UNIVERSIDAD DE LA REPÚBLICA
Facultad de Ciencias Económicas y de Administración
Instituto de Estadística

Satisfacción Estudiantil: análisis desde una perspectiva multivariante.

Ramón Álvarez-Vaz; **Elena Vernazza**

Diciembre, 2017

Serie Documentos de Trabajo

DT (17/3) - ISSN : 1688-6453

Forma de citación sugerida para este documento:

Álvarez-Vaz, Ramón y Vernazza, Elena (2017). Satisfacción Estudiantil: análisis desde una perspectiva multivariante. [en línea].
Serie Documentos de Trabajo, DT (17/3). Instituto de Estadística, Facultad de Ciencias Económicas y de Administración, Universidad de la República, Uruguay.

Satisfacción Estudiantil: análisis desde una perspectiva multivariante.

Ramón Álvarez-Vaz [1]; Elena Vernazza [2]
Departamento de Métodos Cuantitativos, Instituto de Estadística, Facultad de Ciencias Económicas y de Administración, Universidad de la República

RESUMEN

En este trabajo se estudian las principales características de la construcción de la Satisfacción Estudiantil, en los cursos de grado de la Facultad de Ciencias Económicas y de Administración, Universidad de la República, Uruguay, a través de la utilización y comparación de dos técnicas de análisis de datos multivariantes: Análisis Jerárquico de Clusters y Análisis de Clases Latentes.

Los datos utilizados para la aplicación presentada en este trabajo provienen de una encuesta aplicada sobre una muestra de estudiantes de grado de la Facultad, en el año 2009. Dicho cuestionario, presenta una estructura de bloques de preguntas; el primero contiene las variables que permitirán realizar una caracterización sociodemográfica de los estudiantes. Por otra parte, se presentan las variables del modelo ECSI (European Customer Satisfaction Index) que serán las utilizadas para la caracterización de la Satisfacción Estudiantil.

Las variables manifiestas consideradas como insumo para la construcción y caracterización de la Satisfacción Estudiantil son las siguientes seis: expectativas (E) de los estudiantes al ingresar al centro de estudios, la imagen (I) que tienen de éste, la calidad de la enseñanza recibida (CSA) y de los servicios brindados (CSF), las necesidades y deseos personales con respecto a la Facultad (ND) y el valor percibido (VP).

Los resultados presentados surgen, por un lado, de considerar que efectivamente existe una variable que refiere a la Satisfacción Estudiantil y que ésta queda definida, a partir de la interacción de las 6 variables manifiestas, por cuatro clases latentes.
Por otra parte, en lo que refiere a los clusters, se propone agrupar a los estudiantes en tres grupos, a partir del análisis de los resultados que surgen de clusterizar a través del método Ward.

Se presenta, además, la comparación de los resultados obtenidos con ambas técnicas.

[1] *email:* ramon@iesta.edu.uy, ORCID: 0000-0002-2505-4238
[2] *email:* evernazza@iesta.edu.uy, ORCID: 0000-0003-3123-2165

Palabras clave: Clases latentes, clusters, independencia condicional, probablidad a posteriori, satisfacción estudiantil.

CÓDIGOS JEL:
C18, C38 **Clasificación MSC2010**:62H17,62P25,62H30

Student Satisfaction: Analysis from a multivariate perspective

Ramón Álvarez [1]; Elena Vernazza [2]
Departamento de Métodos Cuantitativos, Instituto de Estadística, Facultad de Ciencias Económicas y de Administración, Universidad de la República

ABSTRACT

In this work we study the main characteristics of the construction of Student Satisfaction, in the undergraduate courses of the Facultad de Ciencias Económicas y de Administración, Universidad de República, Uruguay, using and comparing two analysis techniques of multivariate data: Hierarchical Analysis of Clusters and Analysis of Latent Classes.

The data used for the application presented in this work comes from a survey applied to a random sample of bachelor students in 2009. This questionnaire presents a structure of blocks of questions; the first contains the variables that will allow a sociodemographic characterization of the students. On the other hand, we present the variables of the ECSI model (European Customer Satisfaction Index) that will be used for the characterization of Student Satisfaction.

The manifest variables considered as input for the construction and characterization of the Student Satisfaction are the following six: expectations (E) of the students when entering the study center, the image (I) that they have of this, the quality of the received education (CSA) and of the services provided (CSF), the personal needs and desires with respect to the school (ND) and the perceived value (VP).

The results presented, on the one hand, to consider that there is indeed a variable that refers to Student Satisfaction and that it is defined, from the interaction of the six manifest variables, by four latent classes.

[1] *email:* ramon@iesta.edu.uy, ORCID: 0000-0002-2505-4238
[2] *email:* evernazza@iesta.edu.uy, ORCID: 0000-0003-3123-2165

On the other hand, as regards the clusters, it is proposed to group students into three groups, based on the analysis of the results that arise from clustering through the Ward method.

It also presents the comparison of the results obtained with both techniques.
Key words: Latent classes, clusters, conditional independence, posteriori probability, student satisfaction
JEL CODES: C18, C38
Mathematics Subject Classification MSC2010: 62H17,62P25,62H30 .

1. Introducción

Conocer el nivel de satisfacción de los clientes, con un determinado servicio que se les brinda resulta fundamental como insumo en la toma de decisiones que tengan como objetivo primordial mantener o mejorar, en caso de que sea necesario, aquellos aspectos que se entiende determinan la *Satisfacción*.

Vinculando esta idea con la educación universitaria, se toma lo propuesto por Alves y Raposo (Alves y Raposo, 2004), quienes plantean: "Sólo con la satisfacción de los alumnos se podrá alcanzar el éxito escolar, la permanencia de los estudiantes en la institución y, sobre todo, la formación de una valoración positiva boca a boca. En este sentido, es extremamente importante encontrar formas fiables de medir la satisfacción del alumno en la enseñanza universitaria, permitiendo así a las instituciones de enseñanza conocer su realidad, compararla con la de los otros competidores y analizarla a lo largo del tiempo".

En este trabajo se considera a los estudiantes universitarios que concurren a la Facultad de Ciencias Económicas y Administración, FCEA, Universidad de la República, como "clientes" y se determina que el "servicio" que se les brinda es el de la educación de nivel terciario.

La información necesaria para poder establecer cómo se construye el concepto de *Satisfacción*, se obtiene a través de la aplicación de un cuestionario formado por apartados de preguntas que conforman el modelo ECSI (European Customer Satisfaction Index). Sobre este instrumento, y a través del Análisis de Clases Latentes y del Análisis de Clusters, se analiza cómo se construye la *Satisfacción Estudiantil* y cómo se agrupan los estudiantes en función de las variables consideradas (Vernazza, 2013), (Álvarez-Vaz *et al.*, 2016).

El presente trabajo se estructura en cuatro secciones. En primera instancia se presenta la metodología utilizada, haciendo referencia a las principales características de las técnicas empleadas. La sección 3 comienza con una descripción de los datos, se presenta el diseño muestral y las variables utilizadas. A continuación se exponen los principales resultados obtenidos a través de un Análisis de Clases Latentes y de un Análisis de Cluster.

Dichos resultados son presentados, en forma conjunta y comparada, en la sección 6. Por último, en la sección 7, se plantean algunas consideraciones finales y propuestas de líneas de trabajo e investigación a futuro.

2. Metodología

En esta sección se presentan las principales características de las dos técnicas estadísticas utilizadas: Análisis de Clases Latentes y Análisis Clusters.
En lo que refiere a la primera técnica, se expone brevemente su principal objetivo y se plantea que puede ser vista desde diferentes enfoques estadísticos: modelado, descripción multivariante e incluso como un problema de clasificación no supervisada. Por último, se define formalmente del modelo, su forma de estimación y de validación.
En cuanto al Análisis de Clusters se presenta la técnica haciendo especial énfasis en las alternativas jerárquicas, en particular, se presenta en detalles la metodología Ward.

2.1. Análisis de Clases Latentes

Existen muchas técnicas para trabajar con datos categóricos multivariados dentro de las que se encuentran, como una de las más destacadas desde una perspectiva de modelado, los Modelos Log-Lineales. Por otra parte, dentro de la estadística descriptiva multivariante, el Análisis de Correspondencias Múltiples se presenta como la herramienta más utilizada para el tratamiento de grandes tablas con atributos categóricos.

