset{g \in G}{\operatorname{argmin}}\, L(f, g, \pi_x) + \Omega(g), \quad (8)$$

where the explanation model for sample x is the model g (e.g. decision three) that minimizes loss L (e.g. cross-entropy), which measures how close the explanation is to the prediction of the original model f (e.g. an random forest model), while the model complexity $\Omega(g)$ is kept low (e.g. select more parsimonious models). G is the family of possible explanations, for example all possible classification models. The proximity measure π_x defines the size of the neighborhood around sample x.

In this work, we used the classifier Random Forest (RF) how the interpretable model, this method was developed by Breiman (2001). RF is a learning method based on the assembly and combination of predictions, that is, it is a strategy that aggregates many predictions to reduce variance and improve the robustness and accuracy of the results (Polikar, 2012; Friedman et al., 2001). The use of the random forest for classification is a powerful statistical tool and has an excellent performance (Díaz-Uriarte and De Andres, 2006).

3.3.3. Network Analysis

Networks analysis in marketing science has so far typically been used to analyze relationships between relatively high-level representations of nodes, such as networks of consumers and firms (Kakatkar and Spann, 2019). The presentation of the relationships in a network diagram allows you to see connections between entities. In particular, the use of networks gave us insight into student flows from high schools to majors.

To build the networks we use the Python programming language and the network analysis software Gephi (Bastian et al., 2009).

3.3.4. Spatial Analysis

Geographic data visualization tools come under the banner of geographic information systems (GIS), which are systems that allow users to visualize, explore, and annotate visual data (France and Ghose, 2019). In this work, we used GIS to know the spatial distribution of the clusters found by utilizing the SOM and k–means algorithms; this was achieved by associating each student with their respective high school of origin.

Spatial data visualizations were generated using Quantum GIS 2.0.1 Team et al., 2018.

4. Results

In this section, we show the empirical results of the application of the proposed methodology to the data set concerning engineering students. We seek to answer the following research questions: 1) Can ML be employed to segment engineering students?. 2) How do project the segmented groups to high school dynamics?.

4.1. Segmentation

First, to obtain an arrangement and grouping of the students, a SOM was trained. Also,

the training process was carried out considering a toroidal structure, thus ensuring that all nodes in the network have access to their neighboring nodes.

To define the number of nodes, we used (6) obtaining 440 nodes; however, we utilized 400 nodes to reduce information dispersion, which produced a 20×20 rectangular grid. Additionally, the processing time was low enough to carry out the study using a laptop, being this one of the main advantages of SOM (Kohonen, 2013).

The ordering of the information provided by SOM made it possible to group students who share similar characteristics, and on average, each node was made up of 20 (± 7) students. Figure 2 shows, for each node, a normalized visualization of the average values of the some numerical variables under study. We averaged the records of all the students grouped within the respective node. Moreover, proper normalization allows us to see the importance of each variable within the node and how its interaction influences the order obtained. Subsequently, to complement the analysis, the respective heatmaps were constructed on the original scale of the variables, as shown in Figure 3.

In the case of the variable associated with the weighted score obtained in the SAT, there is a small area with high scores that gradually fades away until there is a transverse area of low values, given by the toroidal structure of the model. On the other hand, the score associated with high school qualifications (HSQ) shows a more homogeneous distribution of high scores within the map.

The variable that captures the number of credits approved in the first year shows a behavior that is quite counter–intuitive, given that in the upper part there is a strip of nodes that groups students who passed very few credits in their first year of college (even though many had good grades in college), while the lower strip shows relatively high average values.

In terms of correlations, both the SAT weighted score and the HSQ score correlate with the number of credits passed.

An in–depth analysis of students grouped into nodes with high HSQ and a low number of credits approved did not show evident patterns explaining why students with excellent academic performance in high school, had difficulties during their first year.

Table 1: ANOVA results comparing cluster means by k–means analysis.

	Statistics	
	F–score	P–Value
SAT HSQ	803	< 0.0001
SAT Languaje	1598	< 0.0001
SAT Mathematics	2510	< 0.0001
SAT Science	2496	< 0.0001
SAT Weighted	4678	< 0.0001
Credits Approved	1035	< 0.0001
Distance	10.9	0.0001

Now, using the ordering obtained with the SOM, applying the algorithm k–means algorithm (Wehrens et al., 2007), we identify four student clusters. Next, the characterization of each cluster was carried out from the covariates defining the students that compose each cluster. Thus, the partition obtained is a subset of the students' universe (Ghosh et al., 2008).

Figure 4 shows the clusters obtained according to SOM ordering. There is a clear separation, except for some nodes, which is explained by the inherent variability of the data.

To study the effectiveness of the k–means method, an *analysis of variance of a factor* (Montgomery, 2017) was performed with each numerical variable, and the cluster defined factors. Then, we used Tukey's post hoc test when the ANOVA hypothesis test turned out to be significant. The results allow us to infer that there is strong statistical evidence validating the segmentation obtained (see Table 1).

Considering the descriptive analysis at the cluster level given in Table 1 (see, Appendix), the heatmaps in Figure 3 and the results presented in the Table 1, we proceeded to specify the four clusters obtained, which is detailed below:

- Cluster 1 (Green): This conglomerate contains 17.3% of the total number of students considered in this study. The dis-

