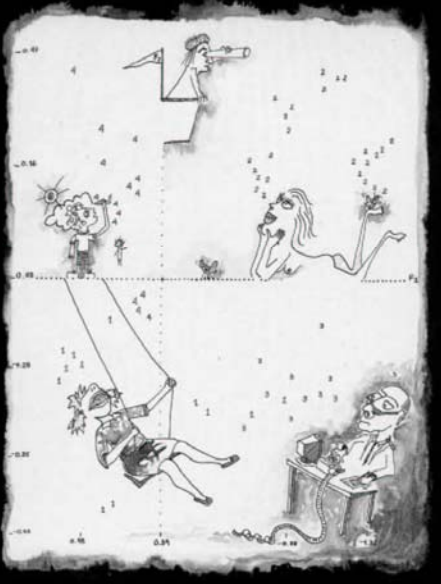
NORA MOSCOLONI

Las Nubes de Datos

Métodos para analizar la complejidad

Prólogo de Alain Morineau



UNR EDITORA

Moscoloni, Nora
Las nubes de datos : métodos para analizar la complejidad / Nora Moscoloni ; con prólogo de Alain Morineau. - 1a ed. - Rosario : UNR Editora. Editorial de la Universidad Nacional de Rosario, 2011.
E-Book.

ISBN 978-950-673-929-4

1. Estadísticas. I. Morineau, Alain, prolog. II. Título.
CDD 310.4

Fecha de catalogación: 20/10/2011

Ilustración de tapa: Elisa Kolodziej

NORA MOSCOLONI

Las Nubes de Datos

Métodos para analizar la complejidad

Prólogo de Alain Morineau

PRÓLOGO

En el curso de una larga y amigable colaboración con la autora, hemos apreciado su amplitud de miras así como el rigor que necesita quien va a hacer hablar a las cifras, muy particularmente en el dominio de las ciencias sociales.

Con el surgimiento del cálculo intensivo y el almacenamiento masivo de los datos, utilizando las computadoras omnipresentes y muy a menudo conectadas entre ellas, las actitudes, así como los objetivos del tratamiento de la información han evolucionado mucho: menos hipótesis a priori (que son a menudo hipótesis de comodidad), más enfoques gráficos, modelizaciones más flexibles, y el cuidado de caracterizar las estructuras reveladas por los datos mismos antes de modelizar los fenómenos.

En el sentido amplio del término, los datos son sinónimos de informaciones. Estas informaciones constituyen en nuestra época una nueva materia prima cuya explotación debe ser fuente de conocimiento para la comunidad y los organismos que la administran.

La obra quiere asegurar la difusión de estas soluciones que extraen ventaja de nuevos recursos aplicados a nuevos problemas. Sea que estos problemas pongan en juego datos voluminosos o complejos (ver los datos textuales no estructurados o aún los datos “simbólicos”). Sea que ellos recurran a métodos históricos, que permanecen de uso corriente, pero son a menudo revisitados en las perspectivas innovadoras.

El gran mérito de esta obra es de haber ubicado justamente en primer lugar el análisis de las relaciones simultáneas entre las “variables” (descriptores observables) como herramienta de exploración y de comprehensión de los fenómenos sociales.

Creo que el manual puede muy bien dirigirse a todo investigador en ciencias sociales enfrentado al tratamiento de datos numerosos y complejos. Será una importante fuente de información y de ideas para los usuarios de cualquier horizonte, ya que el acento está puesto en el buen uso de los métodos y las herramientas.

Esta obra está ciertamente destinada a transformar poco a poco la actitud de los investigadores en ciencias sociales hacia la estadística. Debemos felicitar a Nora y agradecérsela.

Alain MORINEAU
París, febrero 2005.

Agradecimientos

A mi familia, mis colegas y compañeros que me alentaron en la realización de este trabajo y a mis alumnos, en especial a Adriana Arca, por la grabación de mis clases.
Al Profesor Edmundo Rofman, por su permanente apoyo en este camino.
A mis maestros: Jean Paul Benzécri, Alain Morineau y Edwin Diday.

PRESENTACIÓN

Todo ser humano lleva en sí algo de astrónomo. Cuando se apagan las luces y callan las voces de la tecnología, nuestros ojos enfocan el espacio buscando respuesta a tantos interrogantes. Estrellas, nubes, galaxias, forman parte de nuestra iconografía, poética y estética, siendo el primer objeto de estudio de nuestra ciencia moderna.

Los números han formado parte indispensable en este desarrollo y el nacimiento de la Estadística está profundamente ligado al pensamiento de los astrónomos, tal vez sea por eso que las metáforas cósmicas están presentes en tantas herramientas numéricas.

Cuando hacia fines de los años ’70, en medio de un profundo cuestionamiento a mi propia disciplina, divisé las posibilidades de exploración del Análisis de Datos en el enfoque de la escuela francesa, comprendí que me introducía en un camino sin retorno que me llevaría a otra manera de ver la Estadística.

Por esos años se utilizaban herramientas cuantitativas que presentaban opciones de justificación pero brindaban pocas posibilidades para profundizar la interpretación. En la investigación social, sólo en raras ocasiones se plantean efectivamente instancias de validación experimental que son las situaciones por excelencia a las cuales responde el cuerpo de la estadística clásica. Aún en esos pocos casos, la opción “rechazo - no rechazo” acaba con toda posibilidad de obtener riqueza analítica de los datos tan laboriosamente recogidos.

Todavía la idea de la Estadística como técnica confirmatoria era demasiado fuerte para que pudieran ser aceptados esos nuevos gráficos nebulosos que nadie entendía muy bien y que para colmo se decía que cambiaban con cada procesamiento. Los investigadores clamaban por una probabilidad significativa que les permitiera acceder a la ansiada aceptación de sus trabajos en revistas científicas.

La soledad y la inseguridad producida por no contar con pares que pudieran avalar resultados (Francia estaba demasiado lejos de la Argentina de esa época) me llevó a comenzar temerariamente con el dictado de cursos; en un primer momento a colegas y luego a profesionales de otras disciplinas con los cuales profundicé mi

proceso de contaminación multidisciplinar.
A todos ellos debo gran parte de mi conocimiento en este tema ya que pude comprobar que aprender enseñando es algo más que otro juego de palabras.
Pese a ser la Estadística una disciplina conocida, controvertida y polemizada sobre todo en relación con su aplicación en las Ciencias Sociales, no abunda la bibliografía en español relativa a su devenir histórico. Lo mismo sucede con los métodos de Análisis Multidimensional de Datos (AMD), existe poco material accesible a usuarios no experimentados. Éstas son las razones que me llevaron a la construcción de este trabajo y a su publicación.
En este manual presento en la primera parte, consideraciones históricas, epistemológicas y metodológicas.
En la segunda parte realizo una presentación sencilla de las técnicas de AMD, basada en argumentos de corte geométrico e intuitivo. Ello significa hacer énfasis, no en los algoritmos matemáticos, sino en la lógica constructiva de la significación del dato.
He realizado todas las traducciones de las citas de textos en idiomas extranjeros, excepto cuando se especifica otro autor.
Los capítuos 8 y 10 se basan en cuanto a su desarrollo teórico, en el texto de las conferencias pronunciadas por Edwin Diday en el IRICE, Rosario, los días 15 y 16 de julio de 1993. (Diday, 1997)

PARTE I

CONSIDERACIONES HISTÓRICAS, EPISTEMOLÓGICAS Y METODOLÓGICAS

CAPÍTULO 1

VISIÓN HISTÓRICA Y PARADIGMAS EN LOS CUALES SE INSCRIBIERON LOS MÉTODOS CUANTITATIVOS

Acerca de la Estadística Matemática

Hablar de lo cuantitativo lleva inevitablemente a la consideración de los métodos estadísticos, los cuales han sido a menudo criticados debido a la falta de prudencia en su aplicación. Parece apropiado recordar la cita de Simiand (1922)[1] "A la mejor estadística (como a la menos buena también) no hay que exigirle ni hacerle decir más que lo que dice, y del modo y bajo las condiciones en que lo dice".

Desde los tiempos remotos el hombre se valió de instrumentos para remediar las limitaciones de sus sentidos; utilizó piedras ("calculi" en latín) para facilitar sus cuentas e inventó una técnica adecuada para vencer la incapacidad de su mente de cuantificar fenómenos colectivos. Esta técnica que permite el conocimiento de tales fenómenos, es la Estadística.

Ella nace como método de enumeración, descripción y observación de los hechos, pero emerge como teoría coherente, con el nombre de estadística matemática, entre los años 1885-1925 constituyéndose en el brazo poderoso del positivismo.

El camino hacia la objetivación

Aunque en la actualidad estamos familiarizados con conceptos tales como 'la población de la ciudad de Rosario' o el 'promedio de clasificaciones de un curso', es menos conocido el largo camino de objetivación que fue necesario para llegar a ellos. Es precisamente el camino recorrido por la estadística matemática hasta configurarse como ciencia.

¿Qué significa la objetivación? Podemos pensar que implica la superación de un problema metodológico, el que lleva a la definición de unidad de análisis y un problema ontológico, el que

[1]Simiand F (1922) *Statistique et expérience, remarques de méthode,* M. Rivière, París, p.24. Citado por Bourdieu P y ot. (1999) *El oficio de sociólogo,* Siglo XXI Ed., México, p.61.

determina que la unidad de análisis es una cosa de la realidad. Estos problemas se encuentran a caballo de gran parte de las polémicas epistemológicas de las ciencias sociales.
La reflexión sobre el devenir histórico que llevó a las gestación de tales conceptos ayuda no sólo a la comprensión de los mismos, sino que permite evaluar más adecuadamente las limitaciones de las técnicas, sus supuestos y las concepciones implícitas que conllevan su aplicación, de las cuales no siempre los estadísticos somos conscientes.
La Estadística se configura, a lo largo de los siglos como una serie de herramientas que pretendían dar respuesta a necesidades y problemas provenientes ya sea del Estado, de la Ciencia, o de ambos.
Pueden agruparse estas necesidades en tres grandes bloques, correlativos en el tiempo en cuanto a su origen, pero que en su devenir fueron interceptándose:

I. Enumerar, inventariar, administrar
II. Medir, describir, comparar, resumir
III. Predecir, dominar la incertidumbre

La primera es característica del Estado y dio origen a los primeros recuentos y censos que fueron utilizados por antiguas civilizaciones y estados imperiales. Posteriormente aparecen los registros del estado civil inicialmente en las parroquias y luego en las oficinas públicas.
La segunda nace con la Ciencia moderna pero sus herramientas son utilizadas con otros objetivos por las administraciones gubernamentales, como constatar el estado de los bienes y posesiones. Así surgen los estadísticos alemanes y los aritméticos políticos ingleses, quienes reutilizan las medidas concebidas por los astrónomos.
La tercera, a partir de la probabilidad, surge de los desarrollos matemáticos y de las cuestiones legales producidas entorno de los juegos de azar, más tarde la necesidad de conocer los errores de medición, sobre todo en Astronomía.
A lo largo del tiempo estas herramientas fueron pasando de una disciplina a otra, reutilizándose y combinándose con diferentes propósitos. Hoy en día se adoptan a menudo sin cuestionarlas o analizarlas críticamente, desconociendo su origen, su potencia y sus limitaciones.
Vale la pena entonces realizar un breve pantallazo histórico, partiendo de estos tres grupos de prácticas, para dar cuenta de

esas idas y vueltas, de los grandes protagonistas que supieron encontrar la combinación adecuada que diera solución a un problema determinado, sentando las bases para la futura disciplina.
He tomado las notas históricas principalmente de Benzécri(1982), Droesbeke(1990)[2] y Desrosières(1998).

I.- Inventariar, administrar

Origen de los censos y registros

La aparición de la necesidad estadística de poseer datos numéricos precede a su denominación en varios miles de años. En su origen, era el deseo de los jefes de Estado de conocer los elementos de su poder: población, potencial militar, riquezas, era necesario saber, en resumen, de cuántos hombres se disponía para la guerra y cuál sería el monto posible de la recaudación de impuestos. La noción del censo, o de lista de inventario, aparece entonces de una manera muy natural en la historia, implicando además una idea de precisión de la más alta calidad.
Los primeros censos (del latín censere: evaluar) parecen remontarse a la civilización sumeria, entre 5 y 2000 años AC, de la cual se conocen listas de hombres y de bienes escritas sobre tablas de arcilla. El relevamiento (término que aún mantenemos y que podemos derivar de leva) de hombres y de bienes tuvo lugar regularmente en la Mesopotamia 3000 años antes de nuestra era. Egipto parece ser la primera nación que organizó censos sistemáticos de población, al menos después del 2900 AC, así como también la que los institucionalizó con fines fiscales (2700 a 2500 AC). De la misma forma tiene este país la primacía en establecer el principio de la declaración obligatoria: fue bajo el faraón Amasis II, en el sexto siglo anterior a Jesucristo, en que se determinó que todo individuo estaba obligado a declarar sus fuentes de ingreso y su actividad. Cualquier falta a esta regla era castigada con la muerte. En fin, el mejoramiento del conocimiento cuantitativo del Estado fue característico de muchas civilizaciones, que como la china, hebrea, inca, india y griega, poseyeron un

[2]Droesbeke JJ, Tassi P.(1990) - Un detalle anecdótico nos aporta este autor cuando dice que ya la palabra Historia, que data del Renacimiento, proviene del griego antiguo historiè que en Herodoto, tiene el sentido de "encuesta".

sistema administrativo fuerte.
Por otro lado, las enumeraciones no se efectuaron con regularidad hasta que los romanos comenzaron el recuento de los habitantes de su Imperio. Los censos romanos corrían a cargo de los censores locales. Se efectuaban cada cinco años y al principio sólo se llevaban a cabo en Roma, pero en el año 5 a.C. se extendieron hasta cubrir todo el Imperio. Además de ocuparse del registro de la población y de la recaudación de impuestos, el censor se ocupaba también de mantener la moral pública.
No nos olvidemos que según relata la Biblia el nacimiento de Jesús se produce en Belén ya que sus padres debieron trasladarse a su lugar de origen con el objeto de ser censados[3].
Durante la edad media sólo se realizaron algunos censos exhaustivos en Europa. Los reyes caloringios Pipino el Breve y Carlomagno ordenaron estudios minuciosos de las propiedades de la Iglesia en los años 758 y 762 respectivamente.
En el año 721 el valí Amheser envió al califa de Damasco una detallada descripción numérica de la península Ibérica. En tiempo de Alhakén II y bajo el califato de Abd-el-Numen se realizaron empadronamientos censales. Alfonso VII de Castilla (1105-1157) concedió a los mozárabes de Toledo permiso para la formación de un catastro.
Después de la conquista normanda de Inglaterra, el rey Guillermo I el Conquistador encargó un censo. La información obtenida se recoge en el Domesday Book. Esta encuesta, realizada en 1086, tenía como objetivo inventariar de modo sistemático la riqueza rústica del país y determinar las rentas que los propietarios de las tierras tenían que pagar al rey. Este inventario se realizó a una escala sin precedentes en la Europa medieval. Los sistemas anteriores de tasación eran muy antiguos y habían quedado

[3] Biblia de Jerusalén. Evangelio según San Lucas. Cap. 2.Vers. 1 a 7. "Sucedió que por aquellos días salió un edicto de César Augusto ordenando que se empadronase todo el mundo. Este primer empadronamiento tuvo lugar siendo gobernador de Siria Cirino. Iban todos a empadronarse, cada uno a su ciudad. Subió también José desde Galilea, de la ciudad de Nazaret, a Judea, a la ciudad de David, que se llama Belén, por ser él de la casa y familia de David, para empadronarse con María, su esposa que estaba encinta. Y sucedió que, mientras ellos estaban allí , se le cumplieron los días del alumbramiento, y dio a luz a su hijo primogénito, le envolvió en pañales y le acostó en un pesebre, porque no tenían sitio en el alojamiento".

obsoletos. Al quedar registradas todas las propiedades feudales, tanto de la Iglesia como de los laicos, el Domesday Book hizo posible que Guillermo I fortaleciera su autoridad al exigir un juramento de fidelidad a todos los propietarios de tierras, al igual que a la nobleza y al clero, en cuyas tierras vivían los arrendatarios. La labor fue ejecutada por grupos de funcionarios llamados *legati,* quienes visitaban cada condado y realizaban una encuesta pública. *Domesday* es el vocablo resultante de la deformación de la palabra *doomsday* (el día del Juicio Final) la obra fue llamada de este modo por sus dictámenes relativos a las exacciones y a las tasaciones, que eran irrevocables.

A partir del siglo XIII los datos se multiplican gracias a la proliferación de los registros fiscales. El más célebre de Francia, "el estado de las parroquias y de los hogares (feu des baillages) y senescalados de Francia" constituido en 1328. El "fuego", referencia a un hogar, una familia o habitación, será un elemento esencial para estimar una población que rápidamente se transforma en un dato social más que demográfico.

En el siglo XIV comienzan los registros de las actas del estado civil.

Las actas de las Cortes de Alcalá de 1348 mencionan diversos empadronamientos. Las de Valladolid de 1351 dispusieron que se redactara el Becerro de las Behetrías, especie de catastro de los señoríos de las villas importantes de Castilla. En 1482 los Reyes Católicos realizaron un censo de sus reinos al que siguió otro después de la conquista de Granada. Años más tarde se llevaría a cabo un recuento de hogares en Cataluña, Navarra, Vascongadas y Valencia.

Felipe II, además de los empadronamientos generales de 1587 y 1594, emprendió bajo la dirección de Ambrosio de Morales una gran obra estadística que al cabo de siete años reunió solamente 636 relaciones de los 13.000 pueblos que existían en la península y que se conservan en la biblioteca del monasterio de El Escorial.

En América Latina, en la época de los monarcas borbones Carlos III y Carlos IV, movidos por un impulso de control burocrático y administrativo, se procedió al levantamiento de censos de población en los virreinatos. Constituyen un importante antecedente histórico de la preocupación censal de los gobiernos nacionales tras la independencia, que tuvieron que enfrentarse a grandes carencias y dificultades.

En agosto de 1539 Francisco I hace obligatorio el registro de

nacimientos, bajo Enrique III, el edicto de Blois hace extensiva esta obligación a los matrimonios y los decesos.
Los registros parroquiales estaban dirigidos a constatar la existencia de individuos y sus lazos familiares.
En la alta Edad Media se impuso la ley de "la letra por sobre el testimonio oral" por oposición a la inversa que era una ley mucho más antigua. De esta forma el trabajo estadístico estaba legalizado por cuanto trabajaba con unidades definidas, identificadas y estables.
Durante este período, donde se afirma la estadística administrativa, los censos son muy raros, (censo de París en 1590 - 200.000 habitantes).
Los progresos fundamentales de la estadística van a aparecer durante la segunda mitad del siglo XVII, con la necesidad de los monarcas y sus consejeros de conocer y explicar los fenómenos económicos y sociales.
Se asocia la creación del término 'estadística' proveniente del latín 'statisticum' que se relaciona con el Estado, a la escuela alemana de Gottingue y más particularmente a Gottfried Achenwall (1746). Es más probable sin embargo que este último haya sido solamente el primero en proponer una enseñanza que tratara de estadística. El empleo de la palabra es más antiguo; ya existe una Biblioteca Estadística que data de 1701 y un Microscopium Statisticum de 1672. Remontándose más aún en el tiempo, la palabra 'estadística' pertenece al lenguaje administrativo francés colbertiano (1666 a 1669).

Estadísticos alemanes y aritméticos políticos

El registro de nacimientos y defunciones dio lugar en Inglaterra en 1662, al primer estudio estadístico notable de población, titulado Observations on the London Bills of Mortality (Comentarios sobre las partidas de defunción en Londres). Un estudio similar sobre la tasa de mortalidad en la ciudad de Breslau, en Alemania, realizado en 1691, fue utilizado por el astrónomo inglés Edmund Halley como base para la primera tabla de mortalidad.
Alemanes e ingleses produjeron formas de conocimiento estadístico que reposaron en la interacción entre las dos formas de autoridad: la de la Ciencia y la del Estado. Pero estas dos formas de conocimiento variaron según los Estados: la estadística descriptiva en Alemania y la aritmética política inglesa.
El objetivo de la primera era expresar de manera analítica y

comprehensiva una comunidad humana vista como un todo en sus distintos aspectos: clima, recursos naturales, etc. Se pretendía organizar las distintas ramas del conocimiento de un estado particular, una nomenclatura inspirada en la lógica aristotélica. Al decir de Schlözer: "La estadística es historia sin movimiento, la historia es estadística en movimiento" (Desrosières, 1998:19)
La escuela alemana construyó las primeras tablas cruzadas que cruzaban estados en filas con características en columnas, ello requirió la construcción de espacios de comparación: referentes comunes, adopción de criterios a fin de reducir los objetos descritos perdiendo singularidad. Al inicio estas tablas eran literales y luego fueron transformándose paulatinamente en numéricas. Ellas fueron muy criticadas por los aritméticos ingleses quienes las llamaron estadísticas vulgares. "Estos pobres tontos están divulgando la loca idea de que uno puede entender sobre el poder de un estado simplemente a través de un conocimiento superficial de su población, su ingreso nacional y el número de animales pastando en sus campos"[4]

En el contexto de la Inglaterra de 1660, se originaron una serie de técnicas de registro y cálculo designadas con el término de aritmética política. Desde el punto de vista del origen, estos procedimientos materiales de objetivación comprendían tres etapas básicas: guardar registros escritos, contarlos y reunirlos de acuerdo con una grilla predeterminada y por último interpretarlos en términos de números, pesos y medidas.
Esta escuela precursora de la estadística inferencial, cuyos fundadores son John Graunt y William Petty está guiada por la preocupación de la cuantificación y la investigación de constantes de comportamiento que permitieran las estimaciones y las previsiones: número de niños por mujer, tiempo entre dos nacimientos para una misma madre, número de habitantes por casa, número de individuos por hogar, proporción de muertes, etc.
Las técnicas del multiplicador de la aritmética política van a ser utilizadas en detrimento de los censos, y favorecerán la aparición de las encuestas parciales. Paralelamente, el reemplazo de un conocimiento exhaustivo por una extrapolación fundada en el examen de una parte de la población es una actitud que comienza

[4] Citado por Lazarsfeld, 1970, de *Götingen gelehrteAnzeiger, c.1807,* tomado de John, 1884. Cita de Desrosières, (1998:22)

a encontrar elementos de justificación con la aparición del cálculo de las probabilidades.
La diferencia de los aritméticos políticos con los estadísticos alemanes radicaba en que éstos últimos era académicos teóricos construyendo una descripción panorámica y lógica del Estado en general.
A su vez los aritméticos políticos eran hombres de muy diversos orígenes que habían forjado un cierto conocimiento práctico y lo ofrecían al Gobierno. De esta manera se crea un nuevo rol social: el experto en un campo específico que sugiere técnicas a los gobernantes tratando de convencerlos de que para realizar sus intenciones primero debían consultarlos. Ofrecían un lenguaje precisamente articulado a diferencia de los alemanes, quienes identificándose con el Estado tenían un lenguaje general.
El auge de los estudios por muestreo se debió a la necesidad de producir estimaciones basadas en informaciones parciales. Según Desrosières(1998:29), en particular en Francia, las causas de los artificios algebraicos provocada por la ausencia de datos empíricos deben buscarse en la naturaleza secreta del absolutismo real, mientras que en Inglaterra al signo de las corrientes liberales en el poder, que protegían los derechos de los ciudadanos: dos métodos opuestos de establecer el Estado.

El censo moderno

El primer censo verdadero de una época más reciente fue elaborado en la colonia de Nueva Francia (actual Quebec), donde el recuento de individuos comenzó en 1665. Con el advenimiento de gobiernos democráticos se revela un nuevo aspecto en el proceso censal: en Estados Unidos el censo de 1790 tenía el propósito concreto de determinar la representación en el Congreso de acuerdo con la población. Este fue el primer censo en el que se expusieron con carácter público las listas con la información recogida
Durante el siglo XIX y la primera mitad del XX la práctica del censo se fue extendiendo a todo el planeta. En la actualidad han mejorado los métodos y formas de realización de los censos en la mayoría de los países. Las organizaciones internacionales como las Naciones Unidas han animado a todos los países a adoptar un sistema similar a la hora de efectuar sus censos. Dentro de la larga lista de criterios recomendados por la ONU para elaborar un censo se encuentran: lugar de residencia, estado civil, sexo, edad,

hijos, lugar de nacimiento, empleo o situación laboral, ciudadanía, lengua materna, origen étnico o adscripción religiosa, nivel de estudios, población total, distribución de la misma y características de la vivienda familiar. Los censos modernos se suelen llevar a cabo en intervalos de 5 a 10 años, quedando limitada la frecuencia por el costo y el esfuerzo requeridos para el relevamiento y verificación de los datos. Los cuestionarios son uno de los medios preferidos para recoger información.

Los censos se refieren de forma habitual a un hipotético momento de tiempo, pues algunos de los datos pueden haber cambiado durante el período de recogida de información. El uso de las modernas técnicas mediante computadora permite que la clasificación y la evaluación de los datos sea muy rápida, muchos censos se publican ahora en CD, de ahí que la parte que lleva más tiempo en los censos actuales sea la de relevamiento y procesamiento de los datos.

Los sondeos de opinión de Estados Unidos y en la Rusia pre-revolucionaria

Surgen en Estados Unidos en ocasión de la cobertura de prensa de las elecciones presidenciales. En 1824 el Harrisbourg Pennsylvanian y el Raleigh Star realizan encuestas preelectorales por consulta individual a los electores. A partir de allí distintos periódicos retoman esta práctica muy apreciada por el público.

Sin embargo las muestras no poseían ningún criterio de representatividad. Sus propiedades sólo reposaban en sus tamaños muy elevados, por lo que con el tiempo fueron reemplazados por otros más adecuados.

Los aportes de los rusos en lo que se refiere a los métodos basados en una visión parcial de la población fueron desconocidos en Occidente hasta bien entrado el siglo XX. Sin embargo, las técnicas de muestreo fueron utilizadas muy tempranamente por los estadísticos rusos, en la imposibilidad de observar exhaustivamente los territorios de su competencia. En los 'zemstvos' (gobiernos locales) existían institutos de estadística encargados de recoger las informaciones sobre el estado de la agricultura para ayudar a definir la orientación de la política económica agrícola. Se desarrollaron de esta manera importantes avances metodológicos en técnicas de sondeos y muestreo, de manera que pueden ser considerados precursores al mismo nivel que Kiaer, Bowley y Neyman.

II.- Medir, comparar, describir

Promedio

Una de los primeras herramientas, que se convertiría en siglos posteriores en la más apreciada y la base del proceso de objetivación de los números, fue la *media aritmética o promedio,* utilizada como medida representativa de los valores centrales que resumiera a toda una serie de observaciones. Al igual que en muchos otros conceptos científicos es necesario buscar en la astronomía el origen de este parámetro. El primer indicio cierto de la utilización del valor medio puede encontrarse en la obra de Tycho Brahe (1546-1601) quien se valió del recurso de realizar numerosas observaciones de una misma cantidad para estimar un determinado valor. Fueron estas mediciones las que permitieron luego a Kepler formular sus leyes sobre el movimiento de los planetas. En la obra de Tycho Brahe aparece claramente la utilización del concepto de media aritmética para eliminar los errores de las observaciones.

Variabilidad

Le debemos asimismo a los astrónomos la idea de *variabilidad,* quienes en el siglo XVIII utilizaban las medidas experimentales a fin de determinar la posición de los astros en el cielo. Ellos llegarán así naturalmente a estudiar la distribución de los errores. Un precursor fue Galileo Galilei quien se interesó en este concepto al tratar de determinar la distancia entre la Tierra y una nueva estrella.

Diferentes medidas fueron propuestas hasta llegar en el siglo XIX al término de variancia de la mano de los mínimos cuadrados, y luego al desvío tipo, denominado por Karl Pearson desvío standard (σ).

Gráficos

El origen de los *gráficos estadísticos* se remonta a la aparición del uso de las coordenadas cartesianas (René Descartes, en un apéndice del Discurso del Método titulado La Geometría)

Luego William Playfair en su obra The commercial and political atlas, publicado en Londres en 1786, presenta gráficos de gran calidad, los cuales se refieren a series cronológicas, y uno de ellos es el primer diagrama de barras conocido. En otra de sus obras,

The Statistical Breviary (1801) encontramos la presentación de diagramas de sectores.
En el siglo XIX se desarrollaron rápidamente gran cantidad de herramientas gráficas: el cartograma de C.Dupin, la ojiva de J.B.Fourier, las curvas de mortalidad de A.Quételet, el histograma de A.Guerry, la pirámide de edad de F.Walker, y la célebre superficie de correlación de Galton.
En el siglo XX los gráficos estadísticos se multiplicaron y adquirieron niveles de sofisticación que aún se siguen superando, sobre todo a partir del uso de computadoras.

Índices

La reflexión sobre *los índices* nació alrededor de tres siglos atrás en Inglaterra, con la necesidad de medir con un solo número las variaciones de un conjunto de precios entre dos fechas determinadas. En efecto, aún si los índices son utilizados en numerosos dominios como la producción industrial, la demografía, la agricultura, etc., la historia nos muestra que las primeras aplicaciones se realizaron con el estudio del nivel general de los precios o paralelamente, con el poder de compra. El desafío era combinar las modificaciones relativas de precios y de cantidades para obtener índices interpretables, como medida de una variación del nivel de los precios o como medida de bienestar.
La primera reflexión sobre el tema parece remontarse al Chronicon Preciosum, publicado en 1707 por el obispo W. Fleetwood, en el contexto siguiente: los estatutos de un colegio de Oxford, creado en 1450, exigía que todo estudiante dispusiera de una renta anual máxima de 5 libras, bajo pena de expulsión. ¿Qué significaba esta norma en 1700 teniendo en cuenta la depreciación monetaria? Fleetwood intenta responder a esta cuestión, estudiando, desde 1450, la evolución de los precios del trigo, de la carne, de los pescados, y del paño para vestidos y más precisamente evaluando el precio en 1700 de las cantidades que se podían obtener en 1450. Sus conclusiones fijaron un nivel de renta equivalente a 30 libras para 1700, lo que corresponde a una inflación media de 7,4% por decenio.
Numerosos han sido desde entonces los aportes a la teoría de los números índices que se han ido perfeccionando para tener en cuenta consideraciones relativas a las distintas escalas en los precios y cantidades introduciendo medias y ponderaciones y efectuando de diversas formas las comparaciones entre los

períodos considerados. La reflexión actual sobre los números índices puede situarse en el contexto del análisis económico a través de la teoría de la agregación y la búsqueda sobre las funciones de agregación.
Asimismo puede buscarse en la historia del uso de los índices e indicadores de qué manera ellos contribuyeron a instituir y no sólo reflejar el mundo social. A través de la construcción de estas medidas estadísticas, el Poder logra objetivar, determinándolo, el mundo real (Desrosières,1996)

Ajustes

El problema del ajuste de un conjunto de puntos representados en un sistema de ejes coordenados por una recta (y más generalmente por una curva) ocupa un lugar esencial en el desarrollo de la estadística y se origina asimismo en la astronomía. En el siglo XVIII Leonhard Euler y Johan Tobias Mayer desarrollaron independientemente uno del otro, el método de las medias que permitía ajustar puntos observados por una recta. Se ocuparon asimismo de este problema Roger Boscovich y Pierre Simon, marqués de Laplace, con su "método de situación".
Finalmente hacen mención por primera vez al *método de los mínimos cuadrados* Adrien-Marie Legendre y Carl Friedrich Gauss hacia 1805. Pero será ciertamente con la aparición de la ley normal que este método va a poder ser justificado.

Correlación y regresión

Para pasar del concepto de ajuste a los de correlación y regresión, debemos buscar su origen en la biometría. Auguste Bravais fue considerado el padre de la correlación por Karl Pearson. Pero quien juega un papel fundamental en estos desarrollos fue Francis Galton quien buscaba una medida de correlación en el análisis de dos series bivariadas. Sin embargo él comenzó por el concepto de regresión al que llamó en un comienzo "reversión".
Es interesante el rastreo de este concepto de regresión, que en la estadística actual es utilizado como predicción asociado a una probabilidad.
"Reversión es la tendencia de la media ideal del tipo familiar a partir del tipo parental, revirtiendo de lo que puede ser burda y tal vez imparcialmente descrito como la media del tipo ancestral. Si la variabilidad familiar ha sido el único proceso en la descendencia simple que afectó las características de la muestra, la dispersión

de la raza desde este tipo ideal aumentaría indefinidamente con el número de generaciones, pero la reversión chequea este incremento y lo lleva a una pausa"[5].
Galton expresa el deseo de construir un coeficiente de reversión r que indique la reducción de la variabilidad de la familia, poco después, esta reversión se transformará en regresión. En 1888, en un artículo escribe:
"Es fácil de ver que la co-relación debe ser la consecuencia de las variaciones de dos órganos en parte debida a causas comunes. Si ellas fueron totalmente debidas a causas comunes, la co-relación hubiera sido perfecta, como es aproximadamente el caso con las partes del cuerpo simétricamente dispuestas. Si ellas no fueron en ningún caso debidas a causas comunes la co- relación hubiera sido nula. Entre estos dos extremos hay un interminable número de casos intermedios, y se mostrará como la cercanía de co-relación en algún caso particular admite el ser expresada *con un solo número*" [6].
Este número es el coeficiente de correlación r que se utiliza en nuestros días.
En 1896 Karl Pearson retoma los conceptos de Galton para darles la forma con que los conocemos actualmente.
Otras medidas de asociación nacieron con el siglo XX: el coeficiente de rango introducido por Charles E. Spearman a partir de un estudio psicológico sobre la inteligencia, el coeficiente Tau (τ) de Kendall. Por su parte K. Pearson introduce la estadística "clásica" del chi cuadrado (χ^2) y Serwall Wright el path analysis.
La noción de correlación va a extenderse posteriormente hacia otros dominios como el análisis de series cronológicas y el análisis multivariado.

III.- Predecir, dominar la incertidumbre

La prehistoria de la probabilidad

En su origen la probabilidad estuvo ligada a los juegos de azar, entre los más antiguos estaba el de astrágalo o taba, ancestro del

[5] Galton F.(1877) *Typical laws of heredity in man,* Proc.R.Inst.G.Brit., vol.8, p.282-301. Citado por Droesbeke, ob.cit. En inglés en el original.
[6] Galton F. (1888) "Correlations and their measurement chiefly from anthropometric data", R.S. Proc., vol. 45, p. 135-145. Citado por Droesbeke, ob.cit. p. 25. En inglés en el original. Subrayado mío.

actual juego de dados, que se jugaba con ese pequeño hueso del talón, especialmente del cordero. Existen testimonios de estos juegos en la antigüedad, especialmente en Egipto, bajo la primera dinastía (3500 años AC) El dado más antiguo que se conoce proviene de Mesopotamia y remonta a los comienzos del tercer milenio. Estos juegos respondían seguramente a un aspecto lúdico, pero su uso tenía asimismo fines religiosos importantes ya que debían permitir adivinar la voluntad de los dioses. El nacimiento del cristianismo va a trastornar su evolución, ya que no fueron compatibles con la idea de totalidad del poder divino y su uso fue condenado por la Iglesia.

Estos antecedentes constituyen en realidad la prehistoria de la probabilidad, existiendo además muchos escritos que contienen ideas que son la base del desarrollo de la misma: Fra Luca dal Borgo (Summa de arithmetica, geometria, proportioni et proportionalita, 1494), Galileo (Sulla scoperta dei dadi, 1656), Cardano (Liber de ludo aleoe, 1663, Lyon). Puede agregarse además la visión del azar de San Agustín, donde se encuentra legalizado por la teología al ser tenido en cuenta como juicio decisorio.

Pascal, Fermat, Huyghens, Bernoulli

Se atribuye el nacimiento de la probabilidad a Blaise Pascal y a Pierre Fermat en el siglo XVII. Se pueden rastrear los primeros conceptos en la nutrida correspondencia entre ambos relativa a la Regla de los Repartos. Este problema se originó a partir del encuentro de Pascal, en el curso de su vida mundana, con el caballero de Méré, amante de los juegos de azar, quien le propuso dos problemas.

“El primero era: suponiendo que se juegue varias veces con dos dados, ¿cuántas tiradas como mínimo serán necesarias para poder apostar con ventaja que, después de haber hecho esas tiradas se habrá sacado el seis doble? La solución, que se deduce de la evaluación del número de casos favorables con relación al número total de tiradas, es que hay desventaja en intentar sacar el seis doble en veinticuatro tiradas, porque las probabilidades de perder superan a las de ganar, mientras que en veinticinco sucede lo contrario (…) El segundo problema, mucho más complejo, concierne al caso en que dos jugadores, interrumpiendo de común acuerdo el juego antes de su final, quieren hacer entre ellos un justo reparto de la puesta, o un partido, de acuerdo con la

probabilidad que cada uno tenía de ganar. Por medio de un razonamiento de una gran ingeniosidad, Pascal establece que, si han sido jugadas tres partidas y dos de ellas han sido ganadas por el uno y una por el otro, siendo el azar el mismo y debiendo ser ganado el juego por aquel que gane las tres partidas, corresponden sobre una puesta de 16 pistolas 12 al primero y 4 al otro; luego, de deducción en deducción reduce todos los demás casos a éste. Pascal inventa de esta forma uno de los métodos analíticos del cálculo de probabilidades. El otro, basado en la teoría de las combinaciones, fue ideado al mismo tiempo por Fermat"(B.Pascal, 1981:657).

De esta manera los problemas tratados por Pascal sobre la teoría de las probabilidades sobrepasan largamente una simple cuestión de juegos, para internarse en reflexiones jurídicas importantes que se pueden dar en las decisiones de particionar.

Pascal señala la existencia de un "...tratado totalmente nuevo, de una materia absolutamente inexplorada hasta ahora: la repartición del azar en los juegos (...). Así, uniendo el rigor de las demostraciones de la ciencia a la incertidumbre del azar, y conciliando estas cosas en apariencia contrarias, ella puede (...) arrogarse con todo derecho este título sorprendente: La Geometría del Azar"[7]

Por su parte Christiaan Huyghens hacia 1650 propone el principio de la esperanza matemática e introduce el muestreo con o sin reposición. Su obra tuvo influencia en el siglo siguiente sobre Pierre de Montmort, así como sobre Jacques Bernoulli y Abraham de Moivre.

Los estadísticos descuidamos a veces que los importantes desarrollos en probabilidad pertenecen a muchos Bernoulli: toda una familia. Una generación de cuatro hermanos de los cuales interesan en especial los aportes del mayor, Jacques (1654-1705) y de Jean (1667-1748). El primero escribió un tratado, Ars Conjectandi, en la línea de Huyghens, que sirvió de base al primer teorema límite: la ley de los grandes números, donde se pregunta sobre el cálculo de la probabilidad de obtener cara con un dado mal calibrado. Encuentra una respuesta a esta cuestión demostrando que si se repite un gran número de veces la experiencia de lanzar este dado, la frecuencia de aparición de

[7] Memoria de Blaise Pascal (1654) a la Academia parisiense, escrita en latín. Citada por Droesbeke, ob.cit. p. 25.

cara se acerca a una cantidad susceptible de ser calificada como la probabilidad de obtener ese resultado. Se trata de la forma elemental de la ley citada.
Luego se destacan su sobrino Nicolás (1687-1759) y el hijo de Jean, Daniel (1700-1782). Guiado por sus tíos, Nicolás aplicó algunos de sus principios al dominio jurídico. Daniel es conocido en probabilidad por el problema de la esperanza matemática infinita y por su investigación en una teoría probabilística de los errores de observación en astronomía. Se puede mencionar también a Jean (1744-1807), sobrino de Daniel, que colaboró en cierta manera a la teoría de la probabilidad.

La teoría de los errores

Consiste en describir la ley de probabilidad de los errores (diferencias entre el verdadero valor de una cantidad y las medidas provistas por los valores observados). Fue introducida con el objeto de mostrar la utilidad de tomar la media aritmética de los valores observados para 'estimar' un parámetro.
Simpson introdujo la ley uniforme discreta en 1756. Luego recurriendo a la función generatriz utilizada por Abraham de Moivre que tendía a resolver los problemas ligados a los lanzamientos de dados es que Simpson obtiene, para estas dos distribuciones, la ley de probabilidad de la suma de n errores independientes (y por consecuencia, de la media aritmética de éstos últimos).
Sucesivas proposiciones de funciones por parte de Joseph-Louis, conde de Lagrange, Abraham de Moivre, J.H.Lambert, desembocaron en la expresión "teoría de los errores" como la conocemos en la actualidad. Pero todas estas leyes se referían a variables que tomaban sus valores en un intervalo determinado. Fue Pierre Simon, marqués de Laplace, el primero que propuso una ley donde el dominio de definición de la variable era la recta real. Ella llevaría a un hito importantísimo en la historia de la Estadística.

La ley normal

Si bien la ley normal es habitualmente atribuida a Laplace y Gauss, su origen es más antiguo y reposa en los trabajos de Jacques Bernoulli. Trabajando sobre el problema expuesto por Bernoulli, De Moivre descubre una fórmula aproximada de las "probabilidades binomiales" y puede ser considerado el inventor

del primer teorema central del límite.
Carl F. Gauss fue un matemático y un astrónomo brillante. La ley normal fue introducida por él recurriendo al método de los mínimos cuadrados. Retoma el problema del ajuste ya abordado por Boscovich, Euler, Laplace, Legendre y Mayer motivado ahora por el estudio de las órbitas planetarias. El aporte innovador de Gauss se sitúa en el carácter probabilístico de su enfoque.
En 1810 Laplace introduce su segunda ley de los errores, mostrando que la distribución de las medias aritméticas de n errores independientes, teniendo la misma precisión o precisiones distintas, puede ser aproximada por la distribución normal, con un error de aproximación que tiende a cero cuando n crece indefinidamente. Es por lo tanto a través de un teorema de límite central que Laplace obtiene esta distribución. Esta ley fue asimismo obtenida por Adrain, en 1809, deseoso como Gauss, de justificar el método de los mínimos cuadrados.
Las primera tablas que suministraban los valores de probabilidad normales datan de 1899, y corresponden al físico francés Kramp. Es interesante notar que la distribución fue calificada de "normal" por Pearson, un biometrista, en 1893 y es conocida ahora bajo esta denominación.

Otras leyes de probabilidad

La ley normal fue ampliamente aceptada en el siglo XIX, sin embargo fue posteriormente muy criticada en cuanto a las hipótesis subyacentes y a la universalidad de su aplicación.
Siméon Denis Poisson considera en 1824, la ley que será luego llamada ley de Cauchy, y constata que la media aritmética de n errores independientes distribuidos según esta ley no tienden hacia una ley normal. Asimismo, introduce la llamada luego distribución de Poisson como distribución límite de la ley de Pascal y de la ley binomial, a través de la relación existente entre estas dos últimas.
Entre los trabajos que reposan sobre críticas respecto de la ley normal habría que citar las proposiciones de correcciones introducidas por la escuela escandinava (Gram, Charlier, Thiele...)
Otras leyes merecen ser citadas: Chi-cuadrado, t de Student, de Fisher – Snédecor.

Aproximaciones y leyes límite

La historia de los teoremas asintóticos es extremadamente rica.

Ya mencionamos la ley de los grandes números introducida por Bernoulli en 1713 así como el teorema central del límite que se refiere a la convergencia hacia la ley normal de una suma de un número creciente de variables aleatorias. Este teorema fue tratado y enriquecido sucesivamente por diferentes autores entre los que cabe citar a De Moivre, Laplace, Poisson, Cauchy, Bieynamé, Chebychev. El nombre de teorema central del límite fue propuesto por G. Polya el cual en un artículo aparecido en 1920, habla de "central limit theorem of probability theory".

Teoría moderna de las probabilidades

Los fundamentos de la probabilidad matemática se deben a Kolmogorov (1933), quien desarrolló su teoría axiomática inspirándose en el desarrollo de la teoría de la medida.

En 1881, A.Harnack introduce la noción de conjunto discreto y luego el de "igualdad en general", que "mide" el concepto de igualdad casi segura. La "medida" de un conjunto aparece así en numerosos trabajos; las definiciones son diferentes según los autores (Cantor, Peano, Jordan) Pero fue Emile Borel quien introdujo la noción de conjunto de medida nulo. Trabajando sobre R, y más precisamente sobre [0,1], privilegia la clase de conjuntos abiertos, que "mide" a partir de la longitud de los intervalos y muestra que se puede definir sobre esta clase (que llevará luego el nombre de clase boreliana) una "medida" μ que verifica las propiedades de σ, aditividad y diferencia.

La teoría de la medida de Borel abrió el camino a la integral de Lebesgue, o integral generalizada. En 1901, Lebesgue elabora una teoría de la integración más general que la de Riemann que se usaba hasta entonces, a partir del concepto de medida introducido por Borel. Bajo el impulso de J.Radon, los conceptos de medida y de integración abandonarán rápidamente R y R^n ; así aparecerá la teoría de la medida abstracta definida sobre un conjunto Ω cualquiera provisto de una tribu (1913). Es esta noción de medida abstracta la que ha permitido el desarrollo de la teoría matemática de las probabilidades. Por ejemplo, para definir en todos los casos la esperanza E(X) de una variable aleatoria X, hacía falta la integral de Lebesgue y sus extensiones. El logro de la construcción de las teorías de la medida y de la integración pueden fijarse alrededor de 1930, con los teoremas de descomposición de Lebesgue - Nikodym y la existencia de las densidades.

De esta manera la teoría de la medida y de la integración han sentado las bases de una formalización homogénea y coherente de las probabilidades, introduciendo el rigor del razonamiento matemático. Recordemos que Keynes, en 1921, escribió a propósito de las probabilidades que "los sabios revelan un resto de astrología o de alquimia" y que Von Mises es todavía más directo al afirmar (1919) que "el cálculo de las probabilidades no es una disciplina matemática".
Siendo la probabilidad un caso particular de medida se ve entonces, a partir de los años 30, aplicar una gran parte de los resultados de medida y de integración. A partir de allí muchos autores le conferirán título de nobleza. Puede citarse, en particular, la teoría de la adición de variables aleatorias de Lévy (1937), la de variables aleatorias y distribución de probabilidad de Cramer (1937), las distribuciones límites de sumas de variables aleatorias independientes de Gnedenko y Kolmogorov (1949).

Estimación y tests

La teoría de la estimación es desarrollada por Ronald Fisher; en palabras de Benzécri(1982:30): "La era del pescador, es la era de Sir R.A.Fisher, cuya obra amplia y diversa ha recubierto todos los problemas de la estadística: búsqueda de leyes, estimación, planos de experiencia, discriminación, sin olvidar la genética que había inspirado asimismo a Galton y Pearson"(…)
"*Era Piscatoria.* Cuando en 1936 Karl Pearson deja este mundo, G.U.Yule, (1875-1951) pudo exclamar : "I feel as though the Karlovingian era has come to an end, and the Piscatorial era which succeeds it is one in which I can play no part"[8]. En efecto de las manos de Karl el cetro de la estadística pasó al pescador, es decir a R.A.Fisher (1890-1962). Como K.Pearson, Sir Ronald Fisher fue un pensador original, un doctor de la ciencia que despreciaba la penumbra de los compromisos. En casi todo, hasta lo que pudo, Fisher se opuso a Pearson, pero ahora que el silencio se ha hecho sobre el campo de batalla, se puede afirmar serenamente que aquél fue el continuador de éste".
"Desde el punto de vista del análisis de datos, esta obra es preciosa en cuanto a que ha acostumbrado a los estadísticos a la geometría multidimensional. Pero las reglas de Fisher para la

[8] En inglés en el original, juego de palabras: Karlovingian por Karl, Piscatorial por Fisher (pescador)

inducción fundadas sobre una expresión exacta de las leyes probabilísticas, lo han conducido a nuestro parecer a esperar demasiado del modelo normal. Es por esto que el análisis de datos asistido por computadora que trata tal cual son a las distribuciones empíricas – nube de puntos – se alejó en sus comienzos a menudo con violencia, del método generado por Fisher" (…)

El concepto de *test de hipótesis* y la teoría subyacente aparece más tardíamente que la teoría de la estimación. Los fundadores de la teoría de los tests de hipótesis son Jerzy Neyman y Egon Pearson, hijo de Karl y tiene lugar entre 1926 y 1933. Su aporte esencial es el de elaborar una concepción estadística diferente del modelo de verosimilitud de Fisher a través de un modelo decisional. Ellos presentan en efecto su óptica como un problema de elección entre dos decisiones posibles: aceptar (que luego será no rechazar) o rechazar la hipótesis privilegiada.

Emergencia de la estadística matemática

La ciencia moral

La anterior es una reseña no exhaustiva acerca del origen de las distintas técnicas, provenientes de distintas disciplinas, contextos y necesidades, con las cuales emerge la estadística matemática entre los años 1885-1925.

Uno de los grandes protagonistas de esta gestación fue Quételet, astrónomo belga, considerado como el primero en hacer de la estadística una "ciencia moral", y quien aplica su espíritu enciclopedista tanto a la sociología como a la geometría y a la meteorología. Se asocia a Quételet con la doctrina del hombre medio que no fue unánimemente aceptada. Pero la originalidad de Quételet no fue la de haber calculado las medias en antropometría, sino la de haber considerado atentamente la dispersión de las medidas (como las tallas de una población de hombres) y descubrir que la ley normal (que él como astrónomo conocía bien) ofrecía una descripción aceptable.

Asimismo sus participaciones en los diferentes congresos de estadística le dieron la ocasión de defender el principio de una estadística científica basada en el cálculo de las probabilidades, relacionándolo de esta manera con los elementos en uso de la estadística descriptiva.

Fue Quételet quien entre 1830 y 1840 difundió ampliamente la idea que conectaba la teoría de la probabilidad con las observaciones estadísticas. Este concepto unió el azar impredecible del comportamiento individual con la regularidad (y consecuente predictibilidad) de la sumatoria estadística de estos actos individuales, a través de la noción de *hombre medio*. Éste estaba basado a la vez en la generalidad de la distribución de probabilidad de Gauss (futura ley normal) y en conjuntos de "estadísticas morales" (matrimonios, crímenes, suicidios) desarrolladas por las oficinas de estadística.
Esta forma de argumentación va a proyectar a la teoría de la probabilidad emergiendo desde su lado subjetivo, epistémico, expresado en términos de "razones para creer" hacia su lado objetivo, frecuentista: la regularidad de las medias, opuesta al caos y la impredictibilidad de los actos individuales, proveyendo así una poderosa herramienta de objetivación.

Disciplinas emparentadas

El clima intelectual de fines del siglo XIX estaba fuertemente impregnado de darwinismo, que representaba la modernidad científica. Ya asentadas las bases de la antropometría a partir de Quételet, nace la Biometría fundada por Galton, quien era primo de Charles Darwin, y que tenía por objetivo fundamentar las tesis del mismo sobre pruebas estadísticas y aplicarlas a seres humanos. Es así que asocia "valor cívico" con "valor genético" a través de un indicador de aptitud natural individual.
El espíritu de la escuela biométrica podría ser descrito con palabras de Galton: "Hasta que los fenómenos de cualquier rama del conocimiento no son sometidos a medidas y números, ella no puede asumir el estado y dignidad de una ciencia"[9]. Vemos en esta afirmación la necesidad de justificación de la Estadística como una ciencia, así como el principio que llevó a la cosificación del número.
Galton consideraba a la distribución Normal como una ley de desviación que permitía clasificar a los individuos, más que como una ley de errores a eliminar, tal como había sido para los astrónomos. La eugenesia fue concebida por Galton como una nueva ciencia y una nueva moral que se oponía al oscurantismo de la Iglesia, pero que daría posteriormente pie a las aberrantes

[9] Citado por Egon Pearson, 1948:92

teorías sobre mejoramiento de la raza humana.

Por su parte cabe reconocerle a Karl Pearson, continuador de Galton, la conquista de la autonomía del discurso estadístico, empalmando su crítica filosófica de la causalidad con su nueva idea de correlación.
En Pearson es curioso notar su transformación ideológica: de ser simpatizante del socialismo en su primera juventud, en ocasión de su viaje a Alemania donde toma contacto con las teorías socialistas en auge en ese momento[10], pasa luego a través de la teoría de la evolución, a ser director del Laboratorio de Eugenesia[11]. Considerándose discípulo de Mach, quería combatir el idealismo y el pensamiento metafísico a favor de un positivismo científico. Sin embargo si el proyecto científico de Pearson era la biometría, el proyecto político era la eugenesia.
Los desarrollos de Galton y Pearson en cuanto a la ley normal bivariada, van a encontrar posteriormente otro campo de aplicación con el nacimiento de la Psicometría o análisis factorial, escuela esencialmente americana que se propuso la medición estadística del intelecto. Se atribuye generalmente a Charles Spearman el primer artículo de análisis factorial cuyo título sorprendentemente pretencioso fue: "General intelligence objectively determined and measured" (Amer. Jour. Psychol., 1904). Más tarde, hacia 1930, Louis Thurstone retomará los trabajos de Spearman para desarrollarlos introduciendo el naciente cálculo matricial.
Por su parte cabe atribuir al interés de Yule por los problemas sociales el nacimiento de la Sociometría.

Un edificio sostenible

¿Qué fue necesario para que la Estadística se sostuviera?

[10] "Que el joven Pearson fuese socialista era casi inevitable, puesto que la filosofía de vida que estaba desarrollando en aquellos días vinculaba lo moral con lo social, lo social con el bienestar y felicidad de la mayoría". (Egon Pearson, 1948:30)
[11] "Fue (Pearson) Profesor de Eugenesia, pero la organización de la cual estaba encargado, se ocupó de un campo más amplio; así el Laboratorio Biométrico, sostenido por fondos de la Drapers' Company, y el primitivo Laboratorio Eugenésico de Galton, se vieron incorporados a un nuevo Departamento de Estadística Aplicada" Ibidem, p. 121

Como vimos, la estadística moderna deriva de la recombinación de prácticas científicas y administrativas que inicialmente estuvieron muy separadas.
El vínculo entre los dos mundos, de la ciencia y la práctica, se estableció a través de la tarea de objetivación, de hacer cosas que se sostengan, ya sea porque ellas son predecibles o porque si son impredecibles, su impredictibilidad puede ser manejada hasta cierto punto, gracias al cálculo de probabilidad.
Este rasgo pone en claro la relación entre las probabilidades y la forma en que ellas reflejan el juego del azar y las apuestas, con las descripciones macrosociales de las estadísticas controladas por el Estado.
Como ya vimos estos dos dominios se interceptan constantemente, encontrándose y distanciándose dependiendo del período. La intersección ya ocurrió en el siglo XVII con el uso de las tablas de mortalidad como base de los sistemas de seguros, o con las estimaciones realizadas por Laplace de la población de Francia, basadas en una “muestra” tomada en unas pocas parroquias. Sin embargo volvemos a Quételet para rescatar esta relación a través del hombre medio.
La cosas engendradas por los cálculos de medias están envueltas en una estabilidad producto de la introducción del rigor y los métodos de las ciencias naturales en las ciencias humanas. Podemos entender el entusiasmo que creó esta posibilidad entre los que, entre 1830 y 1860, establecieron las oficinas de estadística y los congresos internacionales organizados para propagar este nuevo lenguaje universal y para unificar los métodos de registro.
El proceso de objetivación, proveyendo sólidos elementos en los cuales basar el manejo del mundo social, resultó de la conjunción de dos dominios diferentes. Por un lado, el pensamiento probabilístico apuntaba a dominar la incertidumbre, por el otro la creación de espacios políticos y administrativos equivalentes permitía registrar un gran número de eventos y resumirlos según normas estandarizadas.
La unión de estos espacios permitió la posibilidad de extraer muestras representativas de urnas, con el objetivo de describir fenómenos socioeconómicos con poco costo, gracias a las encuestas por muestreo. En realidad los sistemas de probabilidad a través de urnas pudieron ser concebidos y utilizados gracias a estos espacios de equivalencia, más prácticos que cognitivos,

construidos políticamente.