Generalmente, al trabajar con datos de esta naturaleza, resulta de interés investigar posibles fuentes de confusión entre las variables observadas, identificar y caracterizar clusters de individuos y aproximar la distribución de las observaciones a través de las variables en estudio (Linzer y Lewis, 2011). Existe una técnica que contempla todas estas situaciones: Análisis de Clases Latentes (ACL) o Modelos de Clases Latentes (MCL) (Lazarsfeld, 1950), (Everitt, 1984), (Bandeen-Roche *et al.*, 1997), (Hagenaars, 2002), (Agresti, 2013).

El ACL busca segmentar la tabla/hipercubo de contingencia creado a partir de las variables observadas/*manifiestas*, por una variable no observada/*latente*, con la característica de que se supone que las respuestas a todas las variables manifiestas son estadísticamente independientes con respecto a los valores de la variable latente. Esta restricción supone lo que se conoce como *independencia local o condicional*. De este manera el modelo asocia, en términos de probabilidad, a cada individuo a una *clase* latente. Se puede predeterminar, por lo tanto, el valor esperado con el que esta observación responde a cada variable observada. Si bien el modelo estimado no estipula el número de *clases* latentes, pueden usarse varios estadísticos de bondad de ajuste (GOF), para poder hacer una evaluación tanto teórica como empírica de la cantidad de clases a considerar.

Estos modelos pueden ser vistos como un tipo de modelo de mezcla finita, teniendo en consideración que la variable no observada es nominal (marca la pertenencia a una clase). En este sentido, las distribuciones de los componentes en la mezcla son las tablas de clasi-

ficación cruzada de igual dimensión que la tabla observada de variables manifiestas, donde el supuesto de independencia condicional permite calcular la frecuencia en cada celda de cada tabla de componentes como el producto de las respectivas frecuencias marginales de las clases condicionales. Los parámetros estimados por el ACL serán la proporción de observaciones en cada clase latente y las probabilidades de observar cada respuesta de cada variable manifiesta, condicionada a la clase latente.

Además, es posible obtener una aproximación (estimación de la densidad) de la distribución de casos a través de las celdas de la tabla observada, como una suma ponderada de estas tablas de componentes, con la característica de que las observaciones con perfiles de respuestas similares, en las variables observadas/manifiestas, tienden a agruparse dentro de la misma clase latente.

Otra ventaja de este método es que puede verse como un modelo de regresión y que, por lo tanto, sería posible incluir variables predictivas para la membresía de cada observación a una clase latente (Bandeen-Roche *et al.*, 1997).

Existen antecedentes de estudios con este tipo de variables en disciplinas como la economía y la psicología. En particular, en el trabajo "Segmentación de la población española según su grado de concienciación ecológica mediante modelos de variables latentes", los autores presentan una segmentación de los hábitos de consumo en función de su grado de concienciación ecológica mediante técnicas estadísticas de ACL (Sánchez-Rivero, 2001).

Por otra parte, en el trabajo "Modelos De Clases Latentes Para Definir Perfiles Conductuales en Niños De 4 y 5 Años" (Castro López y Oliva Zarate, 2011) sus autores elaboran perfiles conductuales en niños de 4 o 5 años de México aplicando ACL sobre los resultados del test de screening "Child Behavior Check List" (CBCL).

En otras áreas de la salud como la epidemiología, se presenta el trabajo "Análisis de clases latentes en tablas poco ocupadas: consumo de alcohol, tabaco y otras drogas en adolescentes" (Carlomagno-Araya y Sepúlveda, 2010), en el que sus autores presentan una segmentación del tipo de consumo de drogas en jóvenes de Costa Rica.

2.1.1. Definición del Modelo

Se considera un modelo en el que se observan J variables categóricas politómicas (variables *manifiestas*) tal que cada una tiene K_j posibles respuestas, para los $i = 1, 2....N$ individuos.

El modelo de clase latente aproxima la distribución conjunta observada de las variables

manifiestas como la suma ponderada (por un número finito R) de las tablas de clasificación cruzada.

Y_{ijk} será el valor observado de las J variables manifiestas para el individuo i, tal que $Y_{ijk} = 1$ si el individuo i da la respuesta k de la variable j y $Y_{ijk} = 0$ en otro caso, con $j = 1....J$ y $k = 1....K_j$ y π_{jrk} representará la probabilidad condicional de que una observación en la clase $r = 1, ..., R$ produzca el *k-ésimo* resultado de la variable *j-ésima*. Dentro de cada clase, para cada variable manifiesta, se cumple:

$$\sum_{k}^{K_j} \pi_{jrk} = 1.$$

Por otra parte, p_r corresponderá a las proporciones a partir de las cuales serán generados los pesos para la suma ponderada de las tablas de clasificación ($\sum_r^{max} p_r = 1$). En este sentido, considerando que estos p_r representan la probabilidad *incondicional* de que un individuo pertenezca a una clase (antes de tomar en cuenta el valor de Y_{ijk}), p_r será denominado probabilidades a *priori* de la membresía a cada clase latente.

La probabilidad de que un individuo i en la clase r genere un conjunto J de resultados en las variables manifiestas, asumiendo independencia condicional de los resultados Y dado la pertenencia a una clase dada, es:

$$f(Y_i; \pi_r) = \prod_{j=1}^{J} \prod_{k=1}^{K_j} (\pi_{jrk})^{Y_{ijk}}. \tag{1}$$

Además, la función de densidad es:

$$P(Y_i \mid \pi, p) = \sum_{r=1}^{R} p_r \prod_{j=1}^{J} \prod_{k=1}^{K_j} (\pi_{jrk})^{Y_{ijk}} \tag{2}$$

De esta manera se tienen 2 parámetros a estimar por el modelo: p_r y π_{jrk}. Dadas $\hat{p}_r$ y $\hat{\pi}_{jrk}$, las probabilidades a *posteriori* de que cada individuo pertenezca a una clase latente, condicionada a los valores observados de las variables manifiesta, queda determinada:

$$\hat{P}(r_i \mid Y_i) = \frac{\hat{p}_r f(Y_i; \hat{\pi}_r)}{\sum_{q=1}^{R} \hat{p}_q f(Y_i; \hat{\pi}_q)} \tag{3}$$

con $r_i = 1, ..., R$.

Hay que tener en cuenta que las $\hat{\pi}_{jrk}$ son estimaciones de las probabilidades de los resultados condicionales en la clase r. También es importante observar que el número de

parámetros independientes estimados aumenta rápidamente con R, J y K_j. Dados estos valores, el número de parámetros a estimar es $R\sum_j(K_j-1)+(R-1)$. Este último resultado puede producir una situación no deseada, ya que cuando este número excede el número total de observaciones, o una menos que el número total de celdas en la tabla de clasificación cruzada de las variables manifiestas, el modelo no puede ser identificado.

2.1.2. Estimación de parámetros

Los modelos de clase latente pueden estimarse mediante máxima verosimilitud, donde la log-verosimilitud es:

$$lnL = \sum_{1}^{N} ln \sum_{1}^{R} p_r \prod_{j=1}^{J} \prod_{k=1}^{K_j} (\pi_{jrk})^{Y_{ijk}}. \tag{4}$$

Dicha verosimilitud será maximizada con respecto a p_r y π_{jrk} a través del algoritmo *EM* (Dempster *et al.*, 1977), (McLachlan, 2000). Como con cualquier modelo de mezcla finita, el algoritmo *EM* se puede aplicar en virtud de que la membresía a la clase de cada individuo es desconocida, por lo que se trata como un problema de datos faltantes. El algoritmo trabaja en forma iterativa en 2 fases a partir de valores iniciales arbitrarios de $\hat{p}_r$ y $\hat{\pi}_{jrk}$, los que se etiquetan como $\hat{p}_r^{anterior}$ y $\hat{\pi}_{jrk}^{anterior}$.

1. En la fase de esperanza (E), se calcula la probabilidad de membresía a la clase *latente* usando la ecuación (3), sustituyendo en los valores $\hat{p}_r^{anterior}$ y $\hat{\pi}_{jrk}^{anterior}$.

2. En la fase de maximización (M) los parámetros estimados se actualizan maximizando la log-verosimilitud dada la probabilidad a *posteriori* ($\hat{P}(r_i \mid Y_i)$).