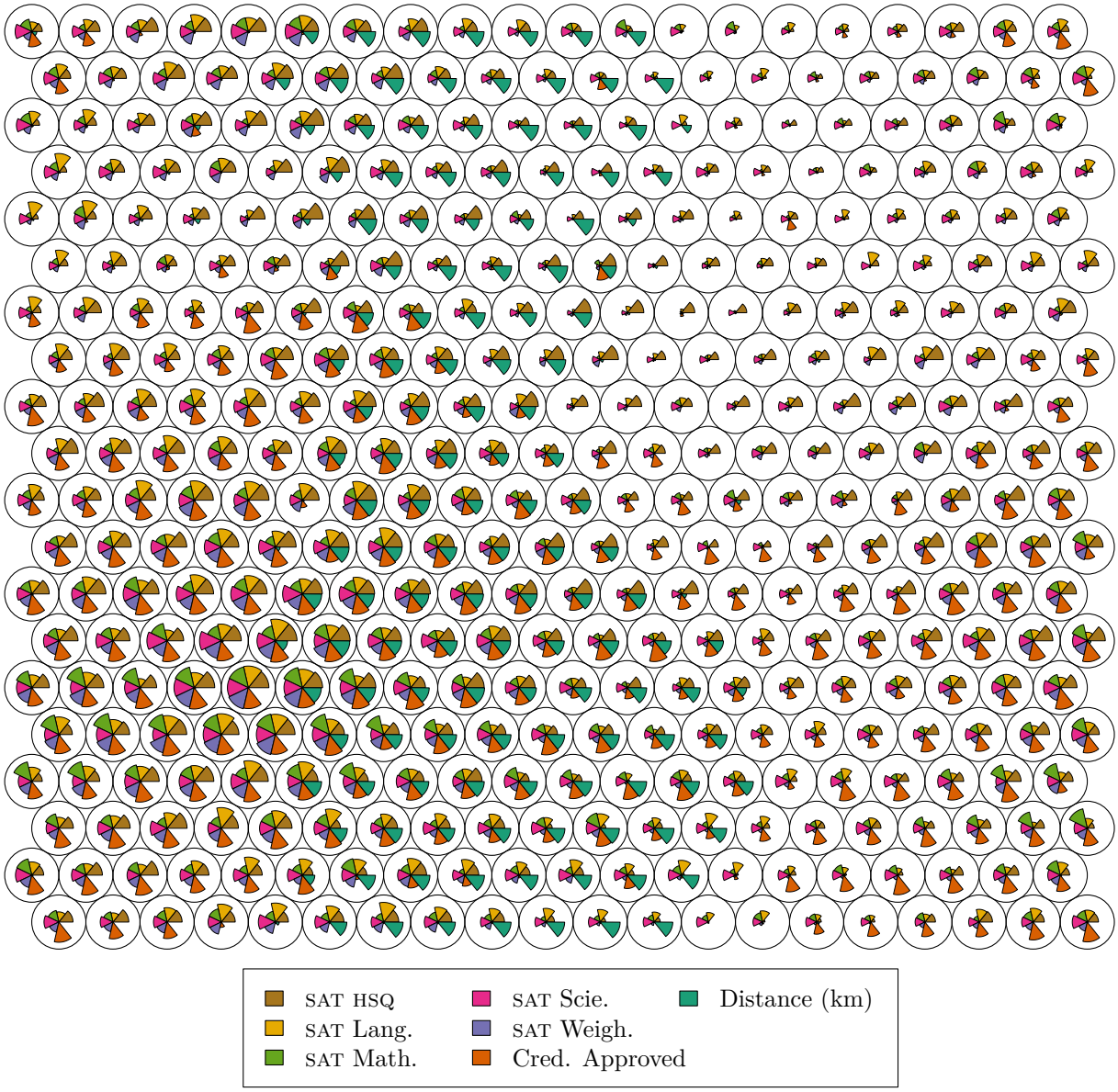

Figure 2: Pie chart of numeric variables at node level.

tribution within the conglomerate according to attended high–school type indicates that 46.6% corresponds to students from private schools, 45.5% to students from subsidized schools and only 7.96% to students who graduated from state–dependent schools. The average distance from schools to the university was 35.1 km. (40.9km). The students who belong to this cluster, on average, have a weighted college selection test score of 725 (32.3), high–school grades score of 724 (52.8), and in their first year of college, these students, on average, pass 30.4 (8.8) credits. The undergraduate majors with more representativeness in the conglomerate are Chemical Engineering (28.5%), Industrial Engineering (17.9%) and Civil Engineering (16.2%).

- Cluster 2 (Yellow): In terms of the number of students, this conglomerate is the largest and represents 47.2% of the population under study. Concerning high–school precedence, 22.8% comes from private schools,

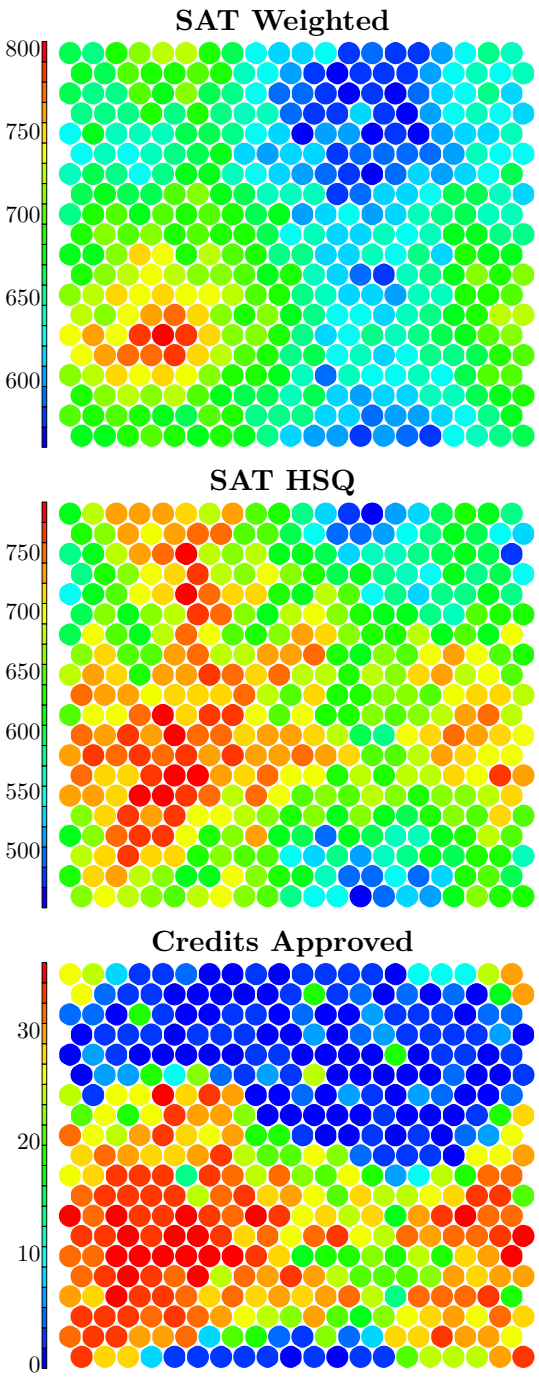

Figure 3: Heatmap of numerical variables in their original scale. The contours show the composition of the clusters.

51.7% comes from subsidized schools, and the remaining 25.5% corresponds to students from state schools. The average levels of the numerical variables are similar to the values in cluster 1. Additionally, the weighted score in the university selection test was

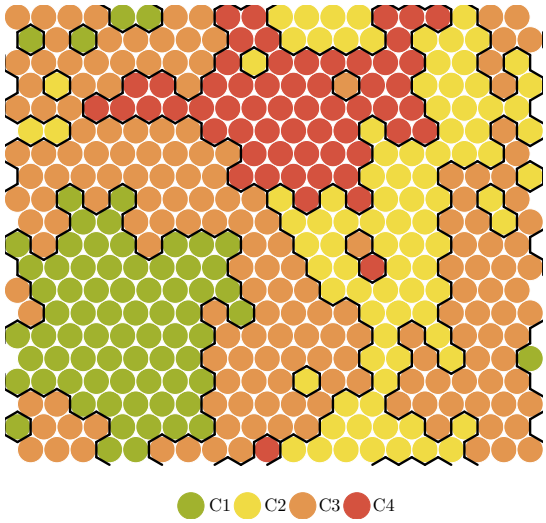

Figure 4: Clusters identified on the SOM from the k–means method

670 (28.7), the score of the high school grades was 680 (59.9), and in their first year, these students pass an average of 22.2 (12.7) credits. The majors with the most students from this conglomerate are Industrial Engineering (14.6%), Metallurgical Engineering (11.6%), Civil Engineering (10.3%) and Mechanical Engineering (10.1%).

- Cluster 3 (Orange): A 23.1% of the students are grouped in this cluster, from which 19.9% are students from private schools, 64.8% graduated from subsidized schools and the other 15.3% from state schools. It is striking that this conglomerate concentrates students whose high–schools are closer to the university; with an average distance of 14.6km. (25.3km). In this conglomerate, the average levels of the numerical variables have a significant decrease; thus, the weighted score in the university selection test was 623 (25.0) points, the score of high school grades was 621 (68.8) points and in the first year of college, these students, on average, pass 10.5 (12.1) credits. The largest number of students are in the following programs: Civil Engineering – Common Plan (23.4%), Biomedical Engineering (12.4%), Telecommunications Engineering (12.0%) and Ma-

