Comparación de perfiles sociales de estudiantes universitarios a través de técnicas de visualización de objetos simbólicos

NORA MOSCOLONI
MARÍA DE LUJÁN BURKE
SILVANA CALVO
GUILLERMINA ISERN
Programa Interdisciplinario de Análisis de Datos,
Universidad Nacional de Rosario, Argentina

Introducción

El conocimiento de las características sociales de los estudiantes universitarios constituye una materia de interés fundamental tanto para la gestión académica como para las investigaciones que refieren a la Universidad como objeto de estudio.

El debate se agudiza especialmente cuando se discuten las condiciones de acceso y equidad en el ingreso a los estudios superiores. Dicho debate se origina en nuestro país en la década de los ochenta con la finalización de la dictadura y los comienzos de la masividad en la universidad. La década de los 90 representó para la Argentina, entre otros países, el inicio de una serie de reformas políticas y económicas que afectaron la esencia de la organización del Estado. Surge el Estado como garante de la calidad de los productos educativos y generador de información suficiente y apropiada para la toma de decisiones. Este nuevo papel impulsó la creación de diversas instancias de control de la calidad y la producción de más y mejores estadísticas sobre el sector (García de Fanelli, 2005).

En el marco de estas reformas el Sistema de Educación Superior es impactado, alterando las relaciones entre Universidad-Estado. La generación de reformas se caracterizó por cambios en los modelos de financiamiento, exigencia de eficiencia a través de la implantación de sistemas evaluativos y presiones por relaciones estrechas con el sector productivo (García Guadilla, 2004).

En este contexto se hace necesario caracterizar el perfil social de los estudiantes que ingresan a la universidad, lo cual implica el estudio del comportamiento de indicadores diferenciales.

Tradicionalmente, para realizar caracterizaciones socioeconómicas de poblaciones como en el caso de estudios de rendimiento académico se recurre a las técnicas de la estadística tradicional con la utilización de variables continuas (Ver por ejemplo Di Grescia, Porto, Ripani, 2002).

Consideramos que estas técnicas, al proponerse desde una óptica mayormente determinística reducen la riqueza de la información. En este trabajo proponemos la utilización de tipologías a través de herramientas que se encuadran en el enfoque del Análisis Multidimensional de Datos. Estas técnicas fueron aplicadas con éxito en la caracterización de las poblaciones de alumnos de las distintas facultades de la UNR como parte del proceso de Autoevaluación Institucional que se desarrolló en esta Casa de Altos Estudios a partir del año 2000.

Análisis Multidimensional de Datos

El Análisis Multidimensional de Datos (AMD) en la versión de la escuela francesa, surge en la década de los 70, planteando fines menos deterministas que los de la Estadística tradicional, su objetivo general es la búsqueda de una estructura presente en los datos, en un contexto de tipo más inductivo que deductivo, que revaloriza el rol del individuo. Su naturaleza, fundamentalmente descriptiva y el acercamiento geométrico asignan un rol muy importante a las representaciones gráficas, sobre todo en una etapa exploratoria.

En el campo de las ciencias sociales, este enfoque se revela como la opción ideal para el procesamiento de la información que, en la generalidad de los casos, es rica en categorías y no en continuos, de naturaleza ambigua, con grandes dificultades de diseño.

Los algoritmos desarrollados en el contexto del análisis multidimensional de datos se adaptan a diferentes niveles de complejidad de la información: datos numéricos, textuales, simbólicos. Es decir que el dato puede ser algo más que un único valor numérico resultado de la asignación de una medida o código a una unidad de análisis: puede ser una palabra, un conocimiento, una posibilidad, una conjunción de valores.

Una novedosa e interesante área de estudio se abre con los desarrollos en Análisis de Datos Simbólicos (Diday,1997). Este método parte de una pregunta: ¿Por qué no se aprovechan en el procesamiento y análisis mismo los valiosos conocimientos de los expertos? La respuesta de la estadística clásica era que no se podían cuantificar. Se plantea en la actualidad el desafío de representarlos por expresiones a la vez simbólicas y numéricas, saber manipularlos y utilizar estas expresiones a los fines de ayudar a decidir, de mejorar el análisis, de sintetizar y de organizar nuestra experiencia y nuestras observaciones respetando más acabadamente su complejidad.