La nueva probabilidad a *priori*, será:

$$\hat{p}_r^{nueva} = \frac{1}{N} \sum_{i=1}^{N} \hat{P}(r_i \mid Y_i) \tag{5}$$

y la nueva probabilidad condicional será:

$$\hat{\pi}_{jr}^{nueva} = \frac{\sum_{i=1}^{N} Y_{ij} \hat{P}(r_i \mid Y_i)}{\sum_{i=1}^{N} \hat{P}(r_i \mid Y_i)} \tag{6}$$

En la ecuación (6), $\hat{\pi}_{jr}^{nueva}$ es el vector de longitud K_j de las probabilidades condicionales para la *j-ésima* variable manifiesta; y por otra parte Y_{ij} es la matriz $N \times K_j$ de resultados para Y_{ijk} para esa variable. Como todo proceso iterativo este algoritmo repite las 2 fases

sustituyendo el valor viejo por el nuevo, hasta alcanzar un máximo o hasta que el incremento que tiene la log-verosimilitud sea menor a un cierto umbral previamente establecido.

La librería *poLCA* (Linzer y Lewis, 2011), utilizada en este trabajo, es actualmente considerada como la mas versátil ya que permite trabajar con variables categóricas politómicas y evaluar modelos de clase latente con covariables, mientras que otras librerías como *e1071* (Meyer *et al.*, 2017), o *randomLCA* (Beath, 2017), solo permiten trabajar con variables dicotómicas.
La librería *poLCA* tiene, además, una característica muy interesante y es que a partir de la naturaleza iterativa del algoritmo *EM*, permite estimar el MCL aún cuando alguna de las observaciones tiene datos faltantes en algunas de las variables observadas. Para determinar el producto en la ecuación (1) y la suma en el numerador de la ecuación (6), la función *poLCA* excluye del cálculo cualquier variable manifiesta con observaciones faltantes. Los probabilidades a *priori* se actualizan en la ecuación (3) usando tantas variables manifiestas como se observan para cada individuo.

Por último, cabe destacar que la aplicación de este método de estimación depende de: los valores iniciales elegidos para $\hat{p}_r^{anterior}$ y $\hat{\pi}_{jrk}^{anterior}$ y de la complejidad del modelo que se estima, por lo que el algoritmo *EM* puede encontrar un máximo local de la función log-verosimilitud, en lugar del máximo global deseado, con lo cual es recomendable estimar más de una vez.

2.1.3. Criterios de selección y validación del modelo

Tal como fuera mencionado previamente, la estimación a través de ACL no estipula una cantidad de *clases* latentes, sin embargo una de las ventajas de esta técnica, a diferencia de varias de las técnicas de clusterización más comúnmente utilizadas, es la variedad de herramientas existentes para determinar dicha cantidad.

En la mayoría de los casos será necesario realizar un análisis exploratorio que permita decidir la cantidad de clases latentes presentes en el problema en estudio. Este proceso comienza presentando el modelo más general posible, es decir, un modelo con independencia completa que determina una sola clase. Una vez estimado dicho modelo, el número de clases se va incrementando de una en una hasta encontrar un modelo que resulte un "modelo adecuado".

Agregar una clase al modelo mejorará el ajuste, pero incorporará ruido y parámetros a estimar, por lo que será necesario tener en consideración un criterio de parsimonia que establezca un equilibrio entre la mejora del ajuste y la cantidad de parámetros que se incorporan al aumentar una clase en el modelo. El criterio de parsimonia utilizado en este

trabajo será el del mínimo BIC (Bayesian information criterion).

Serán utilizados, además, para determinar la cantidad de clases latentes los estadísticos: χ^2 y G^2.

Sea q_c la cantidad de casos observados en la celda c de la tabla de contingencia ($c = 1...C$) con $C = \prod_{K_j}$, la proporción de casos esperados en cada celda de la tabla, será calculado sustituyendo $\hat{p}_r$ y $\hat{pi}_{jrk}$ en la ecuación 2.

Además, sea y_c la secuencia de J respuestas correspondientes a la celda c de la tabla de contingencia, tal que $y_{cjk} = 1$ si la celda c contiene la respuesta k para la variable j y 0 en otro caso, su función de probabilidad queda dada por:

$$\tilde{P}(y_c) = \sum_{r=1}^{R} \hat{p}_r \prod_{j=1}^{J} \prod_{k=1}^{K_j} (\hat{\pi}_{jrk})^{y_{cjk}}$$

El número esperado de casos en la celda c es $\tilde{Q}_c = N\tilde{P}(y_c)$, a partir del cual quedan definidos los dos estadísticos utilizados:

$$\chi^2 = \sum_{c=1}^{C} \frac{(q_c - \tilde{Q}_c)^2}{\tilde{Q}_c}$$

y

$$G^2 = 2 \sum_{c=1}^{C} q_c log(q_c/\tilde{Q}_c)$$

El modelo seleccionado será aquel que cumpla un equilibrio entre mínimo χ^2 o G^2 y cantidad de parámetros a estimar.

2.2. Análisis de Clusters

Existen variados métodos de clusterización que se agrupan, generalmente, en 2 categorías: Jerárquicos y No jerárquicos. Sobre ellos se pueden aplicar distintos tipos de distancias, tomando en cuenta métricas diferentes en función del tipo de variable considerada (Maechler *et al.*, 2016).En este trabajo será utilizado un método de clusterización Jerárquico.

2.2.1. Métodos Jerárquicos - Método de WARD

Los métodos jerárquicos se caracterizan por generar una serie de particiones encajadas y requieren la definición de una distancia. Al comienzo, cada objeto es asignado a su

propio grupo y se inicia un proceso iterativo: en cada etapa se unen los dos grupos "más similares", continuando hasta que solo quede un grupo. En cada etapa las distancias definidas entre las agrupaciones se recalculan por la fórmula disimilitud de Lance-Williams actualizándose de acuerdo al método de agrupación particular que se utilice.

En este trabajo será utilizada una metodología Jerárquica, en particular, el método de Ward. Este consiste en descomponer la variación total en variación dentro de los grupos (*within*) y variación entre los grupos (*between*). Al estar frente a una partición dada, el método unirá aquellos grupos que produzcan el efecto de hacer mínima la variación *within* en la nueva partición.

En formato matricial:

$$T = W + B \tag{7}$$

donde T es la matriz de varianzas y covarianzas del total, W la matriz de varianzas y covarianzas dentro de los grupos y B la matriz de varianzas y covarianzas entre grupos.

A continuación, se presentan algunas de las reglas más utilizadas para determinar con qué cantidad de grupos trabajar (Blanco *et al.*, 2006).

R cuadrado Establece la relación entre la variación explicada y la variación total, tal que la variación explicada es representada por la estructura de grupos hallada en cada nivel.

$$R^2 = 1 - \frac{\sum_{k=1}^{K}\sum_{i=1}^{n_k}\sum_{j=1}^{J}(x_{(ij(k))} - \overline{x}_{kj})^2}{\sum_{i=1}^{I}\sum_{j=1}^{J}(x_{(ij)} - \overline{x}_{j})^2} \tag{8}$$

En cada etapa de particiones encajadas se observa el valor del indicador y el incremento que se produce en él al pasar de k a $k+1$ grupos. Si de un paso al otro el aporte (de variación explicada) deja de ser significativo, se opta por trabajar con k grupos.

Regla de Calinski (*pseudoF*)

$$pseudoF = \frac{R^2/(k-1)}{(1-R^2)/(n-k)} \tag{9}$$

El estadístico *pseudoF* no tiene distribución F pero, empíricamente, se han determinado algunas reglas que contribuyen a su utilización:

- Si el indicador crece monótonamente al crecer el número de grupos $k \Rightarrow$ no se puede determinar una estructura clara.

- Si el indicador decrece monótonamente al crecer el número de grupos $k \Rightarrow$ no se puede determinar claramente la estructura de grupos, pero se puede decir que existe una estructura jerárquica.
- Si el indicador crece, llega a un máximo y luego decrece $\Rightarrow$ la población presenta un número definido de grupos en ese máximo.

Test de Duda-Hart (DH) y pseudo t^2

El *pseudo* t^2, al igual que el *pseudoF*, es un indicador útil para determinar el número de grupos, pero no tiene distribución exacta *t-student*.
Está relacionado con el indicador planteado por Duda-Hart, que compara las trazas de las matrices de varianzas intragrupales G y L con la traza de la matriz de varianzas que surge al unir los grupos G y L.

$$DH = \frac{trW_G + trW_L}{trW_{GL}} \tag{10}$$

Lo que se intenta con estos indicadores es determinar la importancia de fusionar dos grupos considerando, en cada paso, los candidatos a unirse. Se trata de determinar en cada paso si la disminución en la suma de cuadrados residuales (variación intragrupos, o variación en los grupos) como resultado de pasar de k a $k+1$ grupos es significativa o no. Esto significa que el incremento en la heterogeneidad al unir los grupos es muy grande y por tanto no es conveniente su unión.