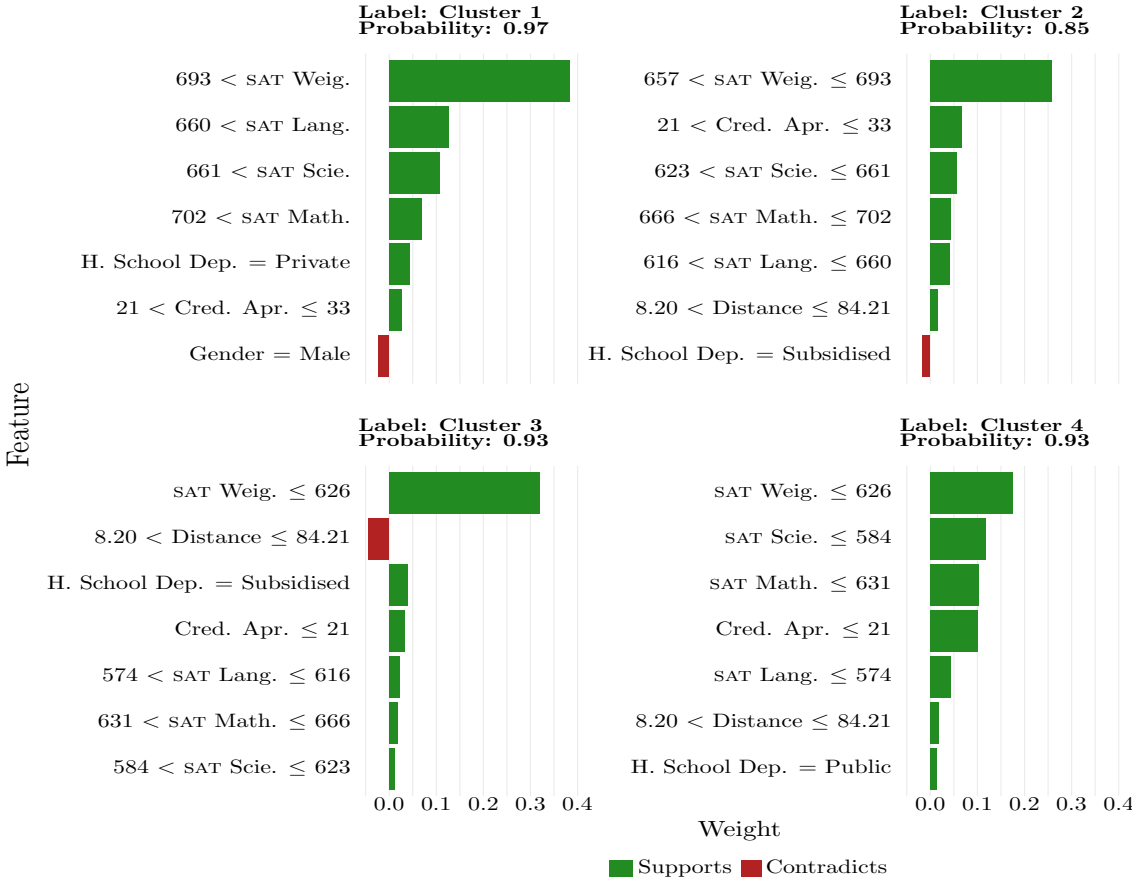

Figure 5: Visual output of LIME for interpretation of predictions made with RF method.

terials Science Engineering (11.6%).

- Cluster 4 (Red): This cluster concentrates 12.4% of the students, being the least numerous. The distribution of students, according to their high–schools, is dominated by those who graduated from state schools (49.6%), followed by students from subsidized schools (41.3%), while only 9.0% come from private schools. In terms of averages, the grade point average is 645 (68.7), the math test scores 606 (33.0), the weighted college selection test scores 606 (31.5) and the number of credits passed does not exceed 3 (6.0). In general, the decrease in performance is considerable, and this conglomerate contains the students whose high–school is geographically far from the University of Concepción. This indicator is, on average, 45.4km. (41.4km).

4.2. Clusters' Dynamics

The combination of SOM and k–means allowed the students to be segmented. With this result, we analyzed different angles of cluster dynamics. First, to classify a new student in a group, a RF was trained, and with LIME, we analyzed the contribution of predictor variables in the prediction. Then we georeferenced the students according to the school of origin. Next, we analyze the flow of students from schools to careers through social networks. Finally, using the squarified treemap algorithm, we visualize the contribution of different schools in enrollment over time.

4.3. Attributes Contribution

MENCIONAR LOS INDIVIDUOS FICTICIOS QUE SE CREARON

We used LIME to complement the multivariate interpretation of the clusters based on the average values of predictor variables of students

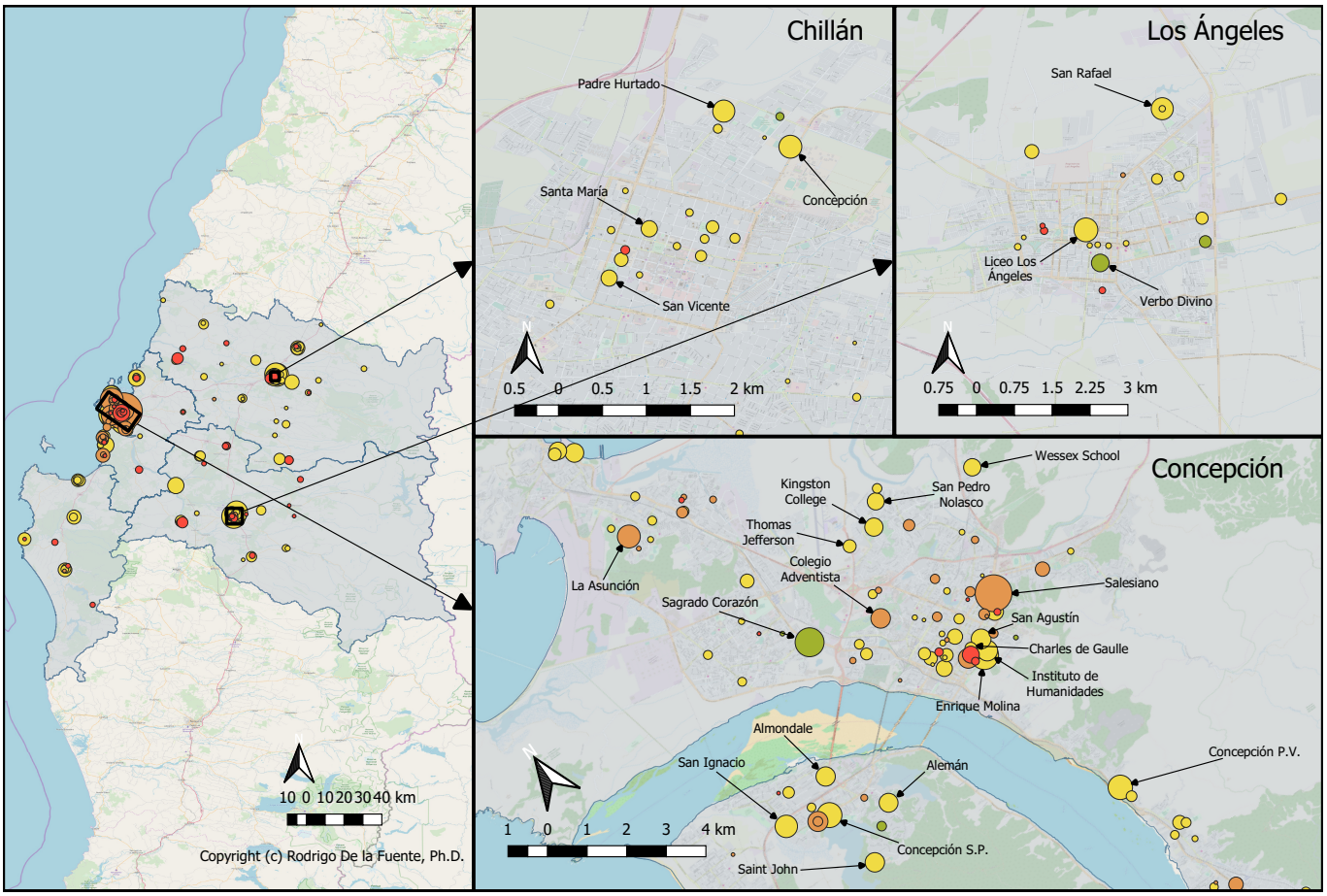

Figure 6: The spatial composition of high–schools indexed by clusters. Bío–Bío Region (left panel), city of Chillán (upper central panel), city of Los Ángeles (upper right panel) and city of Concepción (lower right panel).

that compose each cluster. While the interactions between characteristics are mostly ignored in this way, the visualization casts some light on the black box that is the classifier (Amrit et al., 2017). First, using *cross validation*, a RF model with an 82.2% accuracy was obtained; this means, the probability that a student will be well classified in the correct cluster is high.