Estas técnicas valorizan, sobre todo, el poder de la clasificación como operación interpretadora, tratando de superar con nuevos algoritmos los problemas de descripción de las clases, en especial para los individuos que se encuentren en los bordes de las mismas.

Los conceptos de intención y extensión de una idea aplicados a una clase o grupo son fundamentales para la comprensión del objetivo del Análisis de Datos Simbólicos. Así la intención de una idea se refiere a los atributos que ella contiene y que no pueden ser suprimidos sin destruirla; la extensión de una idea son los sujetos o elementos a los cuales ella se aplica.

En el Análisis Simbólico en lugar de trabajar sobre las extensiones, es decir sobre los individuos, se reemplazan los individuos por las intenciones, aprovechando de esta manera el conocimiento de los expertos.

En el Análisis Multidimensional de Datos clásico o Numérico se estudian conjuntos de objetos individuales representados por elementos atómicos de datos, en el Análisis de Datos Simbólicos se estudian conjuntos de más alto nivel, donde los individuos están constituidos por objetos simbólicos. Se responde con esto a la necesidad de que en muchas situaciones sólo se dispone de objetos simbólicos y que sus propiedades y problemas difieren de los de los objetos individuales.

Estos objetos, que constituyen las filas de una matriz de datos en el Análisis de Datos Simbólicos, permiten representar los individuos complejos o las clases de individuos a través de conjunciones de propiedades o de descriptores pudiendo tomar valores múltiples y ponderados (según diferentes semánticas) y están a veces relacionados entre ellos por relaciones de orden lógico.

El objetivo de los nuevos algoritmos se dirige al desarrollo de herramientas para manipular estos objetos según diferentes grados de complejidad tanto en su composición como en las relaciones que se establecen entre ellos y en el tipo de conocimiento que sobre ellos se tiene.

De la estadística tradicional al ADS

La evolución del enfoque de análisis podría verse de la siguiente manera: la estadística clásica se interesa sobre todo por la modelización de una población vista globalmente; el AMD clásico además de generalizar el análisis univariado, comienza a interesarse por los individuos; el ADS generaliza la noción de individuo ocupándose también de los objetos en sentido general, más en cuanto a objetos que en cuanto a individuos. Esta evolución que parece natural no aparece solamente en AMD sino también de forma general en informática, en inteligencia artificial y en las ciencias del conocimiento.

El AMD comienza por una matriz de datos en la cual hay valores de las variables tomados por las unidades de análisis. El objetivo del AMD es extraer información de una matriz de este tipo y sintetizarla reemplazando números por 'conocimientos' nuevos. Para ello se vale de dos grupos principales de técnicas: la clasificación y el análisis factorial.

El ADS tiene como objetivo reemplazar los individuos del análisis multidimensional de datos tradicional por individuos de más alto nivel, más complejos y aptos para representar conocimientos, porque están definidos en intención, utilizando el poder de la lógica: son los objetos simbólicos (OS). Asimismo se puede decir que las variables son de mayor nivel en el ADS, porque las variables no van a tomar un sólo valor por cada celda, sino que pueden tomar varios valores.

Por ejemplo: cuando se describe una clase o grupo, los individuos de la clase pueden tomar distintos valores. Si se describe una clase de empresas, que tienen beneficios de distinto orden, se puede tomar el beneficio en intervalos para esta clase de empresas. Si una empresa pertenece a una determinada clase, en la variable beneficio tendré un intervalo de valores correspondiente al conjunto de beneficios que poseen las empresas de esta clase.

Se puede decir que los individuos y las variables son de mayor nivel que en la estadística y el AMD clásicos. Es muy importante porque va a plantear todos los problemas teóricos también a un mayor nivel, subiendo un escalón en toda la teoría del AMD.

¿Qué son los objetos simbólicos?

Los objetos simbólicos son especies de átomos de conocimiento, comprenden un campo tan vasto como los conocimientos mismos.

En la práctica los OS se plantean como nuevas unidades de análisis que pretenden resumir grandes cantidades de información almacenada en bases de datos relacionales y describir tanto individuos como grupos.