3. Aplicación

En esta sección se presentan, en forma resumida, los datos con los que se trabajó, describiendo el diseño muestral empleado y el cuestionario utilizado y como a partir de las variables observadas se agregan formando 6 variables del ECSI .

3.1. Diseño muestral

La aplicación que se presentará en este trabajo fue realizada sobre los datos obtenidos mediante la aplicación de un cuestionario sobre una muestra probabilística a estudiantes de los cursos superiores de la FCEA, en el año 2009. En esta sección se presentan las principales características del diseño muestral utilizado.

La muestra fue seleccionada en base a un marco muestral que se construyó a partir de las inscripciones a cursos de FCEA en 2009. El diseño muestral usado fue estratificado

por conglomerados en dos etapas y presentó las siguientes características: en una primera instancia se formaron seis estratos (en base a una clasificación desarrollada en conjunto por investigadores de la cátedra de Metodología de Investigación y del Instituto de Estadística, FCEA, Universidad de la República (IESTA)) que corresponden aproximadamente a cada uno de los cinco años en los en los que podía estar cada estudiante en el 2009. Adicionalmente, se propone un 6to estrato para un grupo reducido de materias que corresponden únicamente a la Licenciatura en Administración. Una vez conformados los estratos, se determina que la muestra total se repartirá en forma proporcional a la matrícula de cada estrato.

Al tener definidas las unidades de muestreo, se procede a seleccionar la muestra, proceso que presentó las siguientes etapas:

1. Se sortean los grupos prácticos de cada materia en cada estrato con probabilidad proporcional a la matrícula de cada grupo (conglomerado)

2. Mediante muestreo aleatorio simple (MAS), se seleccionan la misma cantidad de estudiantes en cada grupo seleccionado en la primera etapa. La cantidad de estudiantes de cada grupo es la misma en los seis estratos.

De esta manera se tiene un diseño muestral que presenta varias ventajas: por un lado, es muy sencillo de implementar en la práctica, ya que no se debe controlar un número diferente de unidades finales de muestreo (estudiantes) en cada grupo y estrato y, por otro, las probabilidades de inclusión que se deben usar para las estimaciones son aproximadamente constantes en los seis estratos, haciendo más sencillos los cálculos.
La muestra finalmente queda conformada por estudiantes que provienen de 60 grupos prácticos (repartidos en forma proporcional en los 6 estratos). Se sortean 12 estudiantes por grupo, lo que determina un tamaño de muestra de 720 estudiantes.

La siguiente tabla (tabla 1) muestra como es la asignación de la muestra de grupos prácticos en los seis estratos.

Tabla 1: Cantidad de grupos prácticos por estrato.

Estrato	1	2	3	4	5	6	Total
# grupos prácticos	21	15	9	9	4	2	60

Con la muestra seleccionada, se procedió a realizar el relevamiento de los datos que culminó con 647 encuestas realizadas. Esto determina una tasa de cobertura de la muestra de $647/720 = 90\,\%$.

En función de esto, al momento de calcular los pesos muestrales, lo primero que se hizo fue analizar el 10 % de estudiantes que quedó sin encuestar, con el objetivo de evaluar si es posible considerarlos como una muestra aleatoria de los 720 estudiantes originales, descartando de esta manera un sesgo de selección. Considerando como variables fundamentales el estrato, la edad y el sexo de los estudiantes, se constató que éstas no estaban asociadas a ese 10 % que quedó sin encuestar, es decir que ninguno de esos 3 atributos estaban sub o sobre representados en la muestra final efectivamente realizada.
Otros dos aspectos a tener en cuenta, previo al cálculo de los expansores, son los siguientes: por un lado se debe tener en cuenta la existencia de multiplicidad en el marco muestral debido a que hay un número diferente de matrículas correspondientes a cada estudiante, lo que impacta en la probabilidad de selección ya que la unidades primarias de muestreo son conglomerados de matrículas y no de estudiantes, es decir, hay estudiantes que están repetidos y pueden ser encontrados en más de una materia. Por último, debe ser tenido en cuenta el hecho de que la distribución por sexo y edad presente en la muestra definitiva no es la distribución poblacional, lo cual genera la necesidad de aplicar un proceso de calibración mediante pos-estratificación.

3.2. Cuestionario utilizado

El cuestionario, aplicado sobre la muestra seleccionada, resulta de una adaptación del cuestionario utilizado por los investigadores Alves y Raposo de la Universidad de Beira Interior (Portugal) (Alves y Raposo, 2004). Éste presenta la siguiente estructura: un primer bloque, claramente diferenciado de los demás, que contiene una serie de variables de carácter sociodemográfico, como sexo, edad y algunas otras variables que caracterizan al estudiante dentro del ámbito de la facultad, como año de ingreso, año y cantidad de materias en curso, entre otras. Los restantes ocho bloques de preguntas (presentados como bloque A - H) presentan todos la misma estructura, se plantea una pregunta general que determina la esencia del bloque y a partir de ella, se establecen una serie de afirmaciones sobre las cuales el estudiante deberá expresar su posición, utilizando una escala Likert que toma valores en el intervalo [1 - 10], donde 1 indicará la mayor discrepancia con lo planteado en la pregunta y 10 el mayor acuerdo.

Los bloques A a H presentan las siguientes características:

- Bloque A - Contiene 12 afirmaciones referentes a las *expectativas* de los estudiantes, previo ingreso a facultad.
- Bloque B - Consta de 6 afirmaciones vinculadas a la *imagen* que tienen los estudiantes sobre la facultad.

- Bloque C - Conformado por 9 afirmaciones asociadas a la *calidad* del servicio que brinda la facultad.
- Bloque D - Contiene 9 afirmaciones asociadas a la *calidad* de los servicios que brinda la facultad con respecto a la biblioteca, bedelía y cafetería, entre otros.
- Bloque E - Conformado por las mismas 9 afirmaciones que el bloque C, pero asociadas a necesidades/deseos actuales
- Bloque F - Presenta 7 afirmaciones que indagan sobre el *valor percibido*.
- Bloque G - Contiene 6 afirmaciones que refieren a la *satisfacción* de los estudiantes con la facultad.
- Bloque H - Conformado por 5 preguntas que pueden dividirse en 2 subgrupos, las 3 primeras referentes a la *lealtad* de los estudiantes con la facultad, y las 2 últimas asociadas al *boca a boca* que se genera entre los estudiantes.

En este trabajo los bloques D y E no serán considerados.

4. Resultados - Análisis de Clases Latentes

La aplicación presentada en este trabajo, toma como insumo las puntuaciones (categorizadas) de las seis variables del ECSI.

Estas variables, y sus respectivas categorías, se presentan a continuación:

- E_R: Expectativas

  Altas (> 90), Medias ($81 : 90$), Bajas (< 81)

- I_R: Imagen

  Alta (> 50), Media ($41 : 50$), Baja (< 41)

- CSA_R: Calidad de los Servicios Académicos

  Alta (> 70), Media ($61 : 70$), Baja (< 61)

- CSF_R: Calidad de los Servicios Funcionales

  Alta (> 65), Media ($56 : 65$), Baja (< 56)

- ND_R: Necesidades/Deseos

Alta (> 70), Media ($61 : 70$), Baja (< 61)

- VP_R: Valor Percibido

  Alta (> 60), Media ($51 : 60$), Baja (< 51)

A modo de simplificar la notación la codificación utilizada será (para i = 1:6):

$$Y_i = \begin{cases} 3 & Alto \\ 2 & Medio \\ 1 & Bajo \end{cases}$$

Se tiene, entonces:

- Tamaño de muestra: n = 470 (sin considerar datos faltantes).
- Una **variable de clases latentes**: Satisfacción estudiantil.
- Seis variables (Y_i) **manifiestas**: p = 6.
- Cada una de las variables manifiestas posee 3 categorías de respuestas posibles: $k_i = 3$ (para i = 1:6).

En la tabla 2 se presentan los seis patrones de respuesta más frecuentes (de las 163 observadas), y sus respectivas frecuencias, para el caso de los 470 estudiantes en estudio. En dicha tabla se puede observar que los dos patrones más frecuentes son los que representan los extremos: niveles *altos de todas* las variables manifiestas y valores *bajos en todas* ellas, respectivamente.