La Estadística entre relativistas y objetivistas

Las medidas o entidades cuantitativas, que son puntos de referencia en el debate entre relativistas y objetivistas, son también ellas mismas sujeto de debate.
Las controversias pueden ser de dos clases, dependiendo de si ellas conciernen sólo a la *medida* o al *objeto* mismo.
En el primer caso la realidad de la cosa medida es independiente del procedimiento de medición, lo que se cuestiona es éste último, ¿son fiables las mediciones que se realizan? ¿son adecuadas las operaciones puestas en juego para llegar a ellas? Se trata de un problema metodológico.
En el segundo caso la existencia y definición del objeto es visto como una convención, motivo de discusión, el problema es aquí ontológico.
La tensión entre estos dos puntos de vista, uno enfocando los objetos a describir como cosas reales, el otro como resultado de convenciones en el trabajo real, ha sido un rasgo en la historia de las ciencias humanas, de los usos sociales que se le han dado, y de los debates que originaron.
“La primera regla, y la más fundamental, es: considerar los hechos sociales como cosas”. Cuando Durkheim formuló su regla del método sociológico en 1894, ubicó a los hechos sociales en una perspectiva de objetivación característica de las ciencias naturales. Sin embargo esta fórmula es ambigua, ya que puede ser leída de dos maneras: como una afirmación de realidad o como una elección metodológica. En la primera se expresaría que los hechos sociales *son* cosas mientras que en la segunda que los hechos sociales deben ser *tratados como si fueran* cosas. Para Bourdieu, sin embargo, la lectura es clara cuando afirma que Durkheim mismo en el segundo prefacio de Las Reglas del Método Sociológico, explicó que se trata de precisar una actitud mental y no de asignar al objeto un status ontológico (P.Bourdieu, 1999:52)
Estas dificultades son análogas a las encontradas, en el curso de la historia, por los creadores del lenguaje estadístico que habilita a establecer hechos sociales como cosas.
Precisamente, el largo camino recorrido por la estadística hasta

lograr la objetivación debe ser tenido en cuenta a los fines de no caer en la ingenuidad de considerar los números como cosas.
En la actualidad esos lenguajes están basados en conceptos claramente definidos y precisamente formalizados de manera sintética: medias, desviaciones estándar, probabilidad, categorías idénticas o equivalencias, correlación, regresión, muestreo, ingreso nacional, estimaciones, tests, máxima verosimilitud, mínimos cuadrados.
El estudiante, investigador o usuario de técnicas estadísticas se encuentra con estos conceptos compactos, encapsulados en concisas fórmulas, sin embargo estas herramientas son el resultado de gestaciones históricas a veces desacompasadas, marcadas por dudas, retrocesos e interpretaciones conflictivas.
Para comprender el alcance y el sentido de estos conceptos, es necesario una reflexión sobre cuestiones que han sido debatidas durante décadas o en algunos casos durante siglos. En este sentido la Historia nos enseña cómo los hechos sociales se transformaron en cosas y en especial para todo el que use técnicas estadísticas.
La historia de la gestación de estas técnicas nos permite, a medida que vamos remitiéndonos a las distintas polémicas, establecer conexiones entre el lenguaje de las mismas y su utilización en el debate social.
"El razonamiento estadístico sólo puede ser reintegrado a una cultura científica reflexiva a condición de que retornemos a estas traducciones y debates, redescubriendo senderos inciertos y momentos de innovación, que siempre conforman nuevos nexos entre viejos esquemas"(Desrosières, 1998:10)
Las herramientas estadísticas reflejan las concepciones a partir de las cuales se fueron realizando distintas descripciones del mundo y de la forma en la cual actuamos en él. De ambas podemos decir a la vez que son reales y que han sido construidas, luego repetidas en otros contextos y han circulado, arrancadas de sus orígenes, con otros propósitos.
Es así como ciertas herramientas creadas para la estadística matemática fueron reformuladas, desde otros enfoques, para ser utilizadas por la escuela francesa del análisis de datos.

CAPÍTULO 2

LA ESTADÍSTICA TRADICIONAL Y LA PULVERIZACIÓN DEL INDIVIDUO. REVALORIZACIÓN DE SU IMPORTANCIA A TRAVÉS DEL ANÁLISIS MULTIDIMENSIONAL DE DATOS

Estadística clásica y análisis multidimensional de datos

El desarrollo de la estadística matemática fue históricamente paralelo al del positivismo. Esta etapa, que comienza a fines del siglo XIX y se prolonga a lo largo del siglo XX, es considerada como de la estadística clásica o tradicional, de neto corte probabilístico. Se caracteriza en líneas generales por:

- *Énfasis en la población y no en el individuo*. Veamos cómo definen la Estadística grandes teóricos como M.G.Kendall y A.Stuart (1963:10) "La noción fundamental en la teoría de la Estadística es la de grupo o agregado, un concepto para el cual los estadísticos usan una palabra especial: "población". Este término será en general empleado para denominar todo conjunto de objetos en consideración, ya sean animados o inanimados; por ejemplo se considerarán poblaciones de hombres, de plantas, de errores en la lectura de una escala, de alturas barométricas en diferentes días y aún poblaciones de ideas. Tales como todas las posibles maneras de repartir un mazo de cartas.
- La noción común de todas estas cosas es la de masa, conjunto.
- La ciencia de la Estadística trata con las propiedades de las poblaciones. Considerando una población de hombres, nosotros no estamos interesados, estadísticamente hablando, en si algún individuo en particular tiene ojos marrones o es un falsificador, sino más bien en cuántos de los individuos tienen ojos marrones o son falsificadores o si la posesión de ojos marrones está relacionada con una propensión a la falsificación en la población.
- Nosotros estamos, para aclararnos, interesados en las propiedades de la población misma. (…) El estadístico, como la Naturaleza, está principalmente interesado en las especies y no en el individuo."
- *Énfasis en la inferencia probabilística.* El cuerpo de problemas a resolver por los métodos estadísticos tuvieron que ver, al compás de los desarrollos en la teoría del cálculo de probabilidad, con la antigua necesidad de realizar pronósticos pero que ahora eran considerados sobre bases científicas y no

metafísicas. Al mismo tiempo, proporcionar un fundamento dentro de los cánones de la ciencia moderna, para la toma de decisiones a través de la medición del error posiblemente cometido.

- *Aceptación de modelos a priori.* La estructura de la teoría de la probabilidad está edificada sobre supuestos fuertes, entre ellos los relativos al ajuste de las distribuciones empíricas a determinadas leyes de probabilidad teóricas, por lo tanto la realización de inferencias probabilísticas implica necesariamente el tácito acuerdo previo acerca de los principios fundamentales de los modelos a aplicar.
- *Utilización de muestras probabilísticas.* Otro aspecto relacionado estrechamente con el cumplimiento del modelo probabilístico fue el desarrollo de las normas de extracción de muestras aleatorias cuyo estricto cumplimiento pudiera garantizar el ajuste de los datos a los supuestos adoptados.
- *Escaso número de variables.* La mayor parte de la teoría estadística fue desarrollada, al igual que otras ciencias, antes del advenimiento de las computadoras, de esa manera gran parte de los avances en las técnicas respectivas se dedicaban a la resolución de problemas con menor costo de cálculo, como consecuencia de ello los métodos apuntaban a minimizar y/o sintetizar el número de variables en juego. Un ejemplo de ello es también la construcción de índices.
- *Tratamiento predominante de variables continuas.* El cuerpo fundamental de la Estadística clásica estuvo dedicado al tratamiento de variables medidas en escalas de razón o intervalo. Evidentemente las observaciones de los fenómenos físicos y naturales eran compatibles con ese tipo de mediciones. Cuando los métodos estadísticos se aplicaron a los fenómenos sociales, los esfuerzos estuvieron puestos más que en desarrollar nuevos métodos, en tratar de forzar la adecuación de las escalas, predominantemente nominales u ordinales en este campo, para lograr su procesamiento como si fueran continuas.
- *Nivel abstracto de los resultados.* Los resultados obtenidos mediante los métodos estadísticos tradicionales requieren un entrenamiento especial, ya que se refieren a entidades con un alto grado de abstracción. Difícilmente podrá un usuario que haya recogido su propios datos comprender los resultados estadísticos en términos de cantidades y magnitudes que

respeten los niveles de medición originales. El grado de abstracción de los resultados lleva a que no puedan interpretarse los mismos sin el conocimiento del encadenamiento lógico que lleva a la postulación de las hipótesis estadísticas, las cuales pueden encontrarse deductivamente alejadas de las hipótesis de investigación. De la misma manera la interpretación de un coeficiente de correlación requiere de un conocimiento especial para llegar a la significación en términos reales, del tipo y la intensidad de la relación entre dos variables.

- *Condición irreversible de los resultados.* Por el hecho de garantizar la objetividad, los resultados son irreversibles, si un experimento resultó no significativo, deberá intentarse otro camino, se considera que existe una gran distancia entre el punto de vista del investigador y los datos.
- *Gráficos de resultados.* El énfasis en el desarrollo de artilugios gráficos estuvo puesto especialmente en la representación de los fenómenos cuantitativos estudiados, sobre todo en la manera que los mismos pudieran captar la atención del lector presentando en forma fidedigna los resultados.

Con la aceptación de otros paradigmas, en especial en las ciencias sociales, las manipulaciones cuantitativas se desprestigian y pasan a ser poco menos que prohibidas en ciertos círculos académicos.

Mientras tanto surgen otras corrientes, algunas parten del tronco común sin abandonar totalmente los presupuestos del modelo probabilístico básico, intentando el desarrollo de otros métodos adaptables a escalas de medidas "inferiores"[12] como la ordinal o nominal o a datos que no se adaptan a las distribuciones de probabilidad conocidas, tal el caso de los métodos no paramétricos. Otros enfoques, como el análisis de datos (o análisis multidimensional de datos) en la versión de la escuela francesa, surge en la década del '70, planteando fines menos deterministas. En el *analyse des données*[13]*,* el objetivo general es

[12]En este contexto se consideran las escalas de medidas inferiores o superiores según la cantidad de operaciones básicas, y medidas estadísticas, susceptibles de realizar con las variables correspondientes a cada nivel de medición. (Ver Baranger, 1992:15)

[13]El término en su idioma original tiene una connotación diferente que su traducción literal, la cual podría interpretarse como un análisis de datos

la búsqueda de una estructura presente en los datos, en un contexto de tipo más inductivo que deductivo, que revaloriza el rol del individuo pero sin dejar de considerarlo como una observación. Su naturaleza fundamentalmente descriptiva y el acercamiento geométrico a los problemas asignan un rol muy importante a las representaciones gráficas sobre todo en una etapa exploratoria. Se nutre de algoritmos adaptados a diferentes niveles de complejidad de la información: datos numéricos, textuales, simbólicos.

Esto es particularmente importante en el campo de las ciencias sociales y humanas, donde el objeto de investigación presenta mayor nivel de complejidad y no admite reducciones simplificadoras. En este sentido el análisis multidimensional de datos posee una mayor capacidad para profundizar el conocimiento del objeto de investigación al permitir la exploración de las diferentes dimensiones del mismo.

Posee gran vinculación con el Exploratory Data Analysis desarrollado por J.Tukey (1977), en el sentido de no trabajar con modelos a priori. Se dice que el análisis exploratorio no es sólo un conjunto de técnicas, sino una actitud, una flexibilidad, una confianza en las representaciones gráficas. Como nexo entre la estadística y distintos campos de aplicación, el análisis exploratorio afronta el desafío de transformar los datos en conocimiento útil. Asimismo valoriza el análisis detallado de los puntos que se escapan a la norma: "es mejor resistir los efectos de los *outliers*[14] en el ajuste de un modelo, aún cuando no comprendamos el motivo por el cual son extremos" P.Velleman y D.Hoaglin (1992)

En la práctica del análisis exploratorio, el estudio de los outliers permite a veces formular hipótesis inesperadas.

El análisis exploratorio, es visto en la tradición anglosajona como una etapa previa, heurística, que deberá ser completada con el 'análisis confirmatorio' es decir de pruebas de hipótesis probabilísticas. Esto no sucede en el análisis de datos a la francesa, donde la búsqueda de la estructura subyacente en los datos es un fin en sí mismo.

Veamos cómo planteaba Benzécri (1976,T.II:3-17) la situación en los '70. Los principios básicos del análisis de datos son los

---

ordinario y no el nombre de una escuela de pensamiento.

[14] Individuos que se escapan a la norma

siguientes:

- *1er. Principio*: "Estadística no es probabilidad. Bajo el nombre de estadística matemática, algunos autores (...) edificaron una pomposa disciplina, rica en hipótesis que no son satisfechas jamás en la práctica. No es de estos autores de quienes hay que esperar la solución de nuestros problemas tipológicos"
- *2do. Principio*: "El modelo debe seguir a los datos y no a la inversa (…) lo que necesitamos es un método riguroso que extraiga las estructuras a partir de los datos"
- *3er.Principio:* "Conviene tratar simultáneamente las informaciones referidas al mayor número posible de dimensiones"
- *4to.Principio:* "Para analizar los hechos complejos y sobre todo los hechos sociales la computadora es indispensable"
- *5to. Principio:* "Abandonar todas las técnicas concebidas antes del advenimiento del cálculo automático"

Para una disciplina donde la noción de modelo ha jugado un papel primordial, estas afirmaciones de Benzécri en esa época, sonaron como un grito de guerra. Posteriormente se fueron atemperando las diferencias con la escuela anglosajona y en la actualidad se generaliza el uso de combinaciones de diversas técnicas de ambas escuelas.
Según E. Diday (1997:24) se puede representar la relación entre la estadística clásica y el análisis de datos, como dos conjuntos en cuya intersección podrían ubicarse la estadística exploratoria y las técnicas de estadística descriptiva, pero sin duda existen en los conjuntos originales dominios de estudio que no se interceptan.
En el campo de las ciencias sociales, el análisis de datos 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.
Para la descripción de estos objetos complejos, la condición de multidimensionalidad así como la posibilidad de consideración de lo contradictorio es sin duda favorable. Las representaciones en planos factoriales permiten la observación de los opuestos en diferentes dimensiones, sin llegar sin embargo a pretender la exhaustividad. Se habla de un mayor o menor porcentaje de variación explicada, de configuraciones multidimensionales más o

menos estables, etc.
Además, la interpretación de los resultados se hace en el terreno de lo real. A pesar de los complicados cálculos que se realizan con los paquetes de programas, en la etapa final de interpretación se produce lo que se llama el 'retorno a la realidad', es decir los resultados se expresan en unidades de medida coherentes con el objeto de estudio.
Estas características hacen que esta corriente capte el interés de investigadores con enfoques no deterministas y permitan un mayor reconocimiento de los objetos en su complejidad.

Tipos de datos

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, en fin, un objeto simbólico.
Resultan muy ilustrativas las palabras de Baudelot[15] sobre la Estadística Textual:
"Con sus gráficos de análisis factorial, J.P. Benzécri ha devuelto los individuos a la estadística: durante mucho tiempo ignorados a fuerza de ser confundidos en vastos agregados o pulverizados en las fórmulas inferenciales, que se interesan en primer lugar por las relaciones entre magnitudes abstractas (ingreso y consumo, salario y escolaridad...), los individuos hacen su ingreso en la escena estadística bajo la forma de puntos en una nube. Las posiciones respectivas que ellos ocupan en el seno de esa nube demuestran en primer lugar que ellos se diferencian unos de otros. Las distancias y las proximidades que ellos mantienen con las modalidades de las variables consideradas permiten a continuación comprender en qué difiere cada uno del otro: por sus gustos, sus opiniones políticas, su edad, su sexo, la marca de su vehículo, la profesión de su padre... pero la estadística es todavía una historia sin palabra.
Una de las contribuciones mayores de la estadística textual es precisamente de animar todos estos gráficos dando la palabra a

[15]Baudelot, Christian. En Prefacio a: Lebart L, Salem A (1994).

cada uno de estos individuos. Gracias a Lebart y Salem, los famosos puntos - individuos no son ya mudos, ellos hablan. Vuela entonces en pedazos la tradicional pero artificial distinción entre lo cuantitativo y lo cualitativo. Los métodos que aquí se presentan permiten poner en relación las propiedades sociales o personales de los individuos tal como los captura la encuesta estadística, con los textos por los cuales estos mismos individuos responden a las preguntas que se le hacen reduciendo al mínimo el mundo de la información".

Se ve así que las características disciplinares de la estadística se transforman no sólo en cuanto a los métodos sino también al objeto de estudio, que pasan de ser las poblaciones univariadas y sus entrecruzamientos, a la complejidad de una población multivariada donde los individuos no pierden totalmente su significación.

Una novedosa e interesante área de estudio se abre en nuestro medio con los desarrollos en análisis de datos simbólicos. Este enfoque parte de una pregunta: ¿por qué no se aprovechan en el 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.

Estos métodos 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 intensión[16] 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 intensión 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.

El Análisis Simbólico consiste en funcionar no sobre las

[16] En la acepción lógica de 'comprensión'

extensiones, es decir sobre los individuos, sino en reemplazar los individuos por las intensiones, aprovechando de esta manera el conocimiento de los expertos.
En Análisis Multidimensional de Datos clásico o Numérico se estudian conjuntos de objetos individuales representados por elementos atómicos de datos, en Análisis de Datos Simbólicos se estudian conjuntos de más alto nivel donde los individuos en estudio 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 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 estar relacionados entre ellos a través de un 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.
Es de desear que el avance sostenido de estas nuevas técnicas, que progresivamente irán superando los obstáculos inherentes a la complejidad del dato, pueda ir acercando las técnicas cuantitativas a las modalidades del trabajo de corte cualitativo.
Sin pretender ubicar al análisis de datos en alguna de las posiciones intermedias entre los enfoques cualitativo versus cuantitativo de las ciencias sociales, ya que está suficientemente demostrado la inutilidad de esta ya antigua polémica, es cierto que es visto con mayor simpatía por los cultores del primer enfoque. En efecto, se citan como características de la metodología cualitativa (G.Pérez Serrano, 1994) “proceso interactivo”, “volver sobre los datos”, “analizarlos y replantear el proceso”, “búsqueda de tendencias, tipologías, regularidades o patrones”, “necesidad de traducir los datos en categorías”, “análisis exploratorio”. Este mismo vocabulario puede encontrarse en los textos del *analyse des données*.

Dos familias de técnicas

En el ámbito del AMD clásico existen dos familias de técnicas que

permiten realizar reducciones sin perder de vista la estructura fundamental de los datos.

Los *métodos factoriales* basados en el álgebra lineal que producen las representaciones gráficas sobre las cuales las cercanías entre los puntos líneas y los puntos columnas traducen las asociaciones estadísticas entre líneas y columnas. A este grupo pertenecen, entre otros, el análisis de componentes principales, el análisis de correspondencias binario y el análisis de correspondencias múltiples. Este último, por trabajar con variables nominales, es el más adecuado para el procesamiento de encuestas.

Los *métodos de clasificación* que realizan los reagrupamientos en clases de las líneas o de las columnas. Con estos agrupamientos, llamados también tipologías o clusters de individuos con características semejantes se puede obtener una visión macroscópica de la información.

Estas dos familias de métodos pueden utilizarse complementariamente con los mismos datos logrando así una síntesis explicativa más accesible al usuario no experimentado.

Antecedentes de las técnicas de análisis multidimensional de datos

La escuela psicométrica norteamericana y el análisis factorial

Contrariamente a lo que habitualmente se cree, los métodos de AMD han sido desarrollados desde hace mucho tiempo: Hotelling, en los años 30, sentó las bases del análisis en componentes principales y del análisis canónico[17], desarrollando los trabajos de Spearman[18] y de Karl Pearson[19] que databan de principios de siglo (Citados por Borouche et al, 1980)

Según Benzécri (1982:75-97), la estadística multidimensional se fortificó con el desarrollo de la biometría como estudio de las

[17]Hotelling H. (1936) *Relations between two sets of variates,* Biometrika, vol.28, 129-149.

[18] Spearman Charles (1904) *General intelligence objectively determined and measured,* Journal of Psychology, vol.15, 201-292.

[19] Pearson Karl (1901) On lines and planes of closest fit to system of points in space, Phil. Mag., vol.2, n°11, 559-572

dimensiones físicas de los seres vivientes, la cual, estimulada por el pensamiento de Darwin, llevó a considerar el estudio de numerosos sistemas de mediciones naturalmente correlacionadas entre ellas. Pero el análisis de datos tal como se entiende en el contexto de la escuela francesa, es decir como la búsqueda, de naturaleza inductiva, de dimensiones ocultas definidas por combinaciones de medidas primarias, se desarrolla sólo a partir de la psicometría.

Esta escuela esencialmente norteamericana, tiene como objeto la medida estadística del intelecto. A partir de 1764 en que Lambert afirmó que el umbral de percepción era proporcional al tamaño del objeto, se pasa rápidamente a las notas elaboradas a partir de tests y a los trabajos de Spearman que más tarde Thurstone retomará para desarrollarlos introduciendo el cálculo matricial.

Thurstone postula un modelo lineal a partir de resultados de un test, en este modelo una observación es, excepto por el error gaussiano, una combinación lineal de factores comunes y de un factor llamado específico. Otros psicometristas siguieron a Thurstone, entre ellos Guttman y Torgerson, quienes desarrollaron la construcción de escalas multidimensionales.

Benzécri, reconociendo los aportes de Thurstone en el desarrollo del análisis factorial, critica sin embargo lo defectuoso de su modelo en lo que se refiere al requerimiento de variables estandarizadas, las cuales mutilan los datos; al considerar sólo matrices de variables continuas, Thurstone descarta los datos más seguros: las tablas de contingencia, cuya importancia ya habían avizorado Galton y luego Pearson, y para las cuales fue concebido el análisis de correspondencias. Asimismo destaca el francés que "sin adoptar el método de Thurstone subrayamos que la dificultad mayor en análisis multidimensional no es rechazar la hipótesis nula, medir una interacción globalmente con un número…, sino la de proveer para esta interacción una expresión matemática interpretable: *no se trata tanto de reconocer la existencia de lazos significativos, lo que hace falta es descifrar la significación*"(Benzécri, 1984:87). Se ve claramente en esta afirmación el punto de escisión que va luego a separar la escuela estadística tradicional de la escuela francesa.

Por otro lado, la escuela de la psicometría ha sido fuertemente criticada, en especial por haber servido de base científica a teorías políticas discriminatorias. Uno de sus críticos es Stephen Gould (1984:246-338), quien, a pesar de confesar su desagrado por el

análisis factorial como técnica, arremete más que nada contra su utilización como instrumento al servicio de la "cosificación" y la "hereditabilidad". Para este autor es inadmisible la idea de que un concepto tan impreciso y tan dependiente del contexto social como la inteligencia pueda identificarse como una "cosa" localizada en el cerebro y dotada de determinado grado de hereditabilidad, en el sentido de fatalidad, de inevitable, y de que pueda medirse con un valor numérico específico permitiendo una clasificación unilineal de las personas en función de la cantidad que cada uno pueda poseer de lo mismo.

El análisis multivariado

A comienzos del siglo XX se obtuvieron los primeros resultados relativos al desarrollo del modelo multinormal, por parte de Ronald Fisher, Hotelling y la escuela india centrada en Mahalanobis.
El análisis en componentes principales (ACP) aparece en 1901 con Karl Pearson y es situado en un contexto multivariado por Thurstone y Hotelling.
Es decir que el análisis multivariado fue ampliamente desarrollado en sus diversas versiones, correlación canónica, análisis discriminante, análisis de la variancia y covariancia múltiple, sin embargo su utilización no será generalizada más que a partir del advenimiento de las computadoras.
Conviene distinguir entre lo que los anglosajones denominan *multivariate analysis* y el *multidimensional scaling*. El análisis multivariado hace referencia enteramente al modelo probabilístico y multinormal, el segundo está más cerca de los principios del análisis de datos como se lo enfoca desde la escuela francesa. Como ya vimos, en este enfoque se olvida al menos en un primer momento, el modelo probabilístico para construir sobre la base de las observaciones un acercamiento más geométrico y descriptivo.
Un gran número de técnicas de análisis multivariado pueden ser presentadas bajo esta óptica, así el ACP puede ser formulado, a partir de la muestra, como la búsqueda del eje o del subespacio de mayor alargamiento de la nube de las observaciones en el espacio de las variables.

Los métodos de clasificación

La clasificación, como otros métodos estadísticos, tuvo su origen en la biología y posee una extensa historia desde que Aristóteles

estableció la primera taxonomía en el siglo IV a.C. reconociendo, dentro de los seres vivos, dos reinos: el Vegetal y el Animal.
Desde el punto de vista estadístico comprende varios aspectos: partición de un conjunto de objetos conteniendo o no clases superpuestas, asignación de objetos a grupos predefinidos, seriación o clasificación, taxonomía y descomposición de una población en árbol. La representación en forma de árbol es una de las más utilizadas en el entorno del AMD y es por otro lado, la elegida por los naturalistas del siglo XVIII (Linneo, Adanson) para clasificar las especies.
El árbol más usual es el árbol jerárquico o dendrograma, asociado a las clases encastradas. Un procedimiento clásico de construcción es el algoritmo ascendiente donde las clases se van reagrupando progresivamente. Partiendo de un índice de distancia entre objetos, se define un índice de agregación entre clases. De allí en más será suficiente reagrupar los dos objetos más próximos y luego iterar el procedimiento con las clases. Existen gran cantidad de algoritmos: tantos como índices de agregación propuestos.
La obra considerada como básica en clasificación, con consideraciones históricas, es la de Sokal y Sneath (1963), así como la de J-P Benzécri (1973) tomo I.
En el contexto de la ciencia, tanto los métodos de la estadística tradicional como los del análisis de datos a la francesa se acercaron a perspectivas epistemológicas más acordes con sus modos de aplicación y con el tipo de resultados a alcanzar.

CAPÍTULO 3

PERSPECTIVAS EPISTEMOLÓGICAS. IMPORTANCIA DE ALGUNOS EJES POLÉMICOS EN LA ESTADÍSTICA Y EL ANÁLISIS MULTIDIMENSIONAL DE DATOS

Explicación y comprensión

Una de las dimensiones importantes que atraviesan la producción de teorías científicas es el papel otorgado a la *comprensión* y/o a la *explicación* por las diversas tradiciones o corrientes epistemológicas.

Mardones(1991), realiza una interesante consideración histórica acerca de las polémicas surgidas en torno de la importancia de tales términos en relación con la fundamentación de las ciencias sociales como ciencia, asociando el término de comprensión a la tradición aristotélica y el de explicación a la tradición galileana.

Así, caracteriza la *explicación teleológica* de la tradición aristotélica, según la cual en toda explicación científica, se pone el acento en la existencia de una causa final o telos referida al fenómeno a explicar. Dicha causa debiera aclarar el fin por el cual ocurre tal fenómeno.

Oponiéndose a ella, la tradición galileana impone el tipo de *explicación causal,* característica de la ciencia moderna, dentro de una concepción del mundo no tanto metafísica y finalista cuanto funcional y mecanicista. La nueva ciencia que reemplazará a la aristotélica considerará como explicación científica de un hecho aquella que venga formulada en términos de leyes que relacionen fenómenos expresados matemáticamente. Dichas explicaciones serán hipótesis causales las cuales deberán ser comparadas con las consecuencias deducidas, mediante la observación de la realidad es decir a través del análisis experimental.

A mediados del siglo XIX, la ciencia natural está asentada sobre los pilares de la tradición galileana y las ciencias humanas con pretensiones científicas. En este contexto la oposición puede considerarse planteada entre el *positivismo* representado típicamente por Auguste Comte y John Stuart Mill y la *hermenéutica* propuesta por Droysen, Dilthey, Simmel, Max Weber, los neokantianos de la escuela de Baden: Windelband y

Rickert, así como fuera de Alemania por el italiano Croce y el inglés Collingwood, de tendencias más idealistas.
Con referencia a la dimensión que estamos considerando, los positivistas se caracterizaron por un modo de explicación científica de tipo causal: es el erklaren en alemán; mientras que a partir de la hermenéutica se comienza a distinguir el concepto de comprensión (verstehen). Según Dilthey, el investigador y la realidad investigada pertenecen al mismo universo histórico. Se da una unidad sujeto- objeto que permite la comprensión desde dentro de los fenómenos históricos- sociales- humanos. La comprensión se funda en esa identidad sujeto- objeto propia de las ciencias del espíritu (Geisteswissenschaften) que defienden el espacio de la creatividad, la producción de cultura del hombre frente a las posiciones deterministas del positivismo.
Se trata de ponerse en el lugar de los actores. La idea de comprensión es la de comprender lo subjetivo del otro, lo cual introduce de alguna manera la cuestión del psicologismo, donde todo lo que es producto del pensamiento debe necesariamente reducirse a términos de estados mentales subjetivos. En la problemática de investigación en las ciencias humanas a veces el problema no consiste tanto en ponerse en los estados mentales del otro sino en cómo emerger para dar cuenta de ellos de manera inteligible y generalizable.
Si la idea básica de Dilthey era a través de los productos ponernos en el lugar de los productores, la de Weber, supera este carácter inmediato o intuitivo acentuando lo metodológico con el fin de bucear en las profundidades de la *acción* (concebida como la conducta subjetivamente significativa), a través del método de los tipos ideales. Ellos son usados como herramientas de comparación entre un sujeto modelo, abstracto, y situaciones reales, a fin de comprender o reconstruir los motivos que puedan tener los individuos para actuar.
Para Mardones esta reacción antipositivista comenzada a fines del siglo XIX fue el comienzo de un debate incesante hasta hoy. En esa época, el exponente de la tradición galileana era Durkheim y el de la sociología comprensiva, Weber.
Hacia 1920, y luego entre las dos guerras mundiales, se produce el movimiento denominado positivismo lógico. Pertenecen a esta corriente B. Russell, el primer Wittgenstein y el neopositivismo del círculo de Viena.
Esta tendencia afirmaba que el análisis de la realidad es científico

sólo si se basa en la teoría de la relación lógico- matemática y en la fase o verificación empírica. Pero los filósofos del positivismo lógico se ocuparon, casi exclusivamente, de cuestiones relativas a los fundamentos de las matemáticas y de las ciencias naturales o exactas. Sólo Neurath y Carnap trataron de fundamentar la sociología desde sus supuestos, uno de cuyos resultados fue rechazar las exigencias metodológicas de la teoría de la comprensión. Ésta quedaría reducida a un nuevo elemento externo, accidental, dentro del proceso científico.

Como una reacción al positivismo lógico surge el racionalismo crítico de Popper que dice que la pretensión de verificar empíricamente todo enunciado científico conduce a la muerte de la ciencia, ya que la mayoría de ellos no son verificables empíricamente. La ciencia tendrá que ser deductiva en su justificación, se debe abandonar el camino inductivo de verificación para adoptar en cambio el falsacionismo.

Se caracteriza por el monismo metodológico: toda explicación científica, incluidas las de las ciencias humanas y sociales, adopta la forma de un esquema lógico básico:

- Explicans o explanans: teoría, leyes generales y condiciones iniciales.
- Explicandum o explanandum: enunciado general que describe el hecho o fenómeno a explicar

El explicandum sería la conclusión de una inferencia lógico-deductiva cuyas premisas están constituidas por el explicans.

Este planteamiento se sitúa, salvo el último Popper, dentro de la tradición positivista y galileana.

En este mismo período se fundó la Escuela de Frankfurt, un instituto de investigación social cuya cabeza fundadora es Horkheimer y sus continuadores, Adorno, Marcuse, Fromm, Loventhal, Polloch, trabajaron las ideas de lo que sería la teoría crítica de la sociedad. Prosiguen la línea hegeliano marxista a la que incorporan los aportes de Freud.

Tratan de analizar la sociedad occidental capitalista y proporcionar las orientaciones para caminar hacia una sociedad buena, humana y racional. Se oponen al positivismo y al racionalismo crítico de Popper. La polémica entre éste y Adorno fue continuada por los discípulos de ambos: Albert y Habermas. Opinan que el positivista no advierte que su ver, percibir, está mediado por la sociedad burguesa capitalista en la que vive. Si renuncia a percibir esta mediación de la totalidad social del momento histórico que

vive, se condena a percibir apariencias.
La teoría crítica no niega la observación pero sí su primacía como fuente de conocimiento. Adorno entiende por crítica algo distinto que Popper. Este último confía en la fuerza de la razón que mostrará si nuestros enunciados pueden mantenerse conforme a los hechos empíricos o no. De esta manera se deposita en los hechos, en lo "dado" (datos) el criterio último de verdad. Adorno habla del momento hermenéutico de la anticipación: sin anticipar un modelo de sociedad, que exprese el ansia emancipadora, racional y de búsqueda del mundo social bueno del hombre, no se puede escapar del anillo mágico de la repetición de lo dado, ni de dar cuenta del todo social que enmarca y da sentido a los hechos sociales concretos. La metodología debe atender los datos de la realidad, pero sin olvidar que hay que ir más allá de lo que aparece para captar el fenómeno en su objetividad. Esto se logra sólo si la razón mantiene una relativa autonomía respecto de los hechos a través del método crítico. Pero la vía crítica no es sólo formal sino también crítica del objeto de estudio, es decir del sujeto y los sujetos vinculados a la ciencia organizada.
Horkheimer y Adorno no rechazan las aportaciones de la lógica científica y del falsacionismo, pero acentúan la peculiaridad de las ciencias humanas y sociales, en las cuales la sociedad no puede considerarse un objeto más, sino que es también algo subjetivo. En mi opinión en esta fase la polémica no estaría planteada tanto en términos de explicación versus comprensión sino que el acento de la oposición al positivismo estaría puesto en la necesidad de destacar el interés emancipador del método crítico.
Desde 1942 la filosofía analítica (positivista) trata de precisar el modelo o teoría de cobertura legal o de explicación por subsunción. Consiste en el modelo de explicación causal de Popper, sólo que aplicado a la historia y especialmente desarrollado por Hempel. Para él una ley es una implicación universal (todos los A son B) o bien una correlación probabilística. Pero en historia nos topamos con la ausencia de referencia a leyes generales. Las respuestas de Hempel y Popper a este argumento son: a) la excesiva complejidad de las leyes en la historia nos obliga a mantenernos al nivel de bosquejos explicativos y b) las leyes históricas son algo familiar y su trivialidad no merece una mención explícita.
W. Dray reacciona contra esta teoría diciendo que las explicaciones históricas no se fundan en absoluto en leyes

generales. Explicar una acción es mostrar que esa acción fue el proceder adecuado o racional en la ocasión considerada. Este autor ha visto que la explicación histórica tiene sus propias peculiaridades pero no plantea el problema en la dirección teleológica.
El aporte de E. Anscombe está dado en la noción de intencionalidad. La conducta intencional lo es a tenor de una determinada explicación y deja de serlo a tenor de otra. Rescata la noción de silogismo práctico, que procede de Aristóteles y representa la explicación teleológica típica de las ciencias humanas y sociales por contraposición al modelo de cobertura legal que representa la explicación causal en ciencias naturales. Consiste en: mencionar en la premisa mayor una meta de actuación, en la premisa menor un medio dirigido a tal fin y en la conclusión el empleo de ese medio para alcanzar el fin pretendido. Von Wright ha tratado de demostrar que el silogismo práctico provee a las ciencias del hombre un modelo explicativo legítimo por sí mismo que constituye una alternativa definida al modelo de cobertura legal teórico subsuntivo.
Quien hará un replanteo de las ideas de Weber será Schutz, al dar un salto importante con la idea de socialización: la construcción de la conciencia individual es social. La sociedad es la totalidad de las perspectivas sociales y la tarea del científico es tratar de reconstruir los pedazos. Sin embargo el mundo de la vida cotidiana tiene distintas reglas que el de la vida científica, por lo tanto "en la ciencia social lo que hay es una construcción de segundo grado con respecto al primer grado de la vida cotidiana". No importan tanto los recursos metodológicos si está presente la idea de comprensión, de que la realidad social es una realidad interpretada. Con este autor puede ya considerarse la existencia de una filosofía de las ciencias sociales y no solamente una dimensión metodológica.
Las posiciones de Ch.Taylor y P.Winch (1990) se centran en los criterios de la acción social, influidos como los anteriores por los aportes del último Wittgenstein y de la sociología comprensiva de Weber. La explicación científica de la conducta social debe servirse del mismo entramado conceptual utilizado por los propios agentes sociales. El investigador debe comprender el significado de los datos que observa si quiere considerar los hechos sociales. Alcanza esta comprensión mediante la descripción (interpretación) de los datos en términos de conceptos y reglas que determinan la

“realidad social” de los agentes estudiados. Sin la comprensión de las reglas de juego no hay comprensión del comportamiento en sociedad.
Con Gadamer se enriquece la perspectiva del concepto de comprensión, por un lado se abandona la idea del dato puro como tal, todo dato adquiere sentido a la luz de una teoría y toda teoría interpreta el dato. Además, el ideal de la comprensión se considera inaccesible, la comprensión es casi sinónimo de *interpretación* y la interpretación que hagamos de la realidad será parte de un *acuerdo* en el cual vendrá incluido el marco teórico. Es decir siempre hay contenidos previos a partir de los cuales me enfrento a un texto, a un sujeto de otra cultura.
Por su parte Davidson (1992) introduce la dimensión pública en la comprensión que es la que nos permite comunicarnos, pone a la experiencia comunicativa como experiencia básica justificable metodológicamente.

Las explicaciones estadísticas

El modelo de cobertura legal desarrollado por Hempel y Oppenheim en el artículo de 1948 “La Lógica de la Explicación” y luego por Hempel en “Aspectos de la Explicación Científica” propone que explicar científicamente un fenómeno equivale necesariamente a responder a la pregunta de por qué se produce el mismo a partir de la mención de las leyes generales implicadas y de las condiciones de la situación considerada. Cuando se cumplen con certeza ambos requisitos estamos en presencia de una explicación nomológico- deductiva. Si bien toda explicación causal es nomológico- deductiva, no toda explicación nomológico-deductiva es explicación causal. Para que se produzca esta última es necesario un orden temporal y la imposibilidad de que causa y efecto puedan permutarse.
En el caso de la correlación estadística, sucede con alguna frecuencia que se la confunde con una explicación causal, cuando en realidad sólo es una coexistencia de propiedades. Recordemos que fue K.Pearson quien sostuvo que debía renunciarse a la idea metafísica de causalidad para reemplazarla por la idea de correlación entre hechos observados, cuyas mutuas conexiones, medidas a través de esta correlación, podían ser reproducidas en el futuro.
Para Hempel toda explicación adecuada es potencialmente una predicción. Independientemente de las objeciones lógicas que se

le hacen a esta afirmación, cabe considerar la importancia en la investigación científica otorgada a la predicción. En el contexto de aplicación de los métodos estadísticos tradicionales una teoría será considerada verosímil cuanto mayor sea la cantidad de predicciones acertadas que haya sido posible realizar. Al mismo tiempo los métodos estadísticos empleados que dieron sustento a tales predicciones ganarán mayor prestigio.

En las ciencias sociales existen temáticas en las cuales la predicción, el cálculo de pronósticos, juegan un papel muy importante, como en las ciencias económicas y empresariales, mientras que en otros tipos de problemas como los antropológicos no resultan relevantes. En éstos últimos sucede generalmente que de nada sirve predecir la ocurrencia de un determinado fenómeno sin haber comprendido o interpretado con profundidad sus formas de aparición, causas, etc. En el caso de la inferencia probabilística el mismo Hempel deja abierta la posibilidad de que no toda predicción incluya potencialmente una explicación (R.Gaeta y ot., 1996:21).

En la práctica científica las explicaciones estadísticas son utilizadas habitualmente. Incorporan en su explanans por lo menos una ley o principio teórico de forma estadística. Se distinguen dos tipos: deductivo- estadísticas, que consisten en la deducción de una ley estadística a partir de otras leyes estadísticas y las inductivo- estadísticas que dan razón de un suceso particular vinculándolo con regularidades estadísticas.

Las del primer tipo conservan el carácter deductivo a pesar de la presencia de leyes estadísticas tanto en el explanans como en el explanandum. Podrían entrar en este tipo la construcción de modelos estadísticos y la deducción de leyes de distribución de probabilidad para determinadas poblaciones.

Las del segundo tipo dan cuenta de la ocurrencia de un hecho particular dando como base su probabilidad de realización. El carácter no concluyente de estas explicaciones dan lugar a la llamada “ambigüedad estadística”, por lo cual Hempel propone el “requisito de máxima especificidad”. Según este requisito se deberá tener en cuenta toda la información científica que resulte relevante para el explanandum y elegir, en caso de conflicto, la explicación que incluya la información más específica alcanzada hasta el momento.

Descubrimiento y justificación

Con respecto al conocimiento científico los filósofos positivistas dieron por sentada una diferenciación tajante entre las condiciones o métodos relativos a su producción, de aquéllas relativas a su validación. Específicamente fue Hans Reichenbach quien en su libro “Experiencia y predicción” de 1938 se refirió a esos conceptos configurando lo que llamó el “contexto de descubrimiento” y el “contexto de justificación”.

El primero se ocupa de los procesos reales del pensamiento y es de carácter socio - histórico- psicológico- empírico. Se tiene en cuenta la producción de una hipótesis o teoría junto con las circunstancias de distinto tipo, que puedan haber influido en su gestación.

El segundo es primordialmente normativo y apunta a teorizar una reconstrucción racional del conocimiento teniendo como base una lógica del mismo. Aborda cuestiones de validación, como saber si el descubrimiento es auténtico o no, si la teoría es justificable y apoyada en la evidencia, si realmente se ha incrementado el conocimiento disponible.

G. Klimovsky (1995:30) considera además un tercer contexto, el de las aplicaciones del conocimiento científico, en el cual se discute la utilidad, beneficio o perjuicio de las mismas en la sociedad. Aunque considera que existe relación entre los contextos y aprecia la importancia del último sobre los criterios para la evaluación de las teorías desde el punto de vista del conocimiento, opina que la distinción de Reichenbach sigue siendo válida y útil.

Esta diferenciación no sólo se fue acentuando entre los seguidores de Reichenbach sino que llevó a pensar que sólo el de validación podía ser objeto de la metodología ya que era el único en el cual existía la lógica, por lo tanto el de descubrimiento sería no discursivo, luego irracional. Asimismo tal distinción se consideraba de carácter temporal, ya que se presentaban uno después del otro en el proceso de investigación.

Para Popper (1967:31), partidario de una diferenciación absoluta entre ambos contextos, el de descubrimiento carece de importancia ya que privilegia el análisis lógico del conocimiento científico. “La cuestión de cómo se le ocurre una idea nueva a una persona (...) puede ser de gran interés para la psicología empírica”. (...) “La tarea de la lógica del conocimiento (...) consiste pura y exclusivamente en la investigación de los métodos

empleados en las contrastaciones sistemáticas a que debe someterse toda idea nueva antes de que se la pueda sostener seriamente" (...) "Una vez presentada a título provisional una nueva idea, aún no justificada en absoluto, (...) se extraen conclusiones de ella por medio de una deducción lógica; estas conclusiones se comparan entre sí y con otros enunciados pertinentes, con objeto de hallar las relaciones lógicas que existan entre ellas".

Kuhn(1989:31) reconoce la importancia y la fuerza de la distinción entre los dos contextos ya que su propia formación como científico se realizó en dicho marco, si bien sus tentativas de aplicarla aunque sea en líneas generales, a las situaciones reales en que se obtienen, se aceptan y se asimilan los conocimientos, le resultó extraordinariamente problemática. Considera que esta diferenciación, en definitiva, forma parte de una teoría propuesta como otras y no constituye una distinción lógica o metodológica fundamental. A diferencia de Popper, para ese autor la experiencia no es un acontecimiento físico que, como tal, ocurre independientemente de un sistema de creencias u opiniones, el científico se acerca a la realidad desde el enfoque de un paradigma. Con respecto a la justificación, en períodos de ciencia normal acepta la aplicación tanto del método inductivo como del hipotético deductivo.

Para Piaget la preocupación fundamental es cómo se origina y estructura el conocimiento en general, que comprende en particular, el conocimiento científico. Dedicó la mayor parte de su obra al estudio de cómo se generan en el individuo los conceptos de la matemática y la física y otros de carácter biológico, psicológico y sociológico. Por ello se lo suele considerar como un epistemólogo muy ligado a los problemas del contexto de descubrimiento. Como Kuhn piensa que los objetos son vistos a través de un paradigma, pero pueden, mediante la sustitución de un paradigma por otro, finalmente, ser presentados en una secuencia que tiende a una suerte de "objeto límite". No es necesario que el mismo se corresponda con un "objeto real". Se trata de una noción constructivista de la realidad en la cual los objetos reales tienen "status" ontológico independiente. Se niega a conceder que el mero pensamiento acerca de un problema permite acceder a su solución y construir el conocimiento. Para comprender qué ocurre con los objetos, con la lógica y con el hombre en actividad psicológica de conocer, en el contexto de

descubrimiento, tendremos que recurrir a la ciencia, como fenómeno histórico y social, pues ella nos dirá qué sucede en cada una de estas tres dimensiones.

Con respecto al contexto de justificación, se pueden encontrar distintas posiciones en cuanto al rol explicativo de la psicología del conocimiento, de la sociología del conocimiento, de la sociología y de la antropología de la ciencia.
G. Bachelard (1982:281), en su consideración de obstáculo epistemológico, desarrolla el contexto de descubrimiento llegando prácticamente a eliminar los bordes con el contexto de justificación al afirmar que las teorías son necesarias desde el principio. “El instrumento de medida siempre termina por ser una teoría y ha de comprenderse que el microscopio es una prolongación del espíritu más que del ojo. De esta manera la precisión discursiva y social hace estallar las insuficiencias intuitivas y personales. Más fina es una medida, más es indirecta. La ciencia del solitario es cualitativa. La ciencia socializada es cuantitativa”. Tal vez cabría interpretar este último término en el sentido de racional más que con un sesgo de imponer la cuantificación como requerimiento para la confrontación. En la actualidad la normatización en las técnicas cualitativas de observación y la consideración de la falsedad de la oposición cualitativo - cuantitativo hacen innecesaria tal distinción.

Con respecto a la consideración del papel de la sociología del conocimiento y su influencia en el contexto de justificación, cabe destacar un programa débil (Marx, Karl Manheim) y un programa fuerte propuesto por la escuela de Edimburgo (B. Barnes, D.Bloor). “Desde David Bloor, la confortable división entre un contexto de descubrimiento y otro contexto de justificación se confunde, y salta por los aires el apacible pacto implícito entre sociólogos y filósofos (epistemólogos, metodólogos, moralistas...) que confinaba a los primeros al estudio de las condiciones de producción del conocimiento científico (la comunidad científica mertoniana) para conceder a los segundos la consideración de los criterios de validación de sus productos (el conocimiento científico). La caja negra que mantenía resguardados los contenidos propios de la ciencia (teorías, leyes, demostraciones, teoremas...) se ha revelado, al abrirse, como una auténtica caja de Pandora.”(Lizcano E y ot. en D. Bloor,1998:14)

La corriente que se deriva de los aportes de Merton, y que se apoya en la sociología de la ciencia, formulando normas para las ciencias y los científicos (comunalismo, desinterés), realizan una reivindicación de la práctica científica por sobre la teoría, mediatizan el tratamiento macro a través de un campo micro-social y plantean perspectivas institucionales.

Por su parte Samaja (1987) dedica un especial esfuerzo en rescatar la dimensión metódica de la dialéctica a fin delinear cuáles son los procedimientos específicos que caracterizan a este método como guía tanto del descubrimiento como de la validación. Este autor postula (1994:206) que el proceso de investigación se va construyendo a través de procedimientos que pueden ser empleados con dos modalidades distintas que son las de descubrir y validar, estos dos modos del método son en realidad como el anverso y reverso de una misma medalla. Para ello formula la categoría "fases y momentos del descubrimiento" con la cual se refiere "a las acciones o tareas en la perspectiva de *su eficacia* para hacer avanzar el conocimiento" y la categoría de "instancias de validación" con la cual propone "un agrupamiento de las mismas acciones (o tareas) de investigación, pero en la perspectiva de su adecuación a los controles de cientificidad o a los patrones normativos, vigentes en cada comunidad científica según que predominen las referencias a normas de validación i. Conceptuales, ii. Empíricas, iii. Operativas, o iv. De exposición."
De esta manera a través de la dialéctica consigue articular armónicamente este par lógico y a través de la definición de fases y momentos, de los cuales reconoce su antecedente embriológico (según el cual ya en la primera constitución de un ser vivo se comienzan a entrever las configuraciones posteriores); se descarta el referente mecánico del concepto de etapa en el proceso de investigación que termina por sugerir un desarrollo lineal del mismo.

Descubrimiento y justificación en la estadística y el AMD

En el campo de los métodos cuantitativos aplicados a las ciencias sociales, entre los cuales la estadística clásica ha jugado un papel preponderante en forma creciente desde fines del siglo pasado, puede decirse que su utilización corrió pareja con el mayor o menor auge del paradigma positivista.

Un primer aspecto involucraría la consideración de que si bien se ha definido a la Estadística como la ciencia de la observación y por lo tanto cabría incluirla en el ámbito del contexto de descubrimiento, sus métodos probabilísticos son de naturaleza lógica y apuntan a la justificación a través del empleo de un razonamiento similar al del modus tollens. Las contrastaciones estadísticas o tests propuestos por Jerzy Neyman y Egon Pearson alrededor de los años 30 con las nociones de hipótesis nula a contrastar contra las hipótesis alternativas implican un amago de verificación indirecta a través del logro del rechazo de un consecuente observacional. El mismo estaría constituido por la hipótesis nula la cual sería rechazada con un alto grado de probabilidad y esto llevaría a rechazar la hipótesis general.

Para Popper(1967:388), el "grado de corroboración, no es sino una medida del grado en que ha sido contrastada una hipótesis *h*, y del grado en que ha salido indemne de las contrastaciones. Por tanto, no ha de interpretársela como grado de racionalidad de nuestra creencia en la verdad de *h*, (...); sino más bien como medida de la racionalidad de *aceptar* provisionalmente una conjetura problemática, sabiendo que es una conjetura - si bien una que ha soportado que se la examine escudriñadoramente".

En el ámbito disciplinar los contrastes o tests probabilísticos, contrariando las ideas de Popper, se conocen como métodos confirmatorios, por oposición a la serie de métodos, que como ya mencionamos, fueron propuestos por Tukey en la versión anglosajona y por Benzécri en la corriente de la escuela francesa, que se denominan de forma muy general como exploratorios.

Una de las características que diferencian ambos tipos de métodos se refiere a que en los primeros se apunta a la verificación de enunciados observacionales en el marco de un modelo probabilístico propuesto, mientras que en los segundos se habla de exploración de la estructura de los datos para sugerir nuevas hipótesis sin postular modelos a priori. Sin embargo en los métodos llamados confirmatorios no estaríamos en condiciones de hablar de justificación como criterios validadores de teoría, si nos atenemos a Popper, a pesar de que actualmente así funciona en la práctica científica.

De la misma manera, tampoco en los llamados exploratorios se podría hablar estrictamente de generación de hipótesis ya que, siguiendo entre otros a Bachelard, la observación ya está influenciada por un marco teórico.

A partir de los ‘70, con el advenimiento de las computadoras, los métodos exploratorios se tornan multidimensionales y adquieren mayor potencia para el tratamiento de datos sociales permitiendo la consideración de variables múltiples y con mayor grado de complejidad. Esto puede tener que ver con la respuesta a uno de los problemas que se plantean a los métodos de justificación en el sentido del cuestionamiento a la lógica como un proceso unívoco y la necesidad de disponer de instrumentos de análisis que contemplen la posibilidad de tener en cuenta hipótesis que vienen “acompañadas”, en el sentido de P.Duhem, cuya justificación no es posible de manera aislada.

Curiosamente, si bien los métodos factoriales fueron conocidos desde principios de siglo en el marco de la psicología experimental en un contexto de tipo justificatorio - confirmatorio cuyo objetivo era la medición de la inteligencia, se retoma su algoritmo matemático para aplicarlo con otro enfoque completamente diferente dando lugar al análisis factorial de correspondencias de corte exploratorio cercano a un contexto de descubrimiento.

De todos modos, tanto los métodos confirmatorios como los exploratorios son usados alternativamente en un mismo proceso de investigación. En el ámbito del AMD, el enfoque probabilístico pierde el carácter confirmatorio que tenía en la Estadística tradicional.

Si bien en un comienzo la escuela francesa a través de Benzécri rechaza de plano la probabilidad, posteriormente comienza a incorporarla con un objetivo diferente de la decisión o la estimación, ello es en la caracterización. Los estadísticos de prueba ya no son valores que permiten el rechazo o no de una hipótesis nula, ellos se transforman, cuando se trata de caracterizar una clase o grupo de un cluster, en elementos ordenadores de las diversas características de una clase.

En el AMD el rol del investigador observador cobra un sentido de compromiso con sus datos, se reconoce su protagonismo en la interpretación dejándole la palabra final a través de la visualización gráfica, ésta es una característica que finalmente es la que separa las escuelas y es que la permite adoptar la posición de aceptación del dato construido.

CAPÍTULO 4

LAS MATRICES DE DATOS. SUS ELEMENTOS. DISTINTOS TIPOS DE MATRICES. LA MATRIZ DE DATOS EN EL ESPACIO

Matrices de datos: de objeto modelo a herramienta operatoria

El análisis de variables es considerado uno de los pilares básicos en el desarrollo de la metodología de la investigación en las ciencias sociales. Sin embargo dentro de esta tradición existen diversas orientaciones, algunas más duras que consideran a las variables sociales por simple homologación de las variables de las ciencias físicas y naturales, y otras que consideran de importancia la construcción del dato, por ejemplo la tradición, que retomando la línea de Lazarsfeld - Galtung pone el énfasis en el diseño de la matriz de datos (Samaja, 1994).

Según esta concepción, en el inicio de un proceso de investigación, se encuentra lo que se define como pre-comprensión modelizante (Ladrière, 1978:23-47), es decir un cúmulo de ideas, todavía desprovistas de una organización formal, que nos impulsan sin embargo en la búsqueda de una contrastación con la realidad. De ese cúmulo deberemos extractar nuestro objeto modelo que va a ser probado a través de la experiencia, provocada o no.

En la formulación de la problemática a estudiar se encuentra el núcleo de la pre-comprensión modelizante, que estructurará nuestra mirada para la elaboración de una hipótesis, teoría o modelo. La pre-comprensión modelizante es deudora de nuestra historia, de nuestros conocimientos previos y presupone un sistema de interpretación de la realidad u ontología. Por eso el modelo no resulta una descripción de los objetos reales, sino una reconstrucción de los posibles comportamientos de los mismos basada en nuestro propio sistema de categorías.

Se trata de descomponer un objeto complejo en sus diversos elementos a los efectos de poder analizarlos. La noción de sistema es lo suficientemente abarcadora para tener en cuenta las variaciones relativas en el tiempo y en el espacio. De esta manera

a través del estudio del estado de un sistema, de la evaluación de sus elementos, será posible compararlo con un estado subsiguiente.
Esta noción de sistema puede relacionarse con la de matrices de datos[20]. En este sentido cada matriz de datos puede constituirse en un elemento de un sistema pero a la vez, en un nivel de integración inferior, ella misma puede ser un sistema siendo sus elementos los componentes del mismo.
El modelo incluye una serie de operaciones o procedimientos de relación entre conceptos. Se consideran cuatro operaciones básicas intrínsecas a la tarea científica que se refieren fundamentalmente a los procedimientos de elección de: las unidades de análisis, las variables, los valores y los indicadores.
La matriz de datos es el instrumento básico para intentar una descripción de esa fase del comportamiento científico que consiste en diseñar la forma de recoger, procesar, analizar e interpretar la información empírica para confrontar los propios marcos teóricos, ya sea que estos actos se realicen sucesiva o simultáneamente.

Enfoque operatorio

Si el proceso de comprensión del significado metodológico del sistema de matrices de datos presenta sus dificultades, las mismas decrecen cuando pasamos al enfoque operatorio visualizando la matriz en estudio con datos reales.
Sin embargo el pasaje del nivel metodológico al nivel operatorio no se realiza habitualmente desprovisto de trabas, ya que ello implica una claridad de conceptos sobre todo en cuanto a la explicitación de variables teóricas, dimensiones e indicadores.
Cuando en un proceso de investigación se ha resuelto el problema de la formulación del diseño del objeto modelo o sistema de matrices de datos y podemos ubicarnos para operar en un determinado nivel, cuando el problema de la construcción del dato ya ha sido explicitado**,** estamos en condiciones de abocarnos a la tarea del procesamiento de la información, de operar sobre una concreta matriz de datos.
La matriz de datos es una forma de organización de los mismos habitual en los tratamientos estadísticos, en especial desde la utilización de computadoras. Los archivos computacionales en

[20] Para sistema de matrices ver Samaja (1994), parte III

formato texto se organizan en filas y columnas donde las primeras representan los individuos y las segundas las variables. En el diagrama n°1 puede observarse la estructura de la misma.

Estructura del dato en estudios extensivos

Las componentes básicas de todo dato son cuatro: unidades de análisis, variables, valores e indicadores y han sido ampliamente desarrolladas por Samaja (1994) y otros autores, en el nivel metodológico y epistemológico.

Según Baranger (1992:3) "un dato es el producto de un procedimiento de medición, y medir supone predicar una propiedad"

Aquí me ocuparé exclusivamente del dato cuantitativo analizando brevemente sus componentes. Hablo de dato cuantitativo en un sentido amplio[21], resultaría más adecuado referirse a datos provenientes de estudios extensivos, donde existe la posibilidad de definir al menos frecuencias de eventos.

Sin embargo no puedo dejar de remitirme al concepto integral de dato, en el sentido de que si bien el dato posee componentes, ellos nunca son independientes. A fuerza de considerar operativamente la matriz de datos, podemos acabar pensando que los datos son los números contenidos en ella.

Una metáfora que me parece interesante es pensar que lo contenido en cada casilla de la matriz está sujeto de tal manera al encabezamiento de su fila y de su columna que si intentáramos extraerlo de la casilla arrastraríamos esos hilos invisibles que atan cada valor a su unidad de análisis por un lado y a su variable con sus distintos niveles de integración (dimensión, indicador) por el otro. En efecto, si analizamos las definiciones de los componentes del dato observaremos que a menudo se expresan unos en función de los otros.