En este sentido de una manera general los objetos simbólicos pueden verse como una representación de conceptos estadísticos que permiten el análisis de datos agregados a partir de la combinación de variables seleccionadas que surgen al analizar grandes matrices de datos. Cada objeto puede representar un grupo de individuos con características comunes que resultan del cruce de variables y se tratan como nuevas unidades de análisis.

En ADS en lugar de tener un conjunto de individuos, tenemos objetos simbólicos que están expresados por propiedades, donde cada propiedad puede ser del tipo probabilístico, booleano, posibilístico o de otra noción. De esta manera se permite a un experto expresar mayor cantidad de conocimiento.

La distinción entre objetos simbólicos y numéricos se establece cuando se considera que un objeto es *numérico* si puede ser representado y utilizado como un punto del espacio R^p considerado como un espacio vectorial provisto de las operaciones habituales y que es *simbólico* si no es el caso. Se deduce de esta definición que el análisis de datos clásico trata, desde hace mucho tiempo, con objetos simbólicos particulares que serían todos los objetos caracterizados por variables nominales u ordinales.

El objetivo del ADS es extender el análisis de datos clásico al estudio de objetos más complejos que se expresan bajo forma de *conjunción* de propiedades aplicadas sobre las variables clásicas: continuas, nominales u ordinales. Ellos se distinguen de los objetos clásicamente tratados en análisis de datos en primer lugar a nivel de su descripción, es decir en cuanto al tipo de variables que los predican o también en cuanto a su manipulación:

a) Cada variable puede tomar valores múltiples para un mismo objeto simbólico, ejemplos:

   [opinión = {regular, mala, regular, indiferente}] para expresar el hecho de que una clase de individuos puede tener opinión regular, mala, o regular e indiferente.

   [edad = [17, 29]] para expresar que los individuos encuestados tienen entre 17 y 29 años.

   En estos dos casos no se transforman estos valores en una modalidad mutuamente excluyente de una variable a los fines de no perder la información contenida en estas descripciones.

b) Como consecuencia de (a) se llegan a expresar diferentes tipos de relaciones entre las variables: cuando una variable toma una modalidad, la otra puede no tener sentido (no se describe el trabajo de una persona que no trabaja) o se debe restringir su campo de valores posibles (si la categoría es estudiante, la edad es entre 6 y 28). Se obtienen así objetos simbólicos provistos de propiedades, se trata de variables llamadas madre-hija, como es el caso clásico de la modalidad 'no se aplica'

c) Un objeto simbólico es una descripción en intención de una clase de objetos elementales de la cual constituyen la extensión. El objeto [categoría = {obrero, empleado}] tiene por extensión todos los objetos elementales en los cuales la categoría es ya sea obrero ya sea empleado.

d) Como consecuencia de (c) se puede generalizar o especializar un objeto simbólico modificando sus propiedades de manera ya sea de extender o de restringir su extensión.

Para generalizar se utilizan las operaciones de unión, de intersección y de complementación traduciendo la semántica del dominio de aplicación (que puede expresarse bajo forma de una taxonomía) Con lo cual se habilita por ejemplo, a generalizar 'cualquiera que beba whisky y agua' y 'cualquiera que beba vino y agua' con: 'cualquiera que beba alcohol y agua' en lugar de simplemente 'cualquiera que beba agua' como se haría en el análisis numérico tradicional.

Construcción y procesamiento de objetos simbólicos

La utilización de OS ha obtenido desarrollo en el marco del proyecto europeo SODAS (Symbolic Official Data Analysis System)

Este *software* provee muy buenas posibilidades de aplicación para la manipulación de bases de datos de estadísticas oficiales.

En nuestro caso trabajamos con la base de ingresantes a la Universidad Nacional de Rosario del año 2005.