Figure 5 shows the visual production obtained by LIME; in this figure, we can see that, there is a high probability of correctly classifying a new observation. The main characteristics influencing the prediction are ordered according to their contribution, including the direction of the influence. We can interpret this as the contribution of the attributes to the result (Amrit et al., 2017). Now, for the average student of the first and second groups, high scores in the university selection tests favor or support classification, with the difference that dependence on the middle school is more important in the first case. For the average student in the rest of the groups, the prediction is mainly based on low values in the selection tests and the small number of credits passed in the first year of university. Additionally, in the third group, the distance from high school to university is a factor that contradicts the prediction, which could be explained by the high variability of this variable within the group.

4.3.1. Georeferencing

REESCRIBIR ESTE PARRAFO DEMASIADO WE USED

Information from the students' cluster was assigned to each high–school to highlight its geographical disposition and characterization. We used a simple majority vote rule, based on the students' assigned cluster, to match schools to a cluster. Initially, only the name of the school-

s was available; for this reason, the inverse georeferencing methodology was used to obtain the geographic coordinates of these institutions.

Figure 6, displays a map showing the schools. The size of each marker depends on the total number of students coming from the school, and color represents the cluster assigned to each of them. We focus on the most densely populated areas in the Bío–Bío Region. The data presented in each panel (from left to right) of Figure 6 are as follows: In the left panel, a general overview of the whole region is given. On the right panel, schools of the city of Chillán are shown, and in the upper right corner, schools located in the city of Los Angeles are displayed. Finally, in the lower right corner panel, schools of the Concepción metro area are represented.

From Figure 6, it can be inferred that, in general terms, there is a broad domain of schools that were assigned to conglomerate C2. In the city of Chillán, few schools fall into conglomerate C3 or C4, and similar behavior can be observed for Los Angeles. On the contrary, in the municipality of Concepción, there is more considerable variability in the classification of schools, which could indicate that perhaps in the latter case the distance from the graduating school to the University of Concepción is not a determining factor of academic performance, or that low–performance students from Chillán and Los Angeles have chances of entering the college of engineering than their Concepcón counterparts due to other non–observed variables, such as, house–hold income, high–school focus, to name some.

4.3.2. Social Network

DESCRIBIR TABLAS DEL APENDICE, NO BASTA CON MENCIONAR QUE ESTÁN

High schools and engineering majors are intrinsically linked as students flow from schools to the various study programs offered by the college of engineering. We defined two large clusters to observe student flow patterns; these clusters arose from the clustering of C1 with C2 and C3 with C4. Thus, Figure 7 show the networks that describe the connections between high schools and majors. The size of the nodes reflects the actual number

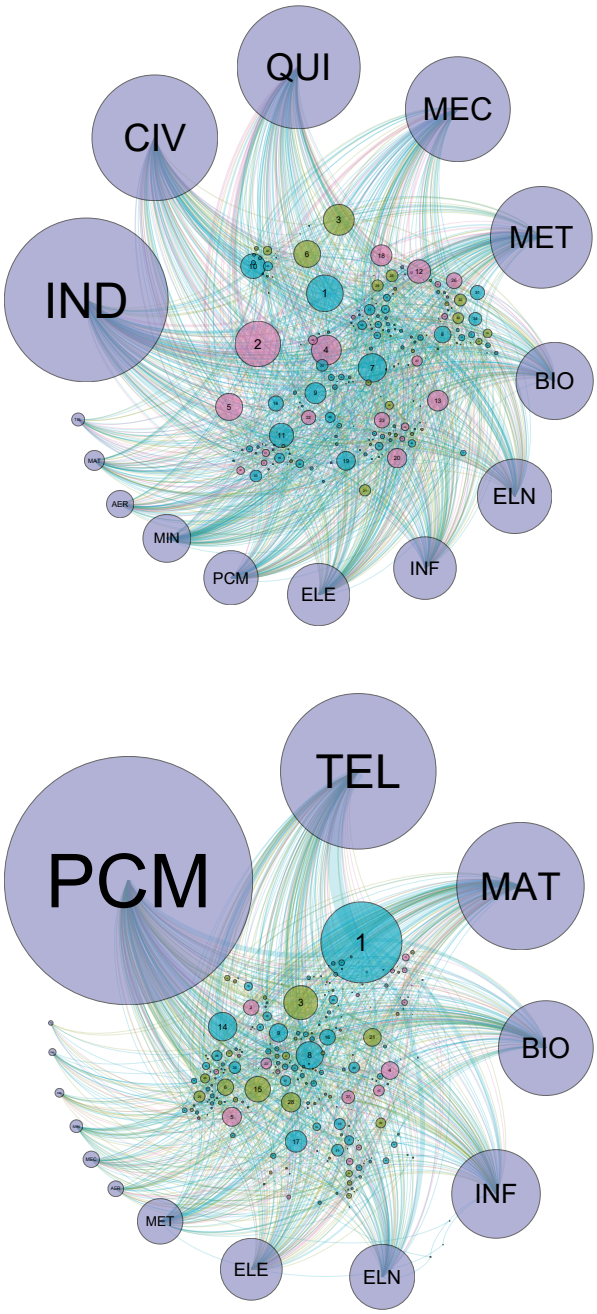

Figure 7: Networking between schools and careers: Clusters 1 and 2 (top), Clusters 3 and 4 (bottom)

of students per node in each group, whereas the thickness of the line indicates the number of students that flows from each school. Also, the color of the nodes identifies the type of school: private (pink), subsidized (light blue), and public (green).

Table 2 (see, Appendix) summarises the distribution of students, at the school level, who are classified within each cluster. Also, in the first

column are the identifiers assigned to each school in Figure 7.

The upper panel of Figure 7 shows that Industrial, Civil, Chemical, Mechanical, and Metallurgical Engineering are fed mainly by students coming from subsidized and private schools. Also, in the composition of schools, private schools dominate in size, while subsidized schools dominate in quantity. Finally, public schools are under–represented in this conglomerate.

Likewise, the bottom panel of Figure 7 shows that the conglomerate formed by C3 and C4 is mainly of students from subsidized and public schools. There is an inverse effect to that described in the previous paragraph, with a higher flow towards the following majors: Common Plan, Telecommunications, Materials, Biomedical, and Computer Engineering. As for the composition of the type of school, there is a vast domain of subsidized schools, followed by public schools and little participation of students from private schools.

4.3.3. Enrollment Over Time

Unfortunately, in Figure 7, it is not possible to appreciate an obvious ordering of schools, and it is even more different to see some temporary behavior of how each type of high education system has evolved through times to majors. To address this problem, we fitted squarified treemaps, ranked by type of high school and maintaining the two conglomerates mentioned above. In addition, to observe temporary effects on the relative contribution of schools, the number of students was divided according to enrollment cohort ($[2005, 2009]$, $[2010, 2013]$ and $[2014, 2017]$).