Para Baranger (1992:9) "una variable no es otra cosa que el conjunto de los valores que la conforman y de las relaciones que éstos mantienen entre sí", aludiendo a Galtung (1968:I,78) para la definición de valor "Dado un conjunto de unidades, un valor es algo que puede predicarse de una unidad, y una variable es un

[21] Como veremos más adelante, a la luz del Análisis de Datos Simbólicos puede ser posible revisar este concepto.

conjunto de valores que forma una clasificación”
Asimismo el mismo autor define (1992:7) “Una unidad de análisis es un sistema definido por presentar determinadas propiedades, algunas de ellas constantes (las que definen su pertenencia a un universo compuesto por todos los sistemas que presentan esas mismas propiedades) y otras variables (las que podrán ser materia de investigación dentro de ese universo)”
Por ejemplo, si estamos estudiando las representaciones sociales de la desnutrición en madres de niños desnutridos, las unidades de análisis podrán ser las madres, cuyas propiedades constantes serán: sexo femenino, con al menos un hijo en situación de desnutrición. Por otro lado sus propiedades variables serán: opiniones acerca de la desnutrición, consideraciones acerca de su historia personal, etc.
¿Cómo se mide? Definiendo variables e indicadores
La definición de variable se origina a partir del concepto de medición o clasificación y se define matemáticamente para el campo de las ciencias exactas.
Entendemos entonces el acto de medir como la asignación de un valor o categoría a una unidad de análisis de acuerdo con un patrón o clasificación establecidos.
Una *variable* es una característica que toma distintos *valores* en los distintos *individuos.* Si considero la variable altura en centímetros de los asistentes a un función de teatro, los valores podrán ser: 177, 156, 165, etc.
El término individuo o unidad de análisis es genérico y puede referirse a un colectivo o a otros objetos según el tipo de matriz que consideremos.
Sin embargo las variables, son conceptos complejos, de manera que para abordar su estudio se hace necesario establecer varias *dimensione*s o aspectos de las mismas. A la vez, para el procedimiento práctico de obtener información que dé cuenta de ellas se definen los *indicadores* que son finalmente las características susceptibles de manejar aritméticamente: en definitiva las variables en lenguaje estadístico.
En la bibliografía estadística no encontraremos el término ‘indicadores’, en su lugar se utiliza de una manera general el término ‘variables’. Ubicándonos entonces en un enfoque operatorio de corte cuantitativo, nos manejaremos por lo tanto en adelante con este término, es decir con las partes observables de las variables teóricas o especulativas.

El proceso de determinación de variables, dimensiones e indicadores está íntimamente relacionado con el diseño del objeto modelo y cada elección que realicemos deberá ser validada en el ámbito del marco teórico y las hipótesis seleccionadas.

¿Qué hay detrás de una matriz de datos?

Se llama *población en estudio* al conjunto de *unidades de análisis*, siendo cada una de ellas un *individuo* al que se le realizan una o varias mediciones. Se entiende que hablamos de individuos en un sentido singular o colectivo, simple o complejo, así como de mediciones en un sentido amplio, de ajuste a un patrón determinado.

Es importante distinguir población en estudio de objeto de estudio, tal como destacan los autores que rechazan "la ingenuidad de los empiristas que toman por objeto científico el objeto real en su totalidad concreta"[22]

Asimismo Baranger (1992) hace la distinción entre "objetos teóricos (por ejemplo el modo de producción capitalista de Marx o el inconsciente de Freud) que son construcciones conceptuales que, aunque estén dirigidas en definitiva a "pensar" la realidad en una investigación empírica no van a funcionar nunca como unidades de análisis (ni tampoco como variables)"

El hecho de tener en cuenta la complejidad del dato remite entonces a evitar ese tipo de "ingenuidad", evita el considerar a un determinado valor numérico como un hecho real e indiscutible, sino por el contrario, como el resultado de un complejo proceso de construcción que implica definir por un lado las unidades de análisis con sus variables implícitas y por otro todo un nivel de desagregación de las variables teóricas o especulativas en consonancia con los objetivos e hipótesis de investigación.

Resulta a veces dificultoso comprender la diferencia, en la práctica, entre variable teórica e indicador, por ejemplo en el caso en que debiéramos diseñar preguntas o ítems de un cuestionario, precisamente por esa tendencia a la simplificación, a soslayar la complejidad de los conceptos teóricos. Un ejemplo que resultó ilustrativo para mis alumnos fue el siguiente.

Cuando el médico debe producir un diagnóstico, realiza

[22] Bourdieu P, Chamborendon JC y Passeron JC (1999) pag. 205, citando a Marx en la Introducción general de 1857 de su Método de la Economía Política

generalmente un primer examen y cuenta con una presunción de enfermedad, digamos que presume una úlcera. Para ir confirmando su diagnóstico necesitará ulteriores elementos de análisis, entre los cuales estará un interrogatorio al paciente. Evidentemente las preguntas que irá construyendo serán 'indicadores' que arrojarán luz sobre el concepto teórico de 'úlcera'. En ningún momento se le ocurrirá preguntar: "¿Ud. tiene úlcera?" como es el caso de algunos cuestionarios realizados por principiantes, donde puede llegar a resultar alguna pregunta como: "¿Cuál es su representación social de la salud?"
En el cuadro que sigue se muestra un ejemplo de una matriz de datos de reducidas dimensiones. Las unidades de análisis, en este caso hogares, se ubican en las celdas de la primer columna y las variables, con sus correspondientes valores para cada individuo, en las columnas siguientes.

Hogares	Número de miembros	Ingreso per capita	Sexo del jefe	Escolaridad del jefe
1	8	120	Masculino	1
2	2	700	Masculino	2
3	4	400	Femenino	3

Podemos considerar esta matriz desde un enfoque reduccionista, que remite a pensar que las variables son lo directamente observable y medible, limitándonos a considerar los números contenidos en la matriz y sus relaciones.
Otro enfoque, el que proponemos, consiste en considerar a estas variables estadísticas, como algunos de los indicadores *posibles* de observar, de algunas dimensiones de una variable teórica. Por ejemplo podría considerarse:

* *Características socioeconómicas* → variable teórica

    - *estructura del hogar* → dimensión 1
        - *número de miembros* → indicador 1.1
        - *sexo del jefe* → indicador 1.2

    - *nivel económico del hogar* → dimensión 2
        - *ingreso per cápita* → indicador 2.1

    - *nivel educacional del hogar* → dimensión 3
        - *escolaridad del jefe* → indicador 3.1

El anterior fue un ejemplo de dimensiones reducidas, tanto de variables como de unidades de análisis, evidentemente cuando se eleva el número de variables la situación se complica pudiendo resultar que algunos indicadores no lo sean sólo de una variable o dimensión sino que 'apunten' en realidad a varias.

Escalas de medición

¿Con qué se mide? Definiendo un patrón o escala de medición.
Tradicionalmente se reconocen cuatro niveles o escalas de medición: nominal, ordinal, de intervalo, de razones, que definen otros tantos niveles de variables.
Recordamos que una variable de nivel *nominal* puede tomar sólo los valores categóricos y surge de una operación clasificadora. Por ejemplo, la categoría ocupacional que toma los valores: obrero, empleado, patrón, etc.
Los valores de una variable de nivel *ordinal* pueden ser ordenados según una determinada jerarquía pero sin poder evaluar la distancia que separa un valor de otro, Por ejemplo, nivel de acuerdo con una determinada propuesta, que toma los valores: muy de acuerdo, medianamente de acuerdo, nada de acuerdo.
En el nivel de *intervalo* los valores de la variable están ordenados en intervalos iguales pero subsiste la limitación debida a la ausencia de un punto cero absoluto. Por ejemplo: temperatura medida en grados centígrados.
Finalmente el nivel *racional* permite ordenar los valores de la variable a intervalos iguales y el valor cero representa realmente la ausencia de la propiedad que se intenta medir. Por ejemplo: altura de las personas.

Al hablar de niveles de medición se está aludiendo a la existencia de una jerarquía en la realización de operaciones aritméticas con las variables. En efecto se habla de niveles inferiores de medición en las variables nominales y superiores en las racionales. El hecho de que en las Ciencias Sociales no se puedan alcanzar los niveles 'superiores' de medición ha llevado a considerar esta situación como de subdesarrollo con respecto a las otras ciencias. Esta matematización de las ciencias se remonta a los orígenes del positivismo y ha llevado con frecuencia a tratar de forzar las escalas naturalmente nominales u ordinales a racionales o de intervalo. Es así que muchas veces encontramos refinadas

construcciones de análisis estadísticos basados en niveles de medición que no responden a los datos y conducen a veces a dudosas interpretaciones.
La selección de la escala de medición implica no sólo el tipo de la misma, sino la determinación de los valores (numéricos o no) representativos de los distintos escalones en el campo de variación de la variable.

Inmediato al proceso de determinación de la escala de medición sigue el de *codificación*, que implicará la asignación de un valor numérico de entre los posibles de la escala elegida, que consideramos adecuado para cada unidad de análisis.
La codificación es la fase numérica del proceso de medición. Por ejemplo medir puede ser asignar la categoría 'empleado' al individuo 'Juan', codificar puede ser asignar el valor 2 al mismo individuo de acuerdo con el código: 2 = empleado. El procedimiento de codificación será el que nos habilite para el procesamiento de la información y el que ha sido malinterpretado por quienes sólo pueden ver el número como cosa en sí misma.

Tipos de variables según escalas de medición

Las variables se clasifican de acuerdo con el tipo de escala elegida para los valores que toma y según veremos, los tipos de procedimientos estadísticos a aplicar están fuertemente condicionados por el tipo de variables.
Una escala de medición nominal da origen a una variable nominal llamada habitualmente cualitativa. En mi opinión es preferible utilizar el término nominal o categórica para no inducir a confusiones. Según J.P.Benzécri(1993:1) "oponer cantidad a cualidad como continuo a discreto (o discontinuo) es un abuso de lenguaje propio de los estadísticos".
Una escala ordinal da origen a una variable de tipo ordinal.
Existen múltiples procedimientos en la estadística clásica para el tratamiento de datos ordinales, en análisis multidimensional de datos podemos considerar este tipo de información como variable nominal o continua según los casos. Un criterio práctico de decisión puede ser la consideración del número de valores. Ejemplo: cuando se consideran notas o resultados de evaluaciones, si el rango es de 1 a 5, la variable se considera como nominal; si en cambio el número de valores posibles es de 1 a 100, como es el caso en algunos tests psicológicos, se podrá

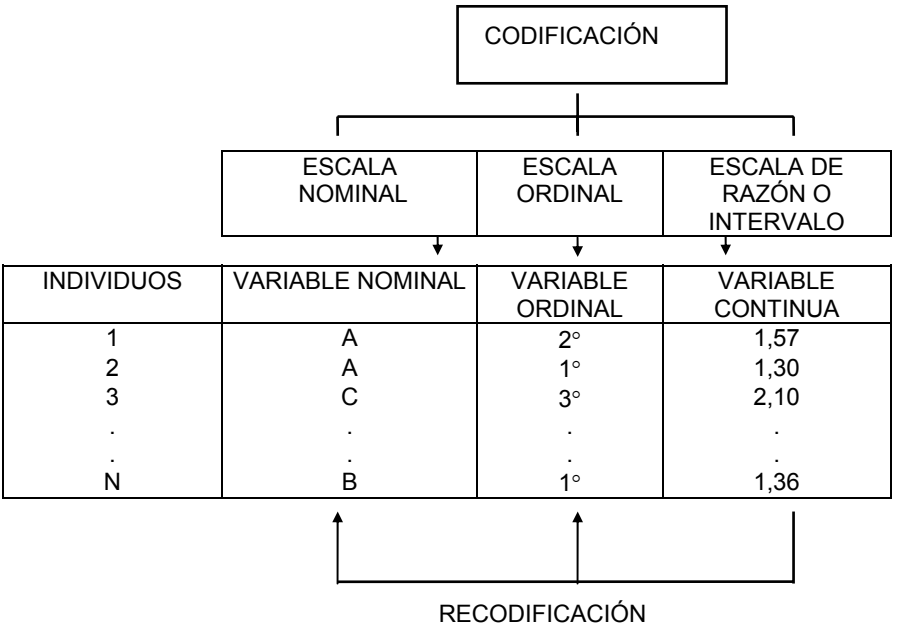
considerar como continua. Este criterio se justifica con el tipo de resultados que se obtienen en el análisis multidimensional de datos, donde como vamos a ver, se manifiesta gráficamente la estructura de los datos y sus escalas de medición.
Las escalas de intervalos y de razones dan origen ambas a variables continuas y su tratamiento estadístico es similar.

En la bibliografía estadística clásica se suelen clasificar las variables como cuantitativas y cualitativas, para distinguir entre las primeras a las discretas de las continuas. Las cualitativas serían las que hemos definido como nominales, y las cuantitativas serían las ordinales y continuas. Las variables discretas son el producto de conteos o frecuencias, por ejemplo número de hijos de un hogar.
Los programas de computación en general no tienen en cuenta el nivel de medición de las variables habilitando procesamientos de manera indiscriminada, por lo que queda librado al conocimiento del usuario la elección de las técnicas adecuadas para cada caso.

Cambios de variables o recodificaciones

La escala de medición y en consecuencia el tipo de variable, debe tenerse presente en todo el procesamiento, sin embargo es habitual que luego de un primer examen de los datos se decida un cambio de escala o reagrupamiento de intervalos. Este proceso se denomina *recodificación***.** Evidentemente el mismo sólo podrá realizarse en un sentido, es decir desde el tipo de escala de intervalo en un extremo, hacia el nominal en el otro.
El análisis multidimensional de datos posee una amplia gama de procedimientos para tratamientos de variables en diferentes escalas, naturalmente de una selección más o menos adecuada dependerán los resultados obtenidos. Contrariamente a los procedimientos de la estadística clásica, es habitual transformar variables continuas en nominales para permitir la emergencia de relaciones no lineales.

Diagrama N°1



Tipos de recodificaciones

De variable continua a ordinal

Consiste en seleccionar valores de una variable continua, que constituirán los extremos de los intervalos de manera tal de cubrir todo el rango de variación de la variable original. Se asignará un código a cada intervalo procediendo luego a codificar cada individuo según este nuevo código.

Existen distintos métodos para la selección de los límites de intervalo, a menudo el investigador los determina de acuerdo con su conocimiento del tema[23].

Deberá tenerse en cuenta que de esta manera se perderá la información referida a la distinción entre los objetos de un mismo intervalo y a la amplitud de la diferencia entre los objetos de dos intervalos diferentes.

Por ejemplo si tomamos la variable edad en años, continua,

[23] Para las opciones estadístico metodológicas se puede consultar Diday, (1992).

podemos recodificarla llevándola a una variable ordinal definida por cuatro valores:

1. niños (de 0 a 12 años)
2. jóvenes (de 13 a 19 años)
3. adultos (de 20 a 69 años)
4. edad avanzada (más de 70 años)

Si bien los valores ordinales expresan conceptos, es claro que los tratamientos estadísticos tomarán en cuenta sólo los valores numéricos asignados en la codificación (de 1 a 4).

Este tipo de cambio de variable suele llamarse “cambio de estructura”(Diday,1982:31)

Como se verá más adelante, las técnicas de análisis multidimensional de datos pondrán luego de manifiesto la relación de orden implícita en la variable.

De variable ordinal a nominal

En algunos casos puede resultar de interés dejar de lado la relación de orden implícita en la variable. Retomando el ejemplo anterior puede ser necesario considerar por separado sólo los adultos (como población económicamente activa) en cuyo caso se podrán transformar los códigos utilizados en los siguientes:

1. niños, jóvenes y edad avanzada
2. adultos

En este caso a los individuos que tenían valores 1, 2 y 4 se les asignará el valor 1, y a los que tenían valor 3 se les asignará el valor 2. Es lo que se denomina un cambio de variable “por codificación”

Por combinación de variables

En determinadas ocasiones, en especial cuando se consideran gran número de variables será importante combinar varias variables en una. En ese caso se deberán reasignar los códigos de todas las variables involucradas simultáneamente. Es un procedimiento fácil de hacer mediante un paquete estadístico de computación. Por ejemplo, si se consideraron las variables edad y nivel de escolaridad, entre otras, se puede construir una nueva variable que podría ser por ejemplo destacar sólo los que tienen escolaridad primaria. En ese caso, agregaríamos la variable escolaridad como variable nominal de tres niveles:

1. sin escolaridad
2. escolaridad primaria

3. escolaridad media o superior

y luego retomando la codificación de edad como variable ordinal, podríamos construir una nueva variable (Edad-escolaridad) que podría ser por ejemplo:

1. niños y jóvenes
2. adultos, sin escolaridad o escolaridad primaria
3. adultos, escolaridad media o superior
4. edad avanzada, sin escolaridad o primaria
5. edad avanzada, escolaridad media o superior

Para ello deberemos recodificar según la siguiente tabla, es decir a un individuo se le asignará el código de la primera columna cuando en la variable edad ordinal tenga los códigos de la segunda columna *y a la vez* en la variable escolaridad los códigos de la tercera columna.

EDAD-ESCOLARIDAD	EDAD ORDINAL	ESCOLARIDAD
1	1 y 2	1 ó 2 ó 3
2	3	1 ó 2
3	3	3
4	4	1 ó 2
5	4	3

Este tipo de recodificaciones puede realizarse combinando varias variables medidas en diferentes escalas, cuidando realizar las correspondientes equivalencias.

Tipos de matrices de datos según nivel de procesamiento

En el ítem anterior se expusieron los elementos básicos de una matriz de datos, en el presente se considerarán los distintos tipos de matrices de datos que se analizan usualmente en la práctica y que se originan de acuerdo con el tipo de unidades consideradas en las filas y columnas.

Para el análisis de datos es importante la consideración de las diferentes variedades de matrices ya que de su composición dependerá el tipo de técnica a aplicar para su análisis.

Matrices diacrónicas y sincrónicas

Cuando se da el caso en que se considera el tiempo como una

variable más, el tipo de matriz de datos se denomina matriz sincrónica y da origen a un análisis estático o transversal, por ejemplo cuando el tiempo es considerado como edad de los individuos.

Individuos	Edad en años
1	12
2	25
3	60

Si en cambio el tiempo es considerado como unidad de análisis el tipo de matriz es diacrónica y da origen a un análisis dinámico con tratamiento estadístico del tipo series de tiempo o estudios de cohortes. En el ejemplo siguiente los “individuos” son años.

Años	PBI
1960	256 789
1965	365.798
...	...
1995	987 603

Matrices de individuos por variables

Matrices de datos continuos

Cuando todas las variables son continuas se puede construir una matriz de datos continuos. Las variables pueden estar o no medidas en la misma unidad, pueden constituir datos absolutos de mediciones, índices, etc. El ejemplo siguiente trata sobre resultados en diferentes exámenes de alumnos universitarios en una escala de 1 a 100.

Individuo	Historia	Psicología
1	85	57
2	32	59
3	98	43

Matrices de frecuencias

Cuando las variables consideradas provienen de un conteo, es decir son variables discretas o frecuencias, podemos obtener matrices como la siguiente, obtenida a partir de una encuesta donde se pregunta el número de veces por semana que se adquieren distintos productos.

Individuos	Marcas de productos			
	I	II	III	IV
1	7	0	5	0
2	4	7	0	0
3	0	3	0	6

Matrices de datos nominales u ordinales

Cuando las variables intervinientes son del tipo nominal y pueden tomar uno y sólo uno de varios valores o modalidades, podemos encontrarnos con matrices como la siguiente, cuyos valores provienen de recodificar la matriz anterior. En este caso cada individuo responde 1, 2 o 3 según su frecuencia semanal de adquisición de un producto. La respuesta 1 significa “nunca”, 2 “a veces”, 3 “siempre”. Nótese que al existir una relación de orden las variables son este caso ordinales. Es decir, la variable *frecuencia de adquisición del producto I,* puede tomar los valores 1, 2 ó 3.

Individuos	Marcas de productos			
	I	II	III	IV
1	3	1	2	1
2	2	3	1	1
3	1	2	1	2

Matrices de datos de preferencias

Este tipo de datos es en realidad un caso especial de valores ordinales, en el cual existe una relación de exclusión en los valores correspondientes al mismo individuo. En el ejemplo que sigue las personas interrogadas marcan con una nota de 1 a 4 el

orden de su preferencia por los cuatro productos. Evidentemente un mismo individuo no puede adjudicar el mismo valor a todas las marcas. Esto no ocurría en el caso anterior donde un individuo podía tranquilamente contestar que adquiría “siempre” o “nunca” todos los productos nombrados al mismo tiempo.

Individuos	Marcas de productos			
	I	II	III	IV
1	4	1	3	2
2	3	4	1	2
3	1	3	2	4

Matrices heterogéneas

Se trata de matrices donde las variables son de diferentes tipos, como en el ejemplo que ya vimos. Los casos típicos son las grandes encuestas y censos.

Hogares	Número de miembros	Ingreso per capita	Sexo del jefe	Escolaridad del jefe
1	8	120	Masculino	1
2	2	700	Masculino	2
3	4	400	Femenino	3

Otros ejemplos de matrices de datos de individuos (filas) por variables (columnas)

- Respuestas de N individuos a las P preguntas de una encuesta.(individuos=N; variables=P)
- Características estructurales y resultados económicos de diferentes empresas de un sector (individuos=individuos; variables= características estructurales y resultados económicos)
- Resultados de diversos análisis clínicos en individuos sanos o que poseen cierta enfermedad (individuos= individuos; variables= resultados de diversos análisis clínicos, condición de salud o enfermedad)
- Características socioeconómicas y resultados electorales en distintas localidades de una provincia (individuos=localidades;

variables=características socioeconómicas y resultados electorales)

- Características de distintos préstamos concedidos por un banco, estos últimos clasificados según las dificultades encontradas en su recuperación (individuos=préstamos concedidos por un banco; variables=características de los préstamos, grado de dificultad en su recuperación)
- Grado de aceptación de diferentes medidas políticas para distintos individuos encuestados (individuos=individuos encuestados; variables=grado de aceptación de diferentes medidas políticas)

Matrices de variables por variables

Estas matrices son obtenidas habitualmente a posteriori de una primera manipulación de la información o cuando se utilizan fuentes secundarias, es decir cuando esta manipulación ya fue efectuada por terceros.

Tabla de contingencia

A partir de dos variables nominales, se puede definir una tabla de contingencia, que cruza todas las modalidades de las variables dos a dos. La casilla de intersección de la fila **i** con la columna **j**, contendrá el número de individuos que eligieron la modalidad i de la primera variable y la modalidad j de la segunda. Es decir, en cada casilla encontraremos siempre frecuencias y no mediciones.

A esta particular tabla de frecuencias, K. Pearson le dio el nombre de tabla de contingencia porque "Todo en el universo ocurre una sóla vez, no existe igualdad absoluta en la repetición. Los fenómenos individuales pueden sólo ser clasificados" [24]

Si dividimos cada casilla de esta tabla por el total de la población, se obtendrá la tabla o *matriz de frecuencias relativas.*
La siguiente tabla de contingencia permite analizar la frecuencia de consumo diario de tabaco según el sexo en una población de estudiantes.

[24] Pearson K. (1912) *La grammaire de la science.* Trans. from the English by Lucien March. Paris:Alcan. U.S. edition: *The Grammar of Science.* New York: Meridian, 1956. Citado por Desrosières (1998),cap.5

Consumo de cigarrillos	Varones	Mujeres
Ninguno	48	55
menos de 5	24	31
entre 5 y 10	14	10
más de 10	5	6

Matriz de frecuencias para datos textuales

Número de veces que aparecen ciertos vocablos elegidos en diferentes poemas de Alfonsina Storni

Palabras	Río de la Plata en negro y ocre	Río de la Plata en gris áureo	Río de la Plata en arena pálido	Río de la Plata en celeste nebiplateado	Río de la Plata en lluvia
AGUA	0	0	2	1	0
AZULES	0	1	1	1	0
CIELO	1	1	2	3	2
CIUDAD	0	1	0	0	2
CUERPO/S	0	1	2	1	1
DE	3	2	6	5	3
GRISES	0	1	0	1	1
HORIZONTE	1	0	0	0	1
HUMO	1	0	1	1	0
NUBES	0	1	1	1	0
TU	0	0	3	1	3

De este tipo de matrices nos ocuparemos en el capítulo referido al análisis factorial de correspondencias. El caso es de especial importancia en el análisis de datos textuales, donde se trabaja con frecuencias de palabras. Se pueden cruzar palabras por textos, respuestas por palabras, etc.

Estos tipos particulares de matrices de contingencia llamadas tablas léxicas, son a menudo tablas de grandes dimensiones “dispersas”, debido a que gran cantidad de casillas pueden resultar vacías (o iguales a frecuencia 0). Como veremos en el capítulo dedicado a análisis textual deben recibir un tratamiento especial.

Tabla de contingencia múltiple

Cuando se cruzan ya no sólo dos variables nominales sino dos

familias de variables nominales definidas sobre una misma población, se obtiene una tabla de contingencia múltiple.
Las dos familias pueden o no ser las mismas, en el caso particular en que se cruce una variable con sí misma se obtendrá un bloque diagonal.
Un caso particular de este tipo de tablas es la *tabla de Burt* a la que nos referiremos en el capítulo correspondiente al análisis factorial de correspondencias múltiples.

Matrices de individuos por individuos

Si se desea evaluar el mayor o el menor parecido entre cada par de individuos se puede construir una *matriz de proximidad o de distancias.* Por ejemplo, un panel de consumidores evalúa el grado de semejanza entre distintos modelos de automóviles existentes en el mercado, marcando una nota de 1 a 10 que mide el parecido de los automóviles.

Automóviles	M1	M2	M3	M4
M1	10	3.7	6.2	1.5
M2	3.7	10	8.7	5.3
M3	6.2	8.7	10	9.4
M4	1.5	5.3	9.4	10

Matrices de datos simbólicos

Cuando las unidades de análisis dejan de ser unidades simples para convertirse en objetos complejos, y/o cuando las variables dejan de medirse en valores mutuamente excluyentes, podemos construir matrices de datos simbólicos, tema del que me ocuparé en el capítulo 10.
Un ejemplo de matriz de datos simbólicos podría ser la siguiente, donde las unidades de análisis son objetos simbólicos, en este caso Facultades de la Universidad Nacional de Rosario y las variables son multinominales, es decir pueden tomar varios valores en cada unidad de análisis.
En este caso los datos proceden de un procesamiento anterior donde las unidades de análisis eran alumnos ingresantes a una Facultad, que trabajaban hasta 20 horas semanales, entre 21 y 36 horas, más de 36 horas o no trabajaban. A la vez la escolaridad

del padre de estos alumnos podía ser: sin estudios, primaria completa, secundaria completa o universitaria completa (los con estudios incompletos se incluyeron en el nivel inferior)
Esa matriz original puede transformarse en otra de mayor nivel de complejidad, como la siguiente, donde nuestro objetivo puede ser operar con Facultades.

Objeto	Horas de trabajo				Escolaridad del padre			
Bioq	No (0.72)	H20 (0.12)	H21 (0.08)	H36 (0.08)	Sin (0.05)	Pri (0.38)	Se (0.36)	U (0.19)
Polit	No (0.66)	H20 (0.09)	H21 (0.12)	H36 (0.13)	Sin (0.09)	Pri (0.35)	Se (0.34)	U (0.22)
Odont	No (0.86)	H20 (0.06)	H21 (0.03)	H36 (0.05)	Sin (0.04)	Pri (0.32)	Se (0.38)	U (0.25)
Medic	No (0.81)	H20 (0.04)	H21 (0.04)	H36 (0.11)	Sin (0.09)	Pri (0.39)	Se (0.33)	U (0.20)
Huma	No (0.50)	Ho20 (0.16)	H21 (0.17)	H36 (0.18)	Sin (0.09)	Pri (0.39)	Se (0.34)	U (0.19)
Psic	No (0.66)	H20 (0.09)	H21 (0.11)	H36 (0.14)	Sin (0.09)	Pri (0.40)	Se (0.34)	U (0.16)
Econ	No (0.69)	H20 (0.07)	H21 (0.08)	H36 (0.15)	Sin (0.07)	Pri (0.42)	Se (0.36)	U (0.15)
Dere	No (0.66)	H20 (0.07)	H21 (0.08)	H36 (0.19)	Sin (0.08)	Pri (0.40)	Se (0.36)	U (0.15)
Agra	No (0.76)	H20 (0.13)	H21 (0.05)	H36 (0.06)	Sin (0.04)	Pri (0.37)	Se (0.41)	U (0.17)
Vete	No (0.82)	H20 (0.06)	H21 (0.05)	H36 (0.07)	Sin (0.10)	Pri (0.43)	Se (0.27)	U (0.19)
Arquit	No (0.78)	H20 (0.06)	H21 (0.08)	H36 (0.09)	Sin (0.06)	Pri (0.30)	Se (0.40)	U (0.23)
Ingen	No (0.77)	H20 (0.08)	H21 (0.05)	H36 (0.10)	Sin (0.04)	Pri (0.31)	Se (0.39)	U (0.25)

La interpretación de la primera fila de esta nueva matriz sería:

En la Facultad de Bioquímica (objeto simbólico)

- la variable multinominal horas de trabajo vale *0.72 no trabaja* (el porcentaje de alumnos que no trabajan, devenido en probabilidad de no trabajar es 0.72), *0.12 trabaja menos de 20 horas, 0.08 trabaja entre 21 y 36 horas y 0.08 trabaja más de 36 horas.*
- la variable multinominal escolaridad del padre vale: *sin estudios 0.05, primaria completa 0.38, secundaria completa 0.36, universitaria completa 0.19*

Para una matriz así construida se han desarrollado métodos de procesamiento tanto de estadística descriptiva, (distribuciones de frecuencia, histogramas) como otros más complejos de clasificación y análisis factorial

La matriz de datos en el espacio

Resulta de utilidad para la comprensión de las técnicas de AMD que se presentarán en los próximos capítulos la visualización de los espacios que se generan a partir del análisis de una matriz de datos.

Estos dos espacios están siempre presentes y son referidos a través de los distintos cálculos que se realizan, por lo tanto es deseable conservar al menos una idea intuitiva de ellos.

Es conveniente recordar que el AMD generaliza el análisis univariado que sería un caso especial del primero, agregando al mismo tiempo otro enfoque.

A la vez el Análisis de datos Simbólicos generaliza al AMD considerando objetos provistos de propiedades en lugar de UA simples. Es decir que la representación geométricamente espacial que presento a continuación de una manera simplificada para permitir su visualización, debiera generalizarse a un número mayor de variables y a objetos de mayor complejidad.

La matriz de datos que representaré en el espacio será la siguiente:

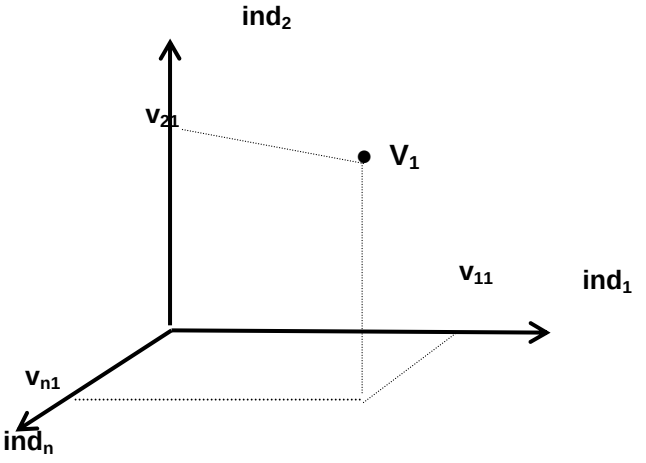
UA	V_1	V_2	…	V_p
ind_1	v_{11}	v_{12}		v_{1p}
ind_2	v_{21}	v_{22}		v_{2p}
…				
ind_n	v_{n1}	v_{n2}		v_{np}

(La fila y columna sombreadas son los puntos representados, es decir los valores v_{11}, v_{12} y v_{1n} correspondientes al ind_1 en el espacio de las variables y los valores v_{11}, v_{21} y v_{n1} correspondientes a la V_1 en el espacio de los individuos)

La condición de multidimensionalidad considera la relación de los dos espacios que se originan a partir de la matriz de datos: el espacio de las n unidades de análisis (o individuos) $\mathbf{R^n}$, y el espacio de las p variables $\mathbf{R^p}$. En ambos espacios “flotan” los

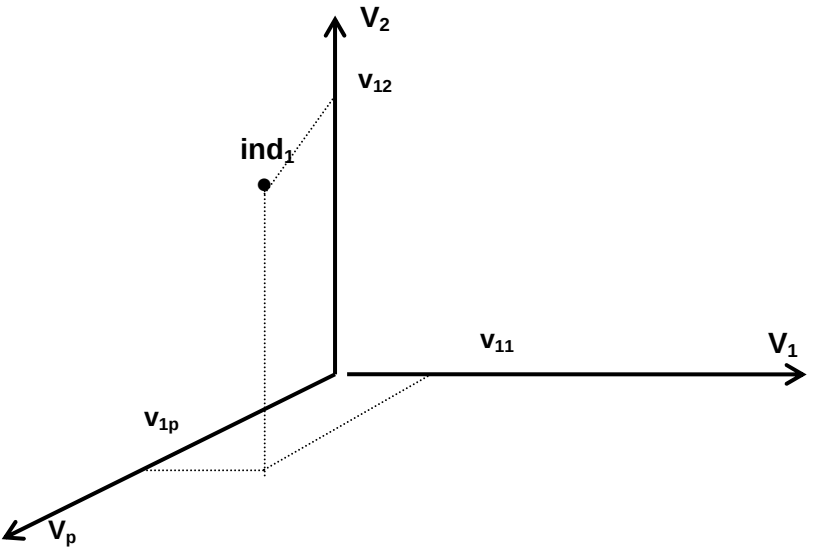
puntos variables e individuos respectivamente, los cuales en el R^n y en el R^p tendrán por coordenadas a los valores de las variables en los individuos.

Espacio de los individuos: R^n



En el espacio de los individuos "flotan" los puntos variables. En la figura se representa un espacio de 3 individuos (que puede ser generalizado a n individuos, pero naturalmente no podemos visualizarlo) definido por 3 ejes: el del individuo 1 ($\mathbf{ind_1}$), el del individuo 2 ($\mathbf{ind_2}$) y el del individuo n ($\mathbf{ind_n}$). Representamos para simplificar una sola variable, el punto de la variable 1 ($\mathbf{V_1}$) cuyas *coordenadas* sobre los ejes de los individuos serán los valores $\mathbf{v_{11}}$ (valor de la variable 1 en el individuo 1), $\mathbf{v_{n1}}$ (valor de la variable 1 en el individuo n y $\mathbf{v_{21}}$ (valor de la variable 1 en el individuo 2).

Espacio de las variables: R_p



Recíprocamente en el espacio de las variables "flotan" los puntos individuos. En la figura hemos considerado un espacio de sólo 3 variables (que puede ser generalizado a un espacio no visualizable de p variables)definido por 3 ejes: el de la variable 1 ($\mathbf{V_1}$), el de la variable 2 ($\mathbf{V_2}$) y el de la variable p ($\mathbf{V_p}$). Se representó para simplificar un sólo punto - individuo ($\mathbf{ind_1}$) cuyas coordenadas sobre los ejes de las variables son: $\mathbf{v_{11}}$ (valor de la variable 1 en el individuo 1), $\mathbf{v_{12}}$ (valor de la variable 2 en el individuo 1) y $\mathbf{v_{1p}}$ (valor de la variable p en el individuo 1).

Obsérvese que en esta representación de los dos espacios existe sólo un valor común: el de la variable 1 en el individuo 1. En cambio en los verdaderos espacios formales de una matriz de datos de n filas y p columnas, deberán encontrarse *todos* (nxp) los valores comunes. Sin embargo deberá observarse que la ubicación de los puntos, aunque hagan referencia a los mismos valores, no será la misma en los dos espacios ya que los ejes de referencia son en un caso las unidades de análisis y en el otro las variables.

Podrá realizarse la objeción de que esta representación es en principio ideal, ya que es imposible de llevar a cabo en las matrices heterogéneas (si las variables no están medidas en la misma escala), sin embargo, como se ve tanto en el Análisis de Correspondencias Múltiples y más aún en el Análisis de Datos Simbólicos, será siempre posible encontrar una transformación de las variables que permita la construcción de los espacios. A este objetivo apuntarán los diversos tipos de técnicas de análisis factorial.

PARTE II

TÉCNICAS PARA DATOS NUMÉRICOS, TEXTUALES, SIMBÓLICOS

CAPÍTULO 5

INTRODUCCIÓN AL AMD. USOS PRINCIPALES EN LAS CIENCIAS SOCIALES.

Acerca del AMD

En el campo de las ciencias sociales, el Análisis Multidimensional de Datos se revela como una opción de gran importancia 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.

Para la descripción de estos objetos complejos, la condición de multidimensionalidad así como la posibilidad de consideración de lo contradictorio es sin duda favorable. En efecto las representaciones en planos factoriales permiten la observación de los opuestos en diferentes dimensiones, sin llegar sin embargo a pretender comprender la totalidad. Se habla de un mayor o menor porcentaje de variación explicada, de configuraciones multidimensionales más o menos estables, etc.

Una ventaja importante de estos métodos es que la interpretación de los resultados se hace en el terreno de lo real, en el sentido de no abstracto. A pesar de los complicados cálculos que se realizan con los paquetes de programas, en la etapa final de interpretación se produce lo que se llama el “retorno a la realidad”, es decir los resultados se expresan en unidades de medida coherentes con el objeto de estudio.

El practicante de análisis de datos no trata de adecuar los datos a sus modelos, sino que busca las estructuras a partir de los datos, por lo que intenta permanentemente dialogar con las otras disciplinas permitiendo así una mayor disponibilidad interdisciplinar. Estas características hacen que esta corriente capte el interés de investigadores con enfoques no deterministas y permitan un mayor reconocimiento de los objetos en su complejidad.

Sin pretender ubicar al análisis de datos en alguna de las posiciones intermedias entre los enfoques cualitativo versus cuantitativo de las ciencias sociales, ya que está suficientemente demostrado la inutilidad de esta ya antigua polémica, es cierto que es visto con mayor simpatía por los cultores del primer enfoque.

En efecto, se citan como características de la metodología cualitativa “proceso interactivo”, “volver sobre los datos”, “analizarlos y replantear el proceso”, “búsqueda de tendencias, tipologías, regularidades o patrones”, “necesidad de traducir los datos en categorías”, “análisis exploratorio”. Este mismo vocabulario puede encontrarse en los textos del Analyse des Données.
Fernando Conde (1987:213-224) en la búsqueda del uso conjunto de las técnicas cuantitativas y cualitativas, plantea el isomorfismo de las dimensiones topológicas entre ambas situando al análisis de datos en el espacio topológico cuantitativo.

Ambitos de aplicación

Procesamiento de encuestas

Las técnicas para el procesamiento y análisis de encuestas tradicional se refieren a la construcción de distribuciones de frecuencias, cruces de variables, cálculos de porcentajes, índices y medidas de tendencia central o de variabilidad.
Se calculan por ejemplo los porcentajes de individuos que eligieron cada ítem de respuesta a una pregunta cerrada, en relación al total de individuos pero también con respecto a otros subtotales que pueden corresponder a los ítems de otra variable. A esto se le denomina cruce de variables o tablas de contingencia sobre las cuales se puede además aplicar un test de independencia como el Chi cuadrado.
Las variables continuas (edad, nº de hijos, ingresos) dan origen al cálculo de medias calculadas sobre el total de la muestra o sobre el conteo de cantidades parciales correspondientes, por ejemplo a las categorías de una variable nominal, o a cálculos de correlaciones midiendo la intensidad del nexo entre dos variables.
Las técnicas de Análisis Multidimensional de Datos (AMD) permiten un análisis más exhaustivo de los datos de una encuesta. Es cierto que es necesario un entrenamiento especial para su puesta en marcha, ya que no son tan simples de utilizar como las técnicas tradicionales. Sin embargo, la rapidez en los cálculos y la percepción casi inmediata de la estructura de los mismos permiten encontrar relaciones que pasarían inadvertidas cuando el usuario está inmerso en una montaña de cifras como es en el caso de los laboriosos análisis de cuadros cruzando

variables de a dos por vez.
La utilización de técnicas de AMD modifica profundamente las primeras fases del tratamiento de datos de encuesta ya que no se trata de complementos refinados que intervienen luego de los métodos tradicionales; por el contrario el AMD cambia radicalmente el encadenamiento de etapas y define una metodología y unos conceptos diferentes.
Con el AMD se puede probar la coherencia global de los datos de una encuesta de una forma rápida y sistemática, da un panorama general, permite criticar la información y orientar los pasos siguientes, elegir las recodificaciones de variables que sean necesarias, agrupamientos, cambios de escala, así como descubrir si una variable ha sido mal definida. Finalmente y como presentación simple de resultados, seleccionar los cruces de variables que más convengan.
Estas operaciones interviniendo al comienzo del procesamiento permiten monitorear los análisis subsiguientes corrigiendo las direcciones de análisis si fuera necesario. En este sentido se encuadran dentro de los procedimientos generales del Análisis Exploratorio.

Entrevistas en profundidad, grupos focales

El desarrollo del análisis textual, como técnica de AMD, ha captado el interés de investigadores en Antropología, Psicología, Ciencias de la Educación que utilizan como herramienta habitual la entrevista semiestructurada, en profundidad o grupos focales. Cuando el corpus textual recogido es muy extenso, los medios informáticos proveen una gran ayuda para su exploración. La aplicación de las técnicas de AMD en este campo se ha visto muy difundida, sin embargo no debe confundirse su utilización con las de otras herramientas informáticas de exploración más al uso de las técnicas cualitativas.
En estos casos las técnicas AMD no deberían utilizarse en todo su nivel de automatización, es importante complementar el procesamiento incluyendo un grado considerable de trabajo artesanal para evitar la supresión de cierta riqueza interpretativa que son de exclusivo patrimonio del trabajo cualitativo.

Investigación de mercados

En el área de investigación de mercados o marketing las técnicas

de AMD encontraron un fecundo campo de aplicación, con diversas modalidades gráficas, que en la jerga del mercadeo se denominan ‘mappings' a los que se agregan los mapas perceptuales.
Algunos de los problemas de marketing comercial o político que se pueden resolver con técnicas de AMD, se refieren a la segmentación del mercado en el marco de una campaña, búsqueda de taxonomías en los programas de audiencia de los medios, percepciones de preferencias en productos cualesquiera, estudios de estructuras de mercado, análisis de la publicidad de distintos productos y/o marcas.
En la actualidad ha tomado gran auge el análisis de las percepciones de los clientes no solamente con respecto a la marca de la campaña sino en relación con el conjunto de marcas. Se enfatiza el interés en desarrollar un ‘mapa de producto’, al cual se lo define como una representación gráfica de las formas en las cuales la gente percibe a los productos en términos de determinados atributos, así como una ayuda para entender sus preferencias. Esto se realiza a través de información que se recoge sobre niveles de productos según algunos atributos y elecciones que se deben realizar entre productos, habitualmente se construyen así matrices de datos de preferencias. De esta manera se pueden encontrar asociaciones entre distintos atributos con varias marcas, productos o servicios y el correspondiente perfil del cliente (Verde, 1999).
Se intensifica el estudio de distintos modelos que son comercializados por diferentes empresas que se dedican al estudio del comportamiento del cliente, la identificación de los factores que componen la satisfacción, el grado de satisfacción de los clientes, de la compañía y/o de la competencia.
Un campo de especial importancia en el actual contexto de recesión económica, lo constituye el diseño de productos (entendiendo a éstos en su sentido más amplio, desde un clásico producto de consumo diario a un producto financiero, o un servicio público) con el fin de ofrecer los productos que mejor puedan satisfacer las necesidades y expectativas de los consumidores. Por eso es de vital importancia incorporar la voz del cliente en el propio diseño del producto. La técnica más adecuada dentro del AMD que permite realizar esto es el Análisis Conjunto (trade-off) cuyo objetivo es determinar aquella combinación de características que hacen “ideal” un producto para un cierto

segmento de consumidores (Aluja Banet, 1994).
Otra posibilidad la constituye por ejemplo obtener un mapa interpretativo del lenguaje publicitario presente en un determinado mercado a través del estudio de los mensajes que se utilizan en publicidad mediante el análisis textual (Balbi, 1999).

Principios básicos de las técnicas de Análisis Multidimensional de Datos

Cada dimensión de una matriz de datos numéricos permite definir las distancias (o proximidades) entre los elementos de la otra dimensión. Así el conjunto de las columnas (que pueden ser variables, atributos) permite definir con la ayuda de fórmulas apropiadas las distancias entre líneas (que pueden ser los individuos, las observaciones). De la misma forma, el conjunto de líneas permite calcular las distancias entre columnas.
Se obtienen así matrices de distancias que se asocian a representaciones geométricas de las similitudes existentes entre las filas y las columnas de las matrices de datos a describir.

El problema es comprender intuitivamente estas representaciones gráficas perdiendo lo menos posible de información.
Creo que este proceso podría ser visto desde el análisis exploratorio como una sucesión de construcciones y de-construcciones de una imagen de la realidad. Tal como en un rompecabezas al que le faltaran piezas y al que en sucesivos armados nuestra intuición y conocimiento del problema permitiera ir agregándolas.

En este tipo de técnicas es más necesario que en otras hacer un uso iterativo de la interpretación como procedimiento asociado del análisis, al respecto me parece muy interesante el concepto de Jesús Ibáñez. “Interpretar es la captación de un sentido oculto: escuchar a la realidad como si la realidad hablara. Analizar es descomponer en sentido en sus componentes sin sentido: silenciar la realidad (porque no dice nada)”(Ibáñez, 1990).

Principales familias de técnicas

En el ámbito del AMD existen dos familias de métodos que permiten realizar reducciones sin perder de vista la estructura fundamental de los datos: los *métodos factoriales* que producen

las representaciones gráficas sobre las cuales las cercanías entre los puntos líneas y los puntos columnas traducen las asociaciones estadísticas entre líneas y columnas y los *métodos de clasificación* que realizan los reagrupamientos en clases de las líneas o de las columnas. Con estos agrupamientos, llamados también tipologías o clusters de individuos con características semejantes se puede obtener una visión macroscópica de la información de manera que el usuario no necesite sumergirse en una marea de datos.
Estas dos familias de métodos se utilizan de manera complementaria y encadenada en etapas logrando así una síntesis explicativa más accesible al usuario.

Evolución del AMD en las últimas tres décadas del siglo pasado

El AMD, entendido como *analyse des données* surge en Francia a inicios de los ’70 con el objetivo de la búsqueda a posteriori de una estructura presente en los datos, en un contexto “más inductivo que deductivo”, que revaloriza el rol del individuo. Su naturaleza prevalentemente descriptiva y el enfoque geométrico de los problemas asignan un lugar importante a las representaciones gráficas, dando gran importancia a los procesamientos computacionales.
En los ’80 se afirmó y se difundió el desarrollo de este enfoque, entrando en contacto con la escuela anglosajona. En 1984 se publican en inglés los libros de Lebart, Morineau y Warwick, así como el del autor sudafricano Greenacre, los cuales contribuyeron en gran medida a acortar las distancias y la incomunicación entre las escuelas estadísticas francesa y anglosajona.
De esta manera, se pasa de un enfoque tal vez demasiado inductivo en el sentido de dar demasiada importancia a los datos en sí mismos para dar la posibilidad de incorporar un lugar para otros conocimientos que no fueran sólo las observaciones. “…el progreso del conocimiento no deriva sólo de conjeturas ni sólo de observaciones, sino de un proceso iterativo que compromete a ambas” (Box, 1979) (En Balbi, 1994)
En la última década se han desarrollado nuevos instrumentos y nuevos campos de aplicación, una de las principales consecuencias de la difusión del AMD ha sido la introducción de numerosos programas en los paquetes estadísticos de mayor difusión, tales como el SAS, el BMDP o el SPSS que se han

interconectado con el paquete especializado, el SPAD, de origen francés.
Entre los nuevos campos de aplicación cabe citar las técnicas de reconocimiento de imágenes y reconstrucción de formas que afrontan problemas de identificación y clasificación.

Asimismo el análisis de datos textuales, cuyas posibilidades de aplicación van desde análisis de las respuestas abiertas en las encuestas al análisis de textos literarios pasando por el procesamiento de textos publicados en internet.
De la misma manera las nuevas herramientas de inteligencia artificial así como de las bases de datos orientadas a objetos han promovido el desarrollo del análisis de datos simbólicos del cual ya hemos hablado.
Una atención especial está recibiendo asimismo la valoración de la estabilidad de los resultados: los métodos de validación empíricos como los cálculos de estabilidad y sensibilidad, cuyas operaciones consisten esencialmente en una verificación de la estabilidad de las configuraciones luego de que se han realizado diversas perturbaciones en la tabla inicial. Las pruebas empíricas de estabilidad se practican en realidad implícitamente de rutina ya que siendo los métodos factoriales de un uso exploratorio, se necesita no sólo de un análisis sino de una serie de análisis: en cada etapa la matriz de datos será modificada por la elección de las variables o de los individuos (agregando o eliminando elementos), las correcciones de eventuales errores, la recodificación de los datos, etc

Otros métodos de validación implican las técnicas de remuestreo. Estas últimas son técnicas de cálculos intensivos que descansan sobre simulaciones de muestras a partir de una sola muestra. Consistirán en este caso en repetir los análisis sobre las diferentes muestras simuladas para luego estudiar las fluctuaciones de los resultados obtenidos, (valores propios, factores o todo otro parámetro estadístico a estimar). Para ello, se evalúa la variabilidad real de un parámetro por intermedio de su variabilidad para el conjunto de estas series de datos.
Los métodos de validación que permiten engendrar de manera diferente muestras artificiales, son numerosos. Los más conocidos son JakKnife, Bootstrap, validación cruzada.

Paquetes de programas

La presentación de las diferentes técnicas de AMD implican la consideración de los correspondientes paquetes de programas que las calculan, ya que como bien decía Benzécri es imposible la práctica del Analyse des Données sin el auxilio de la computadora.

El software de origen francés que más trascendencia tuvo es el SPAD (Système Portable pour l'Analyse des Donnés) (Lebart y ot., 1982). Desarrollado en módulos por distintos autores, los programas que realizaban las rutinas más importantes se incluyeron inicialmente en los libros que presentaban las técnicas (Lebart,1977), escritos en lenguaje FORTRAN o APL.

El primer SPAD grabado en tarjetas perforadas, luego en cintas magnéticas, se ejecutaba en los otrora grandes equipos "mainframe". En esa época (1980) llamaba la atención que mientras los programas estadísticos (Minitab, SAS) que llegaban a nuestro país procedentes de los Estados Unidos sólo podían alquilarse pagando una renta anual y luego se "autodestruían", el SPAD se vendía completo, pero además con el entero listado del programa fuente en FORTRAN, de forma que eventualmente se lo podía reproducir, ampliar o mejorar. Veamos cómo lo definen ellos mismos (Lebart y ot.1983):

…"SPAD es algo más que una simple biblioteca de programas, juega también el rol de permitir la difusión de módulos de cálculo y de gestión relativamente especializados. La normalización y la compatibilidad aseguran la difusión horizontal necesaria a todo trabajo científico, y permiten su instalación en sitios muy variados (países del este, África, América del Norte, del Sur…) Estas cualidades permiten además preparar nuevas versiones para la actualización y la extensión, facilitan la conversión a nuevos materiales presentes o futuros. La informática se descentraliza, y esta evolución parece acelerarse. La autonomía que así se gana tiene el precio de grandes esfuerzos de adaptación y de diversificación de los programas, esfuerzos muy facilitados cuando los programas están escritos de forma transportable y modular"

Los autores de este programa no tuvieron problemas de frontera, en las postrimerías de la guerra fría se trasladaron con sus cajas de tarjetas perforadas organizando congresos tanto en Cuba como en México o en Venezuela (Morineau y ot. 1983). La actitud de esta escuela se ve reflejada en los propios estatutos del CESIA

(Centre de Statistique et d'Informatique Apliquées), la primera institución en comercializar SPAD:
"Art.2: Fines de la asociación.
La Asociación tiene por fines promover la práctica, el estudio y la investigación en el dominio de la estadística aplicada en relación con la informática.
Esta asociación inter-laboratorio tiene igualmente la vocación de ayudar y participar a la creación, la documentación y la difusión de programas para ordenadores con un objetivo exclusivamente científico o pedagógico. Quiere promover una libre circulación de los métodos y los programas, frente a las eventuales retenciones de software que pueden ser practicadas por la empresas privadas con fines comerciales o por la empresas públicas, administraciones o laboratorios a los fines de prestigio"[25]
Puede entenderse que estos conceptos ejercieran cierta fascinación en la que escribe estas páginas, quien en plena época oscurantista de la Argentina. cuando ciertos libros eran bienes preciosos, tuvo la oportunidad de acceder a esta bibliografía y a este nuevo enfoque de pensamiento.
Existen y existieron además, en Francia y luego en otros países europeos que se fueron sumando a esta escuela, variados paquetes de programas ya sea de laboratorios científicos universitarios como de empresas privadas. Entre los primeros podemos citar a MODULAD otro proyecto interlaboratorio creado por iniciativa del INRIA – Rocquencourt; así como el SICLA, desarrollado por el mismo instituto; el IDAMS, un programa internacional de UNESCO. El grupo italiano de la Universidad de Nápoles "Federico II" por su parte desarrolló SLAM, una interfaz amigable para la ejecución de SPAD.N. Todos estos programas estaban destinados para su ejecución en mainframe y por lo tanto eran en general bibliotecas FORTRAN. En la actualidad se reconvirtieron para su utilización en computadoras personales, bajo entorno Windows y algunos de ellos son explotados de forma comercial.
En Estados Unidos sólo a partir de la aparición en inglés de la bibliografía específica, se fueron incorporando las técnicas de análisis de correspondencias en los softwares estadísticos tradicionales SPSS, SAS, SYSTAT, BMDP. Sin embargo la idea básica de complementar los dos grupos de técnicas, análisis

[25] Estatutos del CESIA, creado el 4 de mayo de 1973.

factorial y clasificación a los efectos de interpretar integralmente la estructura de los datos sigue siendo patrimonio del software francés.

Etapas en un AMD clásico

La complementariedad entre las técnicas factoriales y de clasificación se plasma en la práctica de procesamiento como el encadenamiento de una serie de etapas que pueden describirse como sigue:

Etapa 1: Análisis factorial

Elección de una batería homogénea de variables activas.

A pesar de que una característica importante del AMD es la utilización de un gran número de variables, ellas no pueden ser incluidas indiscriminadamente en el análisis esperando que como una caja negra el programa resuelva los algoritmos. Aún cuando es corriente la utilización del término “análisis inteligente” ello no significa que pueda reemplazarse el criterio metodológico de la selección de variables. En los próximos capítulos especificaré con más detalle la definición de variable “activa”, por el momento sólo nominaré como variables activas a aquéllas que intervienen en la determinación de los ejes del análisis factorial. Para que el análisis tenga coherencia será necesario que el conjunto de las variables activas sea homogéneo. En el lenguaje metodológico ello implicará que se trate en realidad de indicadores que midan el mismo tipo de dimensiones y variables. La selección de variables activas implicará asimismo una contraparte: la selección de variables ilustrativas. Éstas últimas no son menos importantes, su inclusión en esta categoría se debe a que constituyen un conjunto de diferente ‘especie’ que las activas, ya sea porque son indicadores de otra dimensión o variable teórica, porque juegan en diferente sentido que las activas o porque pueden distorsionar el análisis.

Descripción gráfica de la población: la proximidad entre individuos es función de la semejanza de las respuestas a las variables activas.

En la misma matriz de datos será equivalente analizar los perfiles filas en el espacio de las columnas que los perfiles columnas en el

espacio de las filas. Esta correspondencia que se establece entre ambos espacios es la base de la interpretación gráfica. La visualización de los puntos individuos en el espacio de la variables deberá corresponder naturalmente al valor de sus coordenadas (valores de las variables) en ese espacio, si el conjunto de valores relativos de un individuo (perfil) es parecido al conjunto de valores relativos o perfil de otro, entonces ambos individuos se representarán como dos puntos cercanos en el espacio de las variables. Esos espacios multidimensionales son reducidos a través del análisis factorial y los puntos podrán ser representados en el nuevo espacio determinado por los ejes factoriales.

Posicionamiento de los elementos ilustrativos sobre los planos factoriales.

Una vez construido el nuevo espacio factorial puede ser visualizado a través del cruce entre dos ejes factoriales, los cuales serán perpendiculares entre sí (o sea estadísticamente independientes). Dos ejes factoriales determinarán un plano factorial de máxima inercia sobre el cual podrán luego proyectarse las variables ilustrativas.

Un ejemplo de esto se puede buscar en el análisis de una encuesta de opinión. Si elegimos como variables activas a los indicadores de opinión, podremos observar en el plano factorial la estructura de respuestas a favor, indiferentes o en contra de un determinado tema. Las variables contextuales que indicamos como ilustrativas estarán en cambio respondiendo a la pregunta: ¿quiénes son los que contestan de una forma u otra?