Tabla 2: Patrones de respuesta - Frecuencias observadas.

E_R	I_R	CSA_R	CSF_R	ND_R	VP_R	Frecuencia
3	3	3	3	3	3	48
1	1	1	1	1	1	45
2	2	2	3	2	2	17
1	2	1	1	1	1	11
3	3	3	3	3	2	11
3	2	3	3	3	2	10

4.1. Estimación del modelo

Los modelos estimados, presentados en esta sección fueron estimados con el paquete *poLCA* (Linzer y Lewis, 2011) del Software libre R-project (R Core Team, 2017).

En el contexto del análisis de variables latentes estimar un modelo consiste, en primera instancia, en determinar cuántas clases latentes existen en el problema en estudio.

Por lo tanto, la hipótesis de partida en la estimación de cada uno de los posibles modelos será:

H_0) El modelo ajustado es el *adecuado*

H_1) El modelo ajustado NO es el *adecuado*

Por *adecuado* se entenderá que la cantidad de clases especificadas es la correcta.

En este caso se han estimado 4 modelos (M = 1, 2, 3, 4).

A continuación (tabla 3), se presenta un resumen de los principales resultados obtenidos al estimar dichos modelos.

Tabla 3: Estimación de 4 modelos - M = 1, 2, 3, 4.

Clases	BIC	χ^2	valor p	G^2	valor p
M = 1	6259.27	11755.38	0	1963.49	0
M = 2	5220.88	1408.42	0	844.92	0
M = 3	4964.77	804.16	0	508.66	0.01
M = 4	4992.73	750.84	0	456.4	0.15

A partir de los resultados presentados en la tabla 3 se observa que siguiendo el criterio de mínimo BIC, el mejor modelo sería aquel que presenta una variable con 3 clases latentes. Sin embargo, tanto en este modelo como en aquellos que proponen una variable con una y dos clases latentes los resultados obtenidos ponen de manifiesto que la hipótesis nula es rechazada[3], por lo que sería necesario un ajuste con más clases.

Para el caso del modelo con una variable con cuatro clases latentes la hipótesis nula no puede ser rechazada, por lo que podría considerarse que ajustar un modelo con cuatro clases latentes es *adecuado*. Además, se verifica que al estimar este modelo no existen problemas de identificabilidad y que en el proceso de maximización se alcanza, al menos, un máximo local (que puede coincidir con el máximo global).

[3]Se considera un $\alpha = 0{,}05$.

4.2. Caracterización de las clases

En función de lo expuesto en la sección 4.1 se decide estimar un modelo de una variable con cuatro clases latentes, cuya caracterización se presenta a continuación.
Tal como se observa en la tabla 4, la probabilidad de pertenecer a la clase 1 es la mayor, mientras que la menor corresponde a la clase 4.

Tabla 4: Probabilidad de cada una de las clases.

Clase	1	2	3	4
P(m)	0.32	0.29	0.24	0.15

La caracterización de cada una de las clases se realiza en función de la probabilidad condicional, de cada una de las categorías de cada variable manifiesta, dada la clase. Tomando como referencia los resultados presentados en la tabla 5, la caracterización de las clases en las que se agrupan a los 470 estudiantes es la siguiente:

Clase Latente 1

Los estudiantes que se encuentran en esta clase presentan un nivel de *expectativas* y una percepción de la *calidad de los servicios funcionales* medio-bajo y niveles medios de *imagen*, percepción de la *calidad de los servicios académicos*, *necesidades y deseos* y *valor percibido*.
En función de la descripción hecha se entiende que los patrones característicos de esta clase son:

$$(E_R,\, I_R,\, CSA_R,\, CSF_R,\, ND_R,\, VP_R) = (1,2,2,1,2,2)$$
$$(E_R,\, I_R,\, CSA_R,\, CSF_R,\, ND_R,\, VP_R) = (2,2,2,2,2,2)$$

Clase Latente 2

Los estudiantes que se encuentran en la clase latente 2, presentan un nivel alto de todas las variables manifiestas. Cabe destacar, además, el hecho de que la probabilidad de que un estudiante que pertenece a esta clase, tenga niveles bajos en su percepción de la *calidad de los servicios académicos*, es 0.
El patrón específico de esta clase es:

$$(E_R,\, I_R,\, CSA_R,\, CSF_R,\, ND_R,\, VP_R) = (3,3,3,3,3,3)$$

Clase Latente 3

En el extremo opuesto a los estudiantes cuya *Satisfacción* se define a partir de la clase 2, se encuentran los estudiantes de esta clase. Éstos presentan un nivel bajo de todas las variables manifiestas. Cabe destacar, además, el hecho de que la probabilidad de que un estudiante que pertenece a esta clase, tenga niveles altos en *imagen* y *necesidades/deseos*, es 0.

El patrón específico de esta clase es:

$$(E_R,\, I_R,\, CSA_R,\, CSF_R,\, ND_R,\, VP_R) = (1,1,1,1,1,1)$$

Clase Latente 4

Por último, los estudiantes que pertenecen a la clase latente 4 se caracterizan por tener nivel medio-alto de *expectativas*, nivel medio de *imagen*, percepción de la *calidad de los servicios académicos*, *necesidades/deseos* y *valor percibido*.

En lo que refiere a la percepción de la *calidad de los servicios funcionales*, los estudiantes que se encuentran en esta clase presentan valores altos.
Además, se destaca que la probabilidad de que un estudiante que pertenece a esta clase, tenga niveles bajos en *imagen*, percepción de la *calidad los servicios académicos* y *necesidades/deseos*, es 0.

Los patrones característicos de esta clase son:

$$(E_R,\, I_R,\, CSA_R,\, CSF_R,\, ND_R,\, VP_R) = (2,2,2,3,2,2)$$
$$(E_R,\, I_R,\, CSA_R,\, CSF_R,\, ND_R,\, VP_R) = (3,2,2,3,2,2)$$

Tabla 5: Probabilidades condicionales $P(Y_i/m)$.

Expectativas (E_R)	1	2	3
m = 1	0.42	0.43	0.15
m = 2	0.04	0.15	0.81
m = 3	0.81	0.17	0.02
m = 4	0.09	0.48	0.43
Imagen (I_R)	1	2	3
m = 1	0.19	0.62	0.19
m = 2	0.01	0.26	0.73
m = 3	0.70	0.30	0.00
m = 4	0.00	0.91	0.09
Calidad de Servicios Académicos (CSA_R)	1	2	3
m = 1	0.30	0.69	0.01
m = 2	0.00	0.05	0.95
m = 3	0.99	0.00	0.01
m = 4	0.00	0.68	0.32
Calidad de Servicios Funcionales (CSF_R)	1	2	3
m = 1	0.38	0.40	0.22
m = 2	0.04	0.16	0.80
m = 3	0.77	0.19	0.04
m = 4	0.03	0.29	0.68
Necesidades/Deseos (ND_R)	1	2	3
m = 1	0.28	0.63	0.09
m = 2	0.01	0.07	0.92
m = 3	0.97	0.03	0.00
m = 4	0.00	0.74	0.26
Valor Percibido (VP_R)	1	2	3
m = 1	0.13	0.57	0.30
m = 2	0.03	0.24	0.73
m = 3	0.75	0.24	0.01
m = 4	0.05	0.95	0.00

Por lo tanto, en lo que refiere a la *Satisfacción Estudiantil*, las clases se podrían categorizar como:

- m = 1: Estudiantes con *Satisfacción Estudiantil* **medio-baja**.
- m = 2: Estudiantes con *Satisfacción Estudiantil* **alta**.
- m = 3: Estudiantes con *Satisfacción Estudiantil* **baja**.
- m = 4: Estudiantes con *Satisfacción Estudiantil* **media-alta**.