The treemaps in Figure 8 capture the temporary effect of students' enrollment from all types of schools. The top row shows that the contribution of students from public schools has been declining in the best clusters. On the other hand, from the lower row, it can be seen that students from private schools have gradually decreased their participation in conglomerates C3 and C4, similar to what happens with students from public schools. We can see that there is clear evidence of the decline of public education, which coincides with the finding of (Paredes and Pinto, 2009). Finally, we noted that subsidized schools had gained ground in both clusters.

5. Conclusions

REVISAR LA CITA DE DILLON (ES 2017 o 2018) Our findings are consistent with the results found by (Sallee et al., 2008) for the school's side of the system: the more schools, the less polarization, and better segmentation. The identification of four different clusters of applicants aligns nicely with the variety of engineering majors provided by Universidad de Concepción. This variety of careers allows applicants to select majors that better match their abilities. This statement is limited in by two facts: first, the straight interpretation of the results requires that students have full information about the University and programs characteristics. Under this assumption, Dillon and Smith (2018) propose that families make the college choice based on the applicant ability and their financial constraints. Second, we only have data from Universidad de Concepción, and our results do not necessarily have external validity. This fact is noteworthy if we consider that there are 102 Engineering and Technology tracks in Chile[3]) and five Universities offering Civil Engineer in Concepción.[4]

The contribution of this paper is to extend the study of the students and university programs matching using unsupervised learning methods. These methods exchange interpretability by prediction accuracy. Given the fact that the causes of the mismatch between students and programs are broadly documented, the increase in accuracy and the identification of student profiles is a novel and significant contribution.

Beyond the visual aid provided by the construction of maps, it was possible to directly correlate student information through visualizations, creating a complete learning experience aligned with the objectives and needs of the organization.

[3] https://www.educaedu-chile.com/carreras-universitarias/ingenieria

[4] Universidad de Concepción, Universidad Católica de la Santísisma Concepción, Universidad del BíoBío, Universidad de las Américas, Universidad San Sebastián.

The combination of SOM with the clustering algorithm allowed the identification of four student segments, allowing us to characterized them through the application of visualization tools.

The variables considered the results obtained in the specific university selection tests (Science, Mathematics, and Language), allowing for better discrimination among the different conglomerates. This seems to indicate that students with an excellent academic performance before entering the university have developed better study methods, specific discipline, and individual responsibility, characteristics valued in the Chilean university system.

This work provides complete radiography of the evolution of the type of students that enter the different majors offered by the college of engineering at the University of Concepción. Geographically speaking, we identified the sectors where they came from and associated their high schools with the characterization of the students that enter each academic program.

The academic information was used to understand the dynamics of segmentation of first–year engineering student at the Universidad de Concepción. Student characteristics at the segment level could be identified, and service strategies can be designed and adapted based on an assessment of these needs.

From this work, the College of Engineering of the University of Concepción will be able to carry out marketing campaigns to attract better students and to improve its current retention levels by providing better information sets to the applicants (Chingos, 2012).

6. Acknowledgements

The authors are grateful to the University of Concepción for providing the data set that allowed this work to be developed. Andrés Riquelme acknowledges support from FONDECYT through grant 11160948.

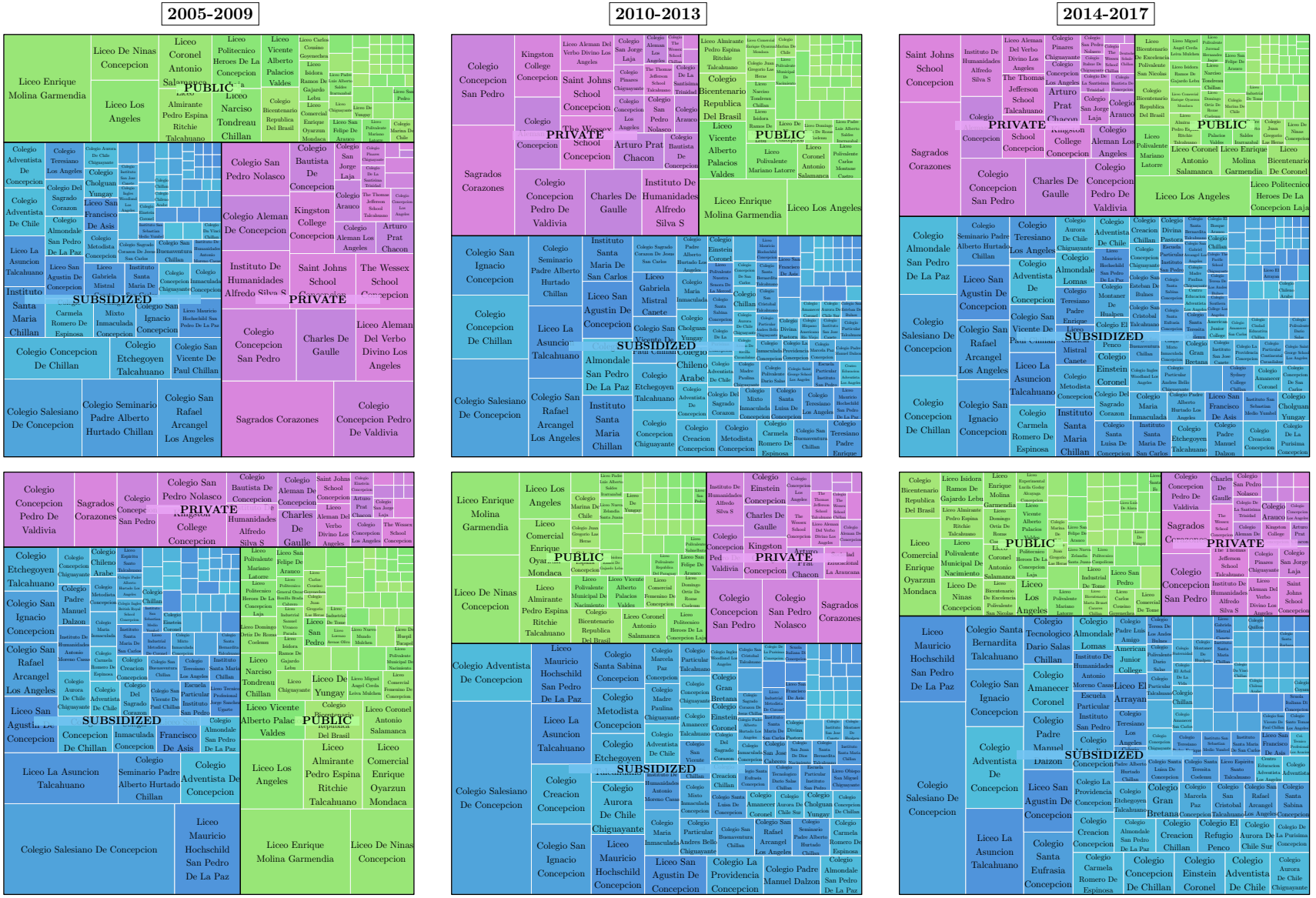

Figure 8: Treemaps of the number of students per school that fall within the defined clusters.

References

Acuña, C., Makovec, M., and Mizala, A. (2010). Access to higher education and dropouts: evidence from a cohort of chilean secondary school leavers. In *Paper presentando en el Primer Congreso Interdisciplinario de Investigación en Educación (CIIE)*.