Etapa 2: Clasificación mixta sobre un subespacio factorial

Las distancias entre los puntos individuos son calculadas en el espacio de los primeros ejes factoriales con la distancia euclidiana usual

Esta etapa marca una diferencia importante con los otros enfoques de análisis multivariado. Los programas computacionales estadísticos incluyeron desde el comienzo técnicas de cluster o clasificación, sin embargo los cálculos siempre se realizaron con los valores ‘crudos’ de las variables, o sea los verdaderos valores que tenía cada individuo en cada una de las variables medidas en su escala original. El enfoque de la

escuela francesa se caracteriza por la complementariedad de las técnicas, lo cual indica que la clasificación se realiza tomando en cuenta no los valores crudos de las variables sino los valores de cada punto individuo respectivamente en cada eje factorial. De esta manera se suaviza la posible distorsión ocasionada por escalas de medidas diferentes y se trabaja con el bloque más importante de variación que contiene las oposiciones fundamentales encontradas en los datos.

Etapa 3: Descripción automática de las clases

A partir de los individuos agrupados en clases se calculan para todas las variables las medidas diferenciales entre los valores de la clase y los totales

Una vez que se clasificaron los individuos de acuerdo con sus cercanías en el espacio factorial, o sea de acuerdo con la similitud de sus respectivos perfiles, se tratará de describir estas clases o grupos de individuos de acuerdo con sus características en común, pero 'retornando a la realidad' de sus verdaderos valores en las variables de origen. Así se calculan los porcentajes de individuos de la clase que poseen una categoría (una determinada opinión por ejemplo) y se los compara con los porcentajes de individuos que la poseen en la población, de esta manera se puede determinar la eventual concentración de esa categoría de opinión en la clase en cuestión. De esta forma pueden asignarse categorías características de ciertas clases facilitando la interpretación de cada uno de los grupos.

Etapa 4: Posicionamiento de las clases en el plano factorial

Esta etapa permite visualizar las posiciones relativas de las clases en el espacio y poner en evidencia trayectorias ocultas por la discontinuidad de las clases.

La posición de cada individuo identificado por el número de su clase permite representar la densidad y la dispersión de las clases en el plano factorial.

La correspondencia entre los espacios permite calcular las coordenadas factoriales tanto de los puntos variables como de los puntos individuos. Podrá observarse que ambas configuraciones son visualmente equivalentes. Ello permite aprovechar esta posibilidad de representación proyectando los puntos individuos

en los primeros ejes factoriales e identificando cada punto con el número (y eventualmente un color) de la clase a la cual pertenece. Esta representación permitirá observar la relativa concentración que presenta cada clase, las posibles superposiciones de clases cuando lo que se está representando a veces son sólo manifestaciones indicadoras de la continuidad de una variable teórica.
Si tenemos en cuenta que las clases construidas en AMD clásico son mutuamente excluyentes, es decir que un individuo pertenece sólo a una clase, se verá que la posibilidad de observar la vecindad de ciertas clases en el espacio factorial proveerá de gran riqueza interpretativa.

En los próximos capítulos expondré a través de ejemplos los principios de funcionamiento de las principales técnicas de AMD. Excluyo la exposición del Análisis de Componentes Principales (ACP), que es habitualmente descrito en primer lugar como técnica de análisis factorial para variables continuas, porque es fácilmente encontrado en los textos habituales de análisis multivariado existiendo además una reciente y excelente versión en español (Aluja y Morineau, 1999).

CAPÍTULO 6

EL ANÁLISIS FACTORIAL DE CORRESPONDENCIAS SIMPLE O BINARIO

El análisis (factorial) de correspondencias, ha sido descubierto y redescubierto por numerosos autores durante decenios, a medida que surgían investigaciones y se utilizaban las técnicas. Ha sido también llamado método de los pesos de Guttman, ACP (análisis de componentes principales) sobre datos cualitativos, teoría de la cuantificación de Hayashi, regresión lineal simultánea, biplot, dual scaling, reciprocal averages, additive scoring, canonical scoring, etc.(Lebart y ot., 1995)

Benzécri (1982:99) lo nombra por primera vez análisis de correspondencias en sus trabajos del otoño del año 1962, puntualizando la distancia distribucional. Con posterioridad, conjuntamente con Brigitte Escofier, (quien demuestra en su tesis doctoral la fórmula baricéntrica o de transición) ponen en evidencia la simetría de filas y columnas en el análisis.

Aunque resulte difícil establecer la paternidad del método ya que diversos caminos fueron recorridos por distintos autores, es en la escuela francesa donde el mismo dio sus mejores frutos. Benzécri (1982:115) puntualiza en su obra: “La prioridad de estos autores es cierta: la única originalidad que pueden reivindicar los investigadores franceses es la de haber conjugado, con un método descubierto independientemente por numerosos autores, unas ideas y unos problemas múltiples para los cuales la síntesis no había sido realizada”.

El Análisis de Correspondencias Simple o Binario se aplica a una tabla de contingencia, estamos por lo tanto en el caso de una matriz de datos especial, producto ya de un procesamiento previo, donde las líneas no son verdaderamente individuos sino categorías de una variable y las columnas las categorías de otra. De allí su nombre de binario, ya que en realidad involucra a sólo dos variables.

El Análisis Factorial de Correspondencias (AFC en adelante) es esencialmente un modo de representación gráfica de las tablas de contingencia que son en general tablas de clasificación cruzadas o grupos de tablas de clasificación cruzadas.

El AFC apunta a agrupar en uno o varios gráficos, generalmente menos de 4 y muy a menudo sólo uno, la mayor parte de la información contenida en la tabla, refiriéndose no a los valores absolutos sino a las *correspondencias* entre los valores relativos.
Desde luego, este método de presentación es tanto más útil cuando la dimensión de la tabla es grande, pues una masa numérica voluminosa tiende automáticamente a diluir los hechos salientes, mientras que una tabla pequeña se interpreta por sí misma.
Como ya puntualizamos, existe el inconveniente de que las reglas de interpretación de estas representaciones no son tan simples como las de la estadística descriptiva elemental. En los años '70 se pensaba que para lograr una provechosa interpretación era necesario algo así como una experiencia clínica: cada matriz de datos era de alguna manera un caso particular. En la actualidad los gráficos de análisis factorial se encuentran lo bastante difundidos como para que puedan explicitarse ciertas reglas generales de interpretación
La paradoja pedagógica consiste en que no se pueden simplificar los ejemplos. Habitualmente en la enseñanza de técnicas estadísticas se parte de un ejemplo con pocos datos para poder comprender la operatoria. En este caso, al tratarse de técnicas que justifican su razón de ser a través de la visualización de la estructura de un gran número de variables, si simplificamos los ejemplos desvirtuamos la lógica de la interpretación.

Principios básicos del AFC a partir de un ejemplo numérico simple

Presento un ejemplo de tabla de contingencia en la cual la simplicidad de su interpretación creo que permitirá un mayor acercamiento a las características del problema.
La tabla siguiente se refiere a datos de inmigración en diversas provincias de la Argentina. Se representan en fila las categorías de la variable provincias ar gentinas y en columna la nacionalidad de los inmigrantes.
Los datos reflejarán naturalmente una asociación resultante de la cercanía geográfica, la cual será interesante de observar cómo se reproduce en la representación gráfica.

CUADRO N° 1

MIGRANTES DE PAÍSES LIMÍTROFES

Provincias	Bolivianos	Brasileros	Chilenos	Paraguayos	Uruguayos	Total
Buenos Aires	27544	33019	59912	117476	91856	329807
Catamarca	100	9	141	14	21	285
Córdoba	1814	2234	1014	756	1453	7271
Corrientes	14	1588	73	3638	1153	6466
Chaco	80	261	149	13570	433	14493
Entre Ríos	54	613	134	515	46672	47988
Chubut	54	42	16080	98	174	16448
Jujuy	40978	48	413	618	37	42094
La Pampa	15	155	714	45	440	1369
La Rioja	28	23	217	14	17	299
Mendoza	3622	2546	9476	259	416	16319
Misiones	116	21106	149	618	37	22026
Neuquén	34	36	11040	36	61	11207
Río Negro	86	136	22453	83	169	22927
Salta	27505	99	1036	1579	86	30305
San Juan	450	559	4037	66	91	5203
San Luis	134	49	209	30	64	486
Santa Cruz	96	21	16122	54	128	16421
Santa Fe	414	3387	800	2469	2013	9083
S.del Estero	88	105	72	64	89	418
T. del Fuego	0	3	3223	9	7	3242
Tucumán	1101	344	471	306	163	2385
Total	104327	66383	147935	142317	145580	606542

El monto de las cifras así como el número de casillas dificulta la interpretación de este cuadro, por lo que se practica una primera reducción que consiste en el cálculo de las frecuencias relativas con respecto al total general.

Dividiendo por el total que es 606.542, podemos realizar un acercamiento comparativo. Estas cifras, multiplicadas por 100 para su mejor visualización, se presentan en el Cuadro Nº2.

CUADRO Nº2

MIGRANTES DE PAÍSES LIMÍTROFES: PORCENTAJES

Provincias	Bolivianos	Brasileros	Chilenos	Paraguayos	Uruguayos	Total
Buenos Aires	4,54	5,44	9,88	19,37	15,14	54,37
Catamarca	0,02	0,00	0,02	0,00	0,00	0,05
Córdoba	0,30	0,37	0,17	0,12	0,24	1,20
Corrientes	0,00	0,26	0,01	0,60	0,19	1,07
Chaco	0,01	0,04	0,02	2,24	0,07	2,39
Entre Ríos	0,01	0,10	0,02	0,08	7,69	7,91
Chubut	0,01	0,01	2,65	0,02	0,03	2,71
Jujuy	6,76	0,01	0,07	0,10	0,01	6,94
La Pampa	0,00	0,03	0,12	0,01	0,07	0,23
La Rioja	0,00	0,00	0,04	0,00	0,00	0,05
Mendoza	0,60	0,42	1,56	0,04	0,07	2,69
Misiones	0,02	3,48	0,02	0,10	0,01	3,63
Neuquén	0,01	0,01	1,82	0,01	0,01	1,85
Río Negro	0,01	0,02	3,70	0,01	0,03	3,78
Salta	4,53	0,02	0,17	0,26	0,01	5,00
San Juan	0,07	0,09	0,67	0,01	0,02	0,86
San Luis	0,02	0,01	0,03	0,00	0,01	0,08
Santa Cruz	0,02	0,00	2,66	0,01	0,02	2,71
Santa Fe	0,07	0,56	0,13	0,41	0,33	1,50
S.del Estero	0,01	0,02	0,01	0,01	0,01	0,07
T. del Fuego	0,00	0,00	0,53	0,00	0,00	0,53
Tucumán	0,18	0,06	0,08	0,05	0,03	0,39
Total	17,20	10,94	24,39	23,46	24,00	100,00

Sin embargo sigue resultando difícil efectuar las comparaciones entre provincias y migrantes, por lo tanto se intenta una nueva simplificación.
En el Cuadro N°3 se representan los *perfiles fila* expresados en porcentajes, ellos son obtenidos dividiendo cada elemento por la suma de la fila correspondiente: el perfil fila de la provincia de Mendoza, por ejemplo, se obtiene dividiendo cada término de la fila del Cuadro N°1 por 16319.

Para mayor legibilidad se multiplica por 100. De esta manera podremos observar que el 22,19% de los migrantes que llegaron a Mendoza son bolivianos, el 15,6% son brasileros, 58% chilenos, 1% paraguayos y 2% uruguayos.
Podemos decir entonces que a Mendoza llegan migrantes fundamentalmente procedentes de Chile.

CUADRO Nº3
MIGRANTES DE PAÍSES LIMÍTROFES: PERFILES FILAS

Provincias	Bolivianos	Brasileros	Chilenos	Paraguayos	Uruguayos	Total
Buenos Aires	8,35	10,01	18,17	35,62	27,85	100,00
Catamarca	35,09	3,16	49,47	4,91	7,37	100,00
Córdoba	24,95	30,72	13,95	10,40	19,98	100,00
Corrientes	0,22	24,56	1,13	56,26	17,83	100,00
Chaco	0,55	1,80	1,03	93,63	2,99	100,00
Entre Ríos	0,11	1,28	0,28	1,07	97,26	100,00
Chubut	0,33	0,26	97,76	0,60	1,06	100,00
Jujuy	97,35	0,11	0,98	1,47	0,09	100,00
La Pampa	1,10	11,32	52,15	3,29	32,14	100,00
La Rioja	9,36	7,69	72,58	4,68	5,69	100,00
Mendoza	22,19	15,60	58,07	1,59	2,55	100,00
Misiones	0,53	95,82	0,68	2,81	0,17	100,00
Neuquén	0,30	0,32	98,51	0,32	0,54	100,00
Río Negro	0,38	0,59	97,93	0,36	0,74	100,00
Salta	90,76	0,33	3,42	5,21	0,28	100,00
San Juan	8,65	10,74	77,59	1,27	1,75	100,00
San Luis	27,57	10,08	43,00	6,17	13,17	100,00
Santa Cruz	0,58	0,13	98,18	0,33	0,78	100,00
Santa Fe	4,56	37,29	8,81	27,18	22,16	100,00
S.del Estero	21,05	25,12	17,22	15,31	21,29	100,00
T. del Fuego	0,00	0,09	99,41	0,28	0,22	100,00
Tucumán	46,16	14,42	19,75	12,83	6,83	100,00
Total	17,20	10,94	24,39	23,46	24,00	100,00

El concepto de *perfil* resulta de gran interés en este enfoque e implica una relativización a los efectos de la comparación. En este sentido podemos comparar las filas, es decir las provincias entre sí, porque están todas en la misma unidad de medida. Así, ponemos en evidencia que el perfil de Mendoza va a ser distinto del de Buenos Aires, comparación que nos resultaba más difícil en las tablas anteriores debido a que Buenos Aires sobrepasaba en mucho a las demás provincias.

La comparación de dos perfiles fila nos va a informar sobre la forma en que dos provincias se asocian con los países de proveniencia de los migrantes.

Cuando dos provincias sean parecidas en su distribución con respecto a la composición por países de los migrantes que reciben, sucederá que los dos puntos que representen a esas provincias en el gráfico de correspondencias se encontrarán

cerca, por ejemplo las provincias de Salta y Jujuy tienen perfiles parecidos, significando que poseen una estructura inmigratoria de los países vecinos más similar entre sí que al resto de las provincias.
Simétricamente podremos simplificar la comparación entre nacionalidad de los migrantes. El cuadro siguiente representa los perfiles-columna. De forma análoga se obtienen dividiendo cada elemento del Cuadro N°1 por el total de cada columna correspondiente y multiplicándolo por 100.

CUADRO Nº 4

MIGRANTES DE PAÍSES LIMÍTROFES: PERFILES COLUMNAS

Provincias	Bolivianos	Brasileros	Chilenos	Paraguayos	Uruguayos	Total
Buenos Aires	26,40	49,74	40,50	82,55	63,10	54,37
Catamarca	0,10	0,01	0,10	0,01	0,01	0,05
Córdoba	1,74	3,37	0,69	0,53	1,00	1,20
Corrientes	0,01	2,39	0,05	2,56	0,79	1,07
Chaco	0,08	0,39	0,10	9,54	0,30	2,39
Entre Ríos	0,05	0,92	0,09	0,36	32,06	7,91
Chubut	0,05	0,06	10,87	0,07	0,12	2,71
Jujuy	39,28	0,07	0,28	0,43	0,03	6,94
La Pampa	0,01	0,23	0,48	0,03	0,30	0,23
La Rioja	0,03	0,03	0,15	0,01	0,01	0,05
Mendoza	3,47	3,84	6,41	0,18	0,29	2,69
Misiones	0,11	31,79	0,10	0,43	0,03	3,63
Neuquén	0,03	0,05	7,46	0,03	0,04	1,85
Río Negro	0,08	0,20	15,18	0,06	0,12	3,78
Salta	26,36	0,15	0,70	1,11	0,06	5,00
San Juan	0,43	0,84	2,73	0,05	0,06	0,86
San Luis	0,13	0,07	0,14	0,02	0,04	0,08
Santa Cruz	0,09	0,03	10,90	0,04	0,09	2,71
Santa Fe	0,40	5,10	0,54	1,73	1,38	1,50
S.del Estero	0,08	0,16	0,05	0,04	0,06	0,07
T. del Fuego	0,00	0.00	2,18	0,01	0.00	0,53
Tucumán	1,06	0,52	0,32	0,22	0,11	0,39
Total	100,00	100,00	100,00	100,00	100,00	100,00

La comparación de dos perfiles-columna nos permite observar la proximidad entre dos categorías de la variable nacionalidad de los migrantes teniendo en cuenta la provincia argentina, representando en cada país de origen a qué provincia se dirigen. Así, podemos observar que de Bolivia se dirigen preferiblemente a Jujuy, Salta y Buenos Aires, mientras que los chilenos optan por la Patagonia y también por Buenos Aires; podemos comparar los

perfiles de los países.
La vecindad de dos puntos que representen dos nacionalidades en el gráfico factorial de análisis de correspondencias, se traducirá en una similitud de los dos perfiles-columna.

Se trata de poner en relación ambos perfiles a la vez: el análisis de correspondencias va a describir simultáneamente las similitudes entre perfiles de filas y perfiles de columnas, y dará una representación esquemática de las informaciones contenidas en los cuadros 3 y 4.

Notaciones

Denotaremos con:
kij término general del cuadro Nº 1 correspondiente a la fila i y la columna j .
(Si i = Mendoza y j = chilenos, entonces kij = 9476)

fij término general del cuadro Nº2 donde sus elementos están divididos por el efectivo total k. También llamado frecuencia relativa. (k = 606.542 en el cuadro Nº1, por lo tanto fij en el caso Mendoza-chilenos será igual a 0.0156)

Luego:

fi. = Σj fij para j variando de 1 a p (p es el total de variables, luego fi. es el total de la suma de columna)

f.j = Σi fij para i variando de 1 a n (n es el total de individuos, luego f.j es el total de la suma de fila)

fi. y f.j son también llamados frecuencias relativas marginales

Llegamos así al primer concepto clave: el de *perfil.*
Se define al perfil de la fila i como el conjunto de p valores (uno para cada columna de la fila i):

(fij/fi.) para j = 1, ... p

El perfil de la columna j es el conjunto de n valores (uno para cada línea de la columna j):

(fij/fj.) para i = 1, ... n

Los elementos de los perfiles, multiplicados por 100, son los porcentajes filas y los porcentajes columnas representados en los cuadros 3 y 4.
En el Cuadro Nº3, el perfil de la fila Mendoza está constituido por los valores: 22.19, 15.60, 58.07, 1.59, 2.55
En el Cuadro Nº4, el perfil de la columna chilenos está constituido por los valores: 40,50 0.10, 0.69, 0.05,...,0.32.
El análisis de correspondencias permite describir las proximidades existentes entre perfiles-fila y perfiles-columna, teniendo en cuenta, sin embargo, la diferencia de cantidades existentes entre estas filas y estas columnas. El análisis de la estructura de los datos puede realizarse de una manera general y no punto a punto, no una determinada fila con una determinada columna, sino en general las filas con las columnas.

Las nubes originadas por la tabla de contingencia

Para el análisis de una tabla de contingencia, razonaremos en términos de perfiles, esto permite la comparación de las modalidades de una misma variable, es decir entre países o entre provincias. Como ya dijimos, las proximidades espaciales se interpretan como similitudes[26].

Nube de las n filas en el espacio de las columnas

El conjunto de los perfiles fila forma una nube de n puntos en el espacio de las p columnas y representa aquí la nube de las 22 provincias. Cada punto fila tiene por coordenadas en el R^p a los 5 componentes del perfil fila: (fij/fi.) para j = 1, ... 5.
Como la suma de estas componentes a lo largo de las 5 columnas es igual a 1, en realidad los n=22 puntos fila van a estar situados en un subespacio de (p-1)= 4 dimensiones.

El centro de gravedad de esta nube es la media de los perfiles fila y corresponde al punto perfil medio, es decir al perfil de países sobre el conjunto de la población: cada uno de sus componentes

[26] Para revisar el concepto de coordenada, ver en el capítulo 4, “La matriz de datos en el espacio”, teniendo en cuenta que en este caso la matriz de datos es una tabla de contingencia.

es f.j , es decir la frecuencia marginal de los países.

Nube de las n columnas en el espacio de las filas

El conjunto de los perfiles columna forma una nube de p puntos en el espacio de las n filas y representa aquí la nube de los 5 países. Cada punto columna tiene por coordenadas en el R^n a los 22 componentes del perfil columna: (fij/f.j) para i = 1, ... 22.
Como la suma de estas componentes a lo largo de las 22 filas es igual a 1, en realidad los p=5 puntos columna van a estar situados en un subespacio de (n-1)= 21 dimensiones.

El centro de gravedad de esta nube es la media de los perfiles columna y corresponde al punto perfil medio, es decir al perfil de provincias sobre el conjunto de la población: cada uno de sus componentes es fi. , es decir la frecuencia marginal de las provincias.

Ajuste de las nubes

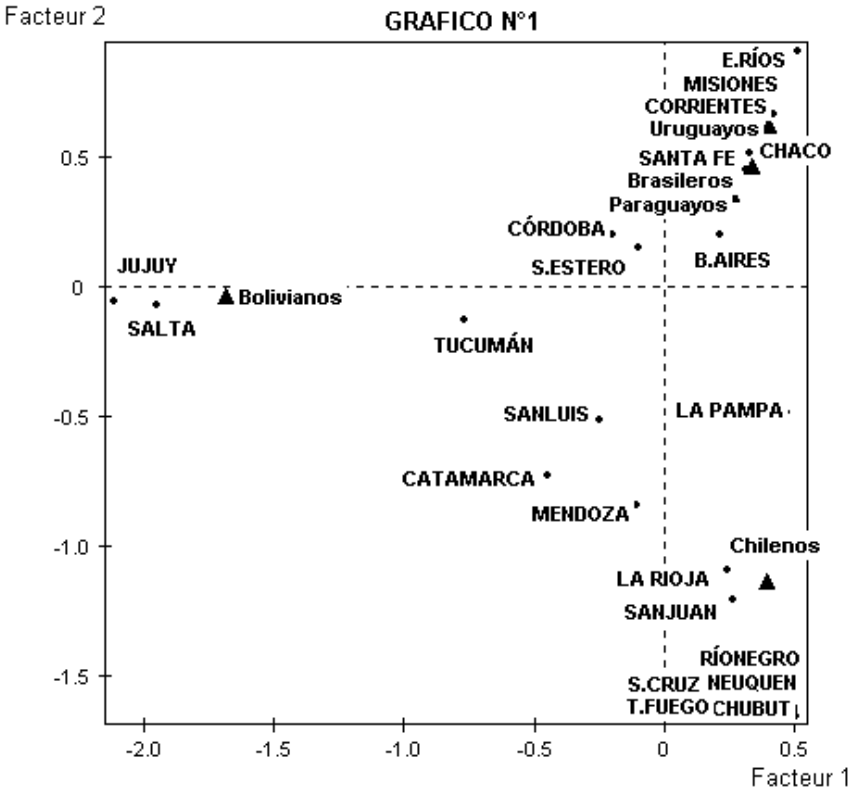
El objetivo que buscamos es el de representar geométricamente las similitudes entre las diferentes categorías de una misma variable, lo cual nos conduce a representar las proximidades entre los perfiles y el perfil medio definido sobre el conjunto de la población.
Una nube de puntos concentrada alrededor de su centro de gravedad es una nube en la cual los puntos perfiles están cerca del perfil medio, por lo tanto ello significará una cierta independencia entre las dos variables nominales.
En la construcción de las nubes en el R^p y en el R^n, la elección de los perfiles (en lugar de valores absolutos) como coordenadas, les da la misma importancia a todas las categorías de países y provincias. Sin embargo se le restituye luego su importancia a través de la masa afectada a cada punto (proporcional a su frecuencia) a los efectos de no privilegiar las categorías con bajas frecuencias y de respetar la distribución real de la población. Esta masa intervendrá por una parte en el cálculo de las coordenadas del centro de gravedad de la nube y por otra en el criterio de ajuste.
Para el cálculo del ajuste, la cantidad que se deberá maximizar será entonces la suma ponderada de los cuadrados de las distancias entre los puntos y el centro de gravedad de la nube (es decir la inercia o variancia de la recta de alargamiento máximo de

la nube) utilizando una distancia entre perfiles que definiremos más adelante.
¿Cómo se calculan los ejes factoriales? El ajuste de los puntos originales consiste en ir encontrando direcciones (o ejes) que acumulen la mayor cantidad de variación de la nube original, con la condición que esos ejes resulten perpendiculares. El cálculo se realiza con lo que se llama diagonalización de la matriz o búsqueda de valores y vectores propios. Se demuestra que el eje de máxima inercia (eje N°1) es el vector propio correspondiente al mayor valor propio de la matriz de datos. El eje N°2, el que le sigue en magnitud de variación, es el vector propio correspondiente al segundo mayor valor propio de la matriz de datos. Así sucesivamente con el resto de vectores y valores propios a encontrar. Para el análisis consideramos habitualmente sólo los 2 primeros.

Gráfico factorial

Las nubes de puntos filas y de puntos columnas van a ser representadas en los planos de proyección formados por los primeros ejes factoriales tomados de a dos.
Es decir, los puntos filas y puntos columnas han sido proyectados sucesivamente en el eje 1 y en el eje 2. El cruce de ambos ejes o factores da lugar al primer plano factorial. Una imagen intuitiva del mismo podría ser la de un plano que atraviesa una nube de puntos por el lugar en la cual ella posee su mayor variación, sobre ese plano se proyectan entoces todos los puntos de la nube.
Veremos a continuación el gráfico factorial resultante de aplicar un análisis de correspondencias a nuestra matriz de datos, a nuestra tabla de contingencia. En el caso del análisis de correspondencias tiene sentido hablar de representación simultánea de ambos espacios: el de filas y el de columnas.

El gráfico Nº 1, representando el plano factorial engendrado por los dos primeros ejes del análisis de correspondencias del cuadro Nº1, da una representación visual de las asociaciones entre filas y columnas de esta tabla de contingencia.



¿Cómo leer el Gráfico nº 1?

Mostraré al mismo tiempo algunas reglas de interpretación y los principios de la técnica.

En este gráfico se representaron los dos ejes factoriales de mayor inercia, es decir de mayor porcentaje de variabilidad, los ejes 1 y 2.

Observamos proyectados en este plano, los puntos correspondientes a los valores de las dos variables consideradas, distinguiéndose en letra minúscula las nacionalidades de los migrantes y en mayúscula los nombres de las provincias.

Hacia abajo, en cuanto a la filas encontramos las provincias de Río Negro, Neuquén, Santa Cruz, Chubut y Tierra del Fuego.

Si dos puntos-fila tienen los perfiles idénticos o vecinos, estarán superpuestos o cerca, en el gráfico. Por ejemplo Jujuy y Salta se encuentran cerca, pero en el extremo izquierdo del gráfico. Igualmente ocurrirá con dos puntos columna.

El origen de los ejes corresponde a los perfiles medios (totales de la tabla de contingencia).
El *perfil fila medio* tiene como componentes:

$$f_{.j} \quad \text{para} \quad j = 1,\ldots p$$

Es decir, sería el total de la fila, sumado a lo largo de las p columnas.

El *perfil columna medio* tiene por componentes:

$$f_{i.} \quad \text{para} \quad i = 1, \ldots n$$

Es decir, sería el total de la columna, sumado a lo largo de las n filas.

En este caso no hay puntos columnas muy semejantes al punto columna medio, y los puntos filas más semejantes al punto fila medio podrían ser: Santiago del Estero, Córdoba, o Buenos Aires.
Los puntos que ocupen las posiciones periféricas tendrán entonces los perfiles más diferentes de los perfiles medios, tal el caso de los puntos: Salta o Jujuy por un lado, Chubut por otro y Entre Ríos por otro extremo.
Una forma práctica de acercarnos a la interpretación del gráfico es analizar las oposiciones en los ejes. El eje 1, en este caso representado como eje horizontal, es el que acumula el mayor porcentaje de variación, ¿qué está oponiendo?
Observamos que los puntos de Salta y Jujuy se encuentran como ya dijimos en el extremo izquierdo, mientras que hacia la derecha están el resto de las provincias.
Sin embargo de estas últimas algunas se encuentran hacia arriba como Entre Ríos, Misiones y Corrientes y otras hacia abajo como Tierra del Fuego, Santa Cruz y Chubut. Estas diferencias de ubicación están reflejando en realidad, las oposiciones relativas al eje 2, representado en este caso como eje vertical.
Por lo tanto observamos que un primer criterio, el de más fuerza de variación, opone Salta y Jujuy contra todo el resto, y un segundo criterio, el que le sigue en magnitud de variación, opone las provincias del noreste contra las de la Patagonia.
¿Cuáles son estos criterios de diferenciación? ¿Qué hace que se diferencien (o se agrupen) las provincias de esta manera? Naturalmente la estructura de nacionalidad de los inmigrantes, que son los datos que estamos estudiando en la segunda variable

de la tabla de contingencia.
¿Qué pasa entonces en las columnas? Los bolivianos se oponen al resto, tienen un comportamiento completamente diferente, emigran casi en exclusividad a las provincias de Salta y Jujuy, mientras que el resto de países se reparten en las otras provincias: los uruguayos, paraguayos y brasileros a las provincias del Litoral y los chilenos a la Patagonia.
¿Qué sucede con los perfiles medios como Buenos Aires? ¿Significa que reciben pocos migrantes? Todo lo contrario. En el centro del gráfico se ubican en general los puntos de mayor frecuencia, por tal razón sucede que no pueden asociarse con algún punto en particular, sino que se encuentran en una posición de indiferencia. En el caso de Buenos Aires recibe inmigrantes de todos los países por igual, es decir respetando las frecuencias relativas de cada país. Por ejemplo llegan a Buenos Aires menor cantidad de brasileros, pero en el total del país la frecuencia de brasileros es también la menor.

¿Qué significación tienen las proximidades?

Ahora tenemos que definir lo que es "cerca" es decir la distancia que se calcula y de la cual estos gráficos planos dan una aproximación.
Los perfiles, es decir esa serie de n o p números según se trate de filas o columnas, permiten definir puntos en los espacios de n o de p dimensiones. Como se trata de perfiles, es decir de proporciones cuya suma vale 1, existe la restricción:
$\sum_{i=1..n} f_{i.} = \sum_{j=1..p} f_{.j} = 1$, es decir la suma de los perfiles medios de filas o columnas es igual a la unidad, los espacios tendrán en realidad n-1 y p-1 dimensiones (la que falta queda determinada, tenemos un grado de libertad de menos).
Las distancias se definirán en estos espacios; el objetivo del análisis de correspondencias será reducir estas dimensiones de forma de obtener una representación visual tratando de deformar lo menos posible estas distancias.

La distancia entre dos puntos - fila i e i'estará dada por:

$$d^2 (i , i ') = \sum_{j=1..p} 1/f_{.j} (f_{ij} / f_{i.} - f_{i'j} / f_{i'.})^2$$

De la misma forma la distancia entre dos puntos - columna j y j´ estará dada por:

$$d^2 (j\, j')^2 = \sum_{i=1..n} 1/f_{i.} \, (f_{ij} / f_{.j} - f_{ij'} / f_{.j'})^2$$

En nuestro ejemplo la distancia entre dos puntos fila, tomemos Mendoza y Salta, estaría dada *por la suma a lo largo de países de*: la inversa de la frecuencia total de bolivianos multiplicada por la diferencia entre: el cociente de la frecuencia de bolivianos en Mendoza sobre el total de Mendoza y el cociente de la frecuencia de bolivianos en Salta sobre el total de Salta, *más* la inversa de la frecuencia total de brasileros multiplicada por la diferencia entre: el cociente de la frecuencia de brasileros en Mendoza sobre el total de Mendoza y el cociente de la frecuencia de brasileros en Salta sobre el total de Salta, *más* la inversa de la frecuencia total de chilenos multiplicada por la diferencia entre: el cociente de la frecuencia de chilenos en Mendoza sobre el total de Mendoza y el cociente de la frecuencia de chilenos en Salta sobre el total de Salta, *más* la inversa de la frecuencia total de paraguayos multiplicada por la diferencia entre: el cociente de la frecuencia de paraguayos en Mendoza sobre el total de Mendoza y el cociente de la frecuencia de paraguayos en Salta sobre el total de Salta, *más* la inversa de la frecuencia total de uruguayos multiplicada por la diferencia entre: el cociente de la frecuencia de uruguayos en Mendoza sobre el total de Mendoza y el cociente de la frecuencia de uruguayos en Salta sobre el total de Salta.

Esta distancia se parece bastante a la distancia euclídea usual (suma de cuadrados de las diferencias entre los componentes de los perfiles) si no fuera que se multiplica por la inversa de la frecuencia de cada término. El hecho de considerar los perfiles y no los valores absolutos hace que se llame también *distancia de Chi-cuadrado*. Veremos más adelante otro punto de contacto con este importante concepto de la estadística clásica.

La distancia euclídea usual entre dos puntos filas expresada en la tabla de contingencia de valores absolutos sería:

$$d^2 (i\, ,\, i') = \sum_{j=1..p} (f_{ij} - f_{i'j})^2$$

Esta distancia solo expresaría las diferencias entre frecuencias de dos provincias. Por el contrario la distancia euclídea usual entre dos perfiles de fila expresa la semejanza o diferencia entre dos provincias eliminando el efecto de los totales de provincias. Esta

distancia sería:

$$d^2 (i , i') = \sum_{j=1..p} (f_{ij} / f_{i.} - f_{i'j} / f_{i'.})^2$$

Sin embargo esta distancia favorece las columnas que tienen una masa $f_{.j}$ importante, es decir los países con frecuencias más altas, por lo tanto para atenuar este efecto, se la pondera con la inversa del total de columna.

Propiedades especiales del AFC

Esta distancia ponderada, así como el rol simétrico jugado por las filas y las columnas de la tabla de contingencia, hacen del análisis de correspondencia una técnica tan particular, asegurando estas propiedades especiales que no tiene el análisis en componentes principales: la equivalencia distribucional y las relaciones de transición.

Equivalencia distribucional

Una propiedad importante de esta distancia es la llamada *equivalencia distribucional* que implica una invariancia de las distancias entre filas cuando se agregan dos columnas que tengan perfiles idénticos y respectivamente entre columnas. Es decir, permite reunir dos modalidades de una misma variable que tengan perfiles idénticos, en una nueva modalidad consistente en la suma de sus masas, sin que nada cambie, ni las distancias entre las modalidades de una cierta variable, ni las distancias entre las modalidades de la otra variable. Pero sobre todo, no varían las distancias entre las columnas. Lo mismo sucede simétricamente con los perfiles columna.

Esta propiedad es fundamental porque garantiza una cierta invariancia de los resultados teniendo en cuenta diferencias en la nomenclatura elegida para la construcción de las modalidades de una variable, siempre que se cumpla la condición de reagrupar las modalidades con perfiles parecidos. Es decir, no se pierde información reuniendo ciertas categorías y no se gana subdividiendo en categorías homogéneas.

En nuestro ejemplo, si tuviéramos distribución homogénea de migrantes en el contexto de las provincias de cada región, sería equivalente realizar el análisis de correspondencias sobre provincias, que sobre regiones. Simétricamente, si tuviéramos los datos de migrantes por regiones de cada país, pero ellos fueran homogéneos con respecto a las provincias argentinas dentro de

cada país, resultaría equivalente realizar un análisis de correspondencias por países que por regiones de cada país.

El concepto de perfil es interesante porque permite comparar el recorrido de una categoría a lo largo de las categorías de la otra variable en la misma unidad de medida, relativizando los valores absolutos.

¿Por qué una representación simultánea? Relaciones de transición

Si bien se representan puntos filas y puntos columnas simultáneamente en el mismo gráfico, no se aconseja interpretar las cercanías entre un punto fila y un punto columna ya que los dos puntos no están en el mismo espacio de partida, lo que es lícito es interpretar la posición de un punto fila con respecto a la totalidad de los puntos del otro espacio.
La principal justificación de esta representación simultánea está dada por las *relaciones de transición* que ligan las coordenadas de un punto en un espacio con todos los puntos del otro espacio. Es decir, para poder llegar a representar los dos espacios simultáneamente tenemos que ponerlos en relación, o sea poder expresar el eje factorial de las filas en función del eje factorial de las columnas.
Si Φ_i designa la coordenada de un punto fila i sobre el eje horizontal del gráfico Nº 1 y Ψ_j designa la coordenada de un punto columna j sobre el mismo eje, entonces:

$$\Phi_i = \beta \sum_{j = 1..p} (f_{ij} / f_{i.}) \Psi_j \qquad (1)$$

$$\Psi_j = \beta \sum_{i=1...n} (f_{ij} / f_{j.}) \Phi_i \qquad (2)$$

¿Qué es β? Un coeficiente positivo, superior a 1 ya que es igual a: $1/\lambda$ siendo λ (valor propio del eje horizontal considerado) un valor comprendido entre 0 y 1.

Si no fuera por este coeficiente, los puntos fila serían exactamente los baricentros de los puntos columna, y recíprocamente. Esta situación ideal es, en general imposible de alcanzar, ya que

implicaría que el intervalo de eje recubierto por el primer conjunto estuviese comprendido en el intervalo de eje recubierto por el segundo, y recíprocamente.
Una representación doblemente baricéntrica no es posible, hace falta entonces este coeficiente 'dilatador' superior a 1. Sin embargo las relaciones de transición justifican la posibilidad de interpretar cada punto fila como el centro de gravedad de todos los puntos columna y cada punto columna como el centro de gravedad de todos los puntos fila.
Se puede presentar entonces al análisis de correspondencias como la búsqueda de los valores Φ_i y Ψ_j correspondientes al más pequeño coeficiente dilatador β.

Las fórmulas (1) y (2) se aplican también a los ejes factoriales subsiguientes (del 3 en adelante).

Otra propiedad de las coordenadas de los puntos fila y de los puntos columna sobre los ejes factoriales es que son centradas:

$$\sum_{i=1...n} f_i . \Phi_i = 0$$

$$\sum_{j=1...p} f_{.j} \psi_j = 0$$

Por lo tanto el origen de los ejes de coordenadas, donde ambos ejes se cruzan es igual a cero.

Cómo leer las salidas del programa

Presento a continuación la salida del programa (SPAD versión 4.5)[27] donde se consignan las coordenadas de las variables en los ejes factoriales 1 a 4, y las contribuciones absolutas y relativas correspondientes al ejemplo que venimos desarrollando.

Frecuencias activas o modalidades de la variable en columnas

Coordenadas

Nos ubicamos en el espacio factorial constituido por los 4 ejes factoriales, para observar el valor de las coordenadas en cada uno

[27] CISIA-CERESTA . Licencia N°00451-1046

de los ejes de ese espacio correspondientes a los puntos países de los cuales provienen los migrantes.

Coordonnées des fréquences actives

Libellé de la variable	Poids relatif	Distance à l'origine	Axe 1	Axe 2	Axe 3	Axe 4
Bolivianos	17.20	2.84929	-1.69	-0.04	0.03	-0.06
Brasileros	10.94	2.63940	0.34	0.47	-1.48	-0.34
Chilenos	24.39	1.44408	0.39	-1.13	0.05	-0.08
Paraguayos	23.46	0.72256	0.26	0.35	0.04	0.73
Uruguayos	24.00	1.06430	0.40	0.62	0.57	-0.44

En la primera columna se expresa la etiqueta de la variable (Libellé de la variable); en la segunda, el peso relativo (Poids relatifs) de cada país, es decir el porcentaje sobre el total correspondiente a cada país (ver totales de columna en el Cuadro N°2).
En la segunda columna la distancia al origen de coordenadas (Distance à l'origine) de cada punto país calculada según lo vimos precedentemente. Evidentemente el punto más alejado del origen es el correspondiente a bolivianos, el cual además tiene un valor negativo como coordenada en el primer eje factorial (columna Axe 1). Si recordamos el gráfico, observamos que bolivianos se encuentra todo hacia la izquierda, mientras que el resto de países tienen valores positivos.
Por otro lado en el eje 2 tenemos que bolivianos y chilenos tienen ambos un valor negativo, porque están por debajo del cero, sin embargo chilenos tiene un valor negativo mucho mayor (está mucho más abajo). De esta manera, observando los valores de las coordenadas se puede ir reconstruyendo el gráfico factorial. Puedo tomar estos valores y diseñar un gráfico factorial mediante otro software o graficador.

Contribuciones absolutas

Las contribuciones absolutas: describen la parte porcentual que le cabe a un elemento (fila o columna) en la construcción de cada eje factorial.
A continuación se presenta la tabla de las contribuciones de cada uno de los puntos columna en cada eje factorial. Las tres primeras columnas son idénticas a las del cuadro anterior. En las 4

columnas restantes se presentan las contribuciones de cada punto columna en cada eje factorial, la suma de los valores en cada eje será igual a 100.

Contributions des fréquences actives

Libellé de la variable	Poids relatif	Distance à l'origine	Axe 1	Axe 2	Axe 3	Axe 4
Bolivianos	17.20	2.84929	82.36	0.05	0.04	0.36
Brasileros	10.94	2.63940	2.09	5.18	75.08	6.70
Chilenos	24.39	1.44408	6.40	68.26	0.16	0.79
Paraguayos	23.46	0.72256	2.61	6.35	0.11	67.46
Uruguayos	24.00	1.06430	6.54	20.15	24.61	24.69

Es evidente que en el eje factorial 1 (Axe 1), el punto *bolivianos* está contribuyendo mucho más que el resto de los puntos. El resto de los países estaban del lado derecho, pero mucho más cerca del origen o sea que lo que está contribuyendo el punto *bolivianos* a la conformación del eje 1 es mayor que lo que están contribuyendo los otros puntos.
En el eje 2 el punto que más contribuye es *chilenos*, en efecto si observamos en el gráfico veremos que es el que más se distancia hacia abajo.

Contribuciones relativas

Asimismo tenemos la tabla de las contribuciones relativas o cosenos cuadrados, ellas miden en cuál de los ejes participa más ese punto fila o columna.

Cosinus carrés des fréquences actives

Libellé de la variable	Poids relatif	Distance à l'origine	Axe1	Axe 2	Axe 3	Axe 4
Bolivianos	17.20	2.84929	1.00	0.00	0.00	0.00
Brasileros	10.94	2.63940	0.04	0.08	0.83	0.04
Chilenos	24.39	1.44408	0.11	0.89	0.00	0.00
Paraguayos	23.46	0.72256	0.09	0.17	0.00	0.74
Uruguayos	24.00	1.06430	0.15	0.36	0.31	0.18

Si sumamos en sentido horizontal las contribuciones de cada país a lo largo de los ejes será igual a 1. Ello significa el aporte de cada eje factorial en la ubicación de cada punto país en el espacio

factorial. De hecho la ubicación del punto bolivianos está completamente dominada por el aporte del eje 1. Por ejemplo para el punto *brasileros* el eje que más está contribuyendo es el eje 3, con 0,83.
Si examinamos el eje 3 nos estaría mostrando otra conformación, otro criterio, por el cual los brasileros se estarían oponiendo a todo el resto. Si puedo encontrar una interpretación coherente a este criterio y me aporta algún otro elemento que agregue significación a mi problema, entonces convendrá analizar este eje en profundidad, de lo contrario no convendrá que arriesgue interpretaciones que en definitiva podrían estar basadas en bajos porcentajes de variación producto de distorsiones en la masa de datos.

Individuos o modalidades de la variable en filas

En el caso del AFC simple o binario, tiene sentido que hagamos el mismo análisis para el caso de las filas, que en nuestro ejemplo se refiere a las provincias.
En la tabla que sigue, se presentan los valores de las coordenadas, en este caso referidas a las provincias, para los 4 ejes factoriales.

Por ejemplo Buenos Aires tiene un peso relativo mucho mayor que el resto, 54,37, que era el perfil medio de Buenos Aires.
En este caso en el eje 1 las provincias que tienen valores negativos son: Catamarca, Córdoba, Jujuy, Mendoza y Salta, es decir las que se encontraban del lado negativo del eje 1, teniendo en cuenta que los valores mayores de las coordenadas correponden a Salta y Jujuy, siempre a la izquierda en el gráfico. Con valores positivos se encuentran todas las otras provincias.

Observar las coordenadas, en este caso en que hay más puntos ya se torna más complicado, por lo tanto se ve claramente la necesidad del gráfico, sin embargo hay ocasiones en que nos interesa revisar los valores exactos de las coordenadas, sobre todo cuando hay superposición de puntos.

INDIVIDUS			COORDONNEES			
IDENTIF.	P.REL	DISTO	1	2	3	4
B. AIRES	54.37	0.13	0.22	0.2	0.06	0.2
CATAMARCA	0.05	0.76	-0.45	-0.73	0.05	-0.16
CORDOBA	1.2	0.52	-0.2	0.2	-0.57	-0.33
CORRIENTES	1.07	1.03	0.39	0.61	-0.42	0.58
CHACO	2.39	2.74	0.33	0.51	0.05	1.54
E. RIOS	7.91	2.94	0.52	0.9	0.95	-0.98
CHUBUT	2.71	2.92	0.5	-1.62	0.08	-0.18
JUJUY	6.94	4.51	-2.12	-0.06	0.04	-0.12
LA PAMPA	0.23	0.67	0.47	-0.48	0.07	-0.45
LA RIOJA	0.05	1.29	0.25	-1.09	-0.08	-0.18
MENDOZA	2.69	0.9	-0.1	-0.84	-0.32	-0.26
MISIONES	3.63	7.39	0.42	0.66	-2.51	-0.71
NEUQUEN	1.85	2.98	0.5	-1.64	0.08	-0.18
RIONEGRO	3.78	2.93	0.5	-1.63	0.07	-0.18
SALTA	5	3.81	-1.95	-0.07	0.04	-0.05
SANJUAN	0.86	1.62	0.27	-1.21	-0.2	-0.23
SANLUIS	0.08	0.38	-0.25	-0.51	-0.08	-0.23
S. CRUZ	2.71	2.95	0.5	-1.63	0.09	-0.18
SANTA FE	1.5	0.83	0.31	0.45	-0.72	-0.08
S. ESTERO	0.07	0.24	-0.1	0.15	-0.41	-0.21
T. FUEGO	0.53	3.05	0.51	-1.66	0.08	-0.18
TUCUMAN	0.39	0.68	-0.77	-0.13	-0.26	-0.07

En la tabla siguiente se consignan los valores de las contribuciones absolutas y relativas (cosenos cuadrados) para cada eje factorial y cada punto individuo o provincia.

De las contribuciones absolutas en el eje 1, la mayor está dada por Jujuy, que era la que se encontraba más hacia la izquierda (coherentemente con el otro espacio: bolivianos) y luego por Buenos Aires.

La contribución relativa mayor para Jujuy está dada asimismo por el eje 1 de la misma manera que para la provincia de Salta.
De esta forma las contribuciones absolutas y relativas nos brindan otros elementos de ayuda a la interpretación.

INDIVIDUS			CONTRIBUTIONS				COSINUS CARRES			
IDENTIF.	P.REL	DISTO	1	2	3	4	1	2	3	4
B. AIRES	54.37	0.13	4.4	4.8	0.7	11.6	0.36	0.31	0.03	0.3
CATAMARCA	0.05	0.76	0	0.1	0	0	0.26	0.7	0	0.03
CORDOBA	1.2	0.52	0.1	0.1	1.2	0.7	0.08	0.08	0.63	0.21
CORRIENTES	1.07	1.03	0.3	0.9	0.6	1.9	0.15	0.36	0.17	0.32
CHACO	2.39	2.74	0.4	1.4	0	30.7	0.04	0.09	0	0.87
E. RIOS	7.91	2.94	3.5	14	22.4	41	0.09	0.28	0.31	0.33
CHUBUT	2.71	2.92	1.2	15.6	0.1	0.5	0.09	0.9	0	0.01
JUJUY	6.94	4.51	52.5	0.1	0	0.5	1	0	0	0
LA PAMPA	0.23	0.67	0.1	0.1	0	0.3	0.33	0.35	0.01	0.31
LA RIOJA	0.05	1.29	0	0.1	0	0	0.05	0.92	0	0.03
MENDOZA	2.69	0.9	0	4.2	0.9	1	0.01	0.8	0.12	0.07
MISIONES	3.63	7.39	1.1	3.5	71.2	9.8	0.02	0.06	0.85	0.07
NEUQUEN	1.85	2.98	0.8	10.9	0	0.3	0.08	0.9	0	0.01
RIONEGRO	3.78	2.93	1.6	21.8	0.1	0.7	0.09	0.9	0	0.01
SALTA	5	3.81	31.9	0.1	0	0.1	1	0	0	0
SANJUAN	0.86	1.62	0.1	2.7	0.1	0.3	0.04	0.9	0.02	0.03
SANLUIS	0.08	0.38	0	0	0	0	0.16	0.69	0.02	0.13
S. CRUZ	2.71	2.95	1.1	15.8	0.1	0.5	0.08	0.9	0	0.01
SANTA FE	1.5	0.83	0.3	0.7	2.5	0	0.12	0.24	0.63	0.01
S. ESTERO	0.07	0.24	0	0	0	0	0.04	0.09	0.68	0.19
T. FUEGO	0.53	3.05	0.2	3.2	0	0.1	0.09	0.9	0	0.01
TUCUMAN	0.39	0.68	0.4	0	0.1	0	0.87	0.02	0.1	0.01

Validez de la representación

Como ya vimos, podemos representar nuestra nube de puntos en un espacio de las filas y en un espacio de las columnas. En cada uno de los espacios vamos a calcular sucesivos ejes factoriales que mejor representen a la nube de puntos, es decir los de mayor inercia; el primer eje factorial va a ser el que corresponda al mayor valor propio, el segundo eje factorial, perpendicular (o sea estadísticamente independiente) va a ser el que corresponda al segundo mayor valor propio y así sucesivamente. Ese cálculo, el de extracción de vectores y valores propios, se denomina diagonalización de la matriz de datos.

Los ejes factoriales en el espacio de las filas van a ser

equivalentes a los del espacio de las columnas. Según las fórmulas de transición antes expuestas, no será necesario calcular un eje factorial para cada espacio.
El procedimiento de extracción de los valores propios es similar al del modelo general del análisis factorial con la diferencia que en este caso la matriz de datos es una tabla de contingencia.
En el cuadro que sigue se presentan los valores propios del análisis de correspondencias de nuestro ejemplo.

Tableau des valeurs propres

Trace de la matrice: 1.55615

Numéro	Valeur propre	Pourcentage	Pourcentage cumulé
1	0.5939	38.16	38.16
2	0.4574	29.40	67.56
3	0.3201	20.57	88.13
4	0.1847	11.87	100.00

Un valor propio cercano a la unidad asegura una buena representación baricéntrica en el eje correspondiente.
Los λ o valores propios, son valores comprendidos entre 0 y 1 que aquí valen 0.5939 para el primer eje y 0,4574 para el segundo acumulando 38,16 y 29,40 % de variación respectivamente, siendo la cuarta columna la del porcentaje acumulado.
La traza t (suma de todos los valores propios) vale aquí 1.55615, hay 4 valores propios no nulos.
Siendo que la traza representa la inercia total de la nube, los valores propios representan las inercias (o variancias) correspondientes a cada eje y los porcentajes de variancia miden la importancia relativa de cada valor propio en la traza.

En nuestro caso si me quedo en la interpretación de los dos primeros ejes factoriales que he representado en el gráfico N°1, estoy trabajando con un aceptable porcentaje de variación, el 67,56%, quiere decir que estoy perdiendo en realidad sólo el porcentaje restante de información (el 32,44%).
Estas medidas que por un lado son indicadores de la inercia total y por el otro de la inercia de los ejes y sus respectivas tasas de inercia, tienen su interés al momento de la interpretación.

Propiedades de la traza t

A diferencia del ACP (análisis de componentes principales), donde trabajamos con una distancia euclídea, en el AFC de una tabla de contingencia (n,p) el producto de la traza t por k (el total general) es igual al χ^2 de Pearson con (n-1) y (p-1) grados de libertad.

La fórmula clásica sería:

$$\chi^2 = k \sum_{ij} (f_{ij} - f_{i.} f_{.j})^2 / f_{i.} f_{.j}$$

En nuestro caso:

k t = 606542 x 1.5562 = 943900,6604 valor test calculado para 84 grados de libertad = (22-1) (5-1)

No necesitamos la tabla de valores del Chi-cuadrado para rechazar la hipótesis de independencia ya que nuestro valor test calculado es a simple vista suficientemente alto. Ello implica que tenemos elementos para justificar una asociación (o más bien rechazar la independencia) entre las variables nacionalidad de los migrantes y provincias argentinas a las que llegan.

Direcciones privilegiadas

La distribución de la inercia total de la nube en los distintos valores propios nos provee un elemento para evaluar la validez de la representación. De todas formas podría darse el caso de una evaluación pesimista a partir de los valores propios y sin embargo que la estructura de los datos representada en los dos primeros ejes fuera de interés para nuestra investigación porque pusiera de relieve por ejemplo, aspectos inesperados o no identificables mediante otros métodos.

El valor de la inercia total no tiene siempre una interpretación interesante. En análisis en componentes principales normado y en análisis de correspondencias múltiples la inercia total depende únicamente del número de variables.

En síntesis, el valor de la inercia es un indicador de la dispersión de la nube y mide la asociación entre las dos variables. Sin embargo, no nos interesamos solamente por la dispersión de la nube sino sobre todo por la existencia de *direcciones privilegiadas en esta nube.*

Consultamos las inercias de cada eje (valores propios) así como las tasas de inercia correspondientes. Este examen nos enseña

acerca de la forma de la nube: forma “esférica” (sin dirección privilegiada) o forma “no esférica” (con dirección/es privilegiada/s). En el diagrama que sigue se presentan configuraciones típicas de las nubes de puntos proyectadas en los dos primeros ejes factoriales, considerando la presencia de una inercia fuerte o débil y la existencia o no de direcciones privilegiadas.

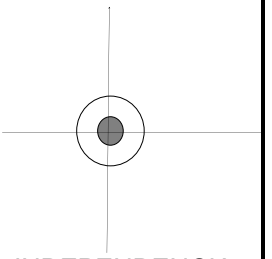
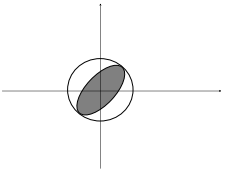
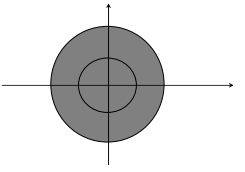
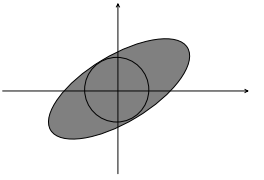
<table>
<tr><td colspan="2"></td><td colspan="2">Direcciones
Tasas de inercia de los ejes</td></tr>
<tr><td colspan="2">Nube</td><td>Forma “esférica”</td><td>Forma “no esférica”</td></tr>
<tr><td rowspan="2">Inercia</td><td>Inercia débil</td><td> 
1.- INDEPENDENCIA
Débil inercia total
No hay dirección privilegiada</td><td> 
2.- DEPENDENCIA
Débil inercia total
Dirección privilegiada</td></tr>
<tr><td>Inercia fuerte</td><td> 
3.- DEPENDENCIA
Fuerte inercia total
No hay dirección privilegiada</td><td> 
4.- DEPENDENCIA
Fuerte inercia total
Dirección privilegiada</td></tr>
</table>

Diagrama tomado de Lebart , Morineau, Piron (1995)

En las situaciones 2 y 4, las nubes tienen tasas de inercia idénticas pero una inercia total diferente. Por otro lado, las situaciones 3 y 4 revelan dos nubes de la misma inercia total y tasas de inercia diferentes.

El test de Chi cuadrado permite detectar estas dos últimas situaciones, pero no permite poner en evidencia la situación 2
En síntesis, la inercia de un factor mide la asociación que el mismo pone en evidencia y ella no puede ser superior a 1. Un valor propio que tienda a 1 indica una dicotomía al nivel de los datos, obtenemos para cada variable dos grupos de modalidades que separan la nube de puntos en dos sub-nubes. Ello puede significar igualmente la existencia de un grupo de puntos aislados de los otros puntos.
Cuando dos valores propios son próximos a 1, obtenemos 3 subnubes y las modalidades de las variables se descomponen en 3 grupos. Si todos los valores propios son cercanos a 1, cada modalidad de una variable tiene una correspondencia casi exclusiva con una sola modalidad de la otra variable.
Sin embargo, el hecho de obtener valores propios débiles (significando que los perfiles son cercanos al perfil medio) no debería impedir una interpretación de los ejes de inercia asociados. Ello podría revelar una estructura interesante y más difícilmente perceptible.

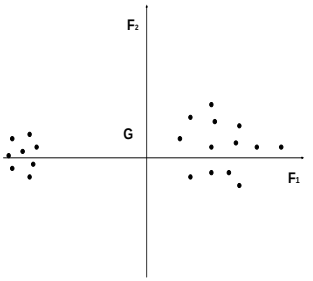
Ejemplos de nubes características

Cuando se presentan determinadas configuraciones en las nubes de puntos proyectadas, pueden realizarse permutaciones de filas y columnas en la tabla de datos de manera de reorganizar los datos y poder así interpretar mejor los gráficos.

Nube de puntos dividida en dos subnubes

En el diagrama siguiente, se muestra el aspecto de la nube en los 2 primeros ejes factoriales y la forma de la matriz de datos.
Para este caso, en que la nube de puntos se encuentra agrupada en dos nubes bien definidas, la matriz de datos puede ser reorganizada ordenando las coordenadas de las filas y columnas del primer factor.

