Departamento de Ciencias Políticas y Sociales de la
UNIVERSIDAD POMPEU FABRA

Facultad de Ciencias Sociales y de la Comunicación

Programa de doctorado: Teoría Política y Social 1ª edición

Desarrollado en el Bienio: 1993-1995

Tesis Doctoral

EL ESTUDIO DEL COMPORTAMIENTO ELECTORAL EN ESPAÑA: SU UBICACIÓN DENTRO DE LA CIENCIA, SU RELACIÓN CON LA ESTADÍSTICA Y LAS NUEVAS POSIBILIDADES DE ANÁLISIS QUE SE OFRECEN AL POLITÓLOGO

Presentada por la doctorando:

Alicia CODURAS MARTÍNEZ

Para optar al título de:
DOCTOR POR LA UNIVERSIDAD POMPEU FABRA

La directora de la tesis ha sido la:

Catedrática: **Dra. Mª Rosa VIRÓS I GALTIER**

Barcelona 1998

3.2 LAS ENCUESTAS ELECTORALES Y LOS NUEVOS TRATAMIENTOS ESTADÍSTICOS.

El diseño de una encuesta pre o post electoral en el momento actual debería renovarse, en nuestra opinión, desde dos puntos de vista:

El temático
El del diseño

La renovación temática se justifica por diversas razones. En primer lugar, porque socialmente se está demostrando que a medida que se iguala el nivel cultural de los distintos colectivos, las variables que tradicionalmente habían tenido cierto poder explicativo, lo están perdiendo a nivel general.

El género, la edad, la confesión religiosa y otras variables ya no tienen la capacidad de discriminación que poseían. La forma de vida está cambiando y con ello el panorama electoral.

Las fuerzas políticas tienen ofertas y maneras de actuar bastante parecidas y la formación del nuevo electorado ya no está tan sujeta a condicionamientos de tipo histórico como antes. La desintegración del comunismo y la estabilidad del capitalismo junto al desequilibrio que supone la situación del tercer mundo afectan al comportamiento social de los países desarrollados y con sistemas democráticos de una forma equiparable. Es como describir una situación de estancamiento en la cual, el seguir una opción u otra puede depender de otros muchos factores que nada tienen que ver con el género, el origen, la edad o variables por el estilo.

Por tanto, desde el punto de vista temático, la encuesta electoral debe profundizar más y buscar nuevos elementos de interés que ayuden a determinar qué influye en los electores cuando escogen su opción.

Por otro lado, el diseño tradicional de una encuesta se puede modificar, no sólo desde el punto de vista temático, sino desde el punto de vista técnico. Si el politólogo ayudado por el estadístico llega a clarificar qué quiere saber y con qué técnicas se va a trabajar para alcanzar el objetivo, las preguntas se diseñarán para que luego puedan ser correctas desde el punto de vista de la técnica. Es mejor tener esto previsto desde el principio que luego quedarse a medias por tener variables medidas en escalas inadecuadas cuando se podría haber evitado perfectamente.

Por tanto, hay que mentalizar tanto a les politólogos y sociólogos como a los técnicos para que el planteamiento sea: quiero averiguar esto, ¿qué técnica o técnicas son las mejores? ¿cómo debo diseñar las preguntas de la encuesta para conseguirlo?.

Los técnicos deben estar preparados y saber algo sobre las materias que les van a consultar y tener en mente asociadas las técnicas para así poder ayudar a los investigadores de otras disciplinas. Son personas "bisagra" que están en el centro de casi cualquier investigación en la actualidad y deben tener una preparación lo más generalista posible.

Los nuevos tratamientos estadísticos, tanto descriptivos como inferenciales ofrecen nuevas posibilidades siempre y cuando las encuestas puedan proporcionar el tipo de datos necesario.

4 ¿QUÉ PUEDEN APORTAR LAS TÉCNICAS ESTADÍSTICAS DESARROLLADAS RECIENTEMENTE AL ÁMBITO DEL ANÁLISIS DE DATOS DE ENCUESTA?

4.1 INTRODUCCIÓN

De todo lo que ha ido apareciendo últimamente en el terreno de las técnicas estadísticas, en esta tesis se desea destacar la posibilidad de efectuar análisis de Componentes Principales con variables cualitativas, cosa que hasta hace poco tiempo no era posible y que conducía a ciertos abusos de las técnicas factoriales con el empleo de variables no apropiadas para ello.

No vale la pena repasar las técnicas tradicionales de tabulación simple y cruzada ampliamente conocidas, documentadas y aplicadas, ni los tipos de gráficos más habituales que se utilizan en la presentación de resultados de estos sondeos.