Alias, U. F., Ahmad, N. B., and Hasan, S. (2006). Student behavior analysis using self-organizing map clustering technique. *learning*, 22:27.

Alkahtani, M., Choudhary, A., De, A., and Harding, J. A. (2019). A decision support system based on ontology and data mining to improve design using warranty data. *Computers & Industrial Engineering*, 128:1027–1039.

Amrit, C., Paauw, T., Aly, R., and Lavric, M. (2017). Identifying child abuse through text mining and machine learning. *Expert systems with applications*, 88:402–418.

Angulo, F., Pergelova, A., and Rialp, J. (2010). A market segmentation approach for higher education based on rational and emotional factors. *Journal of Marketing for Higher Education*, 20(1):1–17.

Aulck, L., Velagapudi, N., Blumenstock, J., and West, J. (2016). Predicting student dropout in higher education. *arXiv preprint arXiv:1606.06364*.

Bao, W., Lianju, N., and Yue, K. (2019). Integration of unsupervised and supervised machine learning algorithms for credit risk assessment. *Expert Systems with Applications*, 128:301–315.

Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. In *Third international AAAI conference on weblogs and social media*.

Beaulac, C. and Rosenthal, J. S. (2019). Predicting university students' academic success and major using random forests. *Research in Higher Education*, pages 1–17.

Becker, G. (1962). Investment in human capital: A theoretical analysis. *Journal of Political Economy*, 70.

Bederson, B. B., Shneiderman, B., and Wattenberg, M. (2002). Ordered and quantum treemaps: Making effective use of 2d space to display hierarchies. *AcM Transactions on Graphics (TOG)*, 21(4):833–854.

Berens, J., Schneider, K., Görtz, S., Oster, S., and Burghoff, J. (2018). Early detection of students at risk–predicting student dropouts using administrative student data and machine learning methods.

Bernardo, A., Cervero, A., Esteban, M., Tuero, E., Casanova, J. R., and Almeida, L. S. (2017). Freshmen program withdrawal: Types and recommendations. *Frontiers in psychology*, 8:1544.

Bholowalia, P. and Kumar, A. (2014). Ebk-means: A clustering technique based on elbow method and k-means in wsn. *International Journal of Computer Applications*, 105(9).

Bloom, J. Z. (2005). Market segmentation: A neural network application. *Annals of Tourism Research*, 32(1):93–111.

Braxton, J. M. and Hirschy, A. S. (2005). Theoretical developments in the study of college student departure. *College student retention: Formula for student success*, 3:61–87.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Bruls, M., Huizing, K., and Van Wijk, J. J. (2000). Squarified treemaps. In *Data visualization 2000*, pages 33–42. Springer.

Camara, W. J. and Echternacht, G. (2000). The sat [r] i and high school grades: Utility in predicting success in college. research notes.

Casidy, R. and Wymer, W. (2018). A taxonomy of prestige-seeking university students: strategic insights for higher education. *Journal of Strategic Marketing*, 26(2):140–155.

Cesarano, A., Ferrucci, F., and Torre, M. (2016). A heuristic extending the squarified treemapping algorithm. *arXiv preprint arXiv:1609.00754*.

Chacon, F., Spicer, D., and Valbuena, A. (2012). Analytics in support of student retention and success. *Research Bulletin*, 3:1–9.

Chen, N., Ribeiro, B., Vieira, A., and Chen, A. (2013). Clustering and visualization of bankruptcy trajectory using self-organizing map. *Expert Systems with Applications*, 40(1):385–393.

Chingos, M. (2012). Graduation rates at america's universities: What we know and what we need to know. *Getting to Graduation: The Completion Agenda in Higher Education*, pages 48–70.

Davari, M., Noursalehi, P., and Keramati, A. (2019). Data mining approach to professional education market segmentation: a case study. *Journal of Marketing for Higher Education*, 29(1):45–66.

Díaz, C. J. (2009). Factores de deserción estudiantil en ingeniería: una aplicación de modelos de duración. *Información tecnológica*, 20(5):129–145.

Díaz-Uriarte, R. and De Andres, S. A. (2006). Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, 7(1):3.

Dillon, E. and Smith, J. (2017). Determinants of the match between student ability and college quality. *Journal of Labor Economics*, 35(1):45 – 66.

Dillon, E. W. and Smith, J. (2018). The consequences of academic match between students and colleges. NBER Working Papers 25069, National Bureau of Economic Research, Inc.

Donoso, S. and Schiefelbein, E. (2007). Análisis de los modelos explicativos de retención de estudiantes en la universidad: una visión desde la desigualdad social. *Estudios pedagógicos (Valdivia)*, 33(1):7–27.

France, S. L. and Ghose, S. (2019). Marketing analytics: Methods, practice, implementation, and links to other fields. *Expert Systems with Applications*, 119:456–475.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer se-

ries in statistics New York, NY, USA:.

Fu, C. (2014). Equilibrium tuition, applications, admissions, and enrollment in the college market. *Journal of Political Economy*, 122(2):225 – 281.

Gallegos, J. A., Campos, N. A., Canales, K. A., and González, E. N. (2018). Factores determinantes en la deserción universitaria. caso facultad de ciencias económicas y administrativas de la universidad católica de la santísima concepción (chile). *Formación universitaria*, 11(3):11–18.

Ghosh, A. K., Javalgi, R., and Whipple, T. W. (2008). Service strategies for higher educational institutions based on student segmentation. *Journal of Marketing for Higher Education*, 17(2):238–255.

Gianoutsos, D. (2011). Comparing the student profile characteristics between traditional residential and commuter students at a public, research-intensive, urban commuter university.

González, L. E. and Uribe, D. (2018). Estimaciones sobre la" repitencia" y deserción en la educación superior chilena. consideraciones sobre sus implicaciones. *Calidad en la Educación*, (17):75–90.

Hajek, P., Henriques, R., and Hajkova, V. (2014). Visualising components of regional innovation systems using self-organizing maps—evidence from european regions. *Technological Forecasting and Social Change*, 84:197–214.

Haselbeck, V., Kordilla, J., Krause, F., and Sauter, M. (2019). Self-organizing maps for the identification of groundwater salinity sources based on hydrochemical data. *Journal of Hydrology*.

Hoxby, C. and Avery, C. (2013). The missing "one-offs": The hidden supply of high-achieving, low-income students. *Brookings Papers on Economic Activity*, 44(1 (Spring)):1–65.

Hung, C. and Tsai, C.-F. (2008). Market segmentation based on hierarchical self-organizing map for markets of multimedia on demand. *Expert systems with applications*, 34(1):780–787.

Johnson, B. and Shneiderman, B. (1991). Tree-maps: A space-filling approach to the visualization of hierarchical information structures. In *Visualization, 1991. Visualization'91, Proceedings., IEEE Conference on*, pages 284–291. IEEE.

Kakatkar, C. and Spann, M. (2019). Marketing analytics using anonymized and fragmented tracking data. *International Journal of Research in Marketing*, 36(1):117–136.

Keivanpour, S. (2019). Adapting treemaps to student academic performance visualization. In *Science and Information Conference*, pages 720–728. Springer.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological cybernetics*, 43(1):59–69.

Kohonen, T. (2013). Essentials of the self-organizing map. *Neural networks*, 37:52–65.

Kohonen, T. and Maps, S.-O. (1995). Springer series in information sciences. *Self-organizing maps*, 30.

Kuncel, N. R. and Hezlett, S. A. (2007). Standardized tests predict graduate students' success. *Science*, 315(5815):1080–1081.

Lewison, D. M. and Hawes, J. M. (2007). Student target marketing strategies for universities. *Journal of College Admission*, 196:14–19.