Puede llegar a ser interesante analizar separadamente las dos sub nubes definidas por las dos tablas correspondientes (I_1 , J_1) y (I_2 , J_2).



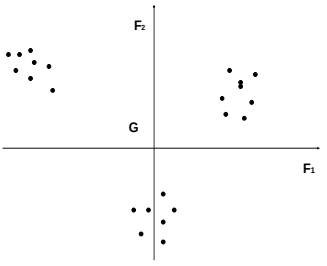
	J_1	J_2
I_1		0
I_2	0	

Diagrama tomado de Lebart , Morineau, Piron (1995)

La nube se descompone en tres subnubes de puntos

Cuando la nube de puntos se descompone en tres subnubes, adoptando una forma de 'triángulo', ello significa que la matriz de datos tiene la forma que ilustra el diagrama que sigue.

Puede reorganizarse la matriz de datos por permutación de las filas y de las columnas. Las tres subnubes pueden igualmente ser objeto de análisis por separado

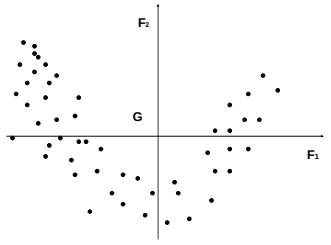


	J_1	J_2	J_3
I_1		0	0
I_2	0		0
I_3	0	0	

Diagrama tomado de Lebart , Morineau, Piron (1995)

Efecto Guttman:

Cuando la nube de puntos presenta una forma de parábola se dice que posee el efecto Guttman. La matriz correspondiente presenta una diagonal principal relativamente cargada.



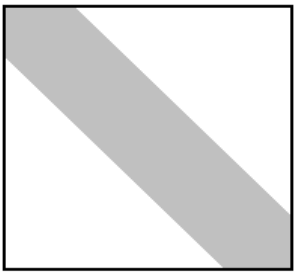


Diagrama tomado de Lebart , Morineau, Piron (1995)

Esta situación traduce una redundancia de las dos variables, ya que del conocimiento de la fila *i* podemos deducir la columna *j* Toda la información está casi exclusivamente contenida en el primer factor. La información provista por los ejes de mayor rango expresa el mismo fenómeno, sin embargo el examen del segundo factor afina la interpretación del primero.

En este caso especial el segundo factor es una función de segundo grado del primer factor, el tercer factor es una función de tercer grado, etc.

Generalmente el efecto Guttman aparece cuando las categorías de las variables presentan un orden jerárquico (variables continuas transformadas en variables nominales). Un eje (a menudo el primero) opone los valores de los extremos y el segundo eje opone los valores extremos contra los valores medios. El efecto Guttman pone a veces en evidencia una estructura trivial que valdrá la pena observar si la forma parabólica no es perfecta. Los puntos de ruptura serán entonces muy interesantes para analizar.

Variables activas e ilustrativas o suplementarias

La posibilidad de introducir variables ilustrativas o suplementarias

es una característica distintiva del análisis multidimensional de datos en el entorno SPAD.
Las variables activas son las que se incluyen en los cálculos de diagonalización de la matriz mientras que las ilustrativas se proyectan a posteriori en el espacio factorial construido con las variables activas. Esto permite, cuando las variables son numerosas, la elección de un determinado punto de vista definido por el investigador para comparar sus unidades de análisis, sin eliminar el resto de variables que servirán para ilustrar o quizás para explicar las similitudes encontradas entre los individuos.

Una vez hallados los ejes factoriales correspondientes a un valor propio, pueden ser aplicadas las fórmulas de transición a filas o columnas suplementarias.
Una fila suplementaria o ilustrativa se calculará:

$$\Phi_{i+} = \beta \sum_{j = 1..p} (f_{i+j} / f_{i+..}) \Psi_j$$

Una columna suplementaria:

$$\Psi_{j+} = \beta \sum_{i=1...n} (f_{ij+} / f_{j+.}) \Phi_i$$

Podemos entonces ilustrar los planos factoriales con informaciones suplementarias que no participaron en la construcción de los ejes, lo que va a tener consecuencias muy importantes a nivel de la interpretación de los resultados.
Las variables activas deben formar un conjunto homogéneo para que las distancias entre individuos u observaciones se interpreten fácilmente.
Los elementos ilustrativos que intervienen a posteriori no tienen necesidad de formar un conjunto homogéneo, esta temática podrá ser mejor apreciada en el capítulo dedicado al Análisis de Correspondencias Múltiples.

CAPÍTULO 7

EL ANÁLISIS DE CORRESPONDENCIAS MÚLTIPLES

El Análisis de Correspondencias Binario o Simple puede generalizarse de distintas maneras para el caso en que se buscan correspondencias entre más de dos variables.

Una de las más simples y la más utilizada es el Análisis de Correspondencias Múltiples (ACM), que permite describir grandes matrices de datos, entre las cuales el ejemplo más característico es el de las encuestas; donde las filas son los individuos, que pueden ser miles y las columnas las modalidades o categorías de variables nominales, en el caso más frecuente: distintas respuestas a preguntas de un cuestionario.

Se trata de una extensión del dominio de aplicación del análisis de correspondencias con procedimientos de cálculo y reglas de interpretación específicas.

En este capítulo expondré de qué manera, mediante la transformación de matrices, se puede presentar el ACM.

"Se le pueden reconocer los principios de este método a Guttman (1941), pero también a Burt (1950) o a Hayashi (1956). Otro tipos de extensión fueron propuestos por Benzécri (1973), Escofier-Cordier (1965) y por Masson (1974) que se apoyó en los trabajos de Carroll (1968), Horst (1961) y Kettenring (1971).

El ACM ha sido desarrollado igualmente bajo el nombre de *Homogeneity Analysis* por el equipo de J.de Leeuw desde 1973 y bajo el nombre de *Dual Scaling* por Nishisato (1980). Una aplicación del análisis de correspondencias a una tabla disyuntiva completa se encuentra en Nakache (1973) y la presentación del conjunto de resultados y demostraciones se hizo en Lebart y Tabard (1973), pero el nombre de *Análisis de Correspondencias Múltiples* figura por primera vez en Lebart (1975:73-96). Además una exposición sintética de estos diversos enfoques fue realizada por Tenenhaus y Young (1985)"(Lebart et al. 1995:108).

La extensión mencionada se basa sobre la propiedad siguiente: supongamos que se han suministrado a n individuos dos preguntas, cada una con p_1 y p_2 respuestas, y se supone que para cada pregunta, tanto las p_1 como las p_2 respuestas son mutuamente excluyentes es decir existe una sola respuesta posible.

Es entonces equivalente realizar el análisis de correspondencias de la tabla de contingencia p_1 x p_2, cruzando las dos preguntas,

que analizar la tabla disyuntiva completa Z de n filas y $p_1 + p_2$ columnas que describen las respuestas (cada fila de Z tiene 2 unos en las columnas correspondientes a las 2 respuestas elegidas y n + p - 2 ceros en las no elegidas).
El análisis de Z es más dificultoso pero más interesante porque se generaliza rápidamente al caso de más de dos preguntas.

Equivalencia de matrices y de análisis

Si retomamos la matriz de datos (tabla de contingencia) utilizada para presentar el Análisis de Correspondencias Binario:

CUADRO N° 1
MIGRANTES DE PAÍSES LIMÍTROFES

Provincias	Bolivianos	Brasileros	Chilenos	Paraguayos	Uruguayos	Total
Buenos Aires	27544	33019	59912	117476	91856	329807
Catamarca	100	9	141	14	21	285
Córdoba	1814	2234	1014	756	1453	7271
Corrientes	14	1588	73	3638	1153	6466
Chaco	80	261	149	13570	433	14493
Entre Ríos	54	613	134	515	46672	47988
Chubut	54	42	16080	98	174	16448
Jujuy	40978	48	413	618	37	42094
La Pampa	15	155	714	45	440	1369
La Rioja	28	23	217	14	17	299
Mendoza	3622	2546	9476	259	416	16319
Misiones	116	21106	149	618	37	22026
Neuquén	34	36	11040	36	61	11207
Río Negro	86	136	22453	83	169	22927
Salta	27505	99	1036	1579	86	30305
San Juan	450	559	4037	66	91	5203
San Luis	134	49	209	30	64	486
Santa Cruz	96	21	16122	54	128	16421
Santa Fe	414	3387	800	2469	2013	9083
S.del Estero	88	105	72	64	89	418
T. del Fuego	0	3	3223	9	7	3242
Tucumán	1101	344	471	306	163	2385
Total	104327	66383	147935	142317	145580	606542

De la matriz anterior, que cruza en filas las provincias argentinas y en columnas las nacionalidades de los inmigrantes, podremos pasar a la matriz disyuntiva completa siguiente que tendrá 606542 filas (tantas como el total de los migrantes) y 27 columnas (22 provincias + 5 nacionalidades).

22 provincias			5 nacionalidades					
BsAs		Tucumán	Bolivia	Brasil	Chile	Paraguay	Uruguay	
1	0	...	1	0	0	0	0	
1	0	0	1	0	0	0	0	
...	.	...	...	...	...	...	...	27544 filas iguales
	.							
1	0	0	0	1	0	0	0	
1	0	0	0	1	0	0	0	
...	.	...	...	...	...	...	...	33019 filas iguales
	.							
...	.	...	...	...	...	...	...	...
	.							
0	0	1	0	0	0	0	1	
0	0	1	0	0	0	0	1	163 filas iguales
...	.	...	...	...	...	...	...	
	.							
...	.	...	...	...	...	...	...	606542 filas en
	.							
...	.	...	...	...	...	...	...	total
	.							
	.							

El Análisis de Correspondencias de la tabla de contingencia que cruza las variables provincia y nacionalidad es equivalente al Análisis de Correspondencias de la matriz disyuntiva completa anterior, donde se puede observar que es sencillo agregar otra variable nominal, bastará simplemente agregar otro bloque de columnas.

Esta segunda matriz es más asimilable al concepto metodológico de matriz de datos ya que sus filas serían verdaderamente individuos pero cada columna correspondería a un valor (categoría) de una variable nominal. Por lo tanto lo que expresa cada celda es la presencia (con 1) o ausencia (con 0) de ese valor en ese determinado individuo.

La verdadera matriz de datos de individuos por variables se alcanza con la matriz R como lo veremos a continuación.

Ejemplo de construcción de una matriz disyuntiva completa

Como ejemplo, en las tablas que siguen se presenta el caso de una matriz R con tres preguntas (variables) que tienen respectivamente 3, 2 y 4 modalidades de respuesta y 12 individuos, de esta matriz se puede pasar a la matriz disyuntiva completa Z, con las mismas dimensiones. Los procedimientos de cálculo utilizan la tabla R que es menos engorrosa, sin embargo es útil mostrar los tres siguientes tipos de tablas a los efectos de generalizar el Análisis de Correspondencias Simple al caso múltiple.

s = 3

	V1	V2	V3
	2	2	4
	2	1	3
	3	1	2
	1	2	4
	1	2	3
R =	2	2	3
(n,s)	3	1	1
	1	1	1
	2	1	2
	2	2	3
	3	2	2
	1	1	4

$\Rightarrow$

p = 9

	V1			V2		V3			
Z=	M1	M2	M3	M1	M2	M1	M2	M3	M4
(n,p)	0	1	0	0	1	0	0	0	1
	0	1	0	1	0	0	0	1	0
	0	0	1	1	0	0	1	0	0
	1	0	0	0	1	0	0	0	1
	1	0	0	0	1	0	0	1	0
	0	1	0	0	1	0	0	1	0
	0	0	1	1	0	1	0	0	0
	1	0	0	1	0	1	0	0	0
	0	1	0	1	0	0	1	0	0
	0	1	0	0	1	0	0	1	0
	0	0	1	0	1	0	1	0	0
	1	0	0	1	0	0	0	0	1

De estas tablas se puede pasar a una matriz de Burt que cruza todas las modalidades, en cada celda se encuentra la frecuencia de individuos que poseen la modalidad fila y columna:

			V1			V2		V3			
			M1	M2	M3	M1	M2	M1	M2	M3	M4
		M1	4	0	0	2	2	1	0	1	2
	V1	M2	0	5	0	2	3	0	1	3	1
		M3	0	0	3	2	1	1	2	0	0
B=	V2	M1	2	2	2	6	0	2	2	1	1
(p,p)		M2	2	3	1	0	6	0	1	3	2
		M1	1	0	1	2	0	2	0	0	0
	V3	M2	0	1	2	2	1	0	3	0	0
		M3	1	3	0	1	3	0	0	4	0
		M4	2	1	0	1	2	0	0	0	3

La tabla Z constará de tres bloques para las tres variables. En

cambio la tabla B es el producto de la tabla disyuntiva Z por su transpuesta: B = Z' Z
Esta última tabla es simétrica y la llamamos tabla de Burt o tabla de correspondencia múltiple, pero sigue siendo en esencia una tabla de contingencia múltiple.
Contiene en este caso nueve bloques: los bloques diagonales son matrices diagonales donde los elementos diagonales son las frecuencias totales de respuestas correspondientes a cada modalidad. Los tres bloques distintos entre los bloques restantes (que son las transpuestas) son las tres tablas de contingencia cruzando las tres preguntas dos a dos.
El análisis de correspondencias de la tabla Z da los mismos ejes factoriales normados que los de la tabla B, es decir finalmente los mismos gráficos de proximidades, de escalas aproximadas. Los valores propios homólogos, en cambio, son diferentes; pero esto no complica su utilización.

En el caso de dos preguntas o variables, la tabla R tiene sólo dos columnas, la tabla Z está formada de dos bloques, y B incluye cuatro.
Para el caso de sólo dos variables el valor propio de la tabla de contingencia (para un eje dado k) es igual a λ_k; de la misma tabla expresada en forma disyuntiva completa será: $\mu_z = (1+\sqrt{\lambda_k})/ 2$, mientras que el valor propio de la misma tabla expresada como tabla de Burt será: $\mu_B = \mu_z^2$
Como los valores propios son diferentes en los tres análisis, las tasas de inercia también lo serán. De manera general, el análisis de tablas disyuntivas y de tablas de Burt conduce siempre a tasas de inercia débiles. Con respecto a las que se obtienen con Análisis de Componentes Principales y Análisis de Correspondencias Binario, estas tasas dan una idea mucho más pesimista de la parte de información extraída.

Validez de la representación

Como consecuencia de lo anterior, es conveniente evitar la utilización de los porcentajes de variancia para caracterizar los ejes; en el caso de correspondencias múltiples no tienen el mismo sentido que cuando se trata de una tabla de contingencia: el código binario introduce una perturbación que reduce la parte de explicación adjudicada a cada valor propio.

En la práctica la estabilidad del plano factorial puede probarse mediante técnicas de simulación (perturbaciones aleatorias de la matriz de datos) o de muestra-test (proyección de individuos que no participaron en la construcción de los ejes).

Reglas de interpretación

El ACM como extensión del Análisis de Correspondencias Simple, conserva el concepto de distancia de Chi cuadrado como asimismo la propiedad de equivalencia distribucional.
Decir que existe afinidad entre respuestas es equivalente a decir que existen individuos que eligieron simultáneamente todas o casi todas las mismas respuestas.
El ACM pone en evidencia los tipos de individuos que tienen perfiles parecidos en cuanto a los indicadores que se eligieron para describirlos. Teniendo en cuenta las distancias entre los elementos de la tabla disyuntiva completa y las relaciones baricéntricas particulares se expresan:

- *la proximidad entre individuos en términos de semejanzas:* dos individuos se parecen si han elegido globalmente las mismas modalidades.
- *la proximidad entre modalidades de variables diferentes en términos de asociación:* estas modalidades corresponden a los puntos medios de los individuos que las eligieron y se encuentran próximas porque están referidas a los mismos individuos o a individuos parecidos.
- *la proximidad entre dos modalidades de una misma variable en términos de semejanza:* por construcción, las modalidades de una misma variable se excluyen. Si dos de ellas se encontraran cercanas, esta proximidad se interpretaría en términos de semejanza entre los grupos de individuos que las eligieron (en cuanto a las otras variables activas del análisis).

Las reglas de interpretación de los resultados (coordenadas, contribuciones, cosenos cuadrados) que se refieren a los elementos activos de un ACM son las mismas que las de un análisis de correspondencias simple. Se calcula la contribución y la calidad de la representación de cada modalidad y de cada individuo (en el caso de que estos últimos fueran identificables para el análisis).
Para realizar una evaluación de la importancia de cada variable a través de sus modalidades, se calcula la contribución de una

variable en el factor α sumando las contribuciones de sus modalidades sobre el mismo factor. Se obtienen así, además de las modalidades más contribuyentes en la construcción de los ejes factoriales, las variables que más participaron en la definición de los mismos.

Principios de recodificación

Para poder considerar como activas en ACM a las variables medidas en escala continua, ellas deben recodificarse como variables nominales.

Pero cuando se trata de agrupar una variable en intervalos o categorías nos encontramos con varios problemas. ¿Cuántas clases elegir y cómo? ¿Dónde colocar los límites de las clases de una variable continua? La observación de la distribución de cada variable e histogramas es indispensable para efectuar estas elecciones.

Algunos principios, deducidos de las propiedades del ACM, pueden ser utilizados para guiar la fase de recodificación: diseñar las modalidades con frecuencias semejantes, categorizar las variables de manera de tener un número comparable de modalidades. Para dar una idea de magnitud, una categorización en 4 a 8 modalidades puede andar bien para la mayor parte de las aplicaciones.

Se trata entonces de encontrar un compromiso entre una categorización técnicamente aceptable según estos principios y una categorización que posea al menos la información que nos interese retener. No conviene en general hacer uso de algoritmos demasiado complicados para elaborar una categorización satisfactoria. Conservaremos por ejemplo, una modalidad de baja frecuencia si ella es importante y tiene coherencia con los objetivos del estudio. Asimismo para seleccionar los límites de las clases de una variable continua, respetaremos uno o más umbrales naturales en el contexto del estudio, o que resulten significativos luego de la lectura del histograma (la categorización en clases de igual amplitud es a veces inadecuada).

Estos principios no son tan rigurosos para las variables ilustrativas, sin embargo aunque ellas no intervengan en la formación de los ejes, a veces puede ser necesario realizar una categorización fina para lograr una mejor interpretación.

La transformación de las variables continuas en variables nominales ocasiona una pérdida de información bruta, pero

presenta ciertas ventajas: explorar simultáneamente variables nominales y continuas en correspondencias múltiples, validar a posteriori los datos permitiendo observar la eventual contigüidad de las clases vecinas, y poner en evidencia las eventuales relaciones no lineales entre variables continuas.

Relaciones casi-baricéntricas o fórmulas de transición

La coordenada factorial del individuo i (activo o ilustrativo) sobre el eje α está dada por:

$\psi_{\alpha i} = 1/(s\sqrt{\lambda_\alpha}) \sum_{j \in p(i)...p} \varphi_{\alpha j}$ donde s es el número de variables activas

Salvo por un coeficiente igual a $1/\sqrt{\lambda_\alpha}$ el individuo i se encuentra en el punto medio de la nube de las modalidades que él ha elegido.

De la misma manera, la coordenada de la modalidad j sobre el eje α está dada por:

$$\varphi_{\alpha j} = 1/(z_j \sqrt{\lambda_\alpha}) \sum_{i \in I(j)...n} \psi_{\alpha i} = \beta \sum_{i \in I(j)...n} \psi_{\alpha i} \quad (1)$$

donde z_j es el número de individuos que eligieron la modalidad j

Elementos ilustrativos o suplementarios

La representación simultánea de las líneas y de las columnas ligadas al análisis de correspondencias no se utiliza en el caso del Análisis de Correspondencias Múltiples. No se marcan los puntos-fila porque, además de ser en general muy numerosos, los individuos son anónimos y no tiene interés su posición individual.

En cambio se proyectan las otras informaciones sobre los individuos como variables ilustrativas de interés.

La fórmula de transición precedente (1) nos muestra que la coordenada φ_j de una columna j (respuesta ilustrativa) se obtiene multiplicando por $\beta = 1/(z_j \sqrt{\lambda_\alpha})$ la media aritmética de las coordenadas de los individuos que han elegido la respuesta j.

De esta manera se proyecta toda categoría ilustrativa en el espacio formado por los ejes construidos con las variables activas.

La utilización de elementos suplementarios o ilustrativos en ACM permite tomar en cuenta toda la información susceptible de ayudar a comprender o a interpretar la tipología inducida por los

elementos activos.
Esto es particularmente interesante cuando el conjunto de las variables se descompone en temas, es decir en grupos de variables homogéneas en cuanto a su contenido.
En sentido metodológico podríamos equiparar cada uno de estos grupos de variables homogéneas con una dimensión de una variable teórica.

En el análisis de la tabla disyuntiva completa, se harán intervenir los elementos suplementarios para:

- Enriquecer la interpretación de los ejes con variables que no hayan participado en su construcción. Se podrán proyectar entonces en el espacio de las variables, los centros de los grupos de individuos definidos por la modalidades de las variables ilustrativas
- Adoptar un enfoque predictivo proyectando las variables ilustrativas en el espacio de los individuos. Estas serán "explicadas" por las variables activas.
- Situar los individuos ilustrativos en el espacio de las variables, para ubicarlos en relación con los individuos activos o en relación a los grupos de individuos activos en un enfoque de análisis discriminante.

Según la naturaleza de las variables ilustrativas, nominales o continuas, se interpreta de manera diferente su posición sobre los ejes factoriales.

Procesamiento de una encuesta mediante ACM: el caso de un sondeo de opinión[28]

La experiencia se refiere a un sondeo de opinión en una localidad pequeña de la Provincia de Santa Fe, con las siguientes características:

- Un municipio pequeño para atender problemas pequeños y medianos.
- La relación Intendente-Político y Elector-Ciudadano es de

[28] Moscoloni N, Costa R (2002) *El Análisis Multidimensional de Datos como Herramienta de Procesamiento en los Sondeos de Opinión.* Trabajo realizado por la cátedra de Análisis de Datos de la Escuela de Comunicación Social de la Universidad Nacional de Rosario

vecindad y cercanía.

- Contar con muestra grande y representativa para disminuir el riesgo de ocultar la opinión de algún sector de la población.
- La encuesta como instrumento de análisis y diseño de políticas de gobierno en el gabinete local, como un aspecto más a tener en cuenta en la toma de decisiones.
- La onda que generan los impactos de las políticas nacionales e internacionales a veces no tiene un efecto inmediato o no llegan con la virulencia de lo macroeconómico.
- La crisis de los valores ideológico-políticos de los grandes partidos políticos argentinos en temas nacionales y provinciales recuperan otras formas, más acotadas y concretas en un municipio.
- El refugio local frente a la crisis de lo nacional.

Los indicadores

Bajo estas premisas construimos un instrumento de recolección de datos, el cual con algunas variantes referidas a las distintas cuestiones que interesaban en los distintos momentos se repitió en tres oportunidades durante el año 2000.
El cuestionario constaba de preguntas referidas a la ubicación contextual del encuestado y de otras relativas a su opinión acerca del desempeño de los distintos órganos del gobierno municipal y de los servicios que el mismo presta a la comunidad.
El ejemplo aquí citado se refiere a una de esas encuestas en la que se incluyó una pregunta a respuesta abierta. Se listan a continuación los indicadores considerados correspondientes a este instrumento. La reiteración aparente de algunas preguntas tuvo el objetivo de metodológico de observar si con diferente formulación cambiaba el perfil de respuestas.
El procesamiento de la respuesta abierta será descrito en el capítulo referido a análisis textual.

N° de variable	Etiqueta	Tipo de variable	Modali-dades
1	Sexo	N	2
2	Edad	C	
3	Punto muestra	N	14
4	Educación	N	10
5	Act. Económica	N	16
6	¿Sostén familiar?	N	4

7	Vivienda	N	4
8	Características vivienda	N	6
9	¿Qué imagen tiene del Gobierno Municipal?	N	4
10	¿Qué imagen tiene del Intendente?	N	4
11	¿Qué imagen tiene del Concejo Municipal?	N	4
12	Calificación Concejo Municipal	N	4
13	Calificación Departamento Ejecutivo	N	4
14	Recolección de residuos	N	7
15	Alumbrado	N	7
16	Riego	N	7
17	Desmalezamiento	N	7
18	Recolección de Ramas	N	7
19	Zanjeo	N	7
20	Barrido y limpieza	N	7
21	Mantenimiento calles de tierra	N	7
22	Atención en Dispensarios	N	7
23	Pregunta abierta	T	

La matriz de datos original está entonces constituida por los individuos encuestados en fila y por las variables (o indicadores) en columna.

Para realizar el ACM nuestro primer problema es, luego de la observación descriptiva de rutina a los efectos de validar los datos, la elección de las variables activas e ilustrativas. En nuestro caso nos interesó conocer la estructura de los indicadores que apuntaban a la imagen del desempeño en general de la labor de las autoridades municipales, por lo tanto elegimos como variables activas de la 9 a la 22, todas ellas nominales como corresponde en un ACM. Ello implica que la matriz de Burt debiera ser en este caso de 83 filas por 83 columnas, es decir tantas como modalidades tenemos de las variables activas.

Quedarán como variables ilustrativas el resto de las variables nominales y la única variable continua, que en este caso es la edad. Las variables ilustrativas darán cuenta en este caso de los indicadores contextuales.

Salidas del programa (SPAD, v.4.5[29] etapa CORMU)

Presento a continuación las salidas de programa más importantes de este análisis de correspondencias múltiples, con las observaciones a tener en cuenta.

[29] CISIA-CERESTA . Licencia N°00451-1046

La primera se refiere a las distribuciones de frecuencias de las variables activas *(Tris à plat des variables actives)*, que siempre son importantes de controlar. El programa realiza una depuración automática de las modalidades con una frecuencia inferior al 2% del total de casos, en este caso el total de casos era de 295, por lo tanto las modalidades con una frecuencia menor a 6 son "ventiladas" *(Ventilée)*, es decir pasadas a modalidades ilustrativas. De esta manera se evita que modalidades con escasa frecuencia produzcan distorsiones en el análisis.
La salida del programa presenta las frecuencias antes de la depuración (*Effectif avant apurement*), su ponderación *(Poids)* o peso (para el caso en que el mismo no sea uniforme e igual a 1 como en este caso) y el efectivo y el peso de cada modalidad después de la depuración (*après apurement)*:

Tris à plat des variables actives (Seuil: 2.0%)

¿Qué imagen tiene del Gobierno Municipal?

Libellé des modalités	Effectif avant apurement	Poids avant apurement	Effectif après apurement	Poids après apurement
Gob.Mun.mala	57	57.00	57	57.00
Gob.Mun.regular	117	117.00	117	117.00
Gob.Mun.buena	106	106.00	106	106.00
Gob.Mun.no sabe	15	15.00	15	15.00

¿Qué imagen tiene del Intendente?

Libellé des modalités	Effectif avant apurement	Poids avant apurement	Effectif après apurement	Poids après apurement
Intendente mala	47	47.00	47	47.00
Intendente regular	61	61.00	61	61.00
Intendente buena	165	165.00	165	165.00
Intendente no sabe	22	22.00	22	22.00

¿Qué imagen tiene del Concejo Municipal?

Libellé des modalités	Effectif avant apurement	Poids avant apurement	Effectif après apurement	Poids après apurement
Cmun.mala	67	67.00	67	67.00
Cmun.regular	99	99.00	99	99.00
Cmun.buena	47	47.00	47	47.00
Cmun.no sabe	82	82.00	82	82.00

Calificación Concejo Municipal

Libellé des modalités	Effectif avant apurement	Poids avant apurement	Effectif après apurement	Poids après apurement
Cmun.mala	61	61.00	61	61.00
Cmun.regular	103	103.00	103	103.00
Cmun.buena	50	50.00	50	50.00
Cmun.no sabe	81	81.00	81	81.00

Calificación Departamento Ejecutivo

Libellé des modalités	Effectif avant apurement	Poids avant apurement	Effectif après apurement	Poids après apurement
Dep.Ej.mala	50	50.00	50	50.00
Dep-Ej. Regular	94	94.00	94	94.00
Dep.Ej. buena	64	64.00	64	64.00
Dep.Ej. no sabe	87	87.00	87	87.00

Recolección de residuos

Libellé des modalités	Effectif avant apurement	Poids avant apurement	Effectif après apurement	Poids après apurement
RS muy mala	5	5.00	Ventilée	
RS mala	13	13.00	15	15.00
RS regular	31	31.00	32	32.00
RS buena	227	227.00	228	228.00
RS muy buena	19	19.00	20	20.00

Alumbrado

Libellé des modalités	Effectif avant apurement	Poids avant apurement	Effectif après apurement	Poids après apurement
AL muy mala	20	20.00	20	20.00
AL mala	58	58.00	58	58.00
AL regular	86	86.00	86	86.00
AL buena	121	121.00	122	122.00
AL muy buena	9	9.00	9	9.00
AL no sabe	1	1.00	Ventilée	

Riego

Libellé des modalités	Effectif avant apurement	Poids avant apurement	Effectif après apurement	Poids après apurement
RG muy mala	10	10.00	10	10.00
RG mala	28	28.00	28	28.00
RG regular	74	74.00	74	74.00
RG buena	96	96.00	96	96.00
RG muy buena	2	2.00	Ventilée	
RG no sabe	24	24.00	24	24.00
RS no corresponde	61	61.00	63	63.00

Desmalezamiento

Libellé des modalités	Effectif avant apurement	Poids avant apurement	Effectif après apurement	Poids après apurement
DZ muy mala	20	20.00	20	20.00
DZ mala	55	55.00	55	55.00
DZ regular	58	58.00	58	58.00
DZ buena	69	69.00	70	70.00
DZ muy buena	1	1.00	Ventilée	
DZ no sabe	16	16.00	16	16.00
DZ no corresponde	76	76.00	76	76.00

Recolección de Ramas

Libellé des modalités	Effectif avant apurement	Poids avant apurement	Effectif après apurement	Poids après apurement
RA muy mala	12	12.00	12	12.00
RA mala	50	50.00	50	50.00
RA regular	66	66.00	66	66.00
RA buena	138	138.00	138	138.00
RA muy buena	9	9.00	9	9.00
RA no sabe	11	11.00	11	11.00
RA no corresponde	9	9.00	9	9.00

Zanjeo

Libellé des modalités	Effectif avant apurement	Poids avant apurement	Effectif après apurement	Poids après apurement
ZJ muy mala	33	33.00	33	33.00
ZJ mala	53	53.00	53	53.00
ZJ regular	51	51.00	51	51.00
ZJ buena	33	33.00	33	33.00
ZJ muy buena	0	0.00		
ZJ no sabe	21	21.00	21	21.00
ZJ no corresponde	104	104.00	104	104.00

Barrido y limpieza

Libellé des modalités	Effectif avant apurement	Poids avant apurement	Effectif après apurement	Poids après apurement
BA muy mala	24	24.00	24	24.00
BA mala	29	29.00	29	29.00
BA regular	73	73.00	73	73.00
BA buena	98	98.00	98	98.00
BA muy buena	6	6.00	6	6.00
BA no sabe	10	10.00	10	10.00
BA no corresponde	55	55.00	55	55.00

Mantenimiento calles de tierra

Libellé des modalités	Effectif avant apurement	Poids avant apurement	Effectif après apurement	Poids après apurement
TI muy mala	57	57.00	57	57.00
TI mala	58	58.00	58	58.00
TI regular	74	74.00	74	74.00
TI buena	29	29.00	29	29.00
TI muy buena	0	0.00		
TI no sabe	15	15.00	15	15.00
TI no corresponde	62	62.00	62	62.00

Atención en Dispensarios

Libellé des modalités	Effectif avant apurement	Poids avant apurement	Effectif après apurement	Poids après apurement
DI muy mala	2	2.00	Ventilée	
DI mala	3	3.00	Ventilée	
DI regular	27	27.00	28	28.00
DI buena	149	149.00	151	151.00
DI muy buena	92	92.00	92	92.00
DI no sabe	22	22.00	24	24.00
DI muy mala	0	0.00		

Los valores propios

Otra salida del programa es el cuadro de los valores propios, que como vimos, representa la inercia contenida en la información de las variables activas, pero que en el ACM en realidad presenta una visión pesimista de esta variabilidad acumulada por los primeros ejes. En efecto, si debiéramos tener en cuenta el monto del primer valor propio para evaluar la calidad de la representación, en este caso concluiríamos que éste es muy bajo (8.58%). “La utilización de las tasas de inercia (o porcentajes de variancia) como herramienta de evaluación global de la calidad de una representación es muy delicada. (…) La variancia bruta inicial no es en general una medida de referencia adecuada, resulta a menudo injustificado hablar de *parte de la información* en relación con las *tasas de inercia*” (Lebart et al, 1995:368)

Tableau des valeurs propres

Trace de la matrice: 4.07143

Numéro	Valeur propre	Pourcentage	Pourcentage cumulé
1	0.3495	8.58	8.58
2	0.2477	6.08	14.67
3	0.2210	5.43	20.09
4	0.2000	4.91	25.01
5	0.1703	4.18	29.19
…	…	…	…
56	0.0128	0.31	99.75
57	0.0102	0.25	100.00

Coordenadas factoriales

Las coordenadas de las modalidades activas que se presentan en la tabla siguiente nos permitirán ya tener una idea de la ubicación de las mismas en el primer plano factorial. Esta salida se utiliza más que nada como control, en el caso en que necesitemos ubicar o controlar la ubicación de algún punto modalidad en el gráfico. Por razones de espacio incluiré sólo los ejes 1 y 2, aunque el programa edita hasta los 5 primeros ejes factoriales. En esta tabla se consignan además la frecuencia relativa de cada modalidad, y el valor de su distancia al origen de coordenadas.

Coordonnées des modalités actives

Libellé	Poids relatif	Distance à l'origine	Axe 1	Axe 2
¿Qué imagen tiene del Gobierno Municipal?				
Gob.Mun.mala	1	4.17544	1.22	-0.36
Gob.Mun.regular	3	1.52137	0.17	0.04
Gob.Mun.buena	3	1.78302	-0.77	-0.01
Gob.Mun.no sabe	0	18.66670	-0.53	1.13
¿Qué imagen tiene del Intendente?				
Intendente mala	1	5.27660	1.21	-0.27
Intendente regular	1	3.83607	0.24	-0.10
Intendente buena	4	0.78788	-0.41	0.03
Intendente no sabe	1	12.40910	-0.15	0.65

¿Qué imagen tiene del Concejo Municipal?

CMun.mala	2	3.40299	0.93	-0.55
CMun.regular	2	1.97980	0.12	0.01
CMun.buena	1	5.27660	-0.79	0.00
CMun.no sabe	2	2.59756	-0.45	0.44

Calificación del Concejo Municipal

CMun.mala	1	3.83607	0.98	-0.65
CMun.regular	2	1.86408	0.18	0.02
CMun.buena	1	4.90000	-0.79	0.09
CMun.no sabe	2	2.64198	-0.48	0.41

Calificación del Departamento Ejecutivo

Dep.Ej. mala	1	4.90000	1.23	-0.52
Dep-Ej.regular	2	2.13830	0.26	-0.04
Dep.Ej. buena	2	3.60938	-0.78	-0.12
Dep.Ej. no sabe	2	2.39080	-0.41	0.42

Recolección de residuos

RS mala	0	18.66670	1.37	-0.63
RS regular	1	8.21875	0.82	0.04
RS buena	6	0.29386	-0.13	0.18
RS muy buena	0	13.75000	-0.83	-1.66

Alumbrado

AL muy mala	0	13.75000	0.59	-0.27
AL mala	1	4.08621	0.69	0.00
AL regular	2	2.43023	0.11	0.16
AL buena	3	1.41803	-0.43	0.11
AL muy buena	0	31.77780	-0.93	-2.44

Riego

RG muy mala	0	28.50000	0.77	-0.36
RG mala	1	9.53571	1.28	0.06
RG regular	2	2.98649	0.31	0.27
RG buena	2	2.07292	-0.35	0.41
RG no sabe	1	11.29170	-0.05	0.83
RS no corresponde	2	3.68254	-0.49	-1.23

Desmalezamiento

DZ muy mala	0	13.75000	1.00	0.19
DZ mala	1	4.36364	0.81	0.22
DZ regular	1	4.08621	0.10	0.31
DZ buena	2	3.21429	-0.64	0.23
DZ no sabe	0	17.43750	-0.19	1.40
DZ no corresponde	2	2.88158	-0.30	-0.96

Recolección de Ramas

RA muy mala	0	23.58330	1.33	-0.18
RA mala	1	4.90000	0.80	-0.08
RA regular	2	3.46970	0.37	0.16
RA buena	3	1.13768	-0.46	0.12
RA muy buena	0	31.77780	-1.44	-2.75
RA no sabe	0	25.81820	0.05	0.97
RA no corresponde	0	31.77780	-0.40	-0.72

Zanjeo

ZJ muy mala	1	7.93939	1.13	0.21
ZJ mala	1	4.56604	0.69	0.25
ZJ regular	1	4.78431	-0.20	0.56
ZJ buena	1	7.93939	-0.85	0.45
ZJ no sabe	1	13.04760	-0.13	1.20
ZJ no corresponde	3	1.83654	-0.32	-0.85

Barrido y limpieza

BA muy mala	1	11.29170	1.19	-0.41
BA mala	1	9.17241	0.69	0.02
BA regular	2	3.04110	0.17	-0.12
BA buena	2	2.01020	-0.67	-0.04
BA muy buena	0	48.16670	-1.54	-3.07
BA no sabe	0	28.50000	0.16	1.19
BA no corresponde	1	4.36364	0.23	0.52

Mantenimiento calles de tierra

TI muy mala	1	4.17544	0.79	0.15
TI mala	1	4.08621	0.74	0.20
TI regular	2	2.98649	-0.27	0.50
TI buena	1	9.17241	-1.11	-0.02
TI no sabe	0	18.66670	-0.38	1.32
TI no corresponde	2	3.75806	-0.49	-1.24

Atención en Dispensarios

DI regular	1	9.53571	1.22	-0.36
DI buena	4	0.95364	0.12	0.29
DI muy buena	2	2.20652	-0.52	-0.50
DI no sabe	1	11.29170	-0.21	0.50

Contribuciones absolutas

En la tabla siguiente se consignan las contribuciones absolutas de cada variable en los ejes factoriales 1 y 2. Puede ser interesante observar la contribución acumulada total de cada variable en cada eje factorial. Ello refleja la importancia de la misma en la formación de cada uno de ellos y es un indicador del poder discriminatorio de cada variable.

En este caso la variable con mayor contribución en el eje 1 es la opinión sobre el Gobierno Municipal y la opinión sobre el Departamento Ejecutivo. Cuando tenemos una contribución total mayor, nos interesa conocer cuáles son las modalidades de esta variable que más aportan a esta contribución. En el caso de estas variables se trata de las modalidades buena y mala. Esto significa que estas dos categorías están muy alejadas ambas (una hacia la izquierda y otra hacia la derecha) del centro de gravedad que es el origen de coordenadas.

Las variables más discriminatorias en el eje 2 son Mantenimiento de calles de tierra, Riego y Zanjeo en los tres casos debido a la alta contribución de la modalidad No corresponde, que refiere a los encuestados que no necesitan de esos servicios municipales.

Las contribuciones acumuladas de las variables activas son mayores cuanto mayor es el número de modalidades de las variables, por lo tanto es un dato a tener en cuenta en esta evaluación.

En nuestro caso la variable 'Imagen del Gobierno Municipal' no tiene mayor número de modalidades que el resto, por lo tanto es lícito considerarla como de peso en la construcción del eje 1.

Contributions des modalités actives

Libellé	Poids relatif	Distance à l'origine	Axe 1	Axe 2
¿Qué imagen tiene del Gob. Municipal?				
Gob.Mun.mala	1	4.17544	5.87	0.71
Gob.Mun.regular	3	1.52137	0.23	0.02
Gob.Mun.buena	3	1.78302	4.30	0.00
Gob.Mun.no sabe	0	18.66670	0.29	1.88
		Contribution cumulée	10.7	2.6
¿Qué imagen tiene del Intendente?				
Intendente mala	1	5.27660	4.77	0.33
Intendente regular	1	3.83607	0.24	0.06
Intendente buena	4	0.78788	1.95	0.01
Intendente no sabe	1	12.40910	0.03	0.92
		Contribution cumulée	7.0	1.3
¿Qué imagen tiene del Con.Municipal?				
CMun.mala	2	3.40299	4.02	1.96
CMun.regular	2	1.97980	0.09	0.00
CMun.buena	1	5.27660	2.05	0.00
CMun.no sabe	2	2.59756	1.13	1.54
		Contribution cumulée	7.30	3.50
Calific.Con.Municipal				
CMun.mala	1	3.83607	4.09	2.52
CMun.regular	2	1.86408	0.23	0.00
CMun.buena	1	4.90000	2.15	0.04
CMun.no sabe	2	2.64198	1.31	1.34
		Contribution cumulée	7.78	3.91
Calific.Dep.Ejecutivo				
Dep.Ej.mala	1	4.90000	5.21	1.30
Dep-Ej.regular	2	2.13830	0.45	0.01
Dep.Ej. buena	2	3.60938	2.73	0.09
Dep.Ej. no sabe	2	2.39080	1.02	1.53
		Contribution cumulée	9.42	2.92

Recolección de residuos

RS mala	0	18.66670	1.95	0.58
RS regular	1	8.21875	1.49	0.01
RS buena	6	0.29386	0.28	0.73
RS muy buena	0	13.75000	0.95	5.42
	Contribution cumulée		4.67	6.74

Alumbrado

AL muy mala	0	13.75000	0.49	0.15
AL mala	1	4.08621	1.93	0.00
AL regular	2	2.43023	0.07	0.21
AL buena	3	1.41803	1.59	0.15
AL muy buena	0	31.77780	0.54	5.24
	Contribution cumulée		4.61	5.75

Riego

RG muy mala	0	28.50000	0.41	0.13
RG mala	1	9.53571	3.17	0.01
RG regular	2	2.98649	0.48	0.51
RG buena	2	2.07292	0.84	1.61
RG no sabe	1	11.29170	0.00	1.62
RS no corresponde	2	3.68254	1.05	9.37
	Contribution cumulée		5.95	13.25

Desmalezamiento

DZ muy mala	0	13.75000	1.38	0.07
DZ mala	1	4.36364	2.53	0.26
DZ regular	1	4.08621	0.04	0.54
DZ buena	2	3.21429	1.97	0.37
DZ no sabe	0	17.43750	0.04	3.08
DZ no corresponde	2	2.88158	0.47	6.79
	Contribution cumulée		6.43	11.11

Recolección de Ramas

RA muy mala	0	23.58330	1.47	0.04
RA mala	1	4.90000	2.20	0.03
RA regular	2	3.46970	0.61	0.16
RA buena	3	1.13768	2.06	0.19
RA muy buena	0	31.77780	1.29	6.63
RA no sabe	0	25.81820	0.00	1.01
RA no corresponde	0	31.77780	0.10	0.46
	Contribution cumulée		7.74	8.52

Zanjeo

ZJ muy mala	1	7.93939	2.92	0.15
ZJ mala	1	4.56604	1.77	0.33
ZJ regular	1	4.78431	0.15	1.58
ZJ buena	1	7.93939	1.67	0.64
ZJ no sabe	1	13.04760	0.02	2.94
ZJ no corresponde	3	1.83654	0.72	7.43
		Contribution cumulée	7.24	13.07

Barrido y limpieza

BA muy mala	1	11.29170	2.36	0.40
BA mala	1	9.17241	0.96	0.00
BA regular	2	3.04110	0.14	0.10
BA buena	2	2.01020	3.04	0.01
BA muy buena	0	48.16670	0.98	5.52
BA no sabe	0	28.50000	0.02	1.40
BA no corresponde	1	4.36364	0.19	1.43
		Contribution cumulée	7.70	8.85

Mantenimiento calles de tierra

TI muy mala	1	4.17544	2.46	0.13
TI mala	1	4.08621	2.18	0.23
TI regular	2	2.98649	0.36	1.83
TI buena	1	9.17241	2.49	0.00
TI no sabe	0	18.66670	0.15	2.55
TI no corresponde	2	3.75806	1.01	9.27
		Contribution cumulée	8.66	14.01

Atenc.Dispensarios

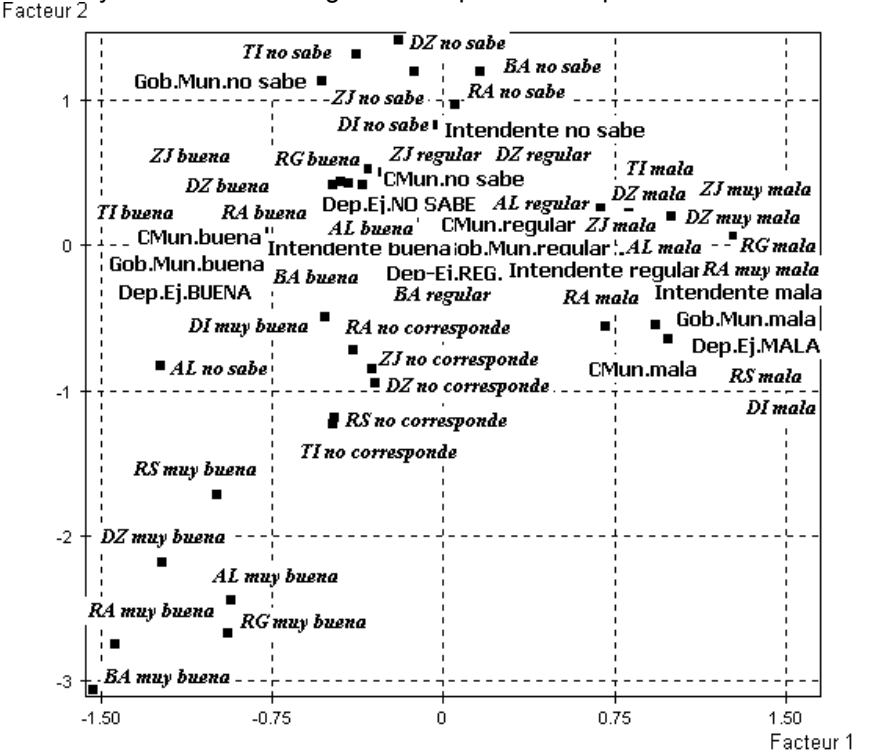
DI regular	1	9.53571	2.89	0.36
DI buena	4	0.95364	0.16	1.25
DI muy buena	2	2.20652	1.71	2.24
DI no sabe	1	11.29170	0.08	0.59
		Contribution cumulée	4.83	4.44

Gráficos factoriales

La salida más importante para el usuario es como ya dijimos el gráfico donde se proyectan las categorías de las variables nominales. En este caso presento el gráfico de las variables activas, correspondientes como vimos, a las categorías de

opinión.
Cada categoría se ubica según su coordenada en el eje 1 (horizontal) y en el eje 2 (vertical). De la misma manera que en el análisis de correspondencias binario, no corresponde interpretar la entidad del valor exacto de la coordenada, sino la posición relativa de cada categoría en relación con las demás estableciendo una síntesis a nivel espacial para la cual confiamos especialmente en nuestra propia capacidad visual, en caso de duda siempre podremos recurrir al listado de valores de las coordenadas que nos dará la ubicación exacta.

Proyección de las categorías de opinión en el plano factorial 1*2



En este gráfico se eliminaron algunos puntos para lograr una mayor legibilidad.
Las preguntas que se refieren a opiniones sobre la gestión del Gobierno, del Concejo Municipal y del Departamento Ejecutivo se extienden en un rango de 4 valores con extremos ‘mala’ – ‘buena’. Las opiniones acerca de los servicios prestados por la Municipalidad poseen 5 valores con extremos ‘muy mala’ – ‘muy

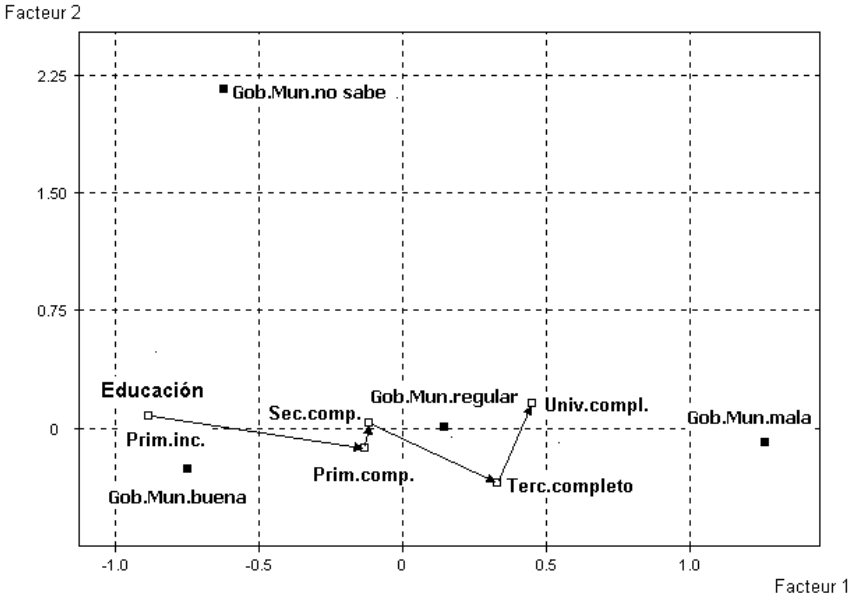
buena'.
El eje factorial 1, que acumula la máxima inercia, opone las categorías 'muy mala' en el extremo derecho, contra las 'muy buena' en el extremo izquierdo, pasando por las intermedias en un continuo ordenado, lo que responde a una correcta validación de los datos. Es de observar que las categorías 'no sabe' se encuentran al medio, lo que implica una frecuencia importante y además están cercanas a las categorías 'regular', indicativas de una situación de indiferencia respecto de la opinión de la gestión del gobierno municipal. En este caso esta categoría resulta importante pues podría estar respondiendo a una situación generalizada de descreimiento en la clase política.
Las categorías 'no corresponde' se refieren a respuestas de encuestados sobre las preguntas de opinión acerca de: desmalezamiento (DZ), recolección de ramas y hojas (RA), mantenimiento de calles de tierra (TI), zanjeo (ZJ). Estos encuestados viven en zonas urbanas y por lo tanto desconocen la efectividad del gobierno municipal con respecto a esos servicios (Ver el gráfico sobre tipos de vivienda). Por otra parte se encuentran cercanas a algunas categorías de gestión 'buena' o 'muy buena', ello significa que estos encuestados respondieron con una mejor imagen general acerca de la gestión municipal.
El eje 2 (vertical) opone las categorías correspondientes a valores medios, más frecuentes, contra las categorías correspondientes a valores extremos, menos frecuentes (especialmente las referidas a opiniones 'muy buena')
Así la categoría 'RA muy mala' se proyecta en el extremo derecho del eje 1 pero en el centro del eje 2, con lo cual tendrá un poder discriminante alto en el primer eje y bajo en el segundo.

La proyección de las variables nominales ilustrativas, referidas al contexto socioeconómico del encuestado sobre el plano factorial resultó en un agrupamiento de las mismas hacia el centro del gráfico.

A los efectos de observar la estructura de las variables ilustrativas, en los gráficos que siguen presento en cada uno, una variable contextual acompañada de la variable activa opinión sobre el Gobierno Municipal. De esta manera se evita la superposición de puntos que impide la visualización de los mismos.

En el gráfico que sigue he destacado la trayectoria de los puntos de la escolaridad de menor a mayor.
De la observación de este gráfico puede interpretarse que las categorías más altas de escolaridad se asocian con una opinión mala sobre el Gobierno Municipal, ya que las categorías universitarias se encuentran a la derecha, cerca del punto *Gob. Mun. mala*.

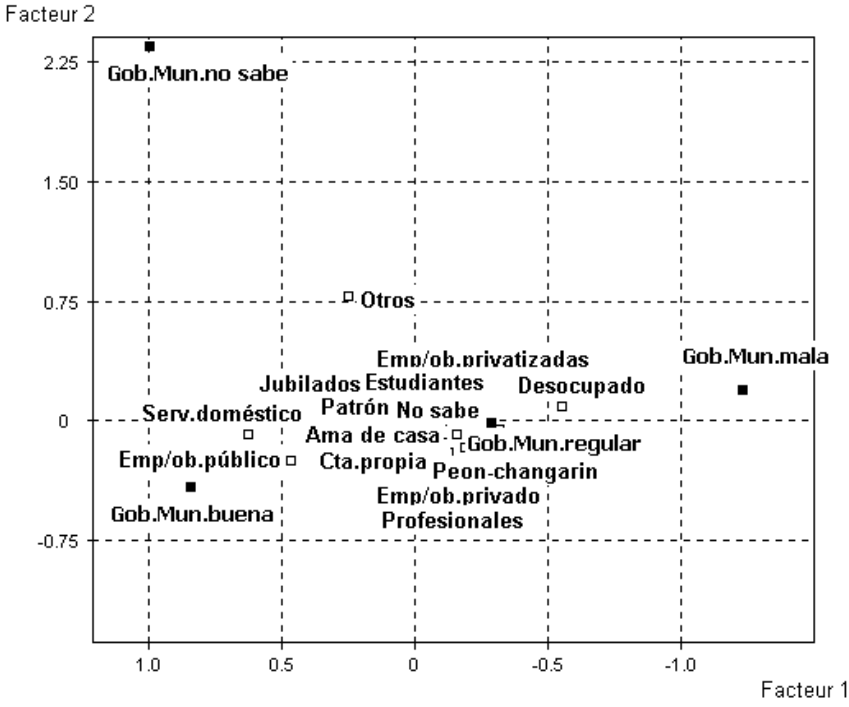
Proyección de las categorías de educación en el plano factorial 1*2



En el gráfico siguiente he proyectado las categorías ocupacionales.
Al igual que en el gráfico anterior, se observan algunas categorías cercanas a los puntos de opinión. Por ejemplo la categoría *Desocupado* cerca de *Gob. Mun. Mala* hacia la derecha y *Jubilados* y *Emp/ob. Público* cerca de *Gob. Mun. Buena* hacia la izquierda. Esto puede tener una interpretación lógica y banal (Los empleados u obreros públicos pueden tener una relación con el Gobierno Municipal y en consecuencia tener buena opinión, mientras que los desocupados serán naturalmente críticos por el peso de su problemática económica)

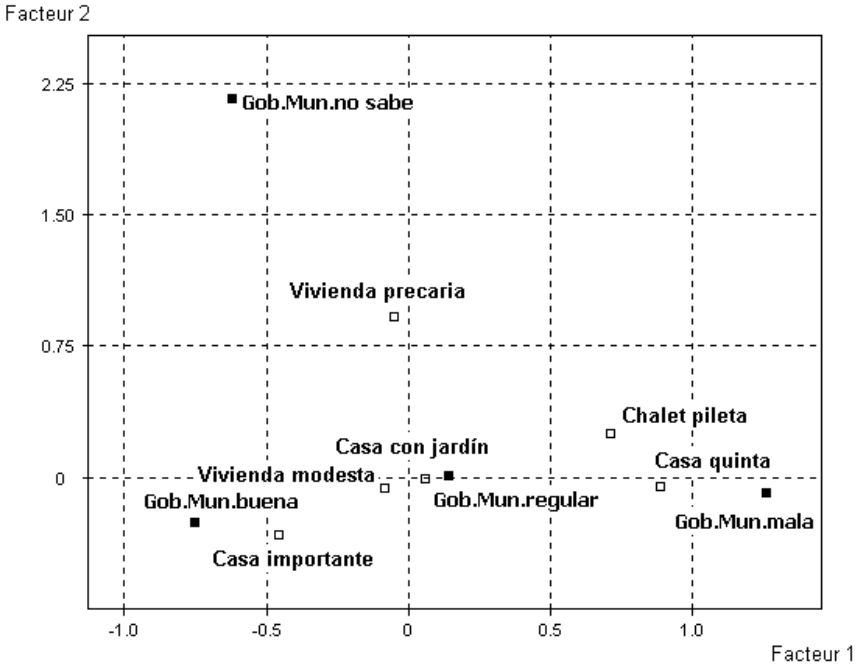
Sin embargo el resto de las categorías ocupacionales no presentan una asociación visible y por lo demás se encuentran muy agrupadas cerca del centro de gravedad, sobre todo teniendo en cuenta que la distancia al mismo de las categorías de opinión es sensiblemente mayor.

Proyección de las categorías ocupacionales en el plano factorial 1*2



En el próximo gráfico las categorías consideradas se refieren al tipo de vivienda. Se observa que las viviendas del tipo fin de semana se encuentran más cercanas a la categoría de mala opinión mientras las ubicadas en la parte urbana se encuentran más cercanas al centro de gravedad.

Proyección de las categorías de vivienda en el plano factorial 1*2



La posición de las variables ilustrativas en los 2 primeros ejes factoriales no aporta suficiente información como para concluir asociaciones significativas, al menos al nivel de los indicadores elegidos, entre ellas y las variables de opinión.
Esta falta de asociación se verá apoyada con los resultados de la clasificación realizada a continuación del ACM y que expondré en el capítulo dedicado a clasificación.
Todo el análisis del ejemplo presentado se concentró sólo en los dos primeros ejes factoriales. Se recomienda incluir el análisis de ejes de más alto nivel sólo cuando el aporte en interpretación sea significativo, caso que no era aplicable en nuestro ejemplo.