Todo investigador sabe que el tratamiento tradicional que se ha dado a los datos de encuestas electorales, desde el punto de vista descriptivo ha comprendido:

Tablas simples de frecuencias o porcentajes que nos han informado de las cantidades de sujetos que han seleccionado las diferentes opciones propuestas en una pregunta.

Tablas cruzadas de frecuencias o porcentajes que nos han informado de las cantidades de sujetos que han seleccionado simultáneamente un par de opciones en dos preguntas, o que han respondido simultáneamente a dos características. Si se ha efectuado el contraste de independencia (X^2), se ha dispuesto, además, de una medida de asociación de las dos variables implicadas en la tabla.

Gráficos: los de barras, sectores y mapas, han sido los más utilizados, junto a los de series temporales y barras cruzadas por ser los más sintéticos para ofrecer información en los medios de comunicación.

Estadísticos: moda, mediana, media, desviación estándar. Son poco utilizados a nivel de encuesta y su potencial desaprovechado, puesto que se pueden aplicar diversas técnicas tradicionales que, muchas veces, por desconocimiento del público en general, no se emplean. Por ejemplo, los contrastes de diferencias de medias y las medidas de dispersión acompañando a las de tendencia central

Un caso que vale la pena comentar referente a esto último es el de la información electoral audiovisual. Cuando, por ejemplo, se proporciona la valoración de un político en términos de una nota media asignada por encuestados no suelen añadirse comentarios tales como:

"La puntuación fue diferente considerando tramos de edad, los jóvenes tendían a valorar mejor a este político que los mayores"

"La puntuación media ha sido "tal", pero el colectivo entrevistado no era homogéneo y esta nota es poco representativa"

El primero es un simple caso de contraste de medias y el segundo hace referencia a la desviación estándar, pero nadie que lea o escuche comentarios como estos tiene porqué saber estadística y sin embargo, la información es más precisa y realista que la que nos suelen proporcionar los medios de comunicación.

Por tanto, llegados a este punto, resulta algo decepcionante saber que la estadística avanza y que ni siquiera se aprovecha aquello que tradicionalmente está más que experimentado.

El politólogo especialista en temas electorales puede aspirar a ofrecer mucha más información de forma inteligible sin necesidad de dominar las técnicas más sofisticadas al nivel de un estadístico. Todo es cuestión de redacción, de extracción de conclusiones e informes a partir de lo que los complejos cálculos nos proporcionan y de colaboración entre la parte técnica y la teórica, entre lo cualitativo y lo cuantitativo.

Por tanto, como aportación a este terreno, profundizaremos en el llamado "Sistema GIFI" que permite el tratamiento multivariable de variables cualitativas sin dificultad y que es muy poco conocido en nuestro país.

El análisis multivariante ha sido y es ampliamente utilizado por investigadores de Comportamiento Electoral a través de todas las técnicas que ofrece: componentes principales, factorial, discriminante, regresión, correlaciones, regresión logístca, cluster, escalogramas multidimensionales, etc. Sin embargo, la búsqueda de dimensiones tales como: izquierda-derecha, nacionalismo-centralismo, participación convencional-participación no convencional, materialismo-postmaterialismo[195] y otras, ha propiciado un empleo masivo de los análisis que trataban el problema de la información redundante y de la dimensionalidad.

Por eso, al profundizar en la bibliografía especializada se ve que una parte importante de los estudios aplicados presentan análisis factoriales (de correspondencias y de correlaciones) y análisis de componentes principales.

La facilidad de diseño de este tipo de aplicaciones, debida en gran parte, a su implementación en las aplicaciones informáticas estadísticas más populares y la agradable presentación gráfica que los acompaña han propiciado su conversión en "clásicos" dentro de la especialidad.

Por tanto, ya desde mediados de los años 70 se aplicaron técnicas multivariables como:

Regresión lineal múltiple: para explicar el número de votos a un partido en función de variables socio-económicas (aunque es más adecuado con datos agregados, también se ha hecho con datos de encuesta)

Componentes Principales: determinación de las variables que sintetizan el comportamiento electoral, situación de fuerzas políticas respecto al electorado, búsqueda de dimensiones izquierda-derecha, etc.

Correspondencias: establecimiento de variables explicativas del comportamiento electoral relacionándolo con lugares de origen (regiones, áreas geográficas, etc.)

Discriminante: explicación del voto indeciso, abstención y otros temas relacionados.

Cluster: establecimiento de perfiles de todo tipo, en especial, de votantes o de zonas geográficas.