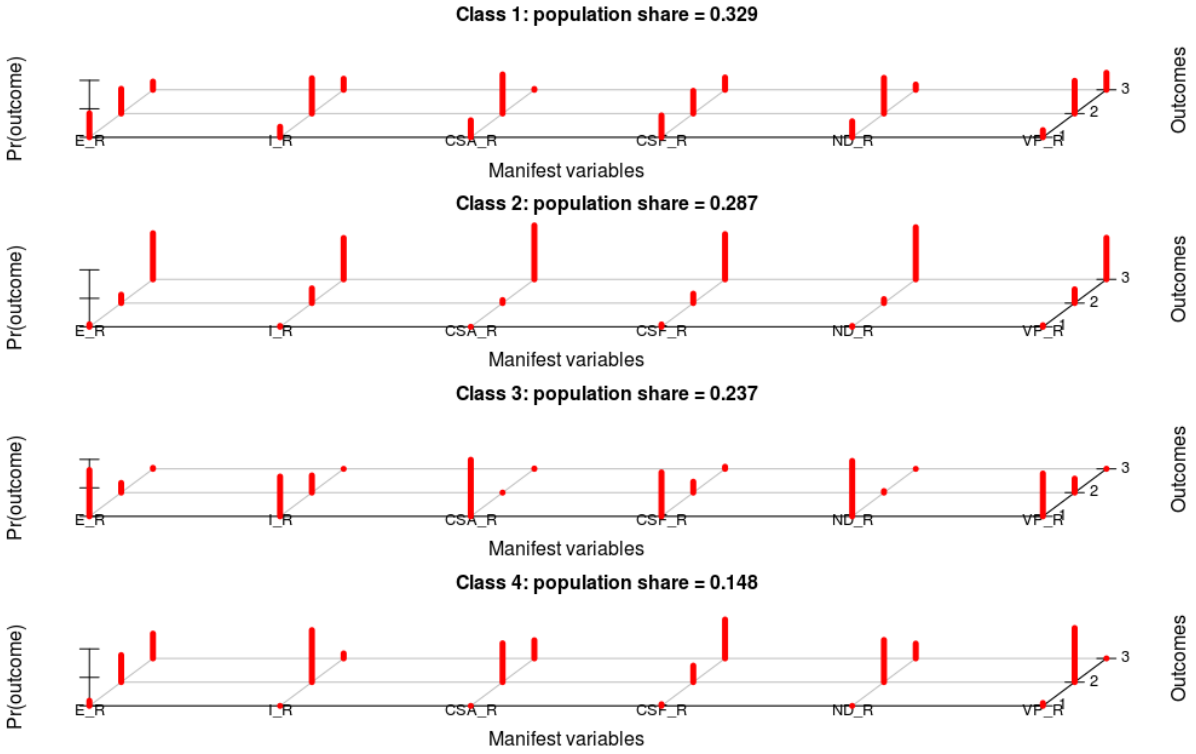

Figura 1: Perfil de las 4 clases latentes en función de las 6 variables manifiestas.

4.3. Probabilidades a posteriori

Los resultados presentados en la tabla 6 hacen referencia, a modo de ejemplo, a las probabilidades a posteriori, para cada uno de los patrones posibles de respuesta (para los 6 patrones más frecuentes, presentados en la tabla 2), y la asignación a cada una de las clases (en función de la máxima probabilidad a posteriori).

Tabla 6: Probabilidades a posteriori según patrones y asignaciones.

E_R	I_R	CSA_R	CSF_R	ND_R	VP_R	$P(m=1)$	$P(m=2)$	$P(m=3)$	$P(m=4)$	Asignación
3	3	3	3	3	3	0.00	1.00	0.00	0.00	2
1	1	1	1	1	1	0.00	0.00	1.00	0.00	3
2	2	2	3	2	2	0.18	0.00	0.00	0.81	4
1	2	1	1	1	1	0.01	0.00	0.99	0.00	3
3	3	3	3	3	2	0.00	0.99	0.00	0.01	2
3	2	3	3	3	2	0.00	0.76	0.00	0.24	2

En función de las probabilidades a posteriori, de cada una de las clases, el total de estudiantes queda distribuido en cada una de ellas en un 30 % (145), 29 % (137), 25 % (118) y 16 % (77) respectivamente.

5. Resultados - Análisis de Clusters

En lo que refiere a los principales resultados obtenidos de la realización de una *clusterización* a partir de un método Jerárquico (en particular, Ward) y considerando distancias euclideas, lo primero que se evalúa es la cantidad de clusters a determinar.

Gráficamente, en la figura 2 se observa que resultaría adecuado considerar una estructura de 3 o 5 clusters.

Tabla 7: Estadísticos.

Paso	Historia	Frecuencia	R^2	psF	psT
467 446	465	62	0.74	149.86	23.43
468 463	460	108	0.73	157.69	30.33
469 451	464	61	0.71	166.10	27.45
470 461	467	136	0.70	182.14	26.35
471 457	455	41	0.68	201.16	22.09
472 468	466	172	0.65	223.95	41.07
473 462	470	203	0.61	249.66	78.64
474 471	469	102	0.57	317.33	39.15
475 472	473	375	0.43	357.74	187.81
476 475	474	477	0.00	NaN	357.74

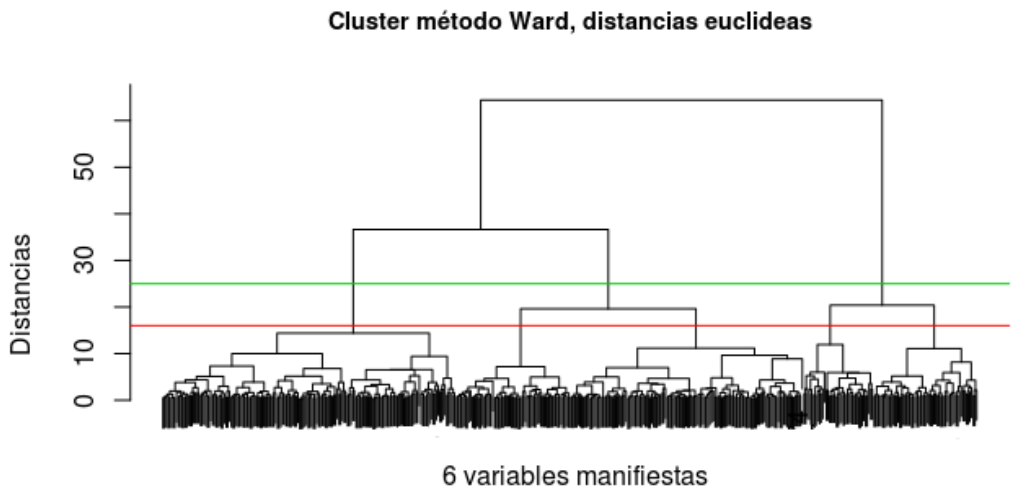

Figura 2: Dendrograma algoritmo de Ward para 6 variables manifiestas.

Si se consideran, además, los estadísticos planteados en la sección 2.2, se tiene (últimos 10 pasos que aparecen en la tabla 8) donde se observa, el mayor incremento en el valor del R^2 se da al pasar de considerar 2 a 3 grupos, lo que reafirmaría una posible estructura de 3 grupos.

Por otra parte, se tiene que si bien el máximo valor del *pseudoF* se da al considerar una estructura de 2 grupos, el decremento de dicho estadístico al considerar 3 grupos (11 % menos), no resulta tan amplio como en los demás pasos (por ejemplo al pasar de 3 a 4 la caída es de más de un 20 %).

Al observar el *pseudo* t^2, se nota una caída muy pronunciada (79 %) al pasar de considerar 2 a 3 clusters por lo que la variabilidad se minimiza al considerar una potencial estructura de 3 grupos. En 5 también hay una caída pero la de 3 resulta más fuerte.

Teniendo en consideración tanto la información gráfica como el análisis numérico de los indicadores, se decide trabajar con 3 clusters.

5.1. Caracterización de los grupos

En primera instancia se presenta la distribución de estudiantes, por cluster. Se observa un 36 %, 43 %, 31 % en el cluster 1, 2 y 3, respectivamente.

Tabla 8: Cantidad de estudiantes por cluster.

Cluster	1	2	3	Total
Estudiantes	172	203	102	477

Cluster 1

Los estudiantes que se encuentran en este cluster se caracterizan por ser un 54 % mujeres y un 46 % hombres.
En lo que refiere a la edad de los estudiantes y al año que están cursando, en la tabla 9 se observa que casi un 30 % se encuentra cursando su primer año en facultad (Año en curso = 1), lo que resulta en concordancia con el hecho de que casi un 40 % se encuentra en el primer grupo de edad (18-20). En particular, la distribución conjunta de esas dos categorías es de casi un 25 % del total.

Tabla 9: Cantidad de estudiantes por Año en curso según Edad - Cluster 1.