CAPÍTULO 8

CLASIFICACIÓN. SU CONCEPTO. CONSTRUCCIÓN DE CLASES DE INDIVIDUOS. DESCRIPCIÓN ESTADÍSTICA DE LAS CLASES. COMPLEMENTARIEDAD CON EL ANÁLISIS FACTORIAL

De los individuos a los objetos

Como ya vimos en capítulos anteriores, uno de los aportes fundamentales del AMD es otorgarle mayor importancia a los individuos. La estadística clásica se interesa más en las variables, ya sea buscando regularidades o leyes de probabilidad y los individuos en sí mismos no son considerados más que en su totalidad.

En el caso de los métodos factoriales los individuos aparecen proyectados en el espacio generado por los ejes factoriales y es analizada su estructura al mismo tiempo que la de las variables. Complementariamente esos mismos individuos son sometidos a un proceso de clasificación.

En la perspectiva de Diday (1997) acerca de la clasificación, el énfasis está puesto más que en los individuos, en los objetos mismos. De esta manera se sube un nivel en generalización. Por ejemplo, cuando se hace una encuesta en marketing, los individuos que son encuestados pueden ser llamados de una forma más general objetos, o en un estudio geográfico donde se observa el comportamiento demográfico o económico de las regiones, éstas pueden ser consideradas también como objetos.

En lenguaje metodológico sería otra manera de denominar a las unidades de análisis, sin embargo el concepto incluye además otra dimensión como veremos a continuación.

Los objetos como clases

El punto fundamental es profundizar cuál es el sentido de esos individuos. Es en este momento cuando interviene la noción de objeto. Cuando se nombra *una silla*, que es un nombre común, se designa a un objeto y cuando se expresa *esta silla*, que es una

silla individual, también se designa a un objeto.
En el primer caso, el concepto de objeto refiere a la idea de clase, cuando se dice 'silla' se está nombrando un sustantivo común que representa el conjunto de todas las sillas individuales. El procedimiento mental que se realiza es el de asignar un objeto observado a una clase, símil a lo que puede ser una inferencia abductiva (dados una regla y un rasgo asignar el caso a la regla)
En el segundo caso en cambio, cuando se dice esta silla se está designando a un individuo determinado y diferente de otro de su misma clase.
Si se profundiza en esta noción de objeto, descripta de esta manera, se puede llegar muy atrás en la historia de la humanidad.
En este sentido entonces, el saber clasificar reside en lo que podemos reconocer como los conocimientos adquiridos por tenacidad, según la terminología de Peirce[30], refiriéndose a los "métodos para obtener conocimientos". Cada uno de estos métodos correspondería a estratos diferentes en la formación del conocimiento humano. El método de la tenacidad "pertenece al trasfondo biótico que existe en el cuerpo orgánico de todo ser humano y que compartimos con los restantes vivientes por el solo hecho de integrar y *actuar* en el seno de la biosfera". En el caso del procedimiento de clasificación es el que nos permite distinguir por ejemplo, en el estadio más primitivo, los alimentos que podemos comer de los que no podemos. Este procedimiento será retomado por todos los sucesivos métodos de fijar creencias que involucran al hombre como ser social.

Se dice entonces que la noción de clase es muy antigua. Cuando el hombre prehistórico dibujó el mamut en su caverna o en su gruta, él representó más bien a una clase, la clase de los mamuts, y no por ejemplo al mamut que se acababa de comer.
Es decir que ya 30.000 años antes de Jesucristo la gente pensaba en términos de clase. En el Génesis de la Biblia, 2000 AC, Dios le dice a Adán que le dé un nombre a los animales de la tierra y a los pájaros del cielo. Esto también de asignar un nombre a los animales, es ya reconocer una noción de clase.
La reflexión sobre la noción de clase se encuentra en los libros de los primeros autores griegos.

[30] Peirce Ch.S. (1988) *El hombre un signo.* Crítica. Barcelona, pp.185 y ss. Citado por: Samaja J. (2000: 151-180)

Sin considerar a los presocráticos, Aristóteles en el siglo IV AC desarrolla ampliamente su preocupación por las clasificaciones.
Nos ilustra Porfirio (1993:5) en su introducción al libro de las Categorías: “§ 39. Por tanto el individuo aparece envuelto por la especie; la especie por el género. El género es un todo, el individuo una parte. La especie es a la vez todo y parte; pertenece como parte a otro que no es ella; y como todo no pertenece a otro, sino que está en otros, porque el todo está en las partes”.
En el libro "De Partibus Animalis" especialmente se interesa por las especies de animales, por los objetos considerados como especies o sea como clases, analizando los diferentes métodos que debieran seguirse para el establecimiento de criterios y normas, descartando por ejemplo, la división dicotómica para llegar a alguna de las “formas finales de animales”. “El orden de la exposición debe ser éste: asentar primeramente los atributos comunes a grupos enteros de animales, intentando luego su explicación” (Aristóteles, 1945:34)

La clase y su descripción

Es necesario distinguir la noción de "clase" de la noción de "descripción de la clase". Aristóteles cuando habla de ciencia de los objetos ya muestra la diferencia entre un objeto y su descripción.
Cuando se habla de clase se piensa en un conjunto en el sentido matemático, un grupo de objetos, pero hay una segunda noción de clase que no es solamente la noción matemática de conjunto, sino que es la descripción de la clase.
Por ejemplo la silla que está allí o la figura de un individuo que está representado en una revista se pueden describir por los valores que toman unas variables, o se puede describir un conjunto de mesas, de sillas, empresas que son productivas o que no lo son. Estas serán clases de empresas que forman un conjunto de objetos pero que también están descriptas por sus propiedades. Estas propiedades o variables podrán ser presentadas de distintas maneras.
Las dos nociones diferentes de clase remiten a dos conceptos fundamentales: las clases definidas en intensión y las definidas en extensión.

Intensión y extensión de una clase

La idea de intensión y extensión se halla muy claramente expresada en los autores Arnault y Nicole de la Escuela de los Lógicos de Port Royal en el siglo XVIII[31]. Ellos afirman que la intensión de una idea son los atributos o las variables que la describen y que no pueden ser suprimidos sin destruirla; la extensión de una idea es el conjunto de los individuos u objetos a los cuales estas propiedades se aplican.

Por ejemplo, si se piensa en la noción de camión esta idea no se encuentra en el espíritu. Hay un modo de descripción de la noción de camión que nos permite reconocerlo tal que cuando se ve pasar a uno de ellos, se dice esto es un camión (y no un auto).

La clase de los camiones, de todos los camiones, se representa en mi espíritu, no como un conjunto de camiones sino como un conjunto de propiedades que describen a los camiones. No es por lo tanto la clase como el conjunto en el sentido matemático sino la descripción de la clase en función de sus propiedades.

No podríamos funcionar si tuviéramos todos los objetos del mundo en la cabeza, no es la teoría de los conjuntos lo que tenemos en la cabeza sino más bien la teoría de las intensiones. No son las extensiones lo que nos permite funcionar sino las intensiones. Esto es la base del Análisis de Datos Simbólico (en adelante el ADS). El ADS va a consistir en trabajar no sobre las extensiones es decir sobre los individuos, sino en reemplazar los individuos por las intensiones.

Representación de una clase. Clases monotética y politética

El primer texto que da noticia sobre este tema es de Jevons de 1877[32]. Luego Beckner[33] en 1959, define exactamente la diferencia entre clase monotética y clase politética.

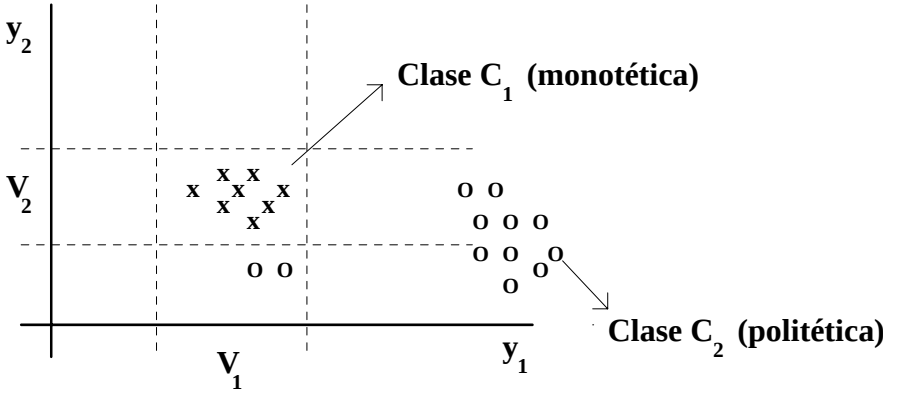
Una clase es monotética si existe un conjunto de propiedades

[31] Arnault A y *Nicole*, P. (1662), *La logique ou l'art de penser,* reeditado *por* Froman, Stuttgart (1965). Citado por Diday (1997)

[32] Jevons, W.S. (1877), *The principles of Science: A treatise on Logic and Scientific Method*, 2da. ed. rev. Macmillan London and New York, 786 p.Citado idem

[33] Beckner (1959) The Biological Way of thought, Columbia University Press, New York, 220 p. Citado idem

necesarias y suficientes satisfechas por los individuos de la clase. Una clase es politética si no hay una condición necesaria y suficiente que haga que un individuo sea miembro de una clase.
Por ejemplo en el gráfico que sigue se representan puntos caracterizados por dos variables y_1 e y_2 . Esos puntos pertenecen a dos clases: la de las cruces C_1, y la de los círculos C_2.
Se puede observar que todos los individuos de la clase C_1 están en el intervalo V_1 con respecto a la variable y_1 y en el intervalo V_2 referido a la variable y_2



Un individuo a_1 pertenece a la clase C_1 si cumple estas condiciones:

$$a_1 = \left[y_1 = V_1\right] \wedge \left[y_2 = V_2\right]$$

Estas condiciones son necesarias y suficientes para que todo individuo de C1 pertenezca a esa clase. Por lo tanto la de las cruces es una clase monotética.
Si el criterio se hubiera restringido a V1 solamente no sería una condición suficiente porque hay otros individuos de otras clases (los círculos que están en ese intervalo y no pertenecen a C1)
De la misma manera si se hubiera tomado sólo V2 también habría otros individuos de otras clases, por lo tanto los intervalos V1 ni V2 no son, tomados de a uno, condiciones necesarias y suficientes. La condición de monoteticidad implica que exista una condición necesaria y suficiente, al menos, que haga que todos los individuos que la satisfagan estén en la clase.
Por el contrario observando la otra clase C2 , la de los círculos, no hay condiciones necesarias y suficientes utilizando y1 e y2 , que hagan que esta población de la clase de los círculos pueda ser

descripta por propiedades, por lo tanto la clase no es monotética sino politética.
¿Cómo describir una clase, cómo representarla y cuál es la intensión de una clase?. Existen tres tradiciones: la de la representación lógica, la de noción de semejanza y la de prototipo.
1) La primera es la llamada aristotélica que da clases monotéticas y que describe las clases por una conjunción lógica de propiedades.
2) La tradición adansoniana (de Adanson[34]) en donde simplemente una clase no está caracterizada por una conjunción de propiedades como lo decía Aristóteles sino por un alto grado de semejanza.
3) Una tercera tendencia que proviene de los psicólogos es la de Rosch[35], que está en desacuerdo con la tradición aristotélica, afirmando que no se puede representar un concepto, por ejemplo el concepto de manzana, mediante una conjunción de propiedades. En nuestro espíritu el concepto de manzana se efectúa, a través de uno o varios prototipos. Por ejemplo el concepto de manzana puede estar representado por el de una manzana roja. O sea una clase puede representarse por ejemplos muy representativos que son llamados prototipos.
Hasta aquí hemos indicado cómo se representan las clases. Ahora debiéramos definir cómo obtenerlas.

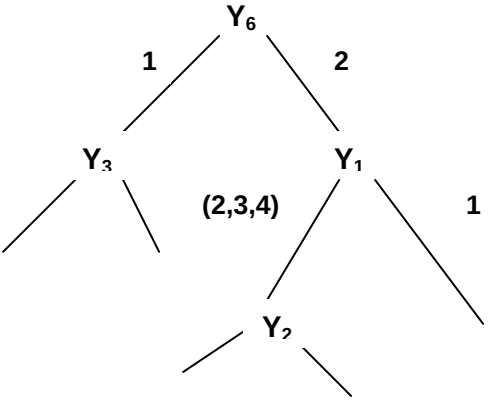
Obtención de una clase

En este caso podemos retomar las tres tendencias precedentes.
1) *En la tradición aristotélica* podemos definir una clase mediante un proceso de arriba hacia abajo que va a permitir elegir las propiedades que caracterizan a cada una de las clases, de la más general a la más específica.

Si se tienen varias variables explicativas y una variable para explicar (Y) el objetivo puede ser buscar cuál es la variable explicativa que explique mejor las variables a explicar.
En el siguiente diagrama encontramos por ejemplo la variable y6

[34] Adanson, M. (1763), *Famille des plants*, Vol.1, Vincent, París. Citado idem .
[35] Rosch, E. (1978) *Principle of categorization* in E. Rosch and B. Lloyd (eds), Cognition and Categorisation pp. 27-48, Hillsdale, N.J.: Erlbaum. Citado idem

que toma dos modalidades obteniéndose dos ramas, una para cada lado.
A continuación entre los individuos que tienen y6 = 1 se busca cuál es la variable que explique mejor, es la Y_3 y así sucesivamente.

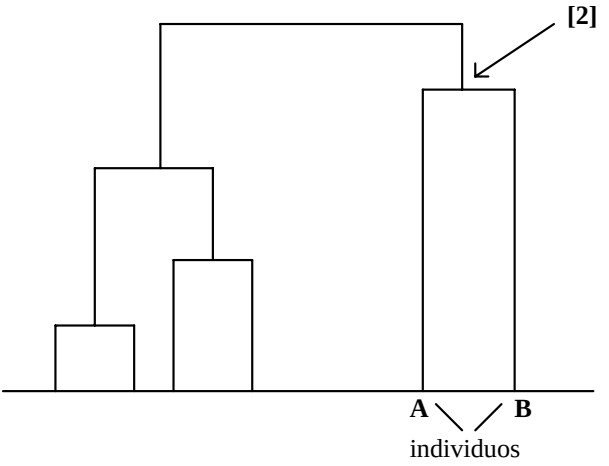


$C_1 : a_1 = [\, y_6 = 2 \,]\ [\, y_1 = 1 \,]\ [\, * \,]$
$C_j : a_j = \hat{\imath}\, [\, y_i = v_i \,]$

Si se toma la rama derecha, se describe la población que toma como valor y6 = 2 e y1 = 1. Esta clase (C1) se hallará descripta por [*]
Si se toma la rama izquierda se describe la población que toma como valor y6 = 1, etc
Esta forma de obtención de clases se halla dentro de la tradición aristotélica ya que se forman clases monotéticas, con condiciones necesarias y suficientes y se llama también método de segmentación. Este método se basa en el principio llamado de subordinación de los caracteres por el cual una clase está caracterizada por una conjunción de propiedades heredadas de sub-clases. Son llamados asimismo árboles construidos de arriba hacia abajo.

2) *El segundo método para obtener clases* es otro tipo de algoritmo de abajo hacia arriba. Se puede nombrar a Adanson (1757) como el creador del primer algoritmo de clasificación ascendente jerárquico que es utilizado hasta nuestros días.

Este método consiste en comenzar con clases reducidas de individuos, reuniendo los que son más parecidos y efectuando nuevamente el procedimiento.



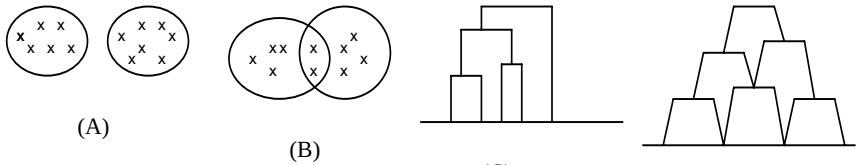
Se toman los dos puntos más cercanos, por ejemplo A y B se reúnen y se forma una nueva clase. Así se continúa sucesivamente.
Como se han utilizado semejanzas, las clases obtenidas son politéticas, pero por supuesto teniendo una clase puedo dar de ella una descripción monotética. Esta descripción tal vez va a cubrirla, es decir, que es la intensión de esta clase, pero no es una condición necesaria y suficiente para los individuos que la forman. Se pueden encontrar individuos exteriores que la satisfagan.
Por ejemplo: los individuos A y B de la figura anterior son rojos o blancos, pero hay individuos rojos o blancos que no están en esta clase, entonces no es una clase monotética.

3) Otra manera de representar a una clase politética, consiste en decir: en esta clase están los individuos que son a menudo rojos y rara vez blancos. Aquí tenemos otra noción que es una descripción pseudológica, que describe la clase pero que no es monotética. Consiste en buscar directamente clases y su representación.
En síntesis: en el primer método se buscaban primero las descripciones y al cabo de una rama se deducía una clase, es decir, la representación eran las ramas.
En el segundo, primero se efectúa una agregación para encontrar

clases, se encuentran los niveles de la jerarquía y luego se hace la representación por una descripción.
El tercer método, consiste en hacer las dos cosas al mismo tiempo, se pueden buscar simultáneamente las clases y su representación.
En esta tercera orientación se encuentra el método de nubes dinámicas que expondremos más adelante.

Espacios de clasificación

Se trata de estructuras interclase que pueden tener propiedades en común y que constituyen espacios de clasificación, ellas pueden ser:
particiones (A) envolturas (B), jerarquías (C) o pirámides (D).



(A) (B) (C) (D)

A) Una partición de un conjunto de individuos es un conjunto de partes no vacías, formadas por intersecciones vacías de elementos tomados dos a dos y donde la reunión forma el conjunto total de individuos. Es decir es la separación de los individuos en clases mutuamente excluyentes.
B) Una envoltura del conjunto de individuos es un conjunto de partes no vacías del conjunto total. Una partición es un caso especial de envoltura. Las envolturas están formadas por clases de individuos no necesariamente excluyentes entre sí, puede haber individuos que pertenezcan a varias clases simultáneamente.
C) Una jerarquía es un conjunto de particiones encastradas. En general una jerarquía es visualizada por un árbol denominado dendrograma.
D) Una pirámide está formada por clases que pueden tener puntos en común.
Las jerarquías son hermafroditas porque cada clase no tiene más que un padre o una madre. En las pirámides cada clase puede tener dos ascendientes, no son forzosamente hermafroditas, de

manera tal que se obtienen clases que se superponen.

Algoritmos de clasificación

En todas las disciplinas científicas las clasificaciones son necesarias. En cada una de ellas encontraremos ejemplos de poblaciones susceptibles de ser ordenadas en categorías.
En las ciencias de la conducta, el uso de tipos psicológicos es de muy vieja data y en biología la clasificación de especies ha sido el fundamento de la disciplina. En botánica por ejemplo, es ampliamente empleada para poner en evidencia las subespecies de una misma variedad a partir de una tabla de datos donde las plantas son caracterizadas por un cierto número de mediciones: los "taxones"[36] de allí que el término taxonomía se utiliza a veces como sinónimo de clasificación automática.
En las ciencias sociales la formación de clases puede ser utilizada como instrumento interpretativo pero de revelada utilidad debido precisamente a que las escalas de medición que en ellas más frecuentemente se utilizan son las nominales o categóricas.

En reconocimiento de formas, se puede utilizar la clasificación automática para obtener tipos de impresiones digitales, de escrituras, de señales electrográficas, de señales de radar. En inteligencia artificial, la clasificación automática se considera como un procedimiento que enseña dando a la computadora una información de orden semántico que no se encuentra en la matriz de datos inicial de forma clara, se habla entonces de aprendizaje sin profesor.
Las técnicas de clasificación automática constituyen un aporte imprescindible para el AMD. La riqueza del punto de vista de la escuela francesa del análisis de datos está puesto precisamente en la complementariedad de ambos grupos de técnicas: análisis factorial y clasificación; difícilmente podría encuadrarse en este enfoque un análisis que involucrara sólo a una de ellas.

En este texto me ocuparé del procedimiento de clasificar individuos construyendo grupos con ellos bajo el principio de que

[36] taxon. (Del gr. Tásso, ordenar.) m. Bot. Entidad sistemática de cualquier rango. Se aplica indistintamente a familia, género, especie, variedad, etc. Diccionario Enciclopédico Espasa-Calpe, tomo 22. Madrid,1981.

esos grupos sean lo más diferentes posible entre ellos y lo más homogéneos posible dentro del grupo.
En este sentido el objetivo de la clasificación está puesto en la posibilidad de ganar poder en interpretación en cuanto a la asociación entre las variables puestas en juego en el análisis más que en asignar un grupo y sólo uno a un individuo. Las clases construidas en este enfoque serán clases de bordes difusos (o politéticas) en el interior de las cuales encontraremos modalidades de variables asociadas por el hecho de presentarse juntas en los mismos individuos.
Formamos clases de individuos a los efectos de interpretar en una perspectiva más cercana a la realidad observable, ya que cada clase será descripta en términos de porcentajes de individuos que poseen una determinada cualidad.

En el ámbito de la estadística clásica la técnica de clasificación mayormente utilizada es el análisis discriminante. El objetivo del mismo es, en un enfoque prescriptivo, encontrar una función para poder predecir la asignación de un individuo a una clase o a otra. El análisis discriminante es de gran aplicación en Medicina en la determinación de diagnósticos específicos, en base a estudios anteriores (los casos) se utilizan síntomas y determinaciones de laboratorio como variables determinantes de un diagnóstico, en base a ellos se construye una función (la regla) de manera que ante un nuevo caso es posible determinar un diagnóstico, es decir asignar el individuo a una clase (abducción).

En nuestro caso el objetivo es otro: considerados los individuos en su totalidad lo que se busca es encontrar cuál es la mejor representación de estos individuos a los efectos de interpretar esa estructura, siempre siguiendo un antiguo criterio estadístico, comparar la variación de las variables dentro del grupo (que debiera ser mínima) con la variación entre los grupos (que debiera ser máxima).
La información aportada por una clasificación se sitúa a nivel semántico, “no se trata de esperar un resultado verdadero o falso, sino sólo aprovechable o no aprovechable”[37] Existen muchas

[37] Lance G. y Williams W (1967)A general theory of classification sorting strategies: 1= hierarchical systems, 2=clustering systems.Computer Journal 9-10, pp. 373-380. Citado por: Diday E. (1992)

razones para considerar esa información aprovechable, ya sea porque pueden aparecer reagrupamientos inesperados que puedan permitir la generación de nuevas hipótesis, reagrupamientos esperados que en realidad no existan lo cual pone de manifiesto un bajo poder separador de las variables utilizadas. Las clases obtenidas y sus imbricaciones aseguran una visión concisa y estructurada de los datos. Las clases significativas implican la eventual definición de funciones de decisión que permiten atribuir un nuevo objeto a la clase más cercana, etc.

La diferencia fundamental entre la clasificación que se realiza en el programa SPAD con los algoritmos clásicos de la escuela anglosajona radica en que en estos últimos los valores que se utilizan en los cálculos son los de las variables originales, por lo tanto están sujetos a las posibles diferencias de las escalas de medición de las mismas, en cambio en el primero se trabaja con los valores de las coordenadas en los ejes factoriales sobre los cuales las variables originales fueron proyectadas según vimos en los capítulos anteriores.

Un individuo se caracteriza por los valores que asume en las variables que he considerado. Es decir en este caso cada individuo no entrará en el algoritmo de clasificación con sus valores originales (códigos de: escolaridad primaria, empleado, etc.) sino con los valores que dicho individuo tiene en cada uno de los ejes factoriales calculados con el análisis de correspondencias precedente, es decir las coordenadas de ese individuo en los ejes factoriales.

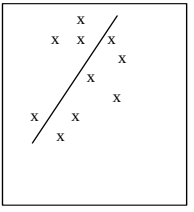
Se podrán tomar tantos ejes como se desee, en general se toman los 10 primeros porque se supone que con ello se acumula un porcentaje suficiente de la información original de los datos. Ello implica que se ha realizado una especie de alisado de la información, nos hemos quedado con lo más representativo, que ya analizamos en el análisis factorial, eliminando el “ruido”, la distorsión. Teniendo en cuenta esos puntos individuos ubicados en el hiperespacio factorial, digamos de 10 dimensiones, el objetivo será agruparlos según su cercanía formando tantos grupos como sea necesario.

En este caso estaríamos trabajando sobre una matriz de datos modificada, donde las unidades de análisis siguen siendo los individuos pero las variables son cada uno de los ejes factoriales. Hemos pasado de un hiperespacio de un gran número de

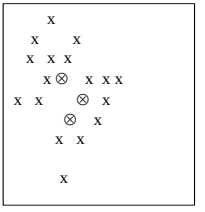
variables a otro más reducido.
Los algoritmos de clasificación que realiza el programa SPAD son numerosos, presentaré el análisis clasificatorio clásico, apto para el procesamiento de encuestas, que se basa en una combinación de técnicas o clasificación mixta que se realizan una a continuación de la otra.

Nubes dinámicas o medias móviles

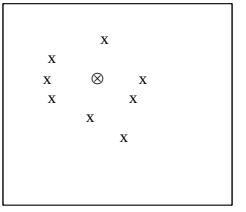
Este algoritmo de clasificación por particiones según lo presenta Diday(1982:117-130), se basa en la definición previa de un modo de representación simbólica de todo grupo de individuos. Dado un grupo de individuos, esta representación simbólica, llamada "núcleo", puede ser por ejemplo, una recta, un grupo de puntos de la población, un centro de gravedad



El núcleo es una recta



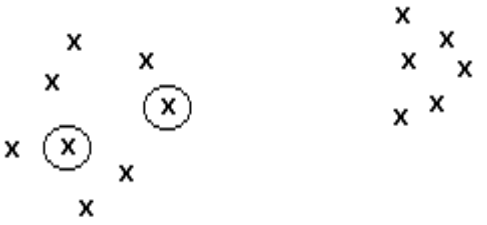
El núcleo es un grupo de puntos tomados en la población



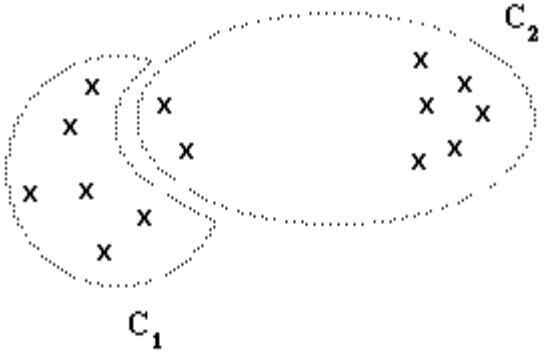
El núcleo es un centro de gravedad

Se parte eligiendo k núcleos estimados o tirados al azar tomados entre una familia de núcleos admisibles llamado “espacio de representación”. Cada punto de la población es luego afectado al núcleo más próximo.
Se obtiene así una partición en k clases en la cual se calculan los núcleos.
Se recomienza el procedimiento con los nuevos núcleos y así sucesivamente.
Se demuestra que, bajo ciertas condiciones, el algoritmo converge hacia una posición estable mejorando, con cada iteración, un criterio matemático.
Asimismo, en lugar de partir de k núcleos, se puede partir de una partición de k clases elegidas al azar o determinadas, por ejemplo, con la ayuda de una clasificación jerárquica.

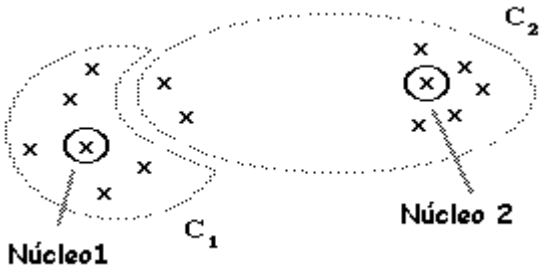
Como primera etapa entonces por ejemplo, quisiera formar una buena partición en 2 clases: se eligen dos puntos al azar llamados núcleos.



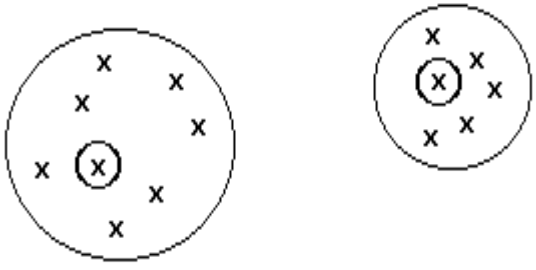
La etapa siguiente consiste en asociar cada punto al núcleo más cercano, cuando se dice el núcleo más cercano se sobrentiende que se está definiendo una distancia. Esta distancia puede ser simplemente la distancia euclidiana del plano, la que puede medirse con una regla una regla. Se asocia cada punto al núcleo más cercano y se van a obtener dos clases: la clase C_1 de los puntos que están más cercanos al Núcleo 1 y la clase C_2 de los puntos que están más cercanos al Núcleo 2.



En la tercera etapa dos nuevos puntos o núcleos son extraídos minimizando la distancia del punto que se busca a todos los otros puntos de su misma clase.



Es decir, en cada clase, para cada punto, se calculan las distancias a todos los otros puntos de la clase y se retiene el que da la menor suma de las distancias a todos los otros puntos.
Las clases C_1 y C_2 fueron obtenidas en la etapa precedente. Se vuelve a la etapa 2, posteriormente a la etapa 3 y así sucesivamente hasta la convergencia. La etapa 2 era la etapa que permitía: conociendo los centros construir clases, por lo tanto se encontraban dos nuevos núcleos y se asociaba cada punto al núcleo más cercano. Se va a ver entonces que un punto que está en una clase puede moverse y pasar a otra clase porque está más cercano de ese núcleo que del otro y así se encuentran dos nuevas clases.



Se vuelven a calcular los centros, otras clases y así sucesivamente.
Este es un ejemplo simple de algoritmo de partición que busca clases y sus representaciones. Se puede decir que es también una búsqueda de prototipos en la tradición de Rosch.
Se pueden tomar otros tipos de núcleos. Por ejemplo como se había enunciado antes se pueden tomar núcleos que no son

puntos, considerar que el núcleo es una recta y en este momento se buscan dos clases y dos rectas tal que la adecuación entre las clases y las rectas sea la mejor posible.
El algoritmo se detiene cuando no es posible mejorar la distancia de cada uno de los puntos a los centros de gravedad, cuando lo que se va incorporando no produce ya una ganancia.
Este método tiene la desventaja de que puede producir particiones diferentes según sean los núcleos iniciales elegidos, por esta razón se complementa con otros dos tipos de algoritmos de clasificación.

Formas fuertes y estabilidad

Habiendo obtenido varias particiones con el método anterior, a partir de diferentes tiradas aleatorias, podemos plantearnos el problema de la búsqueda de clases estables, es decir de individuos que siguen agrupándose cualquiera que sea la tirada de partida. De allí el término de formas fuertes que se utiliza habitualmente en la práctica.
Si por ejemplo contáramos con 21 individuos identificados de la A a la T y hubiéramos obtenido con el algoritmo de nubes dinámicas dos particiones que agruparan ambas a los individuos en 3 clases:
1ra partición:

Clase 1: A B C D	Clase 2: E F G H I J K L	Clase 3: M N Ñ O P Q R S T

2da. Partición:

Clase 1: D F I J K	Clase 2: A B C L M N G	Clase 3: E H O Ñ P Q R S T

Si luego cruzáramos ambas particiones, estaríamos buscando los individuos comunes en las clases de ambas:

2da. Partición	**1ra. Partición**		
	Clase 1: A B C D (4)	*Clase 2*: E F G H I J K L (8)	*Clase 3*: M N Ñ O P Q R S T (9)
Clase 1: D F I J K (5)	D 1	F K I J **4**	0 individuos
Clase 2: A B C L M N G (7)	A B C **3**	G L 2	M N 2
Clase 3: E H O Ñ P Q R S T (9)	0 individuos	E H 2	Ñ O P Q R S T **7**

Nótese que el número de cada clase no tiene necesariamente que coincidir en las dos particiones, lo que importa es el contenido de cada clase. En el ejemplo las formas fuertes están resaltadas en color.

En la 1ra. Partición:

- La clase 1 está constituida por 4 individuos, de los cuales 3 (el A, el B y el C) se vuelven a encontrar juntos en la clase 2 de la 2da. Partición.
- La clase 2 está constituida por 8 individuos, de los cuales sólo 4 (el F, el K, el I y el J) se vuelven a encontrar juntos en la clase 1 de la 2da. Partición
- La clase 3 está constituida por 9 individuos, de los cuales 7 (el Ñ, el O, el P, el Q, el R, el S y el T) se vuelven a encontrar juntos en la clase 3 de la 2da. Partición.

Recíprocamente podríamos buscar la ubicación de los individuos de la 2da. Partición en la 1ra.
La clase 1 de la 1ra. Partición no tiene ningún elemento en común con la clase 3 de la 2da., sólo las celdas resaltadas constituyen las formas fuertes, es decir las configuraciones de individuos cuyas semejanzas son estables cualquiera sea el núcleo de arranque de la partición. Es decir las distancias entre los mismos es lo suficientemente pequeña como para "resistir" cualquier intento de separación.
El resto de los individuos se suelen asignar a una clase residual.

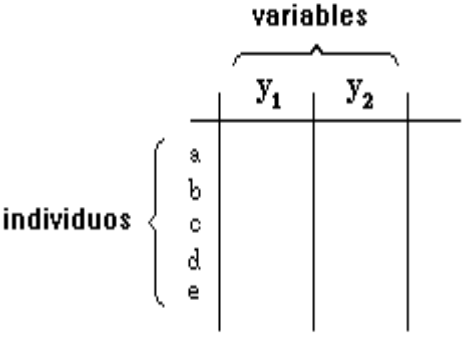
Habitualmente ocurre que se forman gran cantidad de formas fuertes, ya que si tiramos por ejemplo 10 núcleos y cruzamos aunque sea dos particiones ya contamos con 100 casilleros posibles. Aún cuando las formas fuertes no sean la totalidad de ellas significará una gran cantidad de clases, por lo tanto se tratará de conseguir un algoritmo que nos permita seleccionar una cantidad menor de clases.

Agregación jerárquica de las clases y corte del árbol o dendrograma

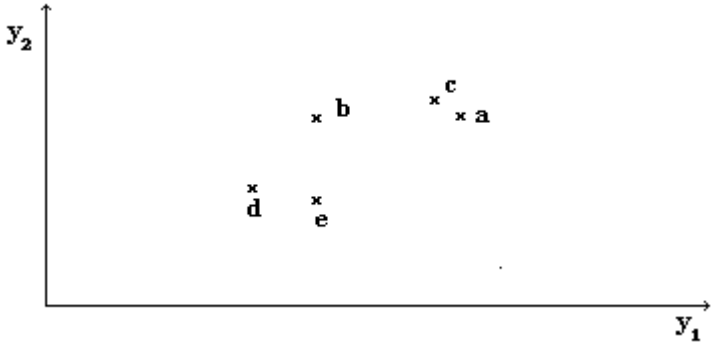
Constituye la tercera etapa de esta técnica mixta de clasificación. Podemos partir de los centros de gravedad de las formas fuertes. Se va estableciendo una jerarquía entre los puntos que depende

del grado de relación entre los mismos medida en términos de distancia.
Las jerarquías, por la comodidad de su interpretación visual, constituyen desde hace mucho tiempo una forma de clasificación muy popular. Las clasificaciones “naturales” de los animales y de los vegetales son jerarquías. Muy a menudo el usuario está interesado en la detección de clases muy significativas, formadas a través de la jerarquía, siendo lo ideal que estas clases formen una partición obtenida por corte de la jerarquía según una línea horizontal bien ubicada.

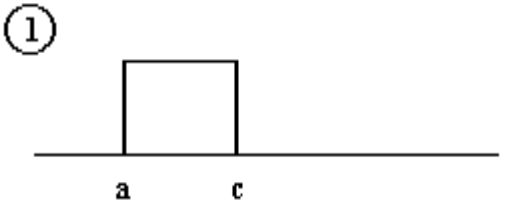
Su principio básico es el siguiente: en cada etapa, se reúnen las dos clases que más se parecen, al inicio cada individuo es considerado como clase y se van agregando luego tantas clases hasta que todos los individuos sean incluidos.
Por ejemplo si tuviéramos una matriz de datos de 5 individuos y 2 variables:



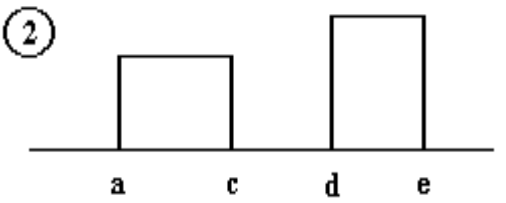
Podrían representarse los individuos en el plano formado por las dos variables:



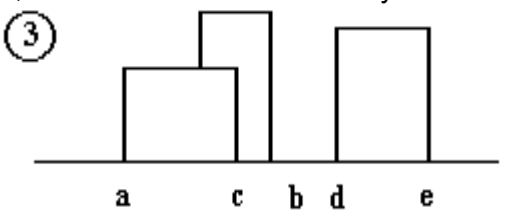
El objetivo es encontrar el par de individuos que estén más cercanos. La primera etapa consiste en buscar los dos puntos que estén más cercanos para formar la primera clase. Observamos que los puntos que se encuentran más cerca son el a y el c



Se determinan los dos puntos más cercanos, y se representa la altura que es la distancia entre estos dos puntos.
Luego se pregunta. ¿Cuáles son las otras dos clases más cercanas? A simple vista en el gráfico los puntos d y e resultan ser los más cercanos.



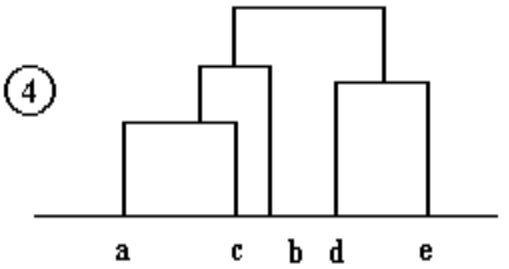
Posteriormente, lo más cercano es b con c y a.

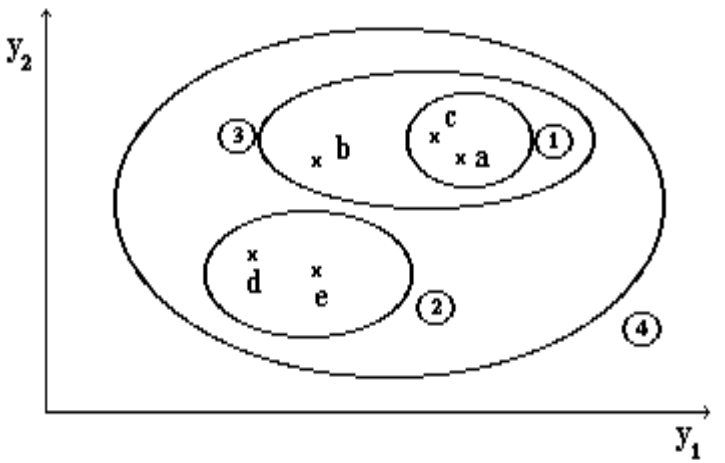


Se calcula una distancia entre b y ca, es decir la suma de las distancias d (b,c) y d (b,a) lo que resulta el tercer nivel. El último nivel se obtiene agregando las otras dos clases.
El nivel máximo de agregación se obtiene cuando junto las dos últimas clases, en ese nivel la heterogeneidad de la clase es máxima, mientras que en la base del árbol hay una completa homogeneidad ya que cada clase (formada por un solo individuo) es igual a sí misma. En este intervalo de agregación deberemos

encontrar una fórmula de compromiso en la cual cortar horizontalmente el dendograma.

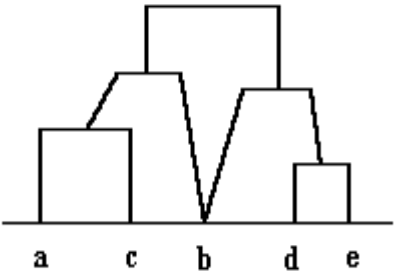
El corte podrá hacerse cuando la cantidad de variación interclase agregada entre un umbral y el siguiente sea máxima.
Se calcula un índice de agregación que responde a un criterio de similitud adentro de la clase, mientras más individuos agregue iré sumando variación. En el nivel de la base del árbol tenemos una similitud máxima, (cada individuo se parece a sí mismo) por lo tanto tengo un nivel de variación intraclase nulo. Al mismo tiempo tengo una variación interclase máxima.
A medida que se van subiendo umbrales en el árbol va aumentando la variación intraclase y disminuyendo la variación interclase. Por lo tanto el índice de agregación va combinando ambos conceptos. ¿Dónde cortar el árbol? Donde el salto producido sea el salto mayor.





Pirámides

Se pueden también hacer pirámides con el mismo procedimiento. Se va agregar primero a y c, luego d y e, luego b se va agregar tanto a a y c como a d y e.

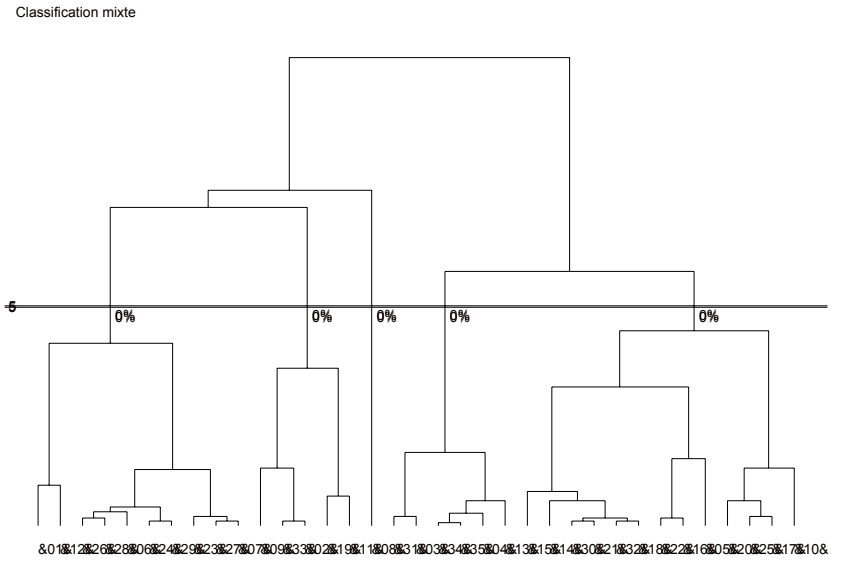


Se observan dos clases que tienen un punto en común y se obtuvo otra representación: (a,c,b), (b,d,e)
Se observa que, contrariamente a las jerarquías, esta representación autoriza la posibilidad de que un punto pertenezca a la vez a dos clases.
Si cortamos una pirámide con una serie de líneas horizontales encontraremos en lugar de clases excluyentes, envolturas encastradas.

Ejemplo de aplicación: clasificación sobre ejes factoriales

Retomando el ejemplo del sondeo de opinión presentado en el capítulo 7, describiré la aplicación de la técnica mixta de clasificación analizando los resultados provistos por el software.
Como complemento del análisis de correspondencias múltiples y con mayor poder de interpretación, se acopla la clasificación o cluster sobre coordenadas factoriales. Es importante destacar que se utilizan los valores de las proyecciones de las variables sobre los ejes factoriales y no los valores originales de las mismas. Ello permite conservar el mayor poder discriminatorio de algunas variables por sobre las otras menos significativas.
El método de clasificación utilizado es mixto, combinando sucesivamente las técnicas recién expuestas, de centros móviles, formas fuertes y árbol jerárquico, de manera de aprovechar las ventajas comparativas complementando los inconvenientes de

cada uno.
En nuestro caso cortamos el árbol jerárquico al nivel de 5 clases o grupos de individuos muy similares dentro del grupo y muy diferentes entre grupos. A cada grupo se le asigna una denominación alfanumérica sin implicar relación alguna de jerarquía entre ellos.
Presento a continuación el árbol jerárquico o dendrograma obtenido en la última etapa.



Se realiza el corte del árbol a la altura en que el índice de agregación es óptimo, es decir, cuando la variancia intraclase es importante pero a la vez el número de clases construidas es suficiente para permitir una provechosa interpretación.

Descripción estadística de las clases

Los elementos de una misma clase se asemejan en cuanto al conjunto de criterios elegidos para describirlos. Queda entonces por precisar cuáles son estos criterios de origen de los agrupamientos observados. Se procede a una descripción automática de las clases que constituye en la práctica una etapa indispensable de todo procedimiento de clasificación y que podría denominarse como una "vuelta a la realidad".

Las ayudas a la interpretación de las clases se basan generalmente en comparaciones de medias o de porcentajes calculados en el interior de las clases con las medias o porcentajes obtenidos sobre el conjunto de los elementos a clasificar (lo cual es equivalente a comparar medias o porcentajes) dentro de la clase y fuera de la clase.
Para seleccionar las variables continuas o las modalidades de las variables nominales más características de cada clase, se mide el desvío entre los valores relativos a la clase y los valores globales. Estas estadísticas pueden ser convertidas en un criterio llamado *valor-test* que permita operar una selección de las variables designando las más características[38].
Entre las variables figuran igualmente aquéllas que no han contribuido a la construcción de las clases pero que pueden participar en su descripción bajo el mismo principio que las variables suplementarias en un análisis factorial. Estas variables permiten *a posteriori* identificar y caracterizar los agrupamientos establecidos a partir de las variables activas.

Valores-test para las variables nominales

Una modalidad (o categoría) de una variable nominal es considerada como característica de la clase si se juzga que su *abundancia* en la clase es significativamente superior a lo que puede esperarse teniendo en cuenta su presencia en la población. La abundancia de una modalidad j se define, en primer lugar, comparando su porcentaje en la K-ésima clase con su porcentaje en la población.
Bajo la hipótesis "nula" [39] de que los individuos de la clase k son extraídos al azar sin reposición en la población de los n individuos, el porcentaje de individuos de la clase k que tienen la modalidad j por una parte, y el porcentaje de individuos que tienen la modalidad j en la población por otra parte, deberán coincidir salvo por fluctuaciones aleatorias.

[38] Morineau A. (1984) Note sur la caracterisation statistique d'une classe et les valeurs-test, *Bull. Techn. du Centre de Statis. et d´Infor. Appl.,*2, p.20-27

[39] Como en el caso de las variables continuas, esta hipótesis nula tiene sentido sólo para las variables nominales suplementarias. Pero los valores-test que se van a calcular pueden todavía jugar el rol de índices de similaridad entre modalidades activas y clases y por lo tanto servir para ordenar estas modalidades por orden de interés para cada clase.

La anterior es la hipótesis de independencia bajo la cual el número N de individuos de la clase k que tienen la modalidad j es una variable aleatoria que sigue una ley hipergeométrica donde los tres parámetros aparecen en los márgenes de la tabla.
Cuanto más esta probabilidad es débil, más difícil es aceptar la hipótesis de una extracción al azar. Nos servimos de esta probabilidad para ordenar las modalidades características de la clase (la más característica corresponde a la más pequeña probabilidad).
Esta probabilidad es a menudo muy débil. Resulta cómodo sustituir ése valor por el de la variable de Laplace-Gauss correspondiente a la misma probabilidad. Es el *valor-test*. El mismo mide el desvío entre la proporción en la clase y la proporción general, en número de desvíos standard de una ley normal. El valor-test, para una modalidad de una variable nominal, es entonces un criterio estadístico asociado a la comparación de los efectivos en el marco de una ley hipergeométrica.

Variables características de una clase

El valor test cumple el objetivo de efectuar un cambio de medida al transformar la probabilidad de una distribución cualquiera en números de desvíos standard de una ley normal centrada reducida.
Ya sea para la búsqueda de las variables continuas o de las modalidades de las variables nominales características de una clase, el valor absoluto del valor-test es el análogo del valor absoluto de una variable normal centrada reducida[40]
Las variables son tanto más interesantes cuanto los valores-test asociados son fuertes en valor absoluto. Se puede entonces ordenar estas variables siguiendo los valores-test decrecientes y no retener más que los elementos más significativos, lo que permite caracterizar muy rápidamente las clases.
Seleccionando, para cada clase, las variables más características, y calculando su media o su porcentaje en la clase, se constituye el "perfil-tipo" de la clase. Recordemos que el valor -test no

[40] En el marco de los tests clásicos, se dirá que será significativo al nivel usual del 5% si él sobrepasa el valor 1,96 : la hipótesis "nula" es rechazada y la media o la proporción de una variable sobre la población global y la de la clase difieren significativamente.

corresponde a un verdadero test de hipótesis[41] a menos que la variable a la cual está asociada sea suplementaria.
Mencionemos en fin, que el hecho de calcular simultáneamente varios valores-test pone al usuario en una situación de "comparaciones múltiples", que impone tomar umbrales de significación más severos que los utilizados en caso de un test único.

Para conocer la descripción de las clases, el programa provee una salida con las modalidades más características de cada grupo o clase.
Explicaré brevemente los encabezados de las columnas:
En la columna 1: **Libellés des variables**, se especifican las etiquetas de las variables
En la 2: **Modalités caractéristiques**, las correspondientes modalidades de esas variables que resultan características en esa clase.
En la 3: **% de la modalité dans l'échantillon**, el porcentaje de la modalidad en el total de la muestra
En la 4: **% de la modalité dans la classe**, el porcentaje de la modalidad en la clase
En la 5: **% de la classe dans la modalité**, el porcentaje de la clase en la modalidad
En la 6: **Valeur-Test**, el valor test que caracteriza a esa modalidad
En la 7: **Probabilité**, la probabilidad N(0,1) correspondiente a ese valor test
En la 8: **Poids**, la frecuencia de esa modalidad en el total de la muestra.
En los cuadros que siguen se consignan sólo las 10 modalidades más características, con valores tests más significativos.

[41] Aquí tenemos la hipótesis de que una variable continua o una modalidad de una variable nominal es independiente de la partición.

Caractérisation par les modalités des classes de la partition

Coupure 'a' de l'arbre en 5 classes

Classe: CLASSE 1 / 5

Libellés des variables	Modalités caractéristiques	% de la modalité dans l'échantillon	% de la modalité dans la classe	% de la classe dans la modalité	Valeur-Test	Probabilité	Poids
¿Qué imagen tiene del Gob. Municipal?	Gob.Mun.buena	35.93	77.91	63.21	9.51	0.000	106
¿Qué imagen tiene del Intendente?	El Intendente buena	55.93	90.70	47.27	8.07	0.000	165
Califica.Con.Municipal	CMun.buena	16.95	45.35	78.00	7.82	0.000	50
Calific.Dep.Ejecutivo	Dep.Ej.Buena	21.69	51.16	68.75	7.43	0.000	64
¿Qué imagen tiene del Con.Municipal?	CMun.buena	15.93	40.70	74.47	6.95	0.000	47
Mantenimiento calles de tierra	TI buena	9.83	29.07	86.21	6.59	0.000	29
Zanjeo	ZJ buena	11.19	31.40	81.82	6.53	0.000	33
Desmalezamiento	DZ buena	23.39	48.84	60.87	6.25	0.000	69
Riego	RG buena	32.54	59.30	53.13	6.05	0.000	96
Recolec.Ramas	RA buena	46.78	74.42	46.38	6.05	0.000	138

Classe: CLASSE 2 / 5

Libellés des variables	Modalités caractéristiques	% de la modalité dans l'échantillon	% de la modalité dans la classe	% de la classe dans la modalité	Valeur-Test	Probabilité	Poids
Riego	RS no corresponde	20.68	79.66	77.05	11.43	0.000	61
Mantenimiento calles de tierra	TI no corresponde	21.02	77.97	74.19	10.96	0.000	62
Desmalezamiento	DZ no corresponde	25.76	84.75	65.79	10.89	0.000	76
Zanjeo	ZJ no corresponde	35.25	94.92	53.85	10.79	0.000	104
Alumbrado	AL muy buena	3.05	13.56	88.89	4.21	0.000	9
Recolec.Ramas	RA muy buena	3.05	13.56	88.89	4.21	0.000	9
Atenc.Dispensarios	DI muy buena	31.19	52.54	33.70	3.70	0.000	92
Recol.residuos	RS muy buena	6.44	16.95	52.63	3.07	0.001	19
Barrido y limpieza	BA muy buena	2.03	8.47	83.33	2.99	0.001	6
Recolec.Ramas	RA no corresponde	3.05	10.17	66.67	2.79	0.003	9

Classe: CLASSE 3 / 5

Libellés des variables	Modalités caractéristiques	% de la modalité dans l'échantillon	% de la modalité dans la classe	% de la classe dans la modalité	Valeur-Test	Probabilité	Poids
Desmalezamiento	DZ no sabe	5.42	66.67	75.00	7.46	0.000	16
Zanjeo	ZJ no sabe	7.12	72.22	61.90	7.35	0.000	21
Riego	RG no sabe	8.14	72.22	54.17	7.02	0.000	24
Mantenimiento calles de tierra	TI no sabe	5.08	61.11	73.33	7.00	0.000	15
Recolec.Ramas	RA no sabe	3.73	33.33	54.55	4.31	0.000	11
Barrido y limpieza	BA no sabe	3.39	27.78	50.00	3.72	0.000	10
Atenc.Dispensarios	DI no sabe	7.46	27.78	22.73	2.48	0.007	22
¿Qué imagen tiene del Gob. Municipal?	Gob.Mun.no sabe	5.08	22.22	26.67	2.38	0.009	15

Classe: CLASSE 4 / 5

Libellés des variables	Modalités caractéristiques	% de la modalité dans l'échantillon	% de la modalité dans la classe	% de la classe dans la modalité	Valeur-Test	Probabilité	Poids
Calific.Dep.Ejecutivo	Dep-Ej.REG.	31.86	74.19	48.94	7.68	0.000	94
Califica.Con.Municipal	CMun.regular	34.92	74.19	44.66	7.03	0.000	103
¿Qué imagen tiene del Gob. Municipal?	Gob.Mun.regular	39.66	77.42	41.03	6.69	0.000	117
¿Qué imagen tiene del Intendente?	El Intendente regular	20.68	50.00	50.82	5.83	0.000	61
Desmalezamiento	DZ regular	19.66	46.77	50.00	5.48	0.000	58
Recolec.Ramas	RA regular	22.37	50.00	46.97	5.38	0.000	66
Zanjeo	ZJ regular	17.29	38.71	47.06	4.52	0.000	51
Atenc.Dispensarios	DI buena	50.51	70.97	29.53	3.52	0.000	149
Barrido y limpieza	BA no corresponde	18.64	35.48	40.00	3.47	0.000	55
Riego	RG regular	25.08	41.94	35.14	3.17	0.001	74

Classe: CLASSE 5 / 5

Libellés des variables	Modalités caractéristiques	% de la modalité dans l'échantillon	% de la modalité dans la classe	% de la classe dans la modalité	Valeur-Test	Probabilité	Poids
Qué imagen tiene del Gob. Municipal?	Gob.Mun.mala	19.32	57.14	70.18	8.41	0.000	57
Riego	RG mala	9.49	37.14	92.86	8.12	0.000	28
Qué imagen tiene del Intendente?	Intendente mala	15.93	45.71	68.09	7.04	0.000	47
Desmalezamiento	DZ mala	18.64	50.00	63.64	7.04	0.000	55
Zanjeo	ZJ muy mala	11.19	37.14	78.79	7.02	0.000	33
Mantenimiento calles de tierra	TI muy mala	19.32	50.00	61.40	6.81	0.000	57
Calific.Dep.Ejecutivo	Dep.Ej.Mala	16.95	45.71	64.00	6.66	0.000	50
Recolec.Ramas	RA mala	16.95	42.86	60.00	6.00	0.000	50
Qué imagen tiene del Con.Municipal?	CMun.mala	22.71	48.57	50.75	5.47	0.000	67
Califica.Con.Municipal	CMun.mala	20.68	45.71	52.46	5.44	0.000	61

Interpretación

El grupo denominado con 1, está constituido por el 29% de la muestra, agrupa a los encuestados cuya opinión tanto sobre los órganos de gobierno municipales como sobre los servicios es calificada como buena.

El grupo 2, del 20% está formado por los ciudadanos que no poseen servicios de zanjeo, mantenimiento de calles de tierra, recolección de ramas y hojas, riego (su respuesta es un no corresponde, posiblemente porque habitan en la zona más urbanizada) y que poseen muy buena opinión sobre el resto de los servicios pero una opinión indiferente sobre el gobierno municipal.

El grupo 3, del 6% contestan 'no sabe' a todas las preguntas de opinión ya sea sobre servicios como sobre el desempeño de las autoridades municipales.