[195] Ronald Inglehart. Es el que dedica mayor atención a mediados de los 80 al tema del post-materialismo y lo asocia con un ambiente de inseguridad. Trabaja en la hipótesis del cambio social del materialismo al post-materialismo basándose en que se está pasando de un clima de materialismo total en que los máximos valores son la seguridad y el estatus a otro en que comienza a tener importancia la calidad de vida

Desde el punto de vista técnico, el empleo de todos los análisis mencionados, con excepción del de Correspondencias, pasaba necesariamente por la disponibilidad de variables cuantitativas (más fáciles de obtener en el ámbito de datos agregados), que además, en ciertos casos, debían cumplir una serie de propiedades.

En el caso de la regresión múltiple, se podía aplicar un logit para explicar la opción votada a través de la edad, la renta, el nivel educacional (años de educación), ideología política (escalas de 1 a 9), nacionalismo, religiosidad, etc.

Estas variables tenían que venir cuantificadas en la base de datos y el coeficiente de bondad del ajuste debía tener cierta envergadura para dar credibilidad a la contrastación de hipótesis. Sin embargo, la mayoría de las veces este tipo de hipótesis se ha contrastado más en el ámbito de los datos agregados porque se ha pensado poco en el diseño de las escalas de medición de estos y otros conceptos para actuar a nivel desagregado. Por eso, se recomienda encarecidamente estudiar el diseño de la encuesta.

Sin embargo, desde el punto de vista estrictamente estadístico, algunas aplicaciones podrían ser cuestionadas en determinados aspectos. Para hacerse una idea del tipo de críticas a que podrían ser sometidas baste con apuntar algunos problemas comunes a todas las técnicas multivariantes que, aún hoy, son tema de discusión:

¿Debemos estar siempre bajo el supuesto de comportamiento Normal Multivariable de nuestros datos o no?

¿Cómo podemos resumir la intensidad de la asociación bivariable entre variables cualitativas con muchas categorías?

¿Cómo podemos resolver el problema que representan las casillas vacías en grandes tablas de contingencia multidimensionales?

¿Es posible efectuar inferencia en relaciones causales cuando hay ausencia de control o manipulación experimental?

Así, en ocasiones se aplica, por ejemplo, un análisis factorial de correlaciones con presencia de variables nominales dicotómicas, que por su propia naturaleza no pueden presentar nunca una distribución Normal.

En otras, se calcula el coeficiente de correlación entre variables nominales, medida de asociación que no puede tener sentido si las categorías de dichas variables han sido escogidas arbitrariamente, ya que su media no es adecuada como medida de posición central y el coeficiente de correlación, se basa en la media.

También es común hallar casillas vacías en grandes tablas de contingencia, para las cuales, el contraste X^2 de asociación es cuestionable.

La aplicación de análisis de componentes principales a datos nominales y ordinales procedentes de encuestas ha sido una práctica corriente y, finalmente, también se han dado casos de inferencia estadística errónea por estar basados en muestras defectuosas.

Hay que tener en cuenta que el análisis multivariante comienza su desarrollo teórico en los años 50, época en que se le tenía, digamos más respeto.

Con el desarrollo de la informática, a partir de finales de los 70 y, sobre todo, de los 80, la facilidad que comporta su empleo dispara la utilización de estos métodos que a veces se han transmitido de unos a otros sin demasiado rigor. La complejidad de sus algoritmos y la casi imposibilidad de desarrollar ejemplos "a mano", ha llevado a tener una visión un tanto superficial del análisis multivariante que es necesario corregir antes de que se abuse más del mismo.

Por eso, aunque pueda parecer exagerado, no está de más volver la vista atrás y saber qué dijeron algunos de los estadísticos que contribuyeron al desarrollo teórico de este conjunto de técnicas.

En 1957 Roy interpreta el análisis multivariable como un conjunto de técnicas que pueden utilizarse para contrastar un número restringido de hipótesis acerca de las relaciones entre variables correlacionadas.

Más interesante resulta la opinión de Kendall que, también en 1957 define el análisis multivariante como la rama del análisis estadístico dedicada al estudio de las relaciones de conjuntos de variables dependientes. Este autor distingue entre el análisis de dependencia y el de interdependencia, punto muy interesante para establecer un primer punto de separación entre grupos de técnicas multivariables. Así, en el análisis de la dependencia se investiga si y cómo un conjunto de variables depende de otro grupo. El primer grupo es el de las variables dependientes y el segundo el de las independientes. Por eso, en este tipo de análisis hay cierto grado de asimetría: la dirección de la influencia causal es de las variables independientes hacia las dependientes. Por otro lado, en el análisis de la interdependencia los conjuntos de variables son tratados de forma simétrica: no hay distinción entre variables dependientes e independientes.