Presentamos a continuación una parte de la matriz de datos simbólicos construida para esa base:

TABLA 1
Datos simbólicos relativos a las variables sexo, residencia y estado civil

	SEXO	RESIDENCIA	ESTADO_CIVIL
Arquit. Plan. y Dis.	Masculino (0.52), Femenino (0.48)	ReFam (0.81), Relnd (0.19)	Soltero (0.98), Casado (0.01), Separado, viudo (0.01)
Ciencias Agrarias	Masculino (0.79), Femenino (0.21)	ReFam (0.45), Relnd (0.55)	Soltero (0.99), Casado (0.01)
Cs. Bioquím. y Farm.	Masculino (0.29), Femenino (0.71)	ReFam (0.68), Relnd (0.32)	Soltero (0.97), Casado (0.03), Separado, viudo (0.00)
Cs. Econom. y Estad.	Masculino (0.41), Femenino (0.59)	ReFam (0.86), Relnd (0.14)	Soltero (0.96), Casado (0.03), Separado, viudo (0.01)
Cs.Exactas, Ing.y A.	Masculino (0.80), Femenino (0.20)	ReFam (0.85), Relnd (0.15)	Soltero (0.98), Casado (0.02), Separado, viudo (0.00)
Cs. Médicas	Masculino (0.23), Femenino (0.77)	ReFam (0.76), Relnd (0.24)	Soltero (0.83), Casado (0.13), Separado, viudo (0.04)
Ca. Política y RRII	Masculino (0.36), Femenino (0.64)	ReFam (0.76), Relnd (0.24)	Soltero (0.96), Casado (0.03), Separado, viudo (0.01)
Cs. Veterinarias	Masculino (0.45), Femenino (0.55)	ReFam (0.83), Relnd (0.17)	Soltero (0.99), Casado (0.01), Separado, viudo (0.00)
Derecho	Masculino (0.39), Femenino (0.61)	ReFam (0.78), Relnd (0.22)	Soltero (0.88), Casado (0.09), Separado, viudo (0.02)
Humanidades y Artes	Masculino (0.39), Femenino (0.61)	ReFam (0.77), Relnd (0.23)	Soltero (0.84), Casado (0.12), Separado, viudo (0.05)
Odontología	Masculino (0.35), Femenino (0.65)	ReFam (0.52), Relnd (0.48)	Soltero (0.99), Casado (0.01)
Psicología	Masculino (0.21), Femenino (0.79)	ReFam (0.72), Relnd (0.28)	Soltero (0.88), Casado (0.08), Separado, viudo (0.03)

En este caso los objetos simbólicos corresponden a las distintas Facultades de la Universidad y por razones de espacio presentamos sólo tres de las variables consideradas en la base de datos.

La semántica utilizada es la de probabilidades basadas en la frecuencia. Los valores de las variables están indicando por ejemplo que: un alumno en la Facultad de Arquitectura tiene una probabilidad igual a 0.52 de ser de sexo masculino, y de 0.48 de ser de sexo femenino, una probabilidad igual a 0.81 de que su residencia sea con la familia y de 0.19 de que la misma sea individual, etc.

Posibilidades gráficas

La visualización de un OS puede realizarse a través de un gráfico que se denomina Zoom Star. Esta representación está basada en los diagramas de Kiviat donde cada eje representa una variable. En un gráfico de Kiviat, se representan los porcentajes de uso y solapamiento de diferentes componentes de un sistema como una figura geométrica que une diferentes puntos, situados sobre los radios de un círculo, que representan esos porcentajes. Teóricamente, es posible ver de un vistazo el problema que tiene el sistema.

El *software* da la posibilidad de dos tipos de representación, en 2D y en 3D, la primera da una visión más general, mientras que la segunda brinda información con más detalle al proveer simultáneamente las distribuciones de frecuencias de todas las variables.

En 2D, que es el caso que aquí se presenta, los ejes están unidos por una línea que conecta los valores más frecuentes de cada variable. De esta manera se pueden comparar las distribuciones de frecuencias de dos OS, a partir de la forma que toma esta línea de conexión.

El ADS en Autoevaluación Institucional de la Universidad Nacional de Rosario

Desde el año 2000, la Universidad de Rosario lleva adelante el proceso de autoevaluación. El mismo tiene como objetivo tomar conocimiento de la realidad institucional y producir transformaciones con el fin de incrementar la calidad educativa. Para ello se evaluaron las funciones sustantivas de la universidad, docencia, investigación y extensión, y la adjetiva como gestión. Para evaluar las funciones mencionadas, las mismas se desagregaron en propósitos, aspectos e interrogantes.