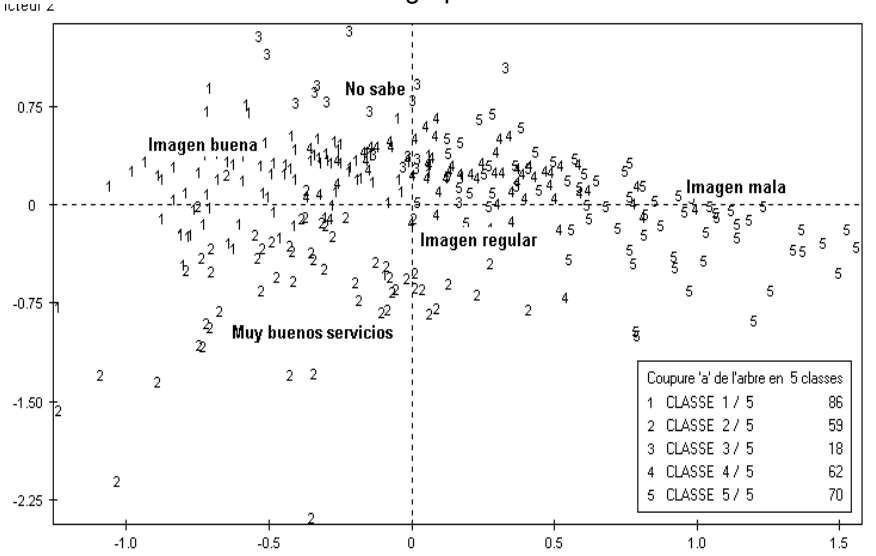
El grupo 4, del 21%, califican con 'regular' la actuación del gobierno y el concejo municipal así como todos los servicios excepto la atención en dispensarios, que consideran buena.

Por último el grupo 5 constituido por el 24% de los encuestados, poseen una mala imagen y calificación (categoría extrema) de las autoridades municipales y muy mala opinión sobre los servicios, excepto la atención en dispensarios que califican como regular.

En el gráfico factorial que sigue se proyecta la contraparte del

presentado en el capítulo anterior sobre análisis de correspondencias múltiples, sobre los mismos ejes factoriales se ubican los encuestados identificados por su número de clase o grupo.

Proyección de los individuos identificados por su número de grupo



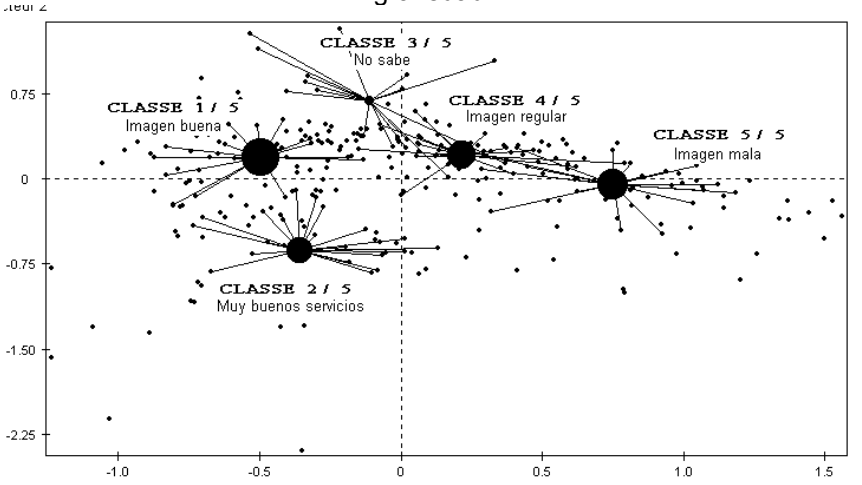
De esta manera se puede completar la interpretación del gráfico factorial de análisis de correspondencias múltiples a través de las características de los grupos de individuos y su ubicación en el plano.
Se observa de izquierda a derecha los grupos de opinión que van de una imagen buena o muy buena hacia una mala o muy mala, pasando en el centro por una opinión indiferente o regular.

Individuos característicos o "parangones"

Una posibilidad interesante que en ocasiones permite profundizar el estudio mediante técnicas cualitativas, consiste en aprovechar la tipología construida eligiendo luego los individuos más característicos de cada clase para estudiarlos en profundidad. Estos individuos son los que se encuentran más cercanos al centro de gravedad de un grupo y que en la terminología francesa son llamados "parangones" o individuos modelo.
En nuestro trabajo no se aplicaron a posteriori técnicas de

profundización, pero a manera de ejemplo se muestran en el gráfico siguiente cómo podrían graficarse esos individuos. En este caso tomamos a los 20 individuos más característicos de cada clase, encontrándose unidos mediante 'rayos' a sus respectivos centros de gravedad. El diámetro de estos últimos está ponderado por la respectiva frecuencia de cada grupo.

Individuos proyectados en los ejes 1 y 2 con referencia a sus centros de gravedad



Complementariedad entre análisis factorial y clasificación

Hemos podido observar en un ejemplo práctico el interés de la utilización complementaria de ambos grupos de técnicas.
Los métodos factoriales son indispensables a la hora de disponer de una visión de conjunto de una gran matriz de datos, sin embargo presentan algunas desventajas que pueden solucionarse con una clasificación complementaria.
Los inconvenientes se refieren especialmente a las características mismas de los gráficos que pueden traducirse a veces en:

a) *dificultades de interpretación*: resulta difícil y delicado interpretar los planos factoriales de orden superior al principal.

b) *compresión excesiva y deformaciones*: las visualizaciones se limitan a dos o pocas dimensiones. Puede implicar distorsiones o superposiciones de puntos que en realidad ocupan posiciones distantes en el espacio
La clasificación se realiza sobre todo el espacio o sobre un sub-espacio definido por los primeros factores más significativos. Las clases toman en cuenta la dimensión real de la nube de puntos y corrigen ciertas deformaciones debidas al operador proyección.
Una clase puede ser típica de un eje de rango elevado y ayudar a la interpretación de este sub-espacio particular difícilmente observable de otra manera

c*) falta de robustez*: las visualizaciones pueden ser engañosas. Un punto – perfil aberrante puede influir notablemente sobre el primer factor y por lo tanto sobre todos los siguientes ya que ellos están ligados al primero por la condición de ortogonalidad de los ejes
La mayor parte de los algoritmos de clasificación son localmente robustos en el sentido que las partes bajas de los dendrogramas (nodos correspondientes a las distancias más pequeñas) son independientes de eventuales puntos marginales aislados

d) *gráficos factoriales indescifrables*: a veces las visualizaciones pueden referirse a centenares de puntos y dar lugar a gráficos recargados o ilegibles
Ello puede solucionarse a través de la simplificación y descripción automática de las salidas gráficas mediante la formación de clases.
La noción de clase es elemental y accesible a la intuición, las clases pueden ser utilizadas para ayudar la interpretación de los planos factoriales e identificar zonas, es más fácil describir clases que un espacio continuo. La descripción de las clases puede basarse en comparaciones de porcentajes o de promedios. Los puntos numerosos pueden reemplazarse por los centros de gravedad de las clases.
Sin embargo los métodos factoriales son necesarios, a pesar de sus desventajas sobre todo teniendo en cuenta las propiedades descriptivas de los ejes y la importancia de poner en evidencia ciertas tendencias o factores latentes continuos. Asimismo la organización espacial de las clases a través del posicionamiento de las mismas sobre los ejes factoriales le confiere a la clasificación un importante poder de visualización.

CAPÍTULO 9

ANÁLISIS DE DATOS TEXTUALES

Aproximación informática al texto

El texto ha sido objeto de estudio de distintas disciplinas con objetivos, métodos y perspectivas muy diferentes: la lingüística, el análisis de discurso, el análisis de contenido, la búsqueda documentaria, la inteligencia artificial.

En la actualidad las dificultades de disponer de los textos en computadora han disminuido considerablemente, no siempre es necesario introducirlos manualmente a través del teclado, los métodos de scaneado (operación que realiza el reconocimiento de los diferentes caracteres) han realizado progresos interesantes, así como la posibilidad de obtención de textos directamente desde internet, han puesto a disposición de los investigadores una abrumadora masa de textos en los cuales es posible aplicar métodos lexicométricos.

La computadora ha permitido disponer de la potencia de la informática al servicio del texto, de su lectura, de su exploración, para satisfacer la curiosidad del investigador, permitiéndole moverse libremente en el interior del corpus, encontrar sistemáticamente las repeticiones, seguir la fluctuación de las significaciones, conocer instantáneamente las características del locutor o de la frase, identificar las asociaciones semánticas fuertes, las influencias susceptibles de explicar la formación del discurso, seleccionar las expresiones típicas, raras o especiales, esas pequeñas frases que a veces dan vida a un informe. Leer, seleccionar pero también anotar, asociar a cada frase o cada unidad de significación el o los temas que los caracterizan para así referir la variedad de un texto a una temática general.

Este trabajo artesanal de lectura, selección, anotación, desglose, archivo, puede ser perfectamente automatizado. La computadora simplemente es más rápida que la tijera, que la lapicera, que el ojo, aún cuando se trate de recorrer rápidamente el texto buscando una palabra que no quiere aparecer.

Asimismo ciertas estructuras muy difícilmente perceptibles mediante la lectura pueden revelarse a través del cruce entre las palabras y las características que identifican a un texto del cual

ellas provienen, por ejemplo, atributos del producto, de la marca o del soporte, cuando se citan términos publicitarios o identidad del entrevistado en el caso de encuestas. La informática permite producir instantáneamente estas tablas cruzadas o tablas de contingencia (asimismo llamadas tablas léxicas) pero también analizarlas automáticamente haciendo aparecer las correspondencias. El retorno al texto, guiado por las palabras más notables, dará cuerpo a las estructuras armadas por el análisis.
La construcción de indicadores léxicos y del discurso (tamaño de un corpus, intensidad de un campo léxico, banalidad) pueden ayudar a la selección y a la interpretación de los contenidos y en ciertos casos llegan a medir los efectos de discurso o de las "intensidades semánticas".

Análisis de contenido: producción del "verbatim"

El uso más común y corriente en los estudios cualitativos consiste en seleccionar una parte de las respuestas en función de su contenido, de su origen o de sus características léxicas, para editarlas luego en un informe, como ilustración, o para apoyar alguna argumentación. Es lo que se llama hacer el "verbatim".
La búsqueda del verbatim puede ser guiada por un recorrido a través de las respuestas a partir de los términos presentes en el léxico. Esto permite explorar muy rápidamente las riquezas del texto analizado, navegando, mediante el software adecuado, con el léxico y el corpus a la vista. La búsqueda se realiza en general a partir de los términos más frecuentes, pero se puede también tener en cuenta las ausencias, las cuales de hecho a veces resultan más interesantes.
Los tipos de búsquedas son numerosos, ellas permiten hacer aparecer en la ventana de texto las respuestas sobre las cuales el investigador quiere concentrar su análisis y su reflexión.
La selección puede hacerse por palabra, por formato fijo, por indicador de frase, por tipología, por proximidad.
Selección por palabra o grupo de palabras: todas las frases de un formato texto determinado incluyendo determinadas palabras.
Selección por formato fijo: todas las frases de un formato texto determinado correspondiente a las modalidades de una o de muchas preguntas cerradas.
Selección por indicador de frase: las respuestas a determinadas

cuestiones respondiendo a un cierto indicador: las más largas, las más cortas, las más banales, las más ricas
Selección por tipología: respuestas correspondientes a las observaciones situadas en cierta parte de un plano factorial.
Selección por proximidad: todas las frases de un determinado formato situadas a una distancia inferior a un conjunto lexical determinado por el usuario.
Cada selección puede a su vez ser objeto de una selección. Estas posibilidades permiten asimismo tomar conocimiento por "sobrevuelo selectivo" del contenido de las respuestas, del navegar de una cuestión a la otra, de un individuo al otro, de una selección a la otra, de imprimir o de guardar en un fichero texto el verbatim así seleccionado.

Análisis de contenido y de discurso asistido por computadora

Luego de haber tomado conocimiento de las respuestas, será necesario analizar el contenido. La recodificación de las preguntas abiertas (cerrar a posteriori) es un trabajo que puede ser muy fastidioso a fuerza de repetitivo: se trata de acelerarlo.
En algunos casos el procedimiento puede consistir en sustituir el texto de una respuesta por ciertas palabras (ya seleccionadas en el léxico) que ellas contienen. Esta técnica se adapta especialmente al caso en que las respuestas son datos del tipo asociaciones de palabras: nombres de productos, de personas o de lugares obtenidos en las preguntas de conocimiento notorio, calificativos que describen ciertos objetos, etc. El contenido de la respuesta se reduce a las palabras que ellas contienen. En todos estos casos la codificación puede hacerse sin riesgo de manera automática a partir del léxico. En otro tipo de respuestas más complejas naturalmente esta técnica no será tan adecuada.

Existen algunos indicadores léxicos cuyo objetivo es caracterizar una respuesta mediante un indicador cuantitativo que puede resumir la forma y el contenido igualmente.
Tamaño del corpus y de las respuestas: se evalúa en número de palabras, corresponde a la longitud del texto total o del de cada respuesta. Este indicador es muy útil para diferenciar los comportamientos de las respuestas según las cuestiones o la identidad de los entrevistados. El tamaño puede ser restringido a

las palabras de un diccionario especializado (los términos del placer, de la insatisfacción, de la personalización…) e indicar entonces una intensidad semántica según que las palabras de los campos considerados sean numerosas o raras. Se tratará entonces de un indicador de contenido.
Riqueza del vocabulario: la riqueza del léxico es igual al número de palabras diferentes que componen el léxico. Más la riqueza es grande, más el vocabulario es abundante y la información producida será más rica.
Banalidad de la respuesta: la banalidad se aplica a una respuesta y es igual a la ocurrencia media de las palabras que componen la respuesta. Más la respuesta está formada de palabras que aparecen muy frecuentemente (de ocurrencia elevada) más banal es la respuesta.
Balance lexical: hacer un balance léxico es establecer globalmente, o según determinadas modalidades de una respuesta cerrada, las características léxicas de las respuestas aportadas. Tamaño, riqueza, intensidad de determinado campo lexical.

Estadística léxica y análisis de datos

La estadística léxica se apoya sobre la teoría de los actos de lenguaje. Consiste en analizar la elección de las palabras como un conjunto de micro decisiones, en general inconscientes pero reveladoras. Mientras que la estadística numérica se refiere al individuo que se expresa, la estadística léxica desciende al nivel de los actos de lenguaje, y aparece así como una micro-estadística. Se refiere al conjunto de decisiones, conscientes o no, que presiden la elección de las palabras que componen el enunciado. Al mismo tiempo los umbrales que condicionan el trabajo estadístico se encuentran desplazados. En efecto, más allá del discurso del interlocutor que puede ser único, es sobre la gran cantidad de sus actos de habla que se debe orientar el estudio.
Cuáles son los términos más utilizados, las asociaciones, los encadenamientos importantes, en qué medida la elección de las palabras depende de la identidad del que la expresa, o del tipo de documento analizado.
Las técnicas de análisis de datos aplicadas a las tablas de frecuencias léxicas permiten revelar ya sea los fenómenos que

permanecen imperceptibles como los que se muestran en la superficie significativa del texto.

Estadística de los orígenes textuales: el objetivo es la identificación en relación con los formatos fijos (preguntas cerradas o numéricas) de las características del estrato de individuos que utilizan un determinado término (distribución en valores absolutos y porcentajes, comparación de la distribución en la muestra total, test de Chi-cuadrado y diagnóstico sobre la importancia de los desvíos)

Estadística de las asociaciones: se trata de analizar para una pregunta abierta dada, las asociaciones de palabras de una misma respuesta, de una misma frase o de una misma unidad de significación.

Estas estadísticas se establecen a partir de las frecuencias de asociación entre una o muchas palabras pivotes o utilizando el análisis factorial de correspondencias. Como veremos más adelante, este procedimiento consiste en analizar una tabla que posee en filas los individuos interrogados y en columnas los términos del vocabulario.

Este análisis conduce a la definición de sub-conjuntos o tipos en los cuales la característica es asociar ciertas palabras de manera privilegiada. Estos tipos pueden también ser el origen de una nueva selección.

Estadística de los encadenamientos: se trata de establecer entre las respuestas a dos o más preguntas abiertas los encadenamientos de términos. Por ejemplo qué encadenamientos se pueden formar entre las respuestas a una pregunta sobre gustos y otra sobre hábitos.

Se trata de buscar las asociaciones entre una pregunta y la otra y no ya al interior de una misma pregunta. Las técnicas estadísticas utilizadas son las mismas: frecuencias de encadenamiento o asociación y análisis factorial de correspondencias.

Existen programas que combinan los procedimientos de navegación del texto con los de análisis factorial.

La idea de tratar textos con métodos estadísticos no es nueva. Pueden citarse las investigaciones lexicométricas del Laboratorio de Lexicología Política de la Escuela Normal Superior de Saint Cloud. Ejemplos del empleo de la estadística para estudiar el léxico de un texto son en Francia los trabajos de P. Guiraud[42] y de

[42] Problèmes et méthodes de la statistique linguistique, París, P.U.F.,

Ch. Muller[43], citados por Maingueneau (1989:29) En los primeros textos se aplicaban los métodos estadísticos elaborados para tratamiento de variables continuas, sin embargo con los métodos de análisis de datos se obtienen resultados más interesantes.

El Análisis de Datos Textuales es una aplicación de los métodos de Análisis Multidimensional de Datos en la perspectiva de la escuela francesa, es decir métodos exploratorios, que son llamados asimismo métodos estadísticos lexicométricos.
Las primeras aplicaciones en este enfoque fueron realizadas por J.P. Benzécri, quien desarrolló el análisis de correspondencias para discutir las tesis de Chomsky sobre la lengua, aunque sus primeros trabajos fueron realizados con datos numéricos. Posteriormente, los aportes de Ludovic Lebart[44], se dirigieron a resolver los problemas de tratamiento de preguntas abiertas en las encuestas sobre condiciones de vida y aspiraciones de los franceses, utilizando métodos más automáticos que la post - codificación manual que hasta ese momento se hacía y que en muchos casos aún se sigue realizando.

El Análisis de Datos Textuales consiste en aplicar los métodos multidimensionales, no sólo el análisis de correspondencias sino también los métodos de clasificación, a tablas específicas creadas a partir de datos textuales. Estos métodos se completan con otros métodos propios del dominio textual como los glosarios de palabras, las concordancias y la selección del vocabulario más específico de cada texto, para así proveer una herramienta comparativa de los mismos.
El ámbito en el cual es más fácil aplicar estos métodos son las preguntas abiertas de las encuestas y en general a numerosos textos cortos, como los textos políticos o las entrevistas en profundidad. Si bien el campo de aplicación es bastante amplio, es deseable contar con textos que presten cierto grado de homogeneidad y de exhaustividad en el tema a estudiar.
El principio fundamental de estos métodos es el análisis a través

---

1960 y Les caractères statistiques du vocabulaire, P.U.F., 1954.
[43] Essai de statistique lexicale : le vocabulaire du théatre de P. Corneille, Larousse, y La statistique linguistique , Hachette, 1974 (colección «Langue-Linguistique-Communication »).
[44] A partir de su trabajo en el CREDOC (Centre de recherche pour l'etude et l'observation des conditions de vie)

de la comparación. Se busca comparar entre sí el discurso de los individuos que han contestado a una encuesta, o de grupos de individuos generalmente formados a partir de la respuesta a una pregunta cerrada. Por ejemplo: el lenguaje de los hombres con el de las mujeres, el lenguaje de los jóvenes con el de los mayores.
En un ámbito literario permite comparar textos a los efectos de identificar estilos. La comparación implica llegar eventualmente a clasificar a los individuos en clases homogéneas en cuanto al vocabulario empleado aunque también puede interesar clasificar palabras.
Estos métodos pueden resumir los textos mediante las palabras más características y mediante las respuestas o frases más características. En este sentido sería otra forma de construir un verbatim.
Un objetivo importante es conectar las repuestas abiertas con toda la información proporcionada por las respuestas cerradas o las variables categóricas relativas a características contextuales de los individuos.
El conjunto de las respuestas abiertas a una pregunta de encuesta o entrevista constituye lo que llamamos, siguiendo a los lingüistas, el “corpus estudiado”. Sin embargo, éste es un corpus particular en el sentido del tratamiento que se hace del mismo.
Los métodos estadísticos lexicométricos se proponen como sistemáticos, en el sentido de que cuentan la presencia de las palabras sin una selección a priori. Son exhaustivos, porque trabajan a partir del texto de todas las respuestas y por lo tanto se dice que son métodos que permiten una mayor ‘objetividad’, aunque en realidad lo que sucede es que posibilitan retrasar la intervención del investigador hasta una fase más tardía del trabajo.
El programa de análisis textual más utilizado en este campo ha sido el SPAD.T (Lebart, Morineau y ot, 1989), que se complementaba con el SPAD.N (numérico). En la actualidad ambos programas se agrupan en un mismo sistema SPAD en la versión 4.5 ya citada.

Un corpus particular: las respuestas a preguntas abiertas en encuestas

Las respuestas a preguntas abiertas, también llamadas respuestas libres, son elementos de información muy específicos,

que pueden desconcertar a la vez a los estadísticos y a los especialistas de estudios textuales. Los primeros pueden decepcionarse por el carácter impreciso y multiforme de estas respuestas, los segundos por su carácter artificial, y su fuerte redundancia global.
El efecto de la repetición y más generalmente el de la frecuencia con la cual se emplean las formas es muy particular. Las frecuencias léxicas observadas son por una parte artificiales, ya que la misma pregunta se hace a centenas o miles de personas, por lo tanto la yuxtaposición de las respuestas constituye un texto redundante por construcción, donde no son raros los estereotipos.
Sin embargo, las preguntas abiertas constituyen una prolongación indispensable de los cuestionarios cuando las encuestas van más allá de un simple sondeo electoral, cuando se trata de explorar y profundizar un objeto complejo o poco conocido.
Repasaremos algunas de las ventajas e inconvenientes de este tipo de instrumento a los efectos de destacar la problemática de su tratamiento estadístico.

Características de las preguntas abiertas

En los manuales de metodología se ha dedicado mucho espacio a la conveniencia o no de la utilización de las preguntas abiertas comparándolas con las preguntas cerradas, recordemos algunas.
Las ventajas son bastante evidentes, condicionan menos al encuestado, permiten explorar, conocer lo que ha entendido cuando se le hizo otra pregunta. Muchas veces una pregunta dada se entiende de forma muy distinta según la formación del encuestado, o su lugar de origen y es necesario entonces saber a qué pregunta contesta en realidad el entrevistado, si a la que entendió o a la que el entrevistador “cree” que hizo.
Las preguntas abiertas son rápidas de contestar; por ejemplo, si se quiere preguntar: ¿Cuáles son las actividades que Ud. realiza los fines de semana?, y si se quiere abarcar todas las posibles actividades, hay que prever largas listas de ítems. Es más agradable para el entrevistado dejarle exponer sus actividades y hacerle perder así menos tiempo. Estas respuestas abiertas permiten al entrevistado expresarse de forma libre y por lo tanto dar mayor riqueza a su expresión.
Sin embargo este tipo de preguntas también presentan desventajas. Una de ellas es que el encuestador puede tener una influencia bastante grande, en particular, en la anotación de la

respuesta.
Asimismo la rapidez que se ganaba al preguntar se puede perder en el tratamiento, ya sea pos-codificando manualmente estas respuestas o tratándolas con lo métodos de análisis textual, que son interesantes pero no especialmente rápidos. Todo esto aumenta el costo del tratamiento.
Finalmente cabe mencionar también que si la pregunta abierta no concierne directamente al entrevistado puede ser que éste malinterprete la pregunta. Evidentemente esta posibilidad no se le ofrece en una pregunta cerrada donde las respuestas están sugeridas.

¿Se pueden comparar las modalidades de respuesta abierta y cerrada?

Se sabe que, en el cuestionario de una encuesta de actitud o de opinión, la *etiqueta* de una pregunta juega un rol fundamental. De hecho, es muy difícil encontrar dos etiquetas distintas, para dos preguntas cerradas con contenidos similares, que den los mismos resultados en términos de porcentajes de las diferentes respuestas posibles. Algunos autores hasta rechazan interpretar estos porcentajes de respuestas en el sentido de votos asignados y no autorizan a interpretar sus variaciones más que por categorías o en el tiempo. Existen muchos estudios comparativos de resultados relativos a una misma pregunta realizada con las modalidades abierta y cerrada.
Los trabajos de Rugg (1941)[45] citados por Lebart (1994:25) mostraron que la respuesta "sí" a la pregunta: "¿Piensa Ud. que los Estados Unidos debieran prohibir los discursos públicos contrarios a la democracia?" obtuvo 21 puntos (sobre 100) de menos que la respuesta "no" a la pregunta: "¿Piensa Ud. que los Estados Unidos deberían permitir discursos públicos contrarios a la democracia?"
Esta ausencia de simetría entre las dos formulaciones, verificada en otros temas, es tanto más fuerte cuando el nivel de instrucción del entrevistado es menor. En un estudio realizado sobre hábito de fumar en jóvenes argentinos, en la ciudad de Rosario, Moscoloni(1988) encontramos que las preguntas de opinión que incluían la palabra "prohibir" eran contestadas con "no" con mayor

[45] Rugg D. (1941) Experiments in wording questions, Public Opinion Quarterly, 5, p.91-92.

frecuencia por los jóvenes con mayor nivel de instrucción.

Es importante saber que la información proporcionada por las preguntas abiertas y cerradas no es comparable, y es fácil verlo a través de estos ejemplos citados por Lebart:
Schuman y Presser (1981)[46] en un estudio sobre la formulación de las preguntas mencionan este ejemplo realizado en Estados Unidos. Se hizo la pregunta siguiente ¿Cuál es el problema más grande que debe afrontar USA? El ítem violencia estuvo mencionado en un 16 % de los entrevistados cuando la formulación se hizo abierta y en un 35 %, es decir más del doble, cuando la formulación se hizo cerrada.
La explicación que dan Schuman y Presser es que la violencia se considera en forma natural como un problema local y no nacional y es por esto que no se piensa tanto en ella cuando la pregunta se hace abierta. Evidentemente cuando se sugiere como respuesta, mediante un ítem cerrado -la violencia-, los entrevistados consideran que esta respuesta es posible y legítima y tienden a emplearla más.
Similarmente, por ejemplo, si se quiere conocer cuáles son las revistas leídas la semana anterior por un entrevistado y si se le hace una pregunta totalmente abierta es posible que olvide ciertos títulos. Esto se debe a que hay un problema de memoria, entonces en este caso no se obtendrán las mismas respuestas en las dos formulaciones.
Otro ejemplo del mismo tipo puede ser encontrado en la encuesta sobre el nivel de vida y las aspiraciones de los franceses (Lebart, 1987:25-26), a propósito de una pregunta relativa al gasto público. En los años 1983, 1984 y 1987 se preguntó de forma abierta "¿cuáles son los rubros en los cuales la sociedad gasta más?" Se hizo la misma pregunta pero cerrada para los años 1985 y 86. En los años en los cuales la pregunta se hizo abierta el porcentaje de entrevistados que citaban a los inmigrantes era de 4, 5, hasta un máximo de 8 %. En los años en los cuales se hizo la pregunta cerrada este porcentaje fue de 28 y 30 %. Estos porcentajes no son comparables; ya que proponer un ítem es también legitimarlo, es decir sugerir esta respuesta.

[46] Schuman H. Presser F (1981) Questions and Answers in Attitude Surveys, Academic Press, New York.

¿Cuándo abrir una pregunta?

Se puede abrir una pregunta para ahorrar tiempo en la entrevista por ejemplo, cuando se trata de explorar las actividades de fin de semana.

Dar detalles sobre la respuesta que se dio a una pregunta cerrada, es una forma de indagar sobre la pregunta real a la que contestan los individuos. Por ejemplo, en una encuesta sobre las relaciones entre los chicos y chicas con las matemáticas se hizo esta pregunta cerrada:

¿Cuál es su sentimiento hacia las matemáticas?

- lo detesto
- me gustan poco
- me gustan bastante
- adoro las matemáticas.

Cuatro ítems cerrados de respuestas posibles y después de la pregunta cerrada una simple pregunta abierta ¿Por qué? Esto permitía entender de forma mucho más rica cuál era el sentimiento hacia las matemáticas.

Este ¿por qué? es útil en particular cuando se quiere saber qué entienden las personas cuando se les hace una cierta pregunta. Puede ser muy útil en las encuestas internacionales en que se deben traducir preguntas y resulta necesario asegurarse de que la pregunta sea realmente la misma en los distintos países.

Finalmente otra utilización importante es para criticar la calidad de la información. Por ejemplo presentar al final de un cuestionario una pregunta abierta ¿qué le ha faltado a este cuestionario?, o bien ¿qué le ha parecido este cuestionario?

Otro caso donde se aconseja abrir una pregunta es cuando se trata de captar una información espontánea por naturaleza. Por ejemplo si se pasa un video publicitario y se quiere conocer qué se recuerda del video, poner ítems cerrados sería realmente sugerir respuestas, cuando por el contrario, lo que se busca es indagar sobre el recuerdo espontáneo.

Problemas que se plantean al cerrar las respuestas abiertas

Se suele cerrar la pregunta abierta mediante una post-codificación que se realiza de la forma siguiente. Se lee un cierto porcentaje de las respuestas, por ejemplo en una encuesta de 2.000 individuos,

se leen 200 respuestas, anotando los ítems o categorías más frecuentes. En este caso el codificador buscará también establecer correspondencias entre palabras que refieran a una misma categoría. Habiendo identificado los principales ítems presentes, se hará corresponder a la pregunta abierta una o varias preguntas cerradas, según la presencia o ausencia de los ítems.
Una post-codificación bien hecha es muy interesante. No necesariamente es superada por un tratamiento de análisis textual, pero el primero es un procedimiento extremadamente costoso, en particular, porque no se le puede pedir a un sociólogo que pase su tiempo post-codificando respuestas abiertas. Finalmente el trabajo suelen hacerlo estudiantes, o en general gente que acepta un pago bajo, ya que en las encuestas los presupuestos suelen ser limitados.
Además hay dos problemas importantes suscitados por esta post-codificación: el olvido o la eliminación de los ítems poco frecuentes, sin conocer a priori si esos ítems son característicos de cierto sub-grupo. Los ítems poco frecuentes en general no son muy interesantes para el estadístico, pero si son mencionados siempre por personas que pertenecen a un mismo grupo en sí poco numeroso, será interesante conservarlos.
Otro problema difícil de resolver es cómo tratar las respuestas complejas. Por ejemplo ante la pregunta: ¿cuáles son los problemas que más le inquietan en lo que concierne al porvenir?; una persona contesta: el trabajo para mis hijos, el resto me importa poco. En este caso la primera parte de la respuesta se codifica muy fácilmente, en cambio el sentido de exclusión de todos los otros ítems que se manifiesta en la última parte, es difícil de conservar.

Selección de las unidades: las formas gráficas

El método estadístico se apoya en medidas y conteos a partir de objetos o unidades de análisis que se quieren comparar. Contar esas unidades y sumarlas significa desde un cierto punto de vista, al menos en el contexto de una experiencia, que ellas son ocurrencias idénticas de un mismo tipo o de una *forma* más general. Para someter una serie de objetos a comparaciones estadísticas hace falta primeramente, definir una serie de nexos sistemáticos entre casos particulares y categorías más generales.

En la práctica, la aplicación de estos principios implica definir una *norma* que permita aislar de la cadena textual las diferentes unidades a cuantificar.
La operación que permite desglosar el texto en unidades mínimas (es decir en unidades que uno no va a descomponer más adelante) se llama *segmentación del texto*. A esta fase, que permite dividir el texto en unidades distintas, sucede una fase de reagrupamiento de las unidades idénticas: la fase de *identificación* de las unidades textuales.
Para un mismo texto, las diferentes normas de procesamiento no conducen a los mismos conteos. Para cada dominio de investigación en particular no presentan el mismo grado de pertinencia, ni las mismas ventajas e inconvenientes en cuanto a su puesta en práctica.
En todas las máquinas capaces de capturar y archivar textos, se dispone de un sistema de caracteres que consisten en general de una centena de elementos. Entre estos caracteres, algunos corresponden a las letras del alfabeto: mayúsculas, minúsculas, con acento o diéresis, etc., propias del idioma de que se trate, otros sirven para codificar las cifras, otros para signos como porcentajes, unidades de monedas, y hay otras que sirven para codificar los diversos signos de puntuación usual.
En la práctica, un conjunto único de normas tipográficas se pone en acto para realizar la operación que se denomina la *codificación* de los textos. Estas normas sufrieron y sufrirán cada vez más los efectos de los progresos tecnológicos en el dominio de la captura de los textos.
El desglose en formas gráficas constituye un medio particularmente simple de formar las unidades textuales a partir de un corpus de textos. Dependiendo de los objetivos del estudio se podrá seguir una secuencia de etapas: verificación de la captura del texto, primera mirada al vocabulario como base de las comparaciones estadísticas a realizar.
Para hacer una segmentación automática del texto en ocurrencias de formas gráficas, es suficiente elegir entre el conjunto de caracteres un subconjunto que se designa con el nombre de *caracteres delimitadores* (en general el blanco, los signos de puntuación). Los otros caracteres contenidos en el formato serán de esta manera considerados como caracteres no – delimitadores. Una serie de caracteres no – delimitadores (en general letras o números) limitados en sus dos extremos por caracteres

delimitadores es una *ocurrencia*. Dos series idénticas de caracteres no – delimitadores constituyen dos ocurrencias de una misma *forma.* El conjunto de las formas de un texto constituye su *vocabulario.*
La segmentación así definida permite considerar el texto como una serie de ocurrencias separadas entre ellas por uno o varios caracteres delimitadores. El nombre total de ocurrencias contenidas en un texto es su *tamaño* o su *longitud*.
Este enfoque supone que a cada uno de los caracteres del texto corresponde un significado y sólo uno, es decir que el texto ha sido depurado de ciertas ambigüedades de significado, como por ejemplo las mayúsculas al principio de las oraciones o los puntos al interior de las siglas (Ej. U.N.R.).
Es decir que todo nos lleva a pensar la unidad estadística con la cual se va a trabajar como la palabra, contar las palabras presentes en el corpus y comparar la frecuencia de cada palabra en los distintos grupos. Sin embargo no se trabaja con las entradas de un diccionario, sino con las palabras tal como vienen escritas, por lo tanto singulares y plurales de un mismo sustantivo son formas distintas así como las distintas inflexiones de un verbo. Por ello es más adecuado hablar de forma gráfica como la unidad básica de recuento que empleamos.
A posteriori se puede *lematizar*, es decir reagrupar las distintas inflexiones de un verbo en el infinitivo; el singular y el plural de un sustantivo en el singular; el masculino y el femenino de un adjetivo en el masculino. Pero suele ser interesante no hacer esta lematización en un primer tiempo para ver por ejemplo qué información contiene la elección de un cierto tiempo de un verbo, o del singular o el plural en un sustantivo.
Por ejemplo en el caso de entrevistas realizadas a docentes la referencia que ellos hacen cuando hablan de “el conocimiento” no es la misma que cuando dicen “los conocimientos”.
Se aconseja entonces dejar la lematización para una fase ulterior. El software SPAD permite una lematización manual, es decir crea listas explícitas de equivalencias.

Tratamiento textual básico

Presentación del corpus ejemplo

Para presentar la aplicación de los métodos retomaré el ejemplo

del capítulo 7, sobre una encuesta realizada en una ciudad de la provincia de Santa Fe, Argentina. Se trataba de indagar la opinión de los ciudadanos acerca de la gestión gubernativa municipal.
El cuestionario constaba de una serie de preguntas cerradas relativas al tema, entre ellas en una se preguntaba "¿Qué imagen tiene Ud. del Concejo Municipal?" y luego mediante una pregunta abierta se indagaba el por qué.
En primer lugar al recibir el corpus, es necesario saber qué dicen globalmente los entrevistados. Contestaron a esta encuesta 295 individuos. El corpus tiene una longitud total de 3108 ocurrencias y está formado por 735 formas distintas.
Generalmente no se trabaja a partir de todas las formas, sino únicamente de las formas repetidas un cierto número de veces; en este ejemplo hemos elegido el umbral de frecuencia igual a 2, es decir conservar las formas empleadas al menos 3 veces por los entrevistados.
Se comparan los individuos, a partir de lo que tienen en común, las formas empleadas una única vez no permiten comparación ninguna. Si se conservan las formas pronunciadas al menos 3 veces se reduce mucho el glosario de palabras. En nuestro caso, nos quedamos con 102 suprimiendo además las formas herramientas, es decir ciertas formas gramaticales, como los artículos, las proposiciones o conjunciones que tienen una frecuencia muy alta pero que en realidad no aportan significado comparativo.

Glosarios

El cuadro siguiente señala resultados habituales. Las formas más frecuentes aquí son*: no, que, de, la, se*, que tienen una frecuencia superior o casi igual a 80 ocurrencias: son palabras herramientas. Luego viene *problemas* y otras palabras que tienen un cierto contenido semántico: intendente, concejo, son palabras que refieren a la pregunta anterior. Es un efecto habitual que los entrevistados recojan las propias palabras del cuestionario para contestar. También llama la atención la alta frecuencia de la negación. Cuando uno trabaja sobre un tema es necesario leer la lista de palabras, el glosario, reordenadas por frecuencia, como es el caso aquí, o bien ordenadas alfabéticamente.
En la tabla siguiente se han suprimido gran cantidad de palabras con distintas frecuencias a los efectos de aligerar el listado.

Lista de palabras según frecuencia		
Palabras empleadas	**Frecuencias**	**N° de letras**
No	252	2
Que	114	3
De	102	2
Se	84	2
concejo	24	7
problemas	22	9
…	…	…
conozco	16	7
tienen	14	6
trabajan	13	8
concejales	10	10
ellos	9	5
yo	9	2
sueldos	5	7
deberían	4	8
…	…	…
quien	3	5
tratan	3	6
…	…	…
cosa	2	4
personal	2	8
…	…	…
televisión	1	10
Quiere	1	6

Estos métodos operan mediante una especie de deconstrucción del texto y éste es el primer resultado; pero es una deconstrucción que permite poner en evidencia signos totalmente transparentes al investigador cuando recorre el texto en su propia lengua. Su utilidad reside en una primera aproximación al texto y si el corpus no es demasiado extenso sirve también de control.
Evidentemente falta aquí el contexto en el cual se emplean las palabras. Las palabras pueden cambiar totalmente de significado según el contexto.

Segmentos repetidos

Una primera manera de acercarse al contexto es mediante el

glosario de los segmentos repetidos, es decir sucesiones idénticas de palabras repetidas en el cuerpo. He extraído algunos de los que sistemáticamente se pueden obtener de la lista de todos los segmentos repetidos con una cierta frecuencia dentro del corpus.
Aquí se ve un poco mejor el contexto de las palabras, el sentido que pueden tener en este corpus. Hay que tener en cuenta que la pregunta abierta "¿por qué?" fue contestada tanto por los que respondieron que tenían una imagen positiva del Concejo Municipal como por los que no lo tenían, por lo tanto las respuestas se referirán a ideas a favor tanto como a ideas en contra de la labor del Concejo Municipal.
Entre las primeras pueden citarse por ejemplo: "se ocupan de la gente", "hacen cosas", "hacen lo que pueden", "nunca tuve problemas", entre las segundas "no hacen nada", "no se ocupan", "no tienen capacidad". Claro que todavía con estos elementos no podemos evaluar el peso real de unas frases y otras y sus relaciones, aunque se podría ir observando que las segundas son más frecuentes.
Para volver sobre el problema del sentido de las palabras, podría haber un segmento que aquí no está, que fuera por ejemplo "son muy eficientes". Es comprensible, dado el contexto general de crisis del país en el momento en el cual fue realizada la encuesta. Se da entonces que las respuestas más positivas tienen que ver con un desempeño de la autoridad municipal en realidad mediocre.

Es curioso ver como las palabras cambian enormemente de sentido y también está claro que cuando leemos un texto aportamos a la lectura todo un conocimiento exterior al texto que nos permite interpretarlo. Es una forma de subrayar los problemas que se plantean en la comprensión del lenguaje natural, a partir del tratamiento de forma automática realizado por los programas computacionales.

Lista de segmentos repetidos por orden de frecuencia	
Frecuencia	**Texto del segmento**
33	no hacen
31	no se
28	porque no
23	no hacen nada
10	de la gente
7	el concejo no

7	no se ocupan
7	no los conozco
5	no se puede
4	hacen lo que
4	no trabajan
3	trabajan bien
3	la desocupación
3	nunca tuve problemas
3	los políticos no
3	porque no hacen nada
2	piensan en sus bolsillos
2	no cumplen su función
2	porque son muchos
2	hacen lo que pueden

Concordancias

Otra herramienta muy útil para entender mejor el sentido de las palabras son las *concordancias*. La forma "concejo" forma parte del tema de la pregunta y es interesante observar entonces su contexto. En la tabla siguiente se consignan las concordancias encontradas.

Concordancias de la palabra concejo		
el	**concejo**	está integrado por malas personas
el	**concejo**	podría trabajar mejor
el	**concejo**	hace política y no soluciona problemas
el	**concejo**	no presta atención a la gente
con la gente y resuelve personalmente los problemas el	**concejo**	no responde
el	**concejo**	trabaja mucho en la cultura del municipio
el	**concejo**	no trabaja para la gente
el	**concejo**	no busca en el pueblo los problemas a solucionar
el	**concejo**	en las sesiones se pelea por cuestiones políticas y no
el	**concejo**	es desordenado el interés particular es mayor que el
en el	**concejo**	traban a concejales y los demás no

del	**concejo**	trabajan algunos concejales
el	**concejo**	no trabaja demasiado para el pueblo
el	**concejo**	no se ocupa de la gente
considera que tiene una función corporativa y no de	**concejo**	
la imagen de un	**concejo**	que no hace nada
no hacen nada el	**concejo**	no existe
el	**concejo**	trataría de que la ciudad volviera a ser la que fue antes
el	**concejo**	trabaja poco
adentro del	**concejo**	algunos hacen algo bien otros lo hacen por política y otros
me gusta el desempeño del	**concejo**	si tuvieran más presupuesto harían más cosas
sueldos muy altos para estar calentando una silla en el	**concejo**	
que sé yo punto no se mucho del	**concejo**	punto bueno no es

Pero también puede ser interesante encontrar las concordancias de otras palabras que pueden añadir significado por oposición, en este caso por ejemplo, era importante examinar las de la palabra "intendente", ya que se tenía la impresión de una diferencia en la percepción de la gente con respecto a la labor de los dos órganos gubernamentales: el ejecutivo y el legislativo.

Concordancias de la palabra: intendente

el	**intendente**	se ocupa de la gente
el	**intendente**	le consigue trabajo a los conocidos
al	**intendente**	no se lo ve en la calle
el	**intendente**	es de perfil bajo recorre el pueblo es humilde se
el	**intendente**	se relaciona con la gente y resuelve personalmente los
hay nepotismo por parte del	**intendente**	
el	**intendente**	trabaja pero no tiene los medios necesarios
estoy vinculado al	**intendente**	
el intendente pago deudas del anterior	**intendente**	a los empleados municipales y trabajan bien
el	**intendente**	es buena persona pero rodeado de mala
el	**intendente**	tendría que recorrer más la ciudad

el	**intendente**	pone muchas condiciones para espectáculos públicos y se
el	**intendente**	trata de bajar la desocupación le busca trabajo a la gente
al	**intendente**	le falta decisión
el	**intendente**	no se hace conocer no recorre la ciudad
el	**intendente**	tiene buenas intenciones pero no hay gente que lo acompañe
el	**intendente**	no se compromete con la gente
no hubo claridad en la reelección del	**intendente**	
el	**intendente**	sí lo hace
creo que al estar este	**intendente**	sus colaboradores son buenos
al	**intendente**	se le fue la mano el control

Comparando ambos grupos de concordancias se ve que las opiniones negativas con respecto al Concejo son prácticamente el doble de las expresadas sobre el intendente. Sería interesante profundizar en el sentido de la cultura política de una comunidad, dando cuenta de la opinión diferencial de la gente en vista a la actuación de una figura política visible, con sus posibles connotaciones demagógicas, y un cuerpo de gobierno del cual no se conoce mucho la función.
Otra forma de explorar estos glosarios, es por ejemplo escoger un aspecto particular o palabras que se correspondan, o ciertas categorías importantes, como por ejemplo elegir verbos y los segmentos que los contienen.

Reagrupamiento de las respuestas en textos artificiales

Una opción de las más interesantes es particionar el corpus en textos artificiales[47], construidos a partir de grupos de respuestas. Estos grupos pueden referirse a categorías de una variable nominal o a combinaciones de ellas.

Por ejemplo, si decidimos reagrupar a las respuestas según la

[47] Se llaman textos artificiales porque no están así construidos en el corpus, sino que son segmentados a partir de las categorías de una variable

edad y el sexo de los entrevistados, obtenemos 8 grupos de respuestas a las que tradicionalmente llamamos textos.

Reagrupamiento según: Sexo * Edad

Etiqueta del grupo	N° de individuos	N° de respuestas
Femenino * Menor 20	9	9
Femenino * 20--30	19	19
Femenino * 30--50	59	59
Femenino * >50	59	59
Masculin * Menor 20	8	8
Masculin * 20--30	30	30
Masculin * 30--50	61	61
Masculin * >50	50	50
Total	295	295

Al reagrupar los individuos y las respuestas según una variable cerrada podemos crear tablas de frecuencia, tablas de contingencias, que serán las tablas sometidas a los métodos estadísticos.
Este tipo de tabla cuenta la frecuencia con la cual se emplea cada forma conservada en cada uno de los textos o categorías de individuos.

Distribución de las formas (Palabras/Segmentos) en los grupos

Etiqueta del grupo	N° de formas	% del total	Media por respuesta	N° de formas distintas	% Formas del grupo	N° de formas conservadas
Femenino * Menor 20	216	5	24	25	11.570	29
Femenino * 20--30	314	7	17	27	8.600	31
Femenino * 30--50	882	20	15	69	7.820	119
Femenino * >50	755	17	13	75	9.930	110
Masculin * Menor 20	95	2	12	12	12.630	15
Masculin * 20--30	364	8	12	44	12.090	65
Masculin * 30--50	991	23	16	68	6.860	120
Masculin * >50	735	17	15	65	8.840	109
Total	4352	100	15			598

Las mujeres jóvenes serían las que expresan respuestas más largas, sin embargo, en el caso de nuestro ejemplo la variable

nominal más interesante para tener en cuenta, será evidentemente la que hace referencia a la imagen positiva o negativa acerca del Concejo Municipal, pero aquí las palabras por respuesta son parejas en las distintas categorías.

Reagrupamiento según: ¿Qué imagen tiene del Concejo Municipal?

Etiqueta del grupo	N° de individuos	N° de respuestas
CMun.mala	67	67
CMun.regular	99	99
CMun.buena	47	47
CMun.no sabe	82	82
Total	295	295

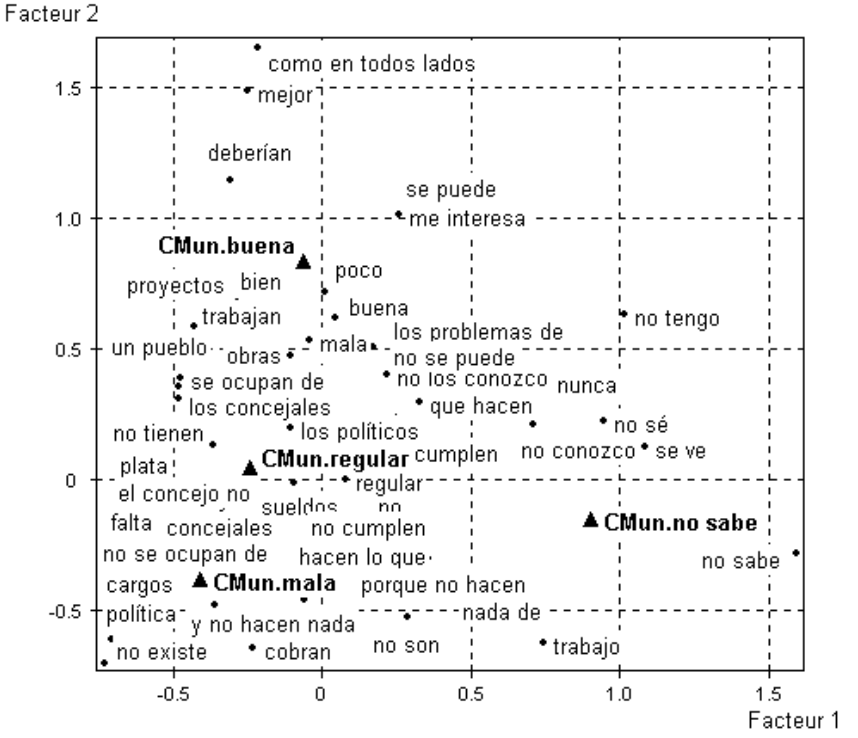
Distribución de las formas (Palabras/Segmentos) en los grupos

Etiqueta del grupo	N° de formas	% del total	Media por respuesta	N° de formas distintas	% Formas del grupo	N° de formas conservadas
CMun.mala	1202	28	18	87.00	7.240	204
CMun.regular	1668	38	17	94.00	5.640	214
CMun.buena	886	20	19	50.00	5.640	86
CMun.no sabe	596	14	7	38.00	6.380	94
Total	4352	100	15			598

Análisis de correspondencias de la tabla léxica

Realizamos una análisis de correspondencias binario de la tabla contingencia que cruza palabras por categorías de respuesta a la pregunta citada y obtenemos el siguiente gráfico factorial.

Proyección de palabras y segmentos en los ejes factoriales 1 y 2



Podemos decir que es la mejor representación plana del contenido de la tabla formas-categorías. Es mejor pero evidentemente no es perfecta.

Lo que se puede decir es que dos categorías próximas en el gráfico emplean más o menos las mismas palabras. A la inversa dos palabras próximas son usadas por las mismas categorías de individuos. Por el contrario dos categorías alejadas emplean un vocabulario muy distinto y dos palabras alejadas vienen empleadas por categorías muy distintas.

En el centro de gravedad encontramos las palabras pronunciadas más o menos con la misma frecuencia por todos los grupos, palabras no diferenciadoras. En la periferia a la inversa, encontramos palabras que diferencian a los grupos de individuos.

De la misma forma, las categorías que están en el centro de gravedad tienen un vocabulario medio, y las categorías más extremas tienen un vocabulario más específico.

Se observa en primer lugar que el eje 1 opone el vocabulario correspondiente a los indiferentes que responden no saber o no

conocer acerca de la actividad del Concejo Municipal, pero a la vez entre los que está presente la preocupación por el 'trabajo', contra los que sí responden con una opinión que se ubica jerárquicamente a lo largo del eje 2 (vertical) buena hacia arriba y regular o mala hacia abajo. Estas dos últimas están casi confundidas y llama la atención que aún los que responden con una imagen buena, sus opiniones no son elogiosas, sino a lo sumo se limitan en todo caso a disculpar el hecho de que "hacen lo que pueden", "como en todos lados".
Es importante destacar aquí la utilidad de la respuesta abierta porque de otro modo hubiéramos podido captar el descontento generalizado con la actuación de la clase política y tal vez hubiéramos interpretado que una proporción de la muestra poseía en realidad una buena imagen del Concejo Municipal.

Palabras y segmentos característicos

Podemos decir que se considera una forma característica de un texto cuando la misma viene sobreampliada en este texto de modo significativo, teniendo en cuenta el modelo hipergeométrico que supone una selección al azar de las palabras.
Si se extrajeran las palabras al azar, la hipótesis nula podría ser que cada categoría emplea más o menos la misma palabra con la misma frecuencia. La hipótesis alternativa sería que hay una selección según las características del individuo y por lo tanto la frecuencia con la cual se observa la palabra en un grupo y en la totalidad de la muestra son totalmente distintas, significativamente distintas.
Según el modelo hipergeométrico a un nivel de significación se asocia una probabilidad.
El nivel de significación que se presenta en las tablas es en realidad una novedad del programa SPAD para evitar tener que leer probabilidades, pues transforma esta probabilidad en el valor que corresponde de una normal central y reducida.
En síntesis, veremos que la diferencia entre las frecuencias es significativa cuando el valor t es mayor que 2 o menor que -2; y también hay que insistir en que esto no es una inferencia, es simplemente una ordenación de valores test a los efectos de individualizar las palabras o segmentos más característicos[48].

[48] Para una exposición más detallada sobre el tema ver: Lebart L. Salem A. (1994) Statistique Textuelle, Dunod, París, pág.172 y sgtes.

En los cuadros siguientes presento las 5 palabras o segmentos más y menos característicos de cada grupo (En la prática se consultan los 10 primeros)

En nuestro ejemplo, las personas que respondieron a la respuesta cerrada con una *calificación mala*, nombraron preferentemente las palabras y/o segmentos: no hacen, no hacen nada.

Vemos que el porcentaje de aparición de 'no hacen' en este grupo es de 2.86 mientras que su porcentaje global en el total del corpus es de 1.36, es por lo tanto un segmento característico del grupo. Estos valores tienen sus correspondientes frecuencias y valores test según el criterio probabilístico que ya expresamos.

Por el contrario las palabras o segmentos menos representativos del grupo de calificación mala, serán los que tengan valores test negativos, siendo el mayor de ellos el término 'no sabe', el cual evidentemente está caracterizando a otro grupo.

Groupe d'individus : CMun.mala

Mots ou segments caractéristiques	Pourcentage interne	Pourcentage global	Fréquence interne	Fréquence globale	Valeur-Test	Probabilité
no hacen	2,86	1,36	17	33	3,147	0,001
no hacen nada	2,18	0,95	13	23	3,065	0,001
nada	3,36	1,81	20	44	2,914	0,002
hacen	3,36	1,98	20	48	2,504	0,006
el	4,71	3,05	28	74	2,476	0,007
lo que	0,17	0,78	1	19	-1,813	0,035
pero	0,34	1,07	2	26	-1,901	0,029
me	0,17	0,86	1	21	-2,035	0,021
sabe	0,17	1,52	1	37	-3,376	0,000
no sabe	0,00	1,40	0	34	-3,786	0,000

Groupe d'individus : CMun.regular

Mots ou segments caractéristiques	Pourcentage interne	Pourcentage global	Fréquence interne	Fréquence globale	Valeur-Test	Probabilité
la	5,02	3,62	48	88	2,815	0,002
trabaja	0,52	0,25	5	6	1,773	0,038
la ciudad	1,15	0,74	11	18	1,619	0,053
realizan	0,31	0,12	3	3	1,545	0,061
respuesta	0,31	0,12	3	3	1,545	0,061
tengo	0,00	0,33	0	8	-2,097	0,018
puede	0,00	0,33	0	8	-2,097	0,018
no	8,57	10,38	82	252	-2,311	0,010
sabe	0,42	1,52	4	37	-3,684	0,000
no sabe	0,21	1,40	2	34	-4,231	0,000

Groupe d'individus : CMun.buena

Mots ou segments caractéristiques	Pourcentage interne	Pourcentage global	Fréquence interne	Fréquence globale	Valeur-Test	Probabilité
acá	1,34	0,37	6	9	2,888	0,002
puede	1,12	0,33	5	8	2,448	0,007
tengo	1,12	0,33	5	8	2,448	0,007
mejor	0,89	0,25	4	6	2,240	0,013
lados	0,67	0,16	3	4	2,023	0,022
mucho	0,00	0,74	0	18	-1,959	0,025
no sabe	0,22	1,40	1	34	-2,374	0,009
sabe	0,22	1,52	1	37	-2,596	0,005
no	6,92	10,38	31	252	-2,670	0,004
nada	0,22	1,81	1	44	-3,013	0,001

Groupe d'individus : CMun.no sabe

Mots ou segments caractéristiques	Pourcentage interne	Pourcentage global	Fréquence interne	Fréquence globale	Valeur-Test	Probabilité
no sabe	7,24	1,40	31	34	9,127	0,000
sabe	7,24	1,52	31	37	8,721	0,000
no	18,22	10,38	78	252	5,419	0,000
opinar	0,70	0,12	3	3	2,546	0,005
no se ve	0,70	0,12	3	3	2,540	0,006
la ciudad	0,00	0,74	0	18	-1,867	0,031
como	0,00	0,74	0	18	-1,880	0,030
con	0,00	0,91	0	22	-2,204	0,014
el	1,17	3,05	5	74	-2,541	0,006
la	1,40	3,62	6	88	-2,799	0,003

Respuestas Características

A partir de la selección de palabras se puede determinar que algunas respuestas son características de cada grupo. Se hace de la forma siguiente.

Para cada una de las respuestas de un grupo o texto, se calcula el valor ‘t’ medio para una palabra de ese texto: cada palabra tiene un valor ‘t’, se suman los valores test de las palabras de cada respuesta, se divide por el número de palabras y se obtiene así el valor ‘t’ medio. A posteriori se reordenan las respuestas desde la que tiene valor ‘t’ más alto, a la respuesta con el valor ‘t’ más bajo, y así se obtienen las respuestas originales más características.

Esto es un retorno al texto que permite entender o precisar de qué forma los individuos hablan en cada grupo. No se trata de decir que la respuesta que está en primer lugar *es la* respuesta característica, sino que es una forma de reordenar las respuestas del grupo y ver en las primeras respuestas cómo se expresan los individuos de este grupo.

Para complementar la interpretación avanzamos con el cálculo de

las 5 (en la práctica se analizan las 10 primeras) respuestas más características correspondientes a cada categoría y que se derivan de las especificidades de palabras o segmentos. Ellas corresponden a respuestas originales y se presentan a continuación.