Edad/Curso	1	2	3	4	5	Total
1	40	23	3	0	1	67
2	6	12	17	15	4	54
3	1	4	8	10	12	35
4	2	2	1	1	4	10
5	0	0	3	1	2	6
Total	49	41	32	27	23	172

En lo que refiere a las 6 variables manifiestas, en la figura 3 se observa que el 75 % de los estudiantes tienen *expectativas* más bajas que el 25 % de los estudiantes del cluster 2.
Además, el 25 % de los estudiantes, tienen *expectativas* más altas que el 75 % de los estudiantes del cluster 3. Un comportamiento similar se da para las restantes 5 variables manifiestas.
Por último, al observar la figura 4, se destaca en primer lugar que existe una correlación positiva entre todas las variables y que la distribución conjunta 2 a 2 de las 6 variables manifiestas para el caso del cluster 1 presenta valores medios, en todos los casos.

Cluster 2

Los estudiantes que se encuentran en el cluster 2 se caracterizan por ser más de un 60 % de mujeres.

En cuanto a la edad de los estudiantes y al año que están cursando, en la tabla 10 se observa que más de un 65 % se encuentra cursando sus primeros años en facultad (Año en curso = 1 y 2), lo que, tal como ocurre en el cluster 1, resulta en concordancia con el hecho de que más de un 75 % se encuentra en los primeros grupos de edad (18-23).
Por último, cabe destacar que en este cluster no existen estudiantes mayores de 27 años, que estén cursando su primer año en facultad.

Tabla 10: Cantidad de estudiantes por Año en curso según Edad - Cluster 2.

Edad/Curso	1	2	3	4	5	Total
1	74	32	1	0	0	107
2	7	13	17	11	2	50
3	1	3	5	8	11	28
4	2	1	2	4	2	11
5	0	2	0	2	3	7
Total	84	51	25	25	18	203

Al analizar las variables manifiestas para el caso particular del cluster 2, se observa que el máximo de todas ellas se da en este cluster. En particular, se destaca que el 100 % de los estudiantes presenta valores más altos que un 25 % de los estudiantes del cluster 3, en las variables *necesidades/deseos* y *calidad de los servicios académicos* (ver figuras 3 y 4).

Cluster 3

El cluster 3 es el cluster con mayor proporción de hombres (52 %). Además, casi un 60 % de los estudiantes de este está cursando su tercer, cuarto o quinto año en FCEA.

Tabla 11: Cantidad de estudiantes por Año en curso según Edad - Cluster 3.

Edad/Curso	1	2	3	4	5	Sin dato	Total
1	16	6	1	0	0	1	24
2	4	4	9	11	3	0	31
3	1	6	3	10	8	0	28
4	0	2	1	3	3	0	9
5	2	0	0	3	5	0	10
Total	23	18	14	27	19	1	102

En cuanto a las variables del modelo ECSI, se observa en la figura 3 que el mínimo de todas

ellas se da en este cluster. Además, el 75 % de los estudiantes de este cluster, presentan valores más bajos que el 25 % de los estudiantes del cluster 1, en todas las variables

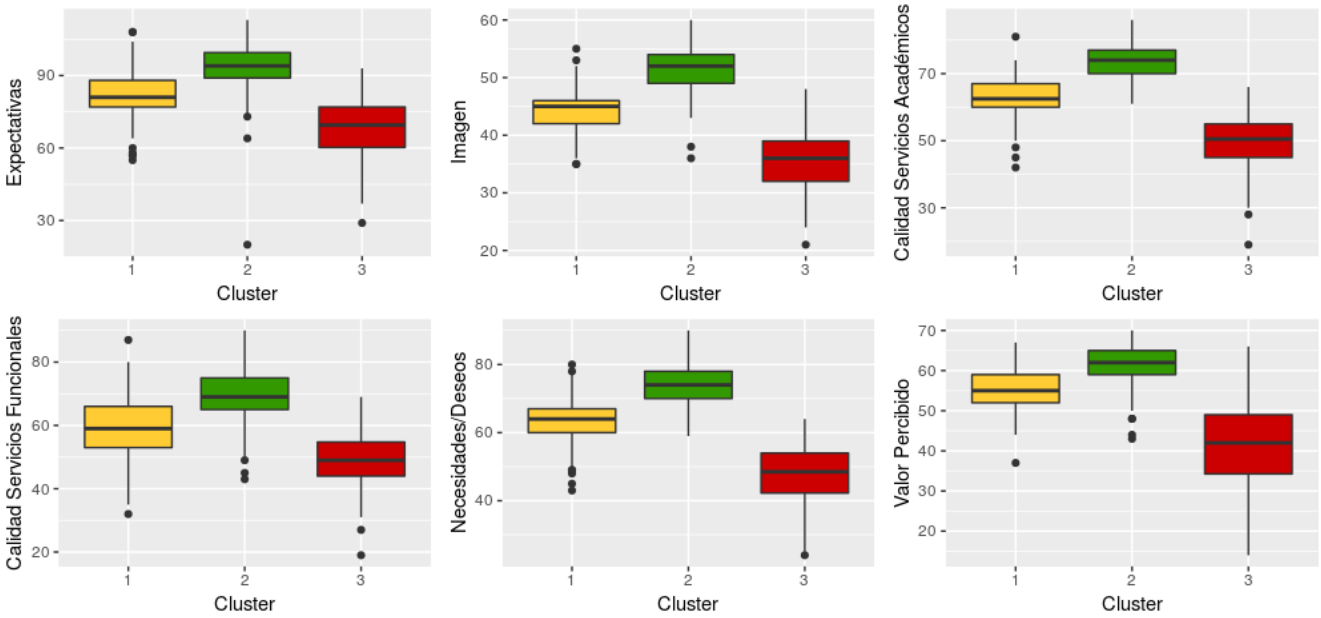

Figura 3: Distribución de las 6 variables manifiestas según cluster.

En función de las caracterizaciones presentadas, se entiende que en cuanto a la *satisfacción* los clusters podrían denominarse:

- cluster = 1: Estudiantes con *Satisfacción Estudiantil* **media**.
- cluster = 2: Estudiantes con *Satisfacción Estudiantil* **alta**.
- cluster = 3: Estudiantes con *Satisfacción Estudiantil* **baja**.

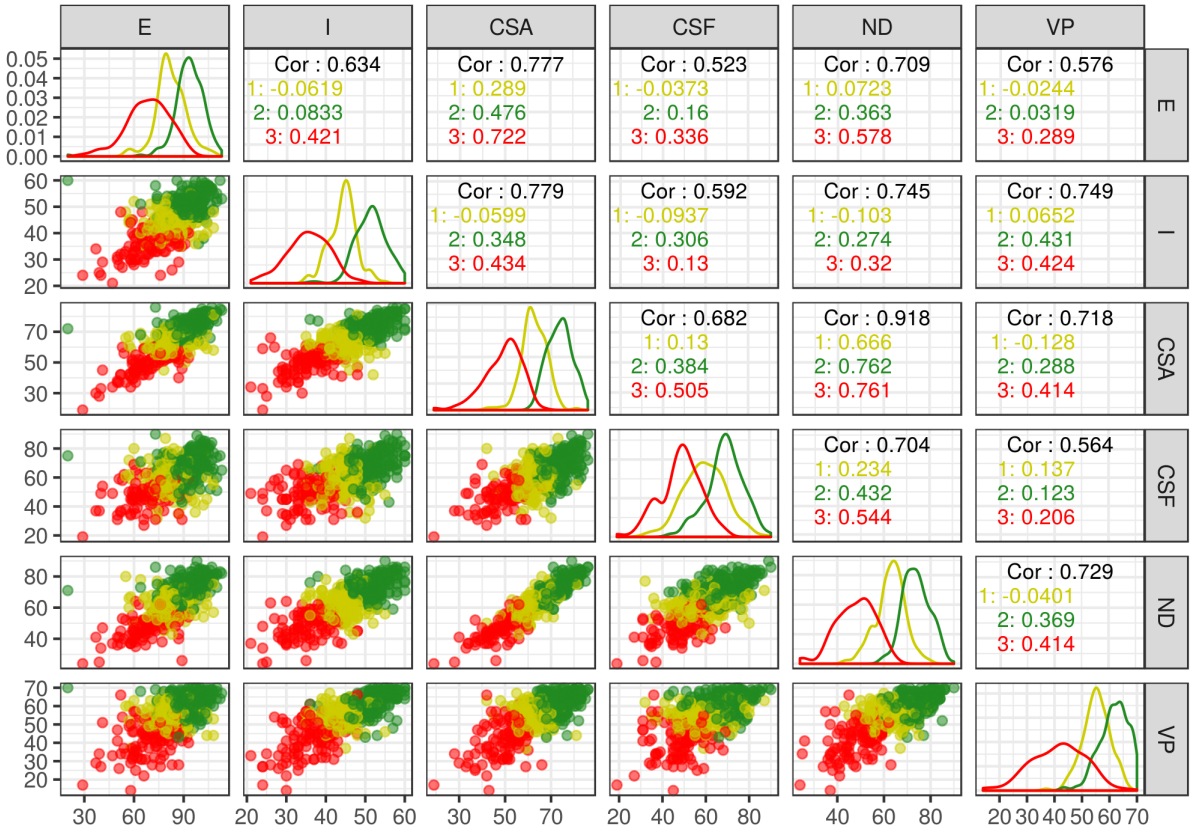

Figura 4: Diagrama de dispersión variables manifiestas.