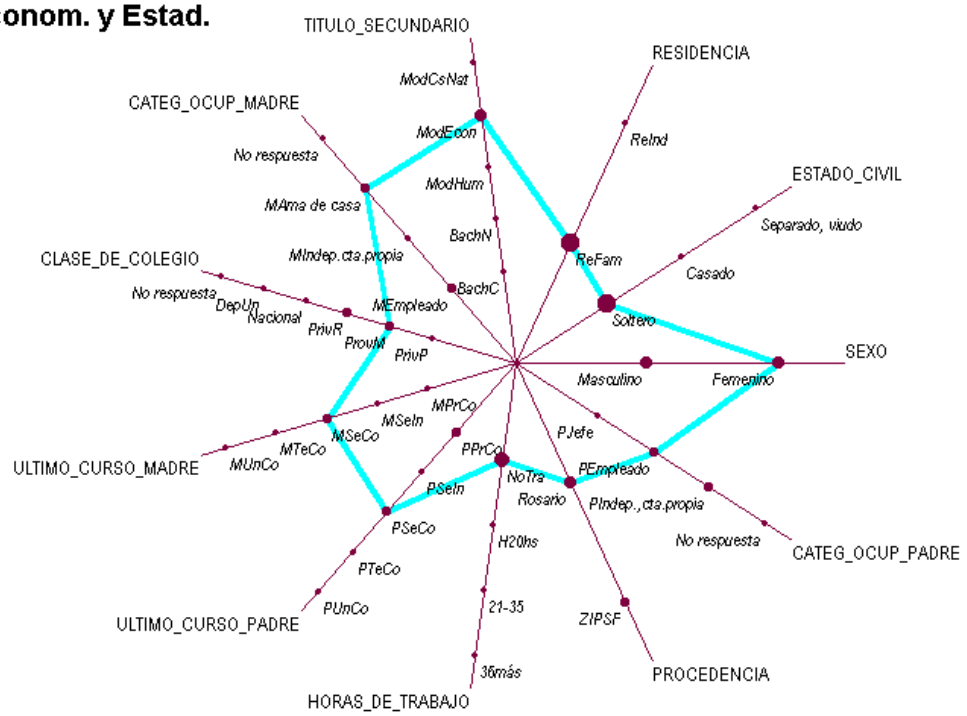
En la función Docencia, uno de los interrogantes planteados fue el perfil socioeconómico de los ingresantes. Para ello se construyeron 12 gráficos Zoom Star, uno por cada Facultad, con el objetivo de proveer una rápida comparación de los perfiles socioeconómicos de los ingresantes a cada facultad, de acuerdo con los indicadores contenidos en el formulario SUR1 de la UNR, procesados en la Dirección de Estadística Universitaria de la UNR.

En esta presentación se optó por comparar dos facultades con perfiles bien diversos: en el primer gráfico Ciencias Económicas y Estadística, con la distribución de horas trabajadas anexa (la distribución de cada variable puede obtenerse deteniendo el cursor, en el entorno del *software*, sobre cada uno de los ejes correspondientes)

El segundo gráfico se refiere a la Facultad de Humanidades y Artes. Pueden observarse rápidamente las características diferenciales de ambas facultades, sobre todo en cuanto a escolaridad de la madre y del padre, categoría ocupacional de la madre.

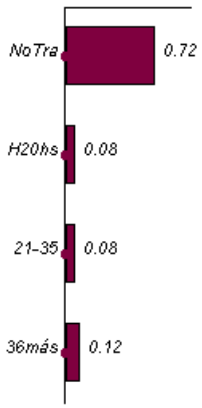
El tercer gráfico se refiere al perfil de los ingresantes del total de la Universidad, que sirve para caracterizar al ingresante más frecuente.

Cs. Econom. y Estad.