Ma, X. and Frempong, G. (2013). Profiles of canadian postsecondary education dropouts. *Alberta Journal of Educational Research*, 59(2):141–161.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

Manski, C. and Wise, D. (1983). *College Choice in America*. Harvard University Press.

Manzi, J., Bravo, D., Del Pino, G., Donoso, G., Martínez, M., and Pizarro, S. (2008). Estudio acerca de la validez predictiva de los factores de selección a las universidades del consejo de rectores, admisiones 2003 a 2006. *Santiago, Chile*.

Marson, F. and Musse, S. R. (2010). Asutomatic real-time generation of floor plans based on squarified treemaps algorithm. *International Journal of Computer Games Technology*, 2010:7.

Marta Ferreyra, M., Avitabile, C., Botero Álvarez, J., Haimovich Paz, F., and Urzúa, S. (2017). *At a crossroads: higher education in Latin America and the Caribbean*. The World Bank.

McFarland, J., Hussar, B., Zhang, J., Wang, X., Wang, K., Hein, S., and Barmer, A. (2019). The condition of education 2019 (nces 2019-144). *US Department of Education*.

Miranda, M. A. and Guzmán, J. (2017). Análisis de la deserción de estudiantes universitarios usando técnicas de minería de datos. *Formación universitaria*, 10(3):61–68.

Molnar, C. (2019). *Interpretable Machine Learning*. https://christophm.github.io/interpretable-ml-book/.

Montgomery, D. C. (2017). *Design and analysis of experiments*. John wiley & sons.

Mostafa, M. M. (2010). Clustering the ecological footprint of nations using kohonen's self-organizing maps. *Expert Systems with Applications*, 37(4):2747–2755.

Nogales, A., García-Tejedor, Á. J., Sanz, N. M., and de Dios Alija, T. (2019). Competencies in higher education: A feature analysis with self-organizing maps. In *International Workshop on Self-Organizing Maps*, pages 80–89. Springer.

Ortega-Zamorano, F., Molina-Cabello, M. A., López-Rubio, E., and Palomo, E. J. (2016). Smart motion detection sensor based on video processing using self-organizing maps. *Expert Systems with Applications*, 64:476–489.

Paredes, R. D. and Pinto, J. I. (2009). ¿ el fin de la educación pública en chile? *Estudios de economía*, 36(1):47–66.

Park, J. S. and Lee, J. (2014). Segmenting green consumers in the united states: Implications for green marketing. *Journal of Promotion Management*, 20(5):571–589.

Polikar, R. (2012). Ensemble learning. In *Ensemble machine learning*, pages 1–34. Springer.

Pridmore, J. and Hämäläinen, L. E. (2017). Market segmentation in (in) action: Marketing and'yet to be installed'role of big and social media data. *Historical Social Research/Historische Sozialforschung*, pages 103–122.

Quesada-Pineda, H., Brenes-Bastos, M., and Smith, R. (2017). Assessing geographic information systems use in marketing applications for the wood products industry. *BioProducts Business*, pages 14–22.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*.

Robert, P. (2010). Social origin, school choice, and student performance. *Educational Research and Evaluation*, 16.

Rolando, R., Salamanca, J., and Lara, A. (2012). Retencioíon de 1er año en educacioón superior: Carreras de pregrado. *Santiago de Chile: MINEDUC*.

Rubio, A. B. (2011). Deserción universitaria en chile: incidencia del financiamiento y otros factores asociados. *Revista CIS*, 9(14):59–72.

Rundle-Thiele, S., Kubacki, K., Tkaczynski, A., and Parkinson, J. (2015). Using two-step cluster analysis to identify homogeneous physical activity groups. *Marketing Intelligence & Planning*, 33(4):522–537.

Saadatdoost, R., Sim, A. T. H., and Jafarkarimi, H. (2011). Application of self organizing map for knowledge discovery based in higher education data. In *2011 International Conference on Research and Innovation in Information Systems*, pages 1–6. IEEE.

Saarenvirta, G. (1998). Mining customer data. *dB2 Magazine*, 3(3):10–20.

Sallee, J., M, R. A., and Courant, P. (2008). On the optimal allocation of students and resources in a system of higher education. *The B.E. Journal of Economic Analysis & Policy*, 8(1):1–26.

Sanchez, A. D., Gutierrez, M. A., and Meneses, C. V. (2005). Using data mining to support the university decision process: A case in a chilean university. *AMCIS 2005 Proceedings*, page 350.

Sarra, A., Fontanella, L., and Di Zio, S. (2018). Identifying students at risk of academic failure within the educational data mining framework. *Social Indicators Research*, pages 1–20.

Segev, A. and Kantola, J. (2012). Identification of trends from patents using self-organizing maps. *Expert systems with applications*, 39(18):13235–13242.

Shapiro, D., Dundar, A., Wakhungu, P., Yuan, X., and Harrell, A. (2015). Completing college: A state-level view of student attainment rates. *Signature Report*, (8a).

Shieh, S.-L. and Liao, I.-E. (2012). A new approach for data clustering and visualization using self-organizing maps. *Expert Systems with Applications*, 39(15):11924–11933.

Shields, R. (2013). Globalization and international student mobility: A network analysis. *Comparative Education Review*, 57(4):609–636.

Shneiderman, B. (1990). Tree visualization with tree-maps: 2-d space-filling approach. *ACM Transactions on Graphics (TOG)*, 11(1):92–99.

Shneiderman, B. and Wattenberg, M. (2001). Ordered treemap layouts. In *infovis*, page 73. IEEE.

SIES (2014). Panorama de la educación superior en chile 2014.

Sirer, M. I., Maroulis, S., Guimera, R., Wilensky, U., and Amaral, L. A. N. (2015). The currents beneath the "rising tide" of school choice: An analysis of student enrollment flows in the chicago public schools. *Journal of policy analysis and management*, 34(2):358–377.

Siri, A. (2015). Predicting students' dropout at university using artificial neural networks. *Italian Journal of Sociology of Education*, 7(2).

Stinebrickner, R. and Stinebrickner, T. (2014). Academic performance and college dropout: Using longitudinal expectations data to estimate a learning model. *Journal of Labor Economics*, 32(3):601 – 644.

Strayer, W. (2002). The returns to school quality: College choice and earnings. *Journal of Labor Economics*, 20(3):475–503.

Team, R. C. et al. (2018). R: A language and environment for statistical computing.

Tu, Y. and Shen, H.-W. (2007). Visualizing changes of hierarchical data using treemaps. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1286–1293.

Ultsch, A. (1990). Kohonen's self organizing feature maps for exploratory data analysis. *Proc. INNC90*, pages 305–308.

Vellido, A., Lisboa, P., and Meehan, K. (1999). Segmentation of the on-line shopping market using neural networks. *Expert systems with applications*, 17(4):303–314.

Venegas-Muggli, J. I. (2019). Higher education dropout of non-traditional mature freshmen: the role of sociodemographic characteristics. *Studies in Continuing Education*, pages 1–17.

Vesanto, J. and Alhoniemi, E. (2000). Clustering of the self-organizing map. *IEEE Transactions on neural networks*, 11(3):586–600.

Vossensteyn, J. J., Kottmann, A., Jongbloed, B. W., Kaiser, F., Cremonini, L., Stensaker, B., Hovdhaugen, E., and Wollscheid, S. (2015). Dropout and completion in higher education in europe: Main report.

Webster, C. M. and Morrison, P. D. (2004). Network analysis in marketing. *Australasian Marketing Journal*

(AMJ), 12(2):8–18.