Según todo lo anterior, la técnica más familiar de análisis dependiente es la regresión múltiple y la de análisis interdependiente el análisis de componentes principales. Cooley (1962) y Lohnes (1971) son de la misma línea que Kendall y también lo será posteriormente el francés Dagnelie (1975).

En 1975, el mismo Kendall establecería los propósitos más importantes de las técnicas multivariantes:

Simplificación estructural
Clasificación
Agrupación de variables
Análisis de dependencia
Análisis de interdependencia
Construcción y contraste de hipótesis

así como los problemas más importantes con que tropieza el desarrollo de las mismas:

En muchas ocasiones no se pueden formular las hipótesis porque la población no está bien definida.

Prácticamente no se pueden usar las técnicas multivariables sin el concurso de un ordenador, pero eso no quiere decir que los programas sean perfectos.

Incluso teniendo muestras aleatorias, a menudo es imposible asumir la normalidad multivariante.

Los gráficos son muy importantes en el contexto del análisis multivariante, pero es casi imposible tener diagramas inteligibles si se pasa de dos dimensiones.

No existe una estadística multivariable no paramétrica comparable a la estadística no paramétrica univariable.

Otros autores como Anderson (1958) y Morrison (1967) hablaban del análisis multivariante como una generalización de la estadística inferencial basada en la Normal para situaciones multinormales. Por tanto, asumen que los individuos analizados constituyen una muestra aleatoria de una población infinita y que el comportamiento de los datos es Normal Multivariante. Por eso, estos autores hablan de análisis estadístico multivariante y en cambio apenas tratan el análisis canónico y el de componentes principales que no tiene tantas restricciones.

Además de los anteriores, hay que destacar a Van de Geer como el representante de aquellos que ven el análisis multivariante (1967) como el arte de descubrir las relaciones entre muchas variables mediante el uso de técnicas matemáticas. Estos autores acostumbran a hablar poco de estadística y mucho de matemáticas (sobre todo de álgebra matricial).

Para ellos, el análisis multivariante es un análisis lineal de matrices de datos cuyos principales propósitos son: la reducción de datos y la representación geométrica.

Green y Carroll (1976) constituyen otro caso de la misma línea y Cilliez y Pagès (1976) su equivalente pero siguiendo la tradición francesa. El análisis de datos multivariante se popularizó en Francia gracias al trabajo de J.P. Benzécri, pero adquirió rasgos diferenciales del modelo anglo-sajón porque el álgebra lineal francesa es más modema y abstracta.

Finalmente, cabe mencionar el enfoque de Dempster (1969), que en nuestra opinión es el que deberían volver a tener presente los investigadores actuales y que distingue entre el análisis de datos y la estadística. En su obra describe ciertos métodos de análisis de datos estadísticos procedentes de muestras multivariantes. El propósito básico de este análisis de datos es el de reducir la gran cantidad de números para disponer de resúmenes claros e interpretables de la información que reside en la muestra. Otro propósito es el de hacer inferencias de la muestra hacia la población. Cooley y Lohnes, además de estar en la línea de Kendall, también descubrieron el análisis de datos (gracias a Tukey, por supuesto) y trabajaron en esa dirección.

Otros autores están entre las dos líneas expuestas: Tatsuoka (1971) se decanta por el análisis estadístico y matemático, pero trata en profundidad el análisis canónico. Harris (1975) defiende la aproximación inferencial, la hipótesis nula y el contraste de significación asociado a la misma, mientras la mayoría de los autores lo consideran obsoleto en ese contexto. Desarrolla métodos de optimización para hallar coeficientes que proporcionen combinaciones lineales óptimas entre las variables analizadas, pero a los analistas sociales, estos coeficientes les resultan poco informativos.

Giri (1977) pone el acento en el contraste de hipótesis, dejando fuera la estimación y su obra puede considerarse como una aplicación de aspectos de la teoría de la decisión a situaciones multivariantes en que se asume que las distribuciones son normales.

Gnanadesikan (1977) se puede alinear con los analistas de datos, pero también está emparentado con Kendall y el estudio de la bondad del ajuste y los outliers.

Kshirsagar (1978) es la continuación de la escuela de Anderson y, en su obra, presenta los nuevos resultados multinormales logrados desde 1958.

Finalmente, Thorndike (1978) representa la continuidad de la línea de Cooley y Lohnes.