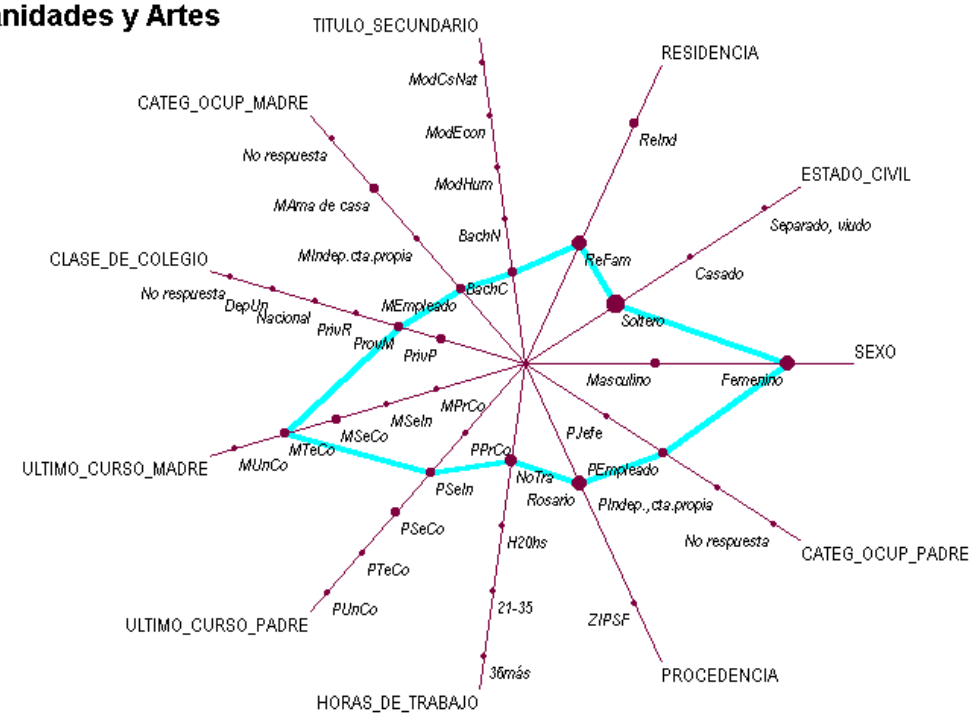


Este gráfico nos permite ver que los ingresantes a esta facultad se caracterizan por ser en su mayoría mujeres solteras que residen con su familia, proceden de Rosario, no trabajan y predomina el título Polimodal en la Modalidad Economía y Gestión de las Organizaciones obtenido en escuelas públicas provinciales o municipales. En relación a los padres son mayormente empleados y tienen título secundario completo. Prevalecen madres amas de casa, con secundario completo.