6. Discusión

Al realizar un análisis comparativo entre los resultados obtenidos con ambas técnicas, se obtiene la información resumida en la tabla 12.

Tabla 12: Cantidad de estudiantes por Clase latente según Cluster.

Cluster/Clase latente	1	2	3	4	Total
1	106	3	23	40	172
2	32	134	0	37	203
3	7	0	95	0	102
Total	145	137	118	77	477

Al analizar la distribución conjunta de ambas categorizaciones se destaca, en primera instancia, que las categorías *alta* de ambas variables prácticamente coinciden (*alta:* cluster = 2 y clase latente = 2). En particular, hay sólo un 2 % de estudiantes de la clase latente *alta* (2), que no corresponden al cluster denominado *alta* (2). Cabe destacar, además, que no existen estudiantes cuya categoría de clase latente sea *alta* (2) y pertenezca al cluster *baja* (3).

En lo que refiere a las categorías *baja* se observa que un 80 % de los estudiantes de la clase latente *baja* (3) pertenecen al cluster *baja* (3). Además, se destaca el hecho de que no existen estudiantes de la clase latente *baja* (3) que pertenezcan al cluster *alta* (2).

Por último, se observa que el 85 % de los estudiantes del cluster *media* corresponden a estudiantes cuya clase latente es *media* (media baja + media alta).

En resumen, se destaca que si bien las técnicas multivariantes utilizadas toman como insumo variables de distinta naturaleza, los resultados no presentan diferencias relevantes en la construcción/caracterización de la *satisfacción estudiantil*.

7. Consideraciones finales

Como consideraciones finales y propuestas a futuro se plantea:

- Evaluar la robustez de la variable latente detectada iterando varias veces para ver el grado de dependencia de los valores iniciales.
- Plantear el uso de las clases latentes detectadas para poder particionar la tabla de datos de manera de volver a estimar mediante modelos de ecuaciones estructurales

para realizar un estudio comparativo con los modelos que ya se probaron con estos datos (Vernazza, 2013), (Álvarez-Vaz y Vernazza, 2013), (Álvarez-Vaz y Vernazza, 2014), (Álvarez-Vaz y Vernazza, 2017).

- Estudiar la creación de variables latentes a través del uso de *Mixture models: latent profile and latent class analysis* (Robertson, 2016) o (Scrucca y Raftery, 2015), trabajando sobre las variables manifiestas en su escala original, es decir, las 6 variables (no categorizadas) que surgen de agregar los 63 ítems según el bloque del cuestionario al que pertenecen.
- Evaluar la asociación entre las categrorías de las variables manifiestas a través de lo que se conoce como *graph model*, donde se construye un grafo entre las categorías observadas y mediante el análisis de redes (SNA) (Højsgaard, 2012).
- Realizar un estudio similar con los datos obtenidos en investigación realizada en 2017 (réplica de la investigación 2009). Comparar resultados.

Referencias Bibliográficas

Agresti, A. (2013). *Categorical data analysis*. Wiley-Interscience, Hoboken, N.J.

Álvarez-Vaz, R. y Vernazza, E. (2013). Aplicación de los modelos de ecuaciones estructurales para el estudio de la satisfacción estudiantil en en los cursos superiores de FCCEEyA. *Documentos de Trabajo -Serie DT IESTA* (13/02), *Universidad de la República. Facultad de Ciencias Económicas y de Administración.*

Álvarez-Vaz, R. y Vernazza, E. (2014). Aplicación de modelos de ecuaciones estructurales en la medición del nivel de satisfacción estudiantil: comparación de tres métodos de estimación. *Documentos de Trabajo -Serie DT IESTA* (14/03), *Universidad de la República. Facultad de Ciencias Económicas y de Administración.*

Álvarez-Vaz, R. y Vernazza, E. (2017). Evaluación de un instrumento de medición del nivel de satisfacción estudiantil a través de la aplicación de modelos de ecuaciones estructurales. *Cuadernos del CIMBAGE*, (19):1–25.

Álvarez-Vaz, R., Freira, D., Vernazza, E., y Alves, H. (2016). Can students' satisfaction indexes be applied the same way in different countries? *Int Rev Public Nonprofit Marketing*, 13(101).

Alves, H. y Raposo, M. (2004). La medición de la satisfacción en la enseñanza universitaria: El ejemplo de la universidade da beira interior. *Int Rev Public Nonprofit Marketing*, 1(1):73–88.

Bandeen-Roche, K., Miglioretti, D. L., Zeger, S. L., y Rathouz, P. J. (1997). Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association*, 92(440):1375–1386.

Beath, K. (2017). *randomLCA: Random Effects Latent Class Analysis*. R package version 1.0-12.

Blanco, J., Camaño, G., Nalbarte, L., y Álvarez-Vaz, R. (2006). *Introducción al Análisis Multivariado*. Instituto de Estadística, FCEA.

Carlomagno-Araya, A. y Sepúlveda, R. (2010). Análisis de clases latentes en tablas poco ocupadas: consumo de alcohol, tabaco y otras drogas en adolescentes. *Revista de Matemática: Teoría y Aplicaciones*, 17(1):25–40.

Castro López, Claudio R.and Montano Rivas, A. y Oliva Zarate, L. (2011). Modelos de clases latentes para definir perfiles conductuales en niños de 4 y 5 años. *Revista Electrónica de Psicología Iztacala.*, 14(1).

Dempster, A. P., Laird, N. M., y Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39:1–38.

Everitt, B. S. (1984). *An Introduction to Latent Variable Models*. Springer Netherlands, Dordrecht.

Hagenaars, J. (2002). *Applied latent class analysis*. Cambridge University Press, Cambridge New York.

Højsgaard, S. (2012). *Graphical models with R*. Springer, New York.

Lazarsfeld, P. (1950). The logical and mathematical foundations of latent structure analysis. *ISA Stouffer (ed.), Measurement and Prediction, pp. 362â412.*

Linzer, D. A. y Lewis, J. B. (2011). poLCA: An R package for polytomous variable latent class analysis. *Journal of Statistical Software*, 42(10):1–29.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., y Hornik, K. (2016). *cluster: Cluster Analysis Basics and Extensions*. R package version 2.0.4 — For new features, see the 'Changelog' file (in the package source).

McLachlan, G. (2000). *Finite mixture models*. Wiley, New York.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., y Leisch, F. (2017). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. R package version 1.6-8.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Robertson, J. (2016). *Modern statistical methods for HCI*. Springer, Switzerland.

Sánchez-Rivero, M. (2001). Segmentación de la población española según su grado de concienciación ecológica mediante modelos de variables latentes l. *Investigaciones Europeas de Dirección y Economía de la Empresa*, 7(3):173–196.

Scrucca, L. y Raftery, A. E. (2015). Improved initialisation of model-based clustering using Gaussian hierarchical partitions. *ArXiv e-prints*.

Vernazza, E. (2013). *Evaluación de un instrumento de medición del nivel de satisfacción estudiantil en los cursos de formación superior de la FCCEEyA de la UDELAR a través de la aplicación de Structural Equation Modelling (SEM)*. Informe de Pasantía, Licenciatura en Estadística - Facultad de Ciencias Económicas y de Administración - Universidad de la República.

Instituto de Estadística

Documentos de Trabajo

Eduardo Acevedo 1139. CP 11200 Montevideo, Uruguay
Teléfonos y fax: (598) 2410 2564 - 2418 7381
Correo: ddt@iesta.edu.uy
www.iesta.edu.uy
Área Publicaciones

Diciembre, 2017
DT (17/3)