Wehrens, R., Buydens, L. M., et al. (2007). Self-and super-organizing maps in r: the kohonen package. *Journal of Statistical Software*, 21(5):1–19.

Wei, J.-T., Lin, S.-Y., Weng, C.-C., and Wu, H.-H. (2012). A case study of applying lrfm model in market segmentation of a children's dental clinic. *Expert Systems with Applications*, 39(5):5529–5533.

Zemsky, R. and Oedel, P. (1983). The structure of college choice.

Zhou, J., Zhai, L., and Pantelous, A. A. Market segmentation using high-dimensional sparse consumers data. *Expert Systems with Applications*, 145(113136).

Zhu, B., He, C., and Jiang, X. (2015). Customer choice prediction based on transfer learning. *Journal of the Operational Research Society*, 66(6):1044–1051.

Zvoch, K. (2006). Freshman year dropouts: Interactions between student and school characteristics and student dropout status. *Journal of education for students placed at risk*, 11(1):97–117.

Appendix

Table 1: Descriptive summary of variables studied, at cluster level.

	Cluster			
	1 ($n = 1332$)	2 ($n = 3631$)	3 ($n = 1773$)	4 ($n = 1130$)
Variable ($\overline{X} \pm$ S.D.)				
HSQ	724 (52.8)	680 (59.9)	621 (68.8)	645 (68.7)
SAT Language	691 (51.7)	625 (51.1)	586 (50.5)	561 (54.0)
SAT Mathematics	738 (53.2)	678 (40.7)	642 (33.5)	606 (33.0)
SAT Science	691 (46.9)	635 (40.7)	590 (42.8)	554 (45.7)
SAT Weigthed	725 (32.3)	670 (28.7)	623 (24)	606 (31.5)
Credits Approved	30.4 (8.8)	22.2 (12.7)	10.5 (12.1)	2.9 (6.0)
Distance (km)	35.1 (40.9)	37.5 (40.4)	14.6 (25.3)	45.4 (41.4)
Sex (*Frequency* (%))				
Female	381 (28.6%)	714 (19.7%)	331 (18.7%)	305 (27.0%)
Male	951 (71.4)	2917 (80.3%)	1442 (81.3)	825 (73.0%)
High school graduation dependency (*Frequency* (%))				
Private ($n = 1903$)	620 (46.6%)	828 (22.8%)	353 (19.91%)	102 (9.1%)
Subsidized ($n = 4099$)	606 (45.5%)	1877 (19.9%)	1149 (64.81%)	467 (41.3%)
Public ($n = 1864$)	106 (7.96%)	926 (25.5%)	271 (15.3%)	561 (49.6%)
Undergraduate Career (*Frequency*)				
Civil Engineering	216	374	10	4
Civil Engineering – Common Plan	11	244	414	341
Aerospace Engineering	36	92	31	17
Biomedical Engineering	76	286	219	70
Materials Science Engineering	2	91	205	185
Mining Engineering	40	185	19	19
Electrical Engineering	32	260	140	48
Electronic Engineering	65	284	170	29
Telecommunications Engineering	8	53	212	263
Industrial Engineering	239	529	15	12
Informatics Engineering	44	248	180	91
Mechanical Engineering	127	366	33	17
Metallurgical Engineering	57	420	105	32
Chemical Engineering	379	199	20	2

Table 2: Distribution of students by cluster in the 50 schools with the highest number of students contributed between 2005 and 2017.

			Cluster				
Id	School	High School Dependency	C1	C2	C3	C4	Total
1	Colegio Salesiano De Concepcion	Subsidised	48	124	202	46	420
2	Sagrados Corazones	Private	119	95	41	6	261
3	Liceo Enrique Molina Garmendia	Public	11	134	62	42	249
4	Colegio Concepcion San Pedro	Private	53	87	48	6	194
5	Colegio Concepcion Pedro De Valdivia	Private	49	78	43	15	185
6	Liceo Los Angeles	Public	22	104	10	40	176
7	Colegio Concepcion De Chillan	Subsidised	33	101	14	14	162
8	Liceo La Asuncion Talcahuano	Subsidised	12	67	69	14	162
9	Colegio San Ignacio Concepcion	Subsidised	25	74	51	5	155
10	Colegio San Rafael Arcangel Los Angeles	Subsidised	26	90	8	26	150
11	Colegio Seminario Padre Alberto Hurtado Chillan	Subsidised	44	70	18	18	150
12	Charles De Gaulle	Private	52	58	18	3	131
13	Instituto De Humanidades Alfredo Silva S	Private	36	61	27	6	130
14	Liceo Mauricio Hochschild San Pedro De La Paz	Subsidised	3	33	57	30	123
15	Liceo De Niñas Concepcion	Public	6	35	49	29	119
16	Liceo San Agustin De Concepcion	Subsidised	24	47	42	6	119
17	Colegio Adventista De Concepcion	Subsidised	10	40	58	8	116
18	Saint Johns School Concepcion	Private	30	69	12	3	114
19	Colegio Almondale San Pedro De La Paz	Subsidised	22	66	22	3	113
20	Colegio Aleman De Concepcion	Private	45	46	13	4	108
21	Liceo Almirante Pedro Espina Ritchie Talcahuano	Public	5	48	25	29	107
22	Kingston College Concepcion	Private	25	42	23	10	100
23	Liceo Aleman Del Verbo Divino Los Angeles	Private	40	33	16	4	93
24	Colegio Etchegoyen Talcahuano	Subsidised	14	41	32	4	91
25	Colegio Bicentenario Republica Del Brasil	Public	6	50	19	15	90
26	The Wessex School Concepcion	Private	26	48	13	3	90
27	Colegio San Pedro Nolasco	Private	13	35	27	12	87
28	Liceo Comercial Enrique Oyarzun Mondaca	Public	0	26	22	37	85
29	Liceo Coronel Antonio Salamanca	Public	4	49	6	22	81
30	Liceo Vicente Alberto Palacios Valdes	Public	6	45	8	22	81
31	Instituto Santa Maria Chillan	Subsidised	29	38	6	5	78
32	Liceo Politecnico Heroes De La Concepcion Laja	Public	4	52	0	22	78
33	Colegio Carmela Romero De Espinosa	Subsidised	17	38	21	1	77
34	Colegio San Vicente De Paul Chillan	Subsidised	15	46	4	12	77
35	Colegio Metodista Concepcion	Subsidised	12	29	24	2	67
36	Colegio Adventista De Chile	Subsidised	10	27	7	16	60
37	Colegio Aurora De Chile Chiguayante	Subsidised	4	24	29	3	60
38	Colegio Creacion Concepcion	Subsidised	5	23	25	6	59
39	Instituto Santa Maria De San Carlos	Subsidised	13	33	7	5	58
40	Liceo Polivalente Mariano Latorre	Public	5	36	1	14	56
41	Colegio Teresiano Los Angeles	Subsidised	16	27	2	10	55
42	Arturo Prat Chacon	Private	11	31	10	2	54
43	Colegio Concepcion Chiguayante	Subsidised	6	31	12	5	54
44	Colegio Padre Manuel Dalzon	Subsidised	3	18	27	6	54
45	Liceo Gabriela Mistral Cañete	Subsidised	12	35	2	3	52
46	Liceo Narciso Tondreau Chillan	Public	3	35	1	13	52
47	Colegio Bautista De Concepcion	Private	14	22	8	7	51
48	Colegio Del Sagrado Corazon	Subsidised	8	27	8	5	48
49	The Thomas Jefferson School Talcahuano	Private	14	22	8	2	46
50	Liceo Isidora Ramos De Gajardo Lebu	Public	4	25	1	15	45