Distribución de frecuencias de horas trabajadas

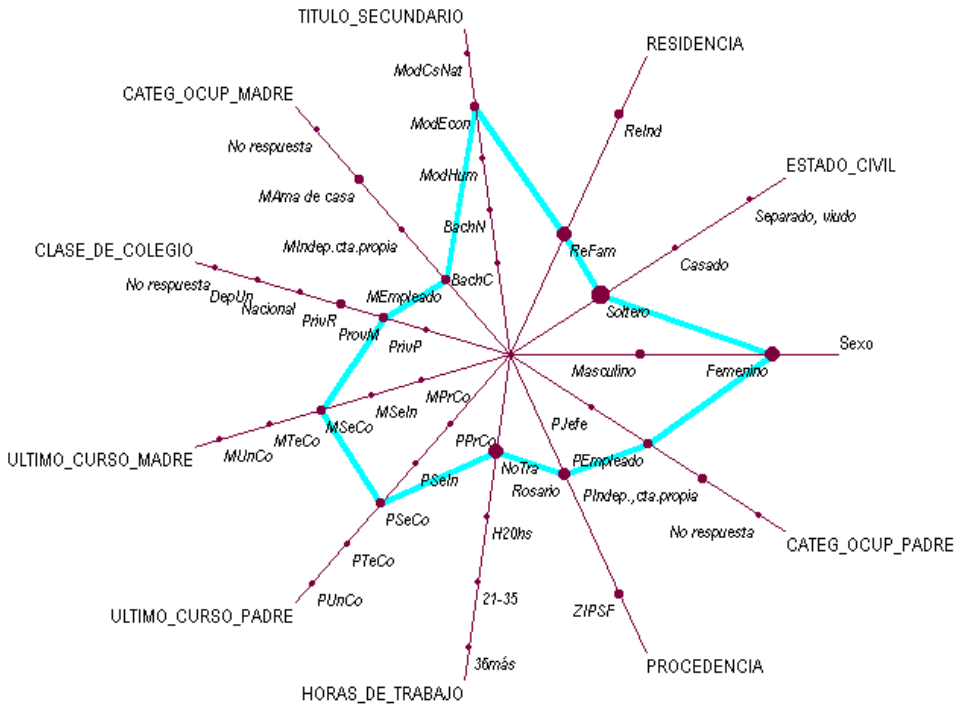


Humanidades y Artes



En este gráfico observamos que prevalecen los ingresantes de sexo femenino, solteras, que residen con sus familias, no trabajan y proceden de Rosario. En su mayoría tienen título Bachiller Comercial obtenidos en escuelas públicas provinciales o municipales. En relación a la categoría ocupacional de padre y madre, ambos son empleados. En cuanto al último curso obtenido, la madre alcanzó terciario completo y el padre secundario completo.

UNR



El ingresante más frecuente de la UNR es mujer soltera que reside con su familia, con título Polimodal en la Modalidad Economía y Gestión de las Organizaciones, de escuela provincial pública, no trabaja y procede de Rosario. Su padre es empleado con secundario completo. En relación a la madre, ésta es empleada y también tiene título secundario completo.

De esta manera, la utilización de las visualizaciones de objetos simbólicos a través de los gráficos Zoom Star permitieron una inmediata comparación de los perfiles de los ingresantes a las distintas facultades. Anteriormente, este trabajo se llevaba a cabo mediante la construcción y análisis de gran cantidad de tablas y gráficos que dificultaban su interpretación.

Bibliografía

ALBATCH, P. (2001) : *Educación superior comparada. El conocimiento, la universidad y el desarrollo.* Cátedra Unesco de Historia y Futuro de la Universidad. Colección Educación Superior. Universidad de Palermo. España.

ALBATCH, P., y KELLY, G. (Comp.) (1990): *Nuevos enfoques en educación comparada*. Editorial Mondadori. Madrid. España.

BENZÉCRI, Jean Paul (1976) : *L´Analyse des données, T.I La taxonomie T.II L´Analyse des correspondances.* Dunod. París.

CHIROLEU, Adriana (1999) : *El ingreso a la universidad. Las experiencias de Argentina y Brasil.* UNR Editora.

DIDAY, Edwin (1992): *Análisis de datos y clasificación automática numérica y simbólica*. EUSTAT, Vitoria-Gasteitz.

— (1997): *Análisis de datos simbólicos*. Ed. IRICE, Rosario.

DIDAY, Edwin, y LECHEVALLIER, Yves Symbolic (1991): *Numeric data analysis and learning*, Versailles, September 18-20. INRIA, Nova Science Publishers Inc. New York.

FERNÁNDEZ AGUIRRE, Karmele: IV International Meeting of Multidimensional Data Analysis (NGUS'97), Bilbao, September 10-12, 1997. Universidad del País Vasco, Bilbao.

GARCÍA DE FANELLI, Ana María (2005): *Acceso, abandono y graduación en la educación superior argentina*. SITEAL, Debate 5. Disponible en internet: http://www.siteal.iipe-oei.org/ [consulta: setiembre 2005].

KROTSH, Pedro (2001) : *Educación superior y reformas comparadas*. Cuaderno universitario n.º 6. Universidad Nacional de Quilmes Editorial.

LEBART, Ludovic; MORINEAU, Alain, y PIRON, Marie (1995*): Statistique exploratoire multidimensionnelle.* Dunod. París.

MOLLIS, Marcela (1993): "La educación comparada de los '80. Memoria y balance", en *Revista Iberoamericana de Educación,* n.º 2: Educación, trabajo y empleo. Organización de Estados Iberoamericanos para la Educación, la Ciencia y la Cultura.

MOSCOLONI, Nora (2005): *Las nubes de datos. Métodos para analizar la complejidad*. UNR Editora, Rosario.

Correo electrónico: piad@sede.unr.edu.ar