Grupo de individuos: CMun.mala

Critère de classement	Numéro	Libellé de la réponse
0.745	1	porque cobran demasiado por lo poco que hacen, que esa plata la inviertan en otra cosa.
0.780	2	se tendrían que ocupar más de la gente con escasos recursos. el intendente tendría que recorrer más la ciudad y ver personalmente los problemas. el intendente pone muchas condiciones para espectáculos públicos y se hace difícil hacerlos. el sector de obras públicas es el que mejor trabaja. los demás trabajan bien también.
0.784	3	ocupan cargos sin hacer nada, acá las cosas están siempre iguales, y ellos se llevan la plata.
0.788	4	son muchos y demasiados sueldos. que esa plata la usen para otra cosa.
0.789	5	no tendrían que existir, se llevan toda la plata, y no hacen nada.

Grupo de individuos: CMun.regular

Critère de classement	Numéro	Libellé de la réponse
0.700	1	tiran proyectos que no cumplen. siempre hablan pavadas.
0.723	2	porque no cumplen su función correctamente y cobran sueldos sin hacer mucho.
0.815	3	en zona céntrica se ven todas las mejoras, pero fuera de ella se lucha por una mejora del barrio, y sin respuesta.
0.817	4	ya tienen una respuesta: regular.
0.854	5	y regular, ni buena ni mala, hacen cosas pero hay muchas más para hacer

Grupo de individuos: CMun.buena

Critère de classement	Numéro	Libellé de la réponse
0.784	1	no estoy de acuerdo con embolsar las ramas para que las lleven. no hay maquinarias para arreglar las calles. la gente no ayuda al gobierno para la limpieza. el intendente trabaja pero no tiene los medios necesarios. el concejo trabaja mucho en la cultura del municipio.
0.797	2	siempre atendieron las quejas.
0.797	3	no tengo quejas, a mí me han atendido siempre muy bien.
0.797	4	no tengo quejas.
0.808	5	yo soy jubilada y con eso te digo que no me alcanza con lo que ganamos, pero el gobierno de acá, de esta ciudad, siempre que precisé algo me ayudó, yo personalmente tengo una buena imagen

Grupo de individuos: CMun.no sabe

Critère de classement	Numéro	Libellé de la réponse
0.600	1	no me interesa la política en sí. esta ciudad carece de muchas cosas
0.704	2	no me interesa.
0.704	3	no me interesa
0.704	4	no tengo tiempo y tampoco me interesa.
0.774	5	el intendente trata de bajar la desocupación, le busca trabajo a la gente personalmente.

A través de la lectura de un grupo reducido de respuestas se puede evaluar el sentido general del conjunto del corpus dividido según las categorías de la pregunta anterior.

CAPÍTULO 10

ANÁLISIS DE DATOS SIMBÓLICOS

Introducción

Recordando lo ya expuesto en el capítulo de clasificación sobre intensión y extensión de una clase, se puede retomar el concepto de 'objeto' como una generalización del 'individuo'.

Por ejemplo, si consideramos que el objetivo de una encuesta sobre consumo de drogas es encontrar clases de adolescentes que tengan el mismo comportamiento, o sea determinar grupos de riesgo; se entrevistan adolescentes provenientes de diferentes categorías sociales, preguntando sobre antecedentes familiares, consumo de cigarrillos, de bebidas alcohólicas, de drogas. En términos del AMD clásico se pueden efectivamente construir las clases de la manera más objetiva y modelizando lo menos posible. Sin embargo se desperdician los conocimientos de los expertos psicólogos, sociólogos, psiquiatras, que no pueden ser plasmados en el cuestionario.

Si nos ubicamos en el punto de vista del Análisis de Datos Simbólicos (ADS en adelante) podemos apoyarnos más en el conocimiento de los expertos. Es decir, pedirles por ejemplo que digan a partir de su experiencia sobre el terreno, si tienen una idea de comportamientos posibles de la población. Estos comportamientos serán descriptos en intensión porque en la cabeza de los expertos hay una idea de clases de comportamientos de adolescentes, pero esta idea de clase no está expresada solamente en términos de individuos: Pablo, Juan, etc. Éstas son clases descritas en intensión por un conjunto de propiedades. Si un experto epidemiólogo que ha trabajado 10 años sobre terreno, provee una descripción en intensión de esta clase, será muy importante y necesario utilizar esta noción.

El ADS va a poder hacerlo porque va a permitir escribir, bajo forma matemática, esta clase que le interesa al experto. Y una vez que se tiene la descripción matemática en intensión se va a poder ver cuál es su extensión en la base de datos y aceptar o rechazar esta hipótesis de clase.

Si los expertos son capaces de describir varias clases en intensión vamos a poder crear un nuevo tipo de matriz de datos donde los

objetos ya no son individuos simples sino que son objetos definidos en intensión, por lo tanto son objetos de primer nivel.
El ADS tiene entonces por finalidad hacer estudios de AMD sobre este tipo de objetos definidos en intensión. Este tipo de análisis contiene un caso particular, el caso donde las clases son individuos. Es entonces una extensión y una generalización del AMD clásico.

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 interesándose también en 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 cifras que se refieren a valores de las variables aplicados a las unidades de análisis. Extraer información de una matriz de este tipo es el objetivo del AMD y se sintetiza en reemplazar números por 'conocimientos' nuevos.
Cuando se construye una gran matriz numérica, el análisis de correspondencias permite organizar individuos en un plano y ver simultáneamente cuáles son los puntos que se parecen, los que se oponen, extraer ejes factoriales que tienen sentido para expresar el tiempo, los perfiles, las trayectorias de modalidades etc. Estos conceptos no estaban visibles, a priori, por sí mismos. Se pueden proyectar sobre el plano factorial clases de objetos obtenidos por diversos métodos de clasificación. Por ejemplo, por medio de jerarquías, pirámides, nubes dinámicas, o por métodos de K-medias que dan particiones. Una partición puede ser llevada sobre el plano factorial para ayudar a interpretarlo.
En el caso de la jerarquía los individuos se hallan en la base, como también en el caso de las pirámides, ya que las clases de individuos están representadas por niveles, por el contrario en el caso de un árbol los individuos son los nodos y allí vemos racimos constituidos por individuos que se parecen.
La clasificación y el análisis factorial constituyen la exploración de

base. El AMD ayuda a encontrar el modelo y eventualmente puede actuar, a través de las técnicas de Análisis Factorial Múltiple por ejemplo, sobre varias matrices de datos simultáneamente.
El ADS tiene como objetivo reemplazar los individuos del análisis de datos tradicional por individuos de más alto nivel, más complejos y más aptos para representar conocimientos, porque están definidos en intensión, utilizando el poder de la lógica: son los objetos simbólicos (OS).
También se puede decir que las variables son de más alto nivel en el análisis de datos simbólicos, 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, 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 un intervalo para esta clase de empresas. Si una empresa pertenece a una determinada clase, en la variable beneficio tendré un intervalo de valores (y no uno sólo).
Uno 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, se va a subir un nivel en toda la teoría del AMD.
En la estadística clásica, cuando se define una variable aleatoria, por ejemplo la variable aleatoria x, toma valores en un determinado intervalo y en este caso cuando se escribe:

$$x = [3, 5]$$

se dice que este suceso correspondiente a esta variable aleatoria se refiere a un conjunto de individuos en el espacio Ω (los individuos cuyos valores en la variable x varían entre 3 y 5).
El objetivo es estudiar las leyes asociadas a estas variables aleatorias. Estas leyes rigen sobre los individuos de Ω, siendo Ω el espacio muestral de los valores obtenidos.
En el ADS, Ω no es un conjunto de valores tomados por los individuos sino es el conjunto de eventos o sucesos. Así es como se sube un grado en la teoría.
Teóricamente se deberá demostrar que los axiomas básicos de la teoría de la probabilidad van a tener que subir un grado, es decir en lugar de aplicarse sobre conjuntos de individuos del espacio muestral, se aplicarán a los eventos o sucesos.
En el caso de la clasificación ocurre lo mismo, en lugar de tener

individuos se tendrán objetos simbólicos que podrán ser considerados como eventos. Las clases podrán representar dos cosas como en clasificación clásica: simplemente un conjunto, pero también una intensión. Recordemos que cuando se habla de clase se hace referencia al mismo tiempo a dos conceptos: el conjunto y lo que permite describirla.
Tenemos por una parte objetos que son O.S. pero además las clases también están representadas por objetos simbólicos.
Por lo tanto tenemos que poder explicitarlo, cada nivel estará descrito por un conjunto de propiedades, de manera que el usuario, apoyándose en un nodo podrá ver las propiedades asociadas al mismo.
Estas propiedades asociadas a cada nodo son propiedades de tipo monotético. Es decir que la extensión de las propiedades asociadas a un nodo son el conjunto de puntos que están allí y la intensión de estos puntos son las propiedades asociadas a dicho nodo.
De esta manera las propiedades asociadas a cada nodo son las condiciones necesarias y suficientes que caracterizan las clases que lo sostienen.
Por supuesto los objetos simbólicos contienen como caso particular individuos, puntos del R_p (espacio de los números reales de dimensión p)

La necesidad de la incertidumbre

En ciertos campos la representación booleana del conocimiento es suficiente. Una variable aleatoria en estadística clásica es una información de orden booleano. Cuando se escribe x = [3,5] uno se refiere a todos los individuos que satisfacen las condiciones de la variable x en ese intervalo.
Es booleano, se está o no se está dentro del intervalo, no cabe otra posibilidad. Pero en muchos casos es necesario hacer intervenir la incertidumbre, por ejemplo, uno puede decir que en la comunidad i el color del cabello de las personas es a menudo negro y rara vez rubio. Este evento observado en estadística clásica sería: 'evento color es igual a negro o a rubio'.
En cambio ahora es expresado de manera incierta. Donde se tenía la aplicación de Ω en {falso o verdadero} o sea Ω [0,1], la aseveración 'color: igual a negro o a rubio', correspondiente a verdadero, ahora pasa a ser:

a_i = [color = 0.9 negro, 0.1 rubio]

El objetivo del ADS es permitir que los conocimientos de los expertos sean expresados en los datos mismos y por lo tanto encontrar una expresión matemática que permita transformar las frases que expresan experiencia en forma de datos siendo estos datos de más alto nivel.
En el momento en que los conocimientos se expresan, es evidente que no son probabilidades, cuando uno habla utiliza otras nociones, como "a menudo", "frecuentemente", "rara vez".
La idea es dar la posibilidad a los expertos de transmitir otras nociones para salir del caso booleano, otra cosa diferente de las funciones que expresan solamente probabilidades en forma de frecuencia, o incluso de probabilidades subjetivas, en los dos casos satisfaciendo los axiomas básicos de probabilidad.
En el ADS se usan funciones características y también funciones probabilísticas, pero además existen otros tipos de conocimiento, como el *posibilístico* que derivan de la teoría de los conjuntos borrosos de Zadeh[49].
Asimismo existen otras teorías que son la teoría de las creencias o de las evidencias y la teoría de la posibilidad que proveen axiomas diferentes. Podremos tener objetos booleanos, probabilísticos, posibilísticos y objetos credibilistas o de creencias.
La probabilidad en el caso de la creencia o de las evidencias se escribe de esta forma :

a_i = [color = 0.4 { negro, castaño}, 0.3 { rubio, rojo}, 0.3 {0}]

En este caso el color puede presentarse en un 40% negro o castaño, en un 30% rubio o rojo y en un 30% se ignora. La ignorancia es un caso particular de la teoría de la evidencia.

Otros tipos de semántica

Las teorías citadas expresan, precisamente, diferentes tipos de semántica.
Por ejemplo, en un caso relativo a clasificación de documentos, éstos están caracterizados por palabras claves que se hallan

[49] Zadeh, L.A. (1971) Quantitative fuzzy semantics. Informations Sciences, pp. 159-176.

presentes en cada texto. Se pueden calcular las tres medidas siguientes:
1) Probabilidades:
Se tendría una matriz de datos con las palabras claves del texto y su correspondiente frecuencia. Si uno utiliza la frecuencia, ésta debe cumplir el axioma de Kolmogorov pudiéndose usar métodos basados en este axioma.

$$Pr (E_1 \cup E_2) = Pr (E_1) + Pr (E_2) - Pr (E_1 \cap E_2)$$

2) Posibilidades:
Supongamos ahora un texto que hable de matemáticas y no use nunca la palabra matemática. En el caso de la probabilidad frecuentista, la palabra matemática tendría un valor nulo. Eso es falso, pues un experto sabe muy bien que el texto se refiere a matemáticas o asuntos muy cercanos a las matemáticas.
Por lo tanto se puede hacer intervenir la noción de posibilidad, a través de un experto que dice que tal palabra no apareció pero tiene una alta posibilidad de aparecer.
Si se analiza la noción de posibilidad, el axioma de Kolmogorov es reemplazado por este axioma que ha sido propuesto por Zadeh y en el que se basa o constituye la teoría de la posibilidad, la teoría de los conjuntos borrosos.

$$Pos (E_1 \cup E_2) = Max (Pos (E_1), Pos (E_2))$$

3) Creencias:
Una tercera noción, puede intervenir de la siguiente manera: un experto puede decir "pienso que esta palabra podría haber aparecido y el argumento que tengo para pensarlo me hace decir que este punto tiene la probabilidad p_1 de aparecer y p_2 de no aparecer".

Si $p_1 + p_2 < 1$ entonces
$1 - (p_1 + p_2)$ expresa su ignorancia

El siguiente axioma reemplaza al axioma de Kolmogorov.

$$Bel (E_1 \cup E_2) \geq Bel (E_1) + Bel (E_2) - Bel (E_1 \cap E_2)$$

La regla de Dempster permite la posibilidad de combinar la creencia de varios expertos.(Para la teoría de la creencia se puede consultar Schafer[50]).
Una manera de intuir la noción de creencia en un ejemplo simple es cuando uno piensa que va a llover, por ejemplo 0.3 de posibilidad, sin embargo también hay sol y creo que hay 0.4 posibilidad de que no va a llover. El resto se expresa en ignorancia.

¿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 un conjunto de objetos simbólicos que están expresados por un conjunto de 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 es clara 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 (por ejemplo si es necesario definir una axiomática que traduzca la semántica propia al dominio de aplicación cuando la axiomática de los números y de los conjuntos usuales no convenga más).

62 Schafer, G. (1990),*Perspectives on the Theory and Practice of Belief functions*, International Journal of Approximate Reasoning, Vol.4, Numbers 5/6.

Resulta de esta definición que el análisis de datos clásico trata desde hace mucho tiempo con objetos simbólicos particulares ya que ése es el caso de 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:

a) Cada variable puede tomar valores múltiples para un mismo objeto "simbólico", por ejemplo:
   [opinión = {regular, mala, regular $\wedge$ indiferente}] para expresar el hecho de que una clase de individuos puede tener opinión regular, mala, o regular e indiferente; o [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 describen las computadoras de una empresa que no las posee) 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'

Los OS se distinguen también a nivel de su manipulación:

c) Un objeto simbólico es una descripción en intensión 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

e) 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"

Esquema del ADS

Para situar al AMD con respecto al Análisis de Datos Simbólicos, podemos decir que hay 4 tipos de análisis según sean las entradas que se posean y según las salidas que se obtengan.

SALIDA	**ENTRADA**	
	objetos - individuos (datos standard)	***objetos simbólicos***
numérico	**(a)**	**(c)**
simbólico	**(b)**	**(d)**

- En (a) tenemos datos clásicos en la entrada y tenemos resultados numéricos en la salida. Este caso es el de AMD clásico.
- En (b) tenemos una entrada de datos clásicos y una salida en que obtenemos resultados de orden simbólico u objetos simbólicos. Por ejemplo si tenemos datos clásicos en la entrada se puede efectuar un análisis factorial clásico, una clasificación clásica y deseamos objetos simbólicos en la salida para interpretar ya sea las clases de un cluster y/o los ejes factoriales.

  ¿Cómo transformar una clase obtenida con un método de clasificación clásico en un objeto simbólico?
  Hay muchas formas de hacerlo, pero una simple es tomar una unión de los valores alcanzados en la clase para cada variable. O bien en términos probabilísticos podemos tomar la frecuencia relativa de cada uno de los valores alcanzados en cada una de las clases a partir de la descripción realizada de las mismas. Se podrán así explicar a través de una conjunción de

propiedades.
En el caso del análisis factorial se pueden obtener también objetos simbólicos buscando para un eje los individuos con mayor contribución en dicho eje y de esa forma construir una clase. Teniendo una clase podemos construir un objeto simbólico, tomando la unión de los valores alcanzados por la clase o calculando las frecuencias en la clase. Por lo tanto podemos obtener una interpretación simbólica de un análisis factorial.

¿Cuál es el interés de tener una interpretación simbólica de un análisis clásico?
El interés primordial es el grado de explicación. Allí es donde el experto tiene el objeto en su mente y dice, yo tengo un conocimiento nuevo. Este eje expresa los individuos más bien de cierta edad, que habitan en tal lugar o barrio de la ciudad, tal como se hacía en AMD clásico, pero con el ADS este conocimiento es expresado por un objeto simbólico *cuya calidad puede ser medida.*
Así tenemos una nueva forma de extracción de conocimientos que se agrega a la cantidad que provee el análisis clásico, lo mismo ocurre con la clasificación. En lugar de decir solamente "esto es una clase", tenemos además una conjunción de propiedades que provee el Objeto Simbólico que tiene también un poder explicativo mayor, interesante, que no reemplaza a los resultados habituales o a los indicadores de la estadística clásica, pero adopta una manera nueva e interesante de interpretar los resultados.
Por ejemplo, si retomo las clases del cluster expuestas en capítulo 8, podría construir un objeto simbólico que resumiera la clase 1, utilizando una semántica de probabilidad frecuentística, una aserción como sigue:

OS “Clase1”(86) = [Imagen Gob.Municipal = {“Mala” (0.035), “Regular” (0.128), “Buena”(0.7791), “No sabe” (0.058)}$\wedge$ Imagen Intendente = {“Mala” (0.023), “Regular” (0.035), “Buena” (0.907), “No sabe” (0.035)} $\wedge$ Imagen Con. Municipal = {“Mala” (0.058), “Regular” (0.198), “Buena” (0.407), “No sabe” (0.337)} $\wedge$ Calificación Dep.Ejecutivo = {“Mala” (0.035), “Regular” (0.07), “Buena” (0.512), “No sabe” (0.384)} $\wedge$ Recolección residuos = {“Muy mala” (0.023), “Mala” (0.0), “Regular” (0.047), “Buena” (0.86), “Muy buena” (0.07), “No sabe” (0.0)} $\wedge$

Alumbrado = {“Muy mala” (0.023), “Mala” (0.116), “Regular” (0.326), “Buena” (0.523), “Muy buena” (0.12), “No sabe” (0.0)} ∧ Riego = {“Muy mala” (0.035), “Mala” (0.012), “Regular” (0.233), “Buena” (0.593), “Muy buena” (0.12), “No sabe” (0.058), “No corresponde” (0.058) }∧ Desmalezamiento = {“Muy mala” (0.023), “Mala” (0.128), “Regular” (0.221), “Buena” (0.488), “Muy buena” (0.0), “No sabe” (0.023), “No corresponde” (0.116) }∧ Recolección ramas = {“Muy mala” (0.012), “Mala” (0.081), “Regular” (0.128), “Buena” (0.744), “Muy buena” (0.012), “No sabe” (0.0), “No corresponde” (0.023) }∧ Zanjeo = {“Muy mala” (0.023), “Mala” (0.128), “Regular” (0.256), “Buena” (0.314), “Muy buena” (0.0), “No sabe” (0.056), “No corresponde” (0.221)}∧ Barrido y limpieza = {“Muy mala” (0.023), “Mala” (0.023), “Regular” (0.209), “Buena” (0.535), “Muy buena” (0.012), “No sabe” (0.023), “No corresponde” (0.174)}∧ Mantenimiento de calles de tierra = {“Muy mala” (0.081), “Mala” (0.14), “Regular” (0.36), “Buena” (0.291), “Muy buena” (0.0), “No sabe” (0.035), “No corresponde” (0.093)} ∧ Atención en dispensarios = {“Muy mala” (0.0), “Mala” (0.0), “Regular” (0.12), “Buena”(0.488), “Muy buena” (0.453), “No sabe” (0.047), “No corresponde” (0.0)} ∧ Sexo = {“Femenino” (0.547), “Masculino”(0.453)} ∧ Educación = “Primaria Inc.”(0.023), “Prim. Comp.”(0.221), “Sec.Incomp”(0.151), “Sec.Comp.” (0.442), “Terc.Incomp.” (0.035), “Terc.comp.” (0.093), “Univ.icomp.” (0.0), “Univ.comp.” (0.023), “Post-univ.incomp.” (0.012)} ∧ Activ.Económica = {“Desocupado” (0.023), “Peón-changarín”(0.07), “Serv.domést.”(0.012), “Emp/ob.público” (0.081), “Emp/ob.privado” (0.093), “Cta propia técnico”(0.047), “Trab.fliar” (0.035), “Cta.propia comercio” (0.14), “Patrón” (0.081), “Profesional” (0-047), “Jubilado” (0.081), “Estudiante” (0.07), “Ama de casa” (0.198), “Otro” (0.023)} ∧ Vivienda = {“Propia” (0.814), “Alquilada” (0.128), “Otra” (0.058)} ∧ Tipo vivienda = {“Casa quinta” (0.012), “Chalet pileta” (0.023), “Casa importante” (0.081), “Casa con jardín” (0.349), “Vivienda modesta” (0.477), “Vivienda precaria” (0.058)}]

- En (c) la entrada son objetos simbólicos y se extraen informaciones de orden numérico.
  Este sería el caso de definir y analizar las distancias entre objetos simbólicos. Luego con estas distancias se pueden efectuar cálculos de escalamiento.
- En (d) tenemos en entrada objetos simbólicos y en la salida también objetos simbólicos. Se puede hacer un análisis factorial de objetos simbólicos y efectuar una interpretación simbólica de los ejes.

El Análisis de Datos Simbólico llena un vacío importante que se sitúa entre la inteligencia artificial y la estadística. Entre los enfoques lógicos, simbólicos y numéricos, abre el camino a un gran campo de aplicación: el del procesamiento de objetos complejos teniendo en cuenta conocimientos no necesariamente de orden puramente numérico.

Construcción y procesamiento de OS

La utilización de OS ha obtenido su máximo desarrollo en el marco del proyecto europeo SODAS (Symbolic Official Data Analysis System)[51]

Efectivamente este software provee muy buenas posibilidades de aplicación para la manipulación de bases de datos de estadísticas oficiales.

En el capítulo 4, presenté un ejemplo de matriz de datos simbólicos que retomaré aquí. El ejemplo se refiere a los alumnos ingresantes a la Universidad Nacional de Rosario en el año 2000 que totalizaron 16.026[52].

Objeto	Horas de trabajo				Escolaridad del padre			
Bioq	No (0.72),	H20(0.12),	H21(0.08),	H36(0.08)	Sin(0.05),	Pri(0.38),	Se(0.36),	U(0.19)
Polit	No (0.66),	H20(0.09),	H21(0.12),	H36(0.13)	Sin(0.09),	Pri(0.35),	Se(0.34),	U(0.22)
Odont	No (0.86),	H20(0.06),	H21(0.03),	H36(0.05)	Sin(0.04),	Pri(0.32),	Se(0.38),	U(0.25)
Medic	No (0.81),	H20(0.04),	H21(0.04),	H36(0.11)	Sin(0.09),	Pri(0.39),	Se(0.33),	U(0.20)
Huma	No (0.50),	H20(0.16),	H21(0.17),	H36(0.18)	Sin(0.09),	Pri(0.39),	Se(0.34),	U(0.19)
Psic	No (0.66),	H20(0.09),	H21(0.11),	H36(0.14)	Sin(0.09),	Pri(0.40),	Se(0.34),	U(0.16)
Econ	No (0.69),	H20(0.07),	H21(0.08),	H36(0.15)	Sin(0.07),	Pri(0.42),	Se(0.36),	U(0.15)
Dere	No (0.66),	H20(0.07),	H21(0.08),	H36(0.19)	Sin(0.08),	Pri(0.40),	Se(0.36),	U(0.15)
Agra	No (0.76),	H20(0.13),	H21(0.05),	H36(0.06)	Sin(0.04),	Pri(0.37),	Se(0.41),	U(0.17)
Vete	No (0.82),	H20(0.06),	H21(0.05),	H36(0.07)	Sin(0.10),	Pri(0.43),	Se(0.27),	U(0.19)
Arquit	No (0.78),	H20(0.06),	H21(0.08),	H36(0.09)	Sin(0.06),	Pri(0.30),	Se(0.40),	U(0.23)
Ingen	No (0.77),	H20(0.08),	H21(0.05),	H36(0.10)	Sin(0.04),	Pri(0.31),	Se(0.39),	U(0.25)

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

Nos limitaremos a considerar este ejemplo donde la semántica

[51] SODAS proyecto N° 20281 de la Comisión Europea, Directorio General III, Industrial RTD, EUROSTAT, programa DOSES. En este programa intervienen centros de investigación, universidades e institutos de estadística oficial de la Unión Europea. La primera etapa del proyecto ha finalizado con la construcción del software SODAS. Citado por: Calvo P. y ot. (2000)

[52] Datos provistos por la Dirección General de Estadística de la UNR.

utilizada es la de probabilidades basadas en la frecuencia. En este caso, los valores de las variables están indicando que un alumno en la Facultad de Medicina por ejemplo, tiene una probabilidad igual a 0.81 de no trabajar, y de 0.11 de trabajar 36 o más horas, una probabilidad igual a 0,20 de que su padre haya cursado estudios universitarios completos, etc.

Matriz de datos simbólicos

La matriz simbólica del ejemplo está compuesta por 22 variables nominales y 16026 individuos con los que se construyeron 12 aserciones. En primer lugar el software provee un listado con las características de las variables, sus categorías y su modo, en este caso probabilístico.

```
 variable Fac_
  nominale {"Agra", "Arqu", "Bioq", "Dere", "Econ",
"Enfe", "Fono", "Huma", "Inge", "Medi", "Musi",
"Odon", "Poli", "Psic", "Vete"}
  multiple,mode=probabilist;
variable SEXO
  nominale {"Feme", "Masc"}
  multiple,mode=probabilist;

.....etc.
```

A continuación se listan las aserciones (OS) construidas. Consigno a modo de ejemplo sólo el primero (Facultad de Ciencias Bioquímicas y Farmacia, con 512 individuos, con todas sus variables: la variable Facultad que toma el valor Bioq con probabilidad igual a 1, **y** (^)la variable sexo que toma el valor Masculino con probabilidad igual a 0.3125 y el valor Femenino con probabilidad igual a 0.6875, **y** la variable Residencia, etc.)

```
os "Bioq"(512) =
   [Fac_ = {"Bioq"(1)}]
  ^[SEXO = {"Masc"(0.3125), "Feme"(0.6875)}]
  ^[RESI =
{"Inde"(0.269531),"Univ"(0.0546875),"Pahe"(0.564453),"
Pafa"(0.0195313), "otro"(0.0546875),
"Flia"(0.0371094)}]
 ^[HSTR = {"H021"(0.0820313), "H020"(0.117188),
"H036"(0.078125), "Nono"(0.722656)}]
  ^[PADR = {"Psei"(0.163386), "Psec"(0.234252), "P-
```

```
no"(0.0137795), "Ppri"(0.0433071), "Pprc"(0.222441),
"Punc"(0.188976), "Puni"(0.133858)}]
  ^[MADR = {"Mpri"(0.023622), "M-no"(0.00787402),
"Muni"(0.0984252), "Munc"(0.226378), "Mprc"(0.194882),
"Msec"(0.311024), "Msei"(0.137795)}]
  ^[PROC1 = {"Pro3"(0.0488281), "Pro5"(0.0273438),
"Pro2"(0.292969), "Pro4"(0.195313), "Pro1"(0.435547)}]
  ^[RAMA6 = {"PiyC"(0.126953), "Prnc"(0.0742188),
"Pcom"(0.285156), "Pagr"(0.0859375), "Ppro"(0.123047),
"Pens"(0.0351563), "Pffa"(0.0195313), "Pvar"(0.0625),
"Pser"(0.150391), "Pban"(0.0332031),
"Pdep"(0.00390625)}]
  ^[CATE6 = {"PCnc"(0.0585938), "Pobr"(0.046875),
"Pger"(0.109375), "Pemp"(0.300781), "Pjef"(0.09375),
"Pjub"(0.0410156), "Pdes"(0.015625),
"Pind"(0.333984)}]
  ^[RAMA7 = {"Mpro"(0.0664063), "Mdep"(0.00976563),
"Ment"(0.0761719), "Mvar"(0.0585938),
"Mrnr"(0.509766), "Mban"(0.0078125),
"Miyc"(0.0117188), "Mens"(0.125), "Mffa"(0.00390625),
"Mcom"(0.126953), "Magr"(0.00390625)}]
  ^[CATE7 = {"Mdue"(0.154297), "Mger"(0.0390625),
"Memp"(0.240234), "Mjub"(0.0351563), "Mrnr"(0.412109),
"Mobr"(0.00390625), "Mdes"(0.0976563),
"Mjef"(0.0175781)}]
```

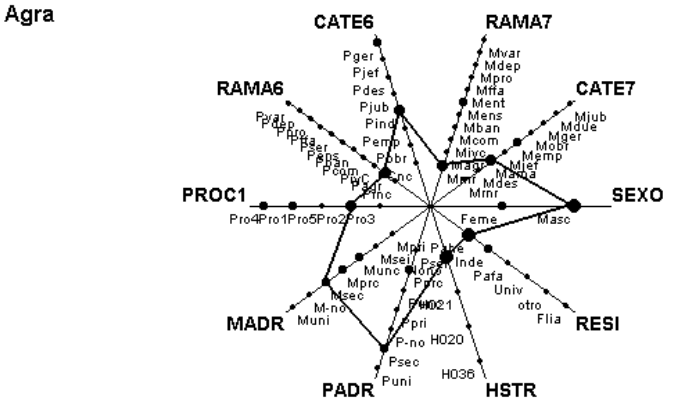
Posibilidades gráficas

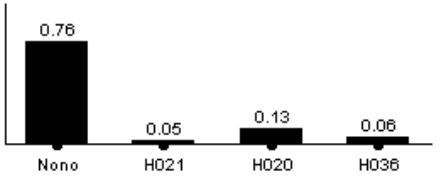
La visualización de un OS puede realizarse a través de un gráfico que se denomina Zoom Star. Esta representación está basada en los diagramas de Kiviat donde cada eje representa una variable. El software da la posibilidad de dos tipos de representación, en 2 dimesiones y en 3 dimensiones, la primera 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 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.

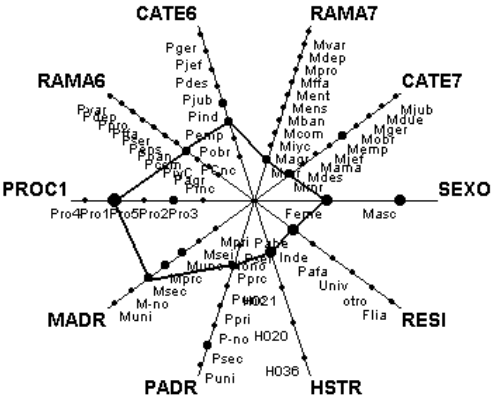
En este caso se comparan dos facultades con perfiles bien diversos: en el primer gráfico Ciencias Agrarias, con la distribución de horas trabajadas en un gráfico más pequeño (la distribución de cada variable puede obtenerse deteniendo el cursor en cada uno

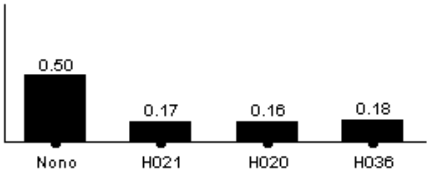
de los ejes correspondientes).
En el segundo gráfico la Facultad de Humanidades. Pueden observarse rápidamente las características diferenciales de ambas facultades, sobre todo en cuanto al sexo, la procedencia (PROC1), la rama y categoría ocupacional del padre (RAMA6, CATE6) y de la madre (RAMA7, CATE7), el nivel de escolaridad del padre y la madre (MADR, PADR).





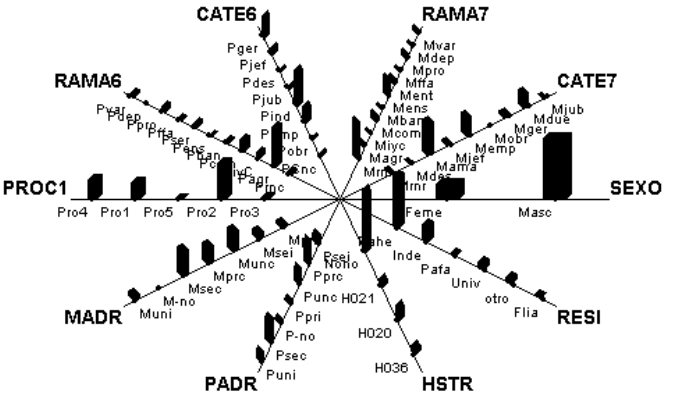
Huma

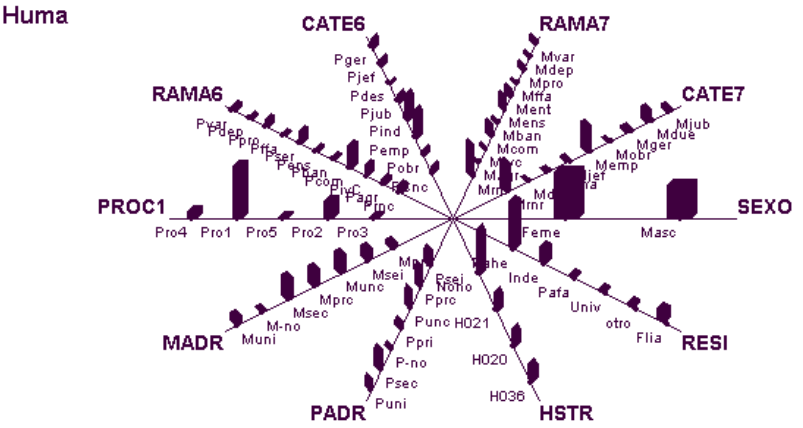




A continuación se presentan los gráficos en 3D para los mismos OS. Como se puede observar se representan al mismo tiempo las distribuciones de frecuencia de todas las variables.

Agra



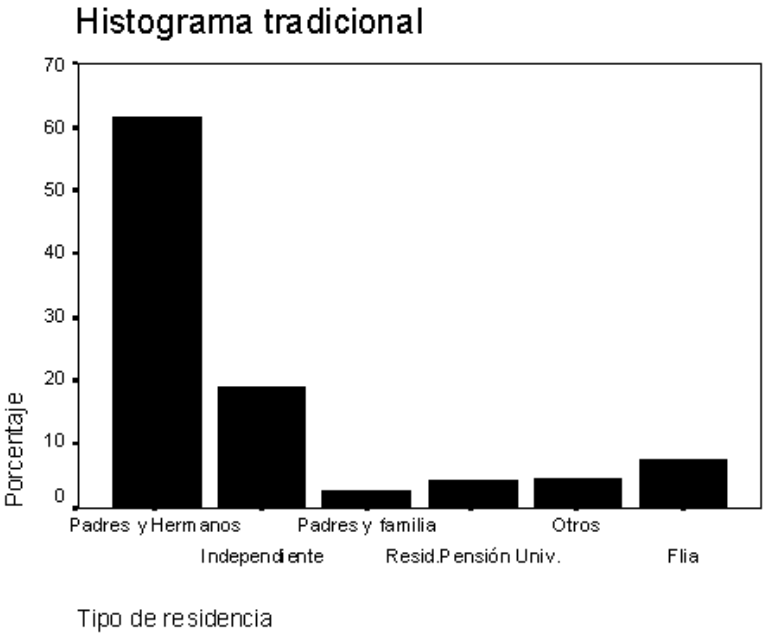


Descripción estadística simbólica

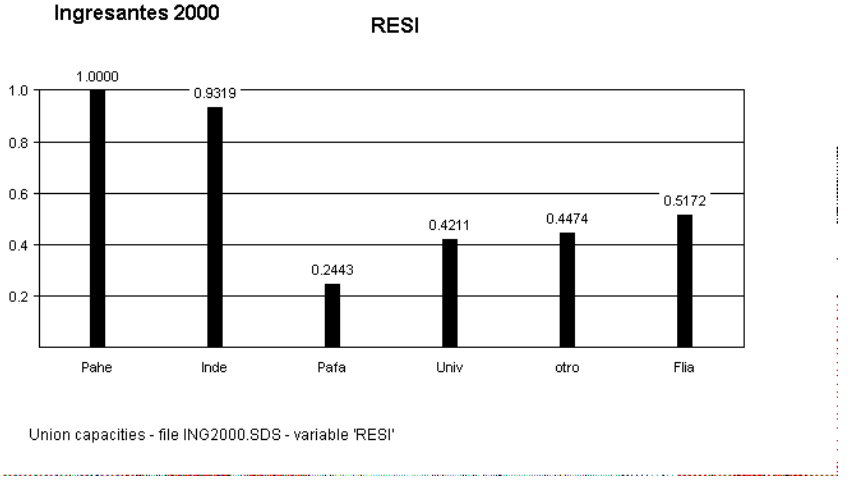
Presentaré a continuación la aplicación al ejemplo citado, de algunos conceptos de estadística descriptiva de objetos simbólicos, con el correspondiente listado provisto por el software.
Recordamos que la matriz está constituida por unidades de análisis que son objetos simbólicos y que en este caso se definieron como las distintas Facultades de la UNR, ya que interesaba que la comparación pasara a través de ellas. Por lo tanto las estadísticas de resumen se referirán a estas unidades.
En el siguiente listado se analizan algunas variables: en la primera columna se consignan las etiquetas de las mismas con sus correspondientes categorías.
En la segunda columna se consignan las 'capacidades', cada una de ellas en este caso representa la probabilidad de encontrar al menos un OS (una Facultad) que presente ese valor de la variable. Podemos notar que tenemos probabilidad igual a 1 de que al menos una facultad presente el sexo femenino, mientras que para el sexo masculino esa probabilidad es muy próxima a 1.
En la tercera columna se presenta el valor mínimo de esa categoría en todos los OS y en la cuarta columna el valor máximo.
En la quinta columna se presenta el valor de probabilidad promedio de la categoría entre todos los OS.

```
---------------------------------------------------------------------
 SODAS - STAT   CAPACITIES

 File:  ING2000.SDS
 Title: Ingresantes 2000
---------------------------------------------------------------------

                                      capa     mini     maxi     mean

  SEXO

   AD01  Feme                       1.0000   0.2026   0.7821   0.5394
   AD02  Masc                       0.9998   0.2179   0.7974   0.4606

  RESI

   AF01  Pahe                       1.0000   0.4996   0.7552   0.6274
   AF02  Inde                       0.9319   0.1340   0.2695   0.1994
   AF03  Pafa                       0.2443   0.0000   0.0348   0.0230
   AF04  Univ                       0.4211   0.0192   0.0908   0.0443
   AF05  otro                       0.4474   0.0122   0.0739   0.0481
   AF06  Flia                       0.5172   0.0043   0.1548   0.0577

  HSTR

   AG01  Notr                       1.0000   0.5002   0.8636   0.7243
   AG02  H021                       0.6267   0.0303   0.1652   0.0780
   AG03  H020                       0.6636   0.0418   0.1591   0.0862
   AG04  H036                       0.7611   0.0492   0.1892   0.1114

  PADR

   AJ01  Psei                       0.8849   0.1034   0.1935   0.1645
   AJ02  Pprc                       0.9413   0.1530   0.2716   0.2098
   AJ03  Punc                       0.9274   0.1500   0.2519   0.1956
   AJ04  Ppri                       0.4722   0.0216   0.0709   0.0517
   AJ05  P-no                       0.2211   0.0038   0.0437   0.0205
   AJ06  Psec                       0.9602   0.1930   0.2974   0.2352
   AJ07  Puni                       0.7930   0.0811   0.1643   0.1227
```

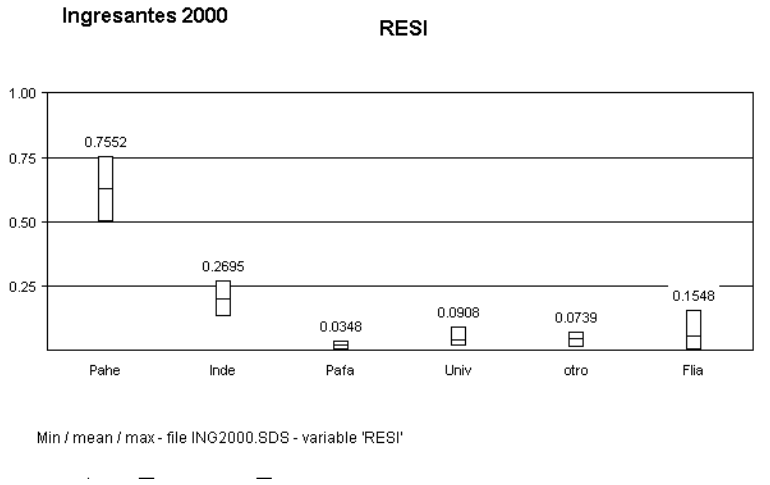
Si consideramos la variable ‘tipo de residencia’ (RESI), la forma de graficar su distribución en la estadística clásica sería un histograma o más específicamente un diagrama de barras como el siguiente. En el mismo la altura de las barras representan las frecuencias relativas de cada categoría con respecto al total general.



En cambio en el siguiente gráfico, la altura de las barras es una 'capacidad' de los OS considerados. La información de ambos histogramas es de diferente tipo.



El diagrama que sigue a continuación, para la misma variable, es similar al gráfico de cajas, donde se consignan los valores promedios, así como mínimos y máximos que toman las probabilidades de cada categoría en el total de OS (Facultades de la UNR)



He presentado algunas de las posibilidades de manipulación de los OS como un nivel superior de generalización en la definición de las unidades de análisis. Ello implica la introducción de un mayor grado de complejidad en las mismas. Existen posibilidades de tratamiento (análisis factorial, clasificación) de estas nuevas unidades que no serán presentados aquí.

El incesante desarrollo de programas informáticos proveerá sin duda nuevas herramientas de visualización, evaluación e interpretación de estos objetos novedosamente construidos.

BIBLIOGRAFÍA

- Aluja Banet Tomás (ed.) (1994) *Diseño del "producto ideal". Ponencias de la Jornada de Marketing y Estadística*. Universidad Politécnica de Cataluña
- Aluja Banet, Tomás y Morineau Alain (1999) *Aprender de los datos: el Análisis de Componentes Principales*. EUB. Barcelona
- Anastex[53] S.J (ed.) (1993) *Actes des Secondes Journées Internationales d'Analyse Statistique de Donées Textuelles,* Montpellier 21-22 octobre 1993, Ecole Nationale Supériere des Télécomunications Enseignement supérieur de France Télécom.
- Aristóteles (1945) *Anatomía de los animales.* Ed. Schapire. Buenos Aires.
- Bachelard Gaston (1982) *La formación del espíritu científico.* Siglo Veintiuno. 10a. edición.
- Balbi Simona (1994) *L'analisi multidimensionale dei dati negli anni '90,* Dipartimento di Matematica e Statistica, Università degli Studi di Napoli Federico II, Rocco Curto Editore, Nápoles
- Balbi Simona (1999) *Lo studio dei Messaggi Pubblicitari con l'Analisi dei Dati Testuali*, Facoltà di Scienze Politiche, Università degli Studi di Napoli Federico II.
- Baranger Denis (1992) *Construcción y análisis de datos* Editorial Universitaria. Cátedra. Posadas.
- Bécue Mónica, Lebart Ludovic, Rajadell Nuria (ed.) (1992) *Jornades Internacionals d'Anàlisi de Dades Textuals (JADT)* Barcelona 1990.
- Benzécri Jean Paul y col.(1976) *L´Analyse des données, T.1 La Taxonomie T.2 L´Analyse des correspondances*. Dunod. París.
- Benzécri Jean Paul y ot. (1980) *Pratique de L'Analyse des Données. 1 Analyse des Correspondances. Exposé Elementaire. 2 Abrége Théorique. Études de Cas Modèle*. Dunod. París.
- Benzécri Jean Paul (1982*) Histoire et Prehistoire de L´Analyse Des Données*, Bordas, París.
- Benzécri Jean Paul (1993) *Cualidad y cantidad en la tradición de los filósofos y en análisis de datos.* IRICE. Rosario. (Traducción de N.Moscoloni del artículo homónimo publicado

[53] Nombre colectivo del comité organizador

en Les Cahiers de l'Analyse des Données, Vol.XIII, 1988, n.1, pp.131-152)
- Bloor David (1998) *Conocimiento e imaginario social.* Gedisa. Barcelona.
- Boudon Raymond, Lazarsfeld Paul (1973) *Metodología de las ciencias sociales* Vol. I y II, Ed. Laia.Barcelona.
- Bourdieu Pierre, Chamboredon J-C y Passseron J-C (1999) *El oficio de sociólogo* Siglo Ventiuno Editores. México
- Bourdieu Pierre (2000) *La distinción. Criterio y bases sociales del gusto*. Taurus.
- Bourdieu Pierre (2001) *Las Estructuras Sociales de la Economía.* Manantial. Buenos Aires
- Bouroche Jean-Marie, Saporta Gilbert (1980) *L'Analyse des Donées*, Colección Que sais-je?, Presses Universitaires de France. París
- Calvo Garrido Patricia y ot. (2000) *Creación de objetos simbólicos a partir de encuestas almacenadas en bases de datos relacionales*, EUSTAT, Instituto Vasco de Estadística, Vitoria-Gasteiz. Disponible en Internet: http://www.eustat.es
- Calvo Garrido Patricia y Pérez Diez Yolanda (2001) *La sociedad de la información analizada mediante objetos simbólicos,* EUSTAT, Instituto Vasco de Estadística, Vitoria-Gasteiz. Disponible en Internet: http://www.eustat.es
- Conde Fernando (1987) *Una propuesta de uso conjunto de las técnicas cuantitativas y cualitativas en la investigación social. El isomorfismo de las dimensiones topológicas de ambas técnicas.* Revista Española de Investigaciones Sociológicas, N°39.
- Cook T D y Reichardt Ch S (1986) *Métodos cualitativos y cuantitativos en investigación evaluativa.* Morata. Madrid.
- Cullen Carlos (1986) *Interdisciplinariedad o la posibilidad de lo epistemológico como ético*, en Reflexiones desde América. T.I. Fundación Ross. Rosario.
- Davidson David (1992) *Mente, mundo y acción,* Paidós. ICE – UAB. Barcelona
- Desrosières Alain (1996) *Reflejar o instituir: la invención de los indicadores estadísticos*, en Methodologica, N°4, Laboratoire de Méthodologie du Traitement des Données, Universitè Libre de Bruxelles, Bélgica.
- Desrosières Alain (1998) *The Politics of Large Numbers.*

Harvard University Press. Cambridge
- Diday Edwin y ot.(1982) *Eléments d'analyse de données,* Dunod, París
- Diday Edwin, Lechevallier Ives (ed)(1991)*Symbolic-Numeric Data Analysis and Learning. Proceedings of the Conference.* Versailles Sept-18-20, 1991. INRIA. Nova Science Publishers. New York
- Diday Edwin (1992) *Analyse des Données et Classification Automatique Numerique et Symbolique* Cuaderno 27, Seminario Internacional de Estadística en Euskadi, Eustat. Bilbao.
- Diday Edwin, Summa M. y ot. (1992) *Analyse Symbolique de Scénarios d'Accidents* Université Paris Dauphine- INRETS, París.
- Diday Edwin (1997) *Análisis de datos simbólicos. Conferencias pronunciadas por el autor.* IRICE. Rosario
- Droesbeke Jean J. Tassi P.(1990) *Histoire de la Statistique*. Presse Universitaires de France, colección Que sais-je?. París
- Elster Jon (1990) *El Cambio Tecnológico,* Gedisa. Barcelona
- Escofier Brigitte, Pagès Jerôme (1992) *Análisis factoriales simples y múltiples.*Universidad del País Vasco. Bilbao.
- Follari Roberto (1982) La posición de Jean Piaget y Anexo. Interdisciplinariedad: espacio ideológico, en Interdisciplinariedad. UAM - Azapotzalco. México.
- Gaeta Roberto y ot. (1996) *Modelos de explicación científica.* EUDEBA. Buenos Aires.
- Galtung Johan (1966) Teoría y métodos de la investigación social. Tomos I y II. EUDEBA. Buenos Aires.
- Gould Stephen (1984) *La Falsa Medida del Hombre,* Antoni Bosch Editor, Buenos Aires.
- Greenacre Michael (1984) *Theory and Application of Correspondence Analysis*, Academic Press. London.
- Hacking Ian (1995) *La domesticación del Azar,* Gedisa. Barcelona.
- Hesse Mary *Teoría y observación,* en Olivé y ot. (1989)
- Ibáñez, Jesús, García Ferrando M. y Alvira, F. (comp) (1990) *El análisis de la realidad social. Métodos y técnicas de investigación.* Alianza Editorial. Madrid.
- Ibáñez Jesús (coord.) (1998) *Nuevos Avances en la Investigación Social,* Proyecto A Ediciones, 2da. ed.

aumentada. Tomos 1 y 2. Barcelona.
- Kendall Maurice y Stuart Alan (1963) *The Advanced Theory of Statistics, Vol.I, Distribution Theory.* Charles Griffin & Co.Ltd., Londres.
- Klimovsky Gregorio (1995) *Las desventuras del conocimiento científico.* AZ editora. 2a.edición. Buenos Aires.
- Kuhn Thomas (1989) ¿Qué son las revoluciones científicas? y otros ensayos. Introducción de Antonio Beltrán. Paidós. Barcelona.
- Ladrière Jean (1978) *El reto de la racionalidad*, Ediciones Sígueme - UNESCO, Salamanca.
- Lebart Ludovic (1975) *L´orientattion du dépouillement de certaines enquêtes par l´analyse des correspondances multiples*. Consommation, 2. Dunod.
- Lebart Ludovic, Morineau Alain, Tabard N (1977) *Techniques de la Description Statistique, Méthodes et Logiciels pour l'Analyse des Grands Tableaux,* Dunod. París
- Lebart Ludovic, Morineau Alain y ot. (1983) SPAD *Systéme Portable pour l'Analyse des Données*. Tomos 1 y 2. CESIA.
- Lebart Ludovic, Morineau Alain, Warwick K.W. (1984) *Multivariate Descriptive Statistical Analysis, Correspondence Analysis and Related Techniques for Large Matrices.* John Wiley. New York.
- Lebart Ludovic, Morineau Alain, Fenelon J.P., *Tratamiento Estadístico de Datos,* Barcelona-México, Marcombo Boixareu Editores, 1985
- Lebart Ludovic (1987) *Conditions de vie et aspirations des francais, evolution et structure des opinions de 1978 a 1986*, Futuribles, sept 1987.
- Lebart Ludovic, Salem André (1988) *Analyse Statistique des Données Textuelles*, Dunod. París.
- Lebart L., Morineau A., Bécue M.(1989) *SPAD.T, Système Portable pour l'Analyse des Données Textuelles. Manuel de l'utilisateur.* CISIA. Paris.
- Lebart Ludovic (ed.ASU) (1992) *La Qualité de l'Information dans les Enquétes,* Dunod. París
- Lébart, L., Morineau,A., Bécue M, Haeusler L, (1993) *SPAD.T Integré® version 1.5 Système Portable pour l'Analyse des Donnés Textuelles*, CISIA. París.
- Lebart Ludovic, Salem André (1994) *Statistique Textuelle,*

Dunod. París.
- Lébart, L., Morineau,A., Lambert, P., Pleuvret, P.(1994) *SPAD.N ® Versión 2.5 Sistema Compatible para el Análisis de Datos,* CISIA. París.
- Lebart Ludovic, Morineau Alain, Piron Marie (1995) *Statistique Exploratoire Multidimensionnelle.* París. Dunod.
- Lébart, L., Morineau,A., y ot. (2000)*Système SPAD, Versión 4.51*, ®CISIA-CERESTA
- Maingueneau Dominique(1989) *Introducción a los métodos de Análisis del Discurso*. Hachette. Buenos Aires
- Mardones, J.M.(1991) *Filosofía de las Ciencias Humanas y Sociales*, Anthropos. Barcelona.
- Morineau Alain, Quidel Patrick (Coord.) *Jornadas de Análisis de Datos e Informática, encuentro Francia-América Latina*, 11-22 de julio de 1983, Caracas.
- Moscoloni Nora, Conti Omar, Tuttolomondo Inés, Meinardi Beatriz (1982) *Perfil Social de los Estudiantes de la Universidad Nacional de Rosario*, UNR Editora. Rosario
- Moscoloni Nora (1988) *Correspondencias múltiples como técnica para analizar encuestas a respuesta cerrada*, Actas del IVº Simposio Interdisciplinario de Metodología de la Investigación en Ciencias Sociales organizado por PROSNE y CONICET, Corrientes.
- Moscoloni, Nora,(1990) *Evaluación de escalas nominales mediante el análisis de corres-pondencias múltiples*. Revista IRICE, Nº 1. Rosario.
- Moscoloni Nora, (1992)*Técnicas de clasificación para la caracterización de poblaciones: alumnos de la Facultad de Ciencias Bioquímicas y Farmacéuticas de la UNR*. Revista IRICE, Nº 3 -4. Rosario.
- Moscoloni Nora, Meinardi Beatriz, Santone Beatriz, Tuttolomondo Inés (1995) *Análisis textual de opiniones de estudiantes de la UNR* en: Bolasco Sergio, Lebart Ludovic, Salem André (eds.) III Giornate Internazionali di Analisi Statistica dei Dati Testuali, vol.II, CNR, Nov.1995. Roma
- Moscoloni Nora (1997) *Reflexiones sobre una experiencia didáctica multidisciplinaria en Análisis de Datos,* en: Fernández Aguirre Karmele (ed.) Actes IV Congrès International D'analyses Multidimensionelles des Données, NGUS'97. Bilbao
- Moscoloni Nora (1999) *Metodología integradora para el análisis*

del proceso de ensenanza aprendizaje, en Actas del 2do. Congreso Mundial de Educación Internacional, UNESCO, UADE, julio de 1999. Buenos Aires.
- Moscoloni Nora, Pallavicini Mercedes, Valdettaro Sandra y otros (1999) *Comunicación: evaluación institucional y curriculum*, UNR Editora, Rosario.
- Moscoloni Nora, Satriano Cecilia (2000) *Importancia del análisis textual como herramienta para el análisis del discurso*, Cinta de Moebio, Electronic Journal for Social Sciences Epistemology [on line] Facultad de Ciencias Sociales, Universidad de Chile, Nº9, Nov. 2000. Disponible en: http//www.moebio.uchile.cl/09/frames08.htm
- Olivé León y Pérez Ransanz, A.R. (comp.) (1989) *Filosofía de la ciencia: teoría y observación*, México, Siglo XXI.
- Pascal Blaise (1981) *Obras. Pensamientos. Provinciales. Escritos científicos. Opúsculos y cartas.* Prólogo, J.L. Aranguren. Traducción y notas C. R. de Dampierre. Ediciones Alfaguara. Madrid.
- Pearson Egon S.(1948) *Pearson, creador de la Estadística Aplicada* Espasa-Calpe Argentina. Buenos Aires
- Pérez Serrano, Gloria (1994) *Investigación cualitativa. Retos e interrogantes* Ed. La Muralla. Madrid
- Popper Karl (1967) La lógica de la investigación científica. Tecnos. Madrid.
- Porfirio *Introducción a las Categorías (La Isagoge)* (1993)En: Aristóteles, Tratados de Lógica (El Organon). Ed. Porrúa. México.
- Samaja Juan (1987) *Introducción a la Epistemología Dialéctica*, Lugar Editorial. Buenos Aires.
- Samaja Juan (1994) *Epistemología y Metodología,* EUDEBA. Buenos Aires
- Samaja Juan (1996) *El lado oscuro de la razón* JVE Episteme.Buenos Aires
- Samaja J. (2000) *Aportes de la Metodología a la Reflexión Epistemológica*, en: Díaz Esther (comp.) La Posciencia. Dédalo, Buenos Aires
- Schuster Félix G. (1992) *El método de las Ciencias Sociales,* Centro Editor de América Latina. Buenos Aires
- Sutcliffe John (1997) *Data Analysis from a Logical Standpoint.* CSNAN: Classification Society of North America Newsletter [on

line] September 1997, Issue #51. Disponible en Internet: http://www.pitt.edu./~csna/news/csna.news51.html#top
- Tukey John W (1977) *Exploratory data analysis,* Reading, MA.Addison Wesley.
- Velleman P Hoaglin D (1992) *Data Analysis*. En Perspectives on Contemporary Statistics, Mathematical Association of America, MAA, Notes Number 21, Chapter 2, Hoaglin, D.C. y Moore, D. Editors.
- Verde Rosanna (1999) *Techniche Multivariate per l'Analisi dei Dati di preferenza,* en Actas de la Jornada Internacional de Análisis Multidimensional de Datos, 1era. Reunión de la Red Lamda, Proyecto Alfa de la Unión Europea, Facultad de Ciencias Médicas de la UNR, Rosario, 13 al 16 de diciembre de 1999.
- Vessereau André (1962) *La Estadística,* EUDEBA. Buenos Aires

INDICE