Por tanto, si bien en apariencia, durante la década de los ochenta las técnicas multivariantes han "servido para todo", lo cierto es que en el fondo, en algunos casos los investigadores se han engañado a sí mismos buscando resultados que ya se habrían puesto de manifiesto con técnicas sencillas como las exploratorias. En otras palabras (y siempre dejando de lado las relaciones espúreas entre variables): si los datos contienen alguna información o estructura relevante, cualquier técnica, desde la más sencilla a la más sofisticada, resaltará esa información, mientras que si en los datos no subyace ninguna estructura, por más que los analicemos, esta no se manifestará.

A pesar de todo, si bien no se puede volver al pasado y repetir muchos de los análisis efectuados, lo que sí que se puede hacer es mirar hacia delante y emplear con propiedad las técnicas estadísticas, atendiendo a las novedades que aparecen y que permiten solventar diversos problemas.

En este sentido, el Análisis Clásico de Componentes Principales (ACCP a partir de ahora) y el Canónico, pueden ser sustituidos por técnicas del sistema GIFI en aquellas ocasiones en que tengamos que trabajar con variables nominales y ordinales.

4.2 EL SISTEMA GIFI[196]: POSIBILIDADES DE APLICACIÓN EN ANÁLISIS DE COMPORTAMIENTO ELECTORAL A TRAVÉS DE DATOS DE ENCUESTA: ANÁLISIS NO LINEAL DE VARIABLES CATEGÓRICAS

4.2.1 PRELIMINARES

Tal y como ya se ha expuesto, la encuesta constituye una de las herramientas básicas para la recopilación de información ligada al estudio del Comportamiento Electoral especialmente por parte de politólogos defensores de la metodología empírica.

Las discusiones previas ya han puesto de manifiesto que existen diversos enfoques metodológicos y que todos tienen justificaciones, ventajas e inconvenientes, que los hacen susceptibles de mejora en cualquier momento.

También ha quedado claro que cualquier politólogo puede suscribirse a cualquier enfoque de los presentados en función de sus intereses particulares sin que exista por ello más o menos problema en todos ellos para cuestionar o dar por válidas sus conclusiones.

Sin embargo, no se puede negar que la compilación de datos mediante encuestas es una práctica asentada y corriente que permite, cuando menos, describir de forma clara y precisa (siempre que el diseño sea correcto y la población no se vea

[196] Albert Gifi es el seudónimo bajo el cual escribe un grupo de investigadores del Departamento de Teoría de la Facultad de Ciencias Sociales de la Universidad de Leiden en Holanda. Entre sus miembros, destacan: John Van de Geer, Jan de Leeuw, Bert Bettonvil, Eeke Van der Burg y otros. No es un grupo estable de personas, pero escogieron este nombre como homenaje al verdadero Albert Gifi, pionero en la investigación en el terreno del análisis multivariante.

afectada por sucesos extraordinarios en el curso del trabajo de campo) el estado de un colectivo respecto a un tema electoral concreto.

Lo que sí que resulta más criticable es la parte de diseño de la encuesta porque, al ir apareciendo nuevas técnicas de análisis estadístico, cada vez tiene más importancia el planteamiento de las preguntas, la batería de opciones facilitadas, las escalas de medición y la definición de los objetivos de la encuesta.

En este sentido se puede afirmar que, si bien el diseño de las muestras está cada vez más perfeccionado y sistematizado para lograr muestras representativas, la fiabilidad de los estudios puede resentirse por la forma en que están confeccionadas las encuestas en estos otros aspectos.

Si lo que se pretende es la mera consecución de porcentajes generales, las preguntas pueden pasar, en muchos casos tal y como se efectúan. En cambio, si lo que se pretende es una descripción más detallada de las relaciones entre los diferentes ítems, a un nivel que puede o no ser inferencial, entonces hay que profundizar mucho más en el diseño de las preguntas.

Para comprender esta necesidad, basta con prestar atención a las aportaciones del sistema Gifi en este terreno y compararlas con las aplicaciones que se han venido efectuando en muchos casos.

Durante muchos años, en la explotación de datos individuales procedentes de encuestas se ha aplicado el ACCP a variables nominales y ordinales, dejando de lado los requerimientos teóricos de que las variables debían ser proporcionales, ya fuera en base a que "no se podía hacer nada más" o apelando al Teorema Central del Límite o a la Aproximación de la Binomial a la Normal.

Sin embargo, los resultados, no dejaban de satisfacer a los investigadores, de forma que se aceptaba que esta técnica no inferencial se adaptaba bien a estos problemas. Lo cierto es que, aparentemente esto es así, pero, en parte, se debe a la idea que se ha apuntado antes de que si los datos contienen alguna estructura, ésta se pone de manifiesto incluso con estas técnicas.

Por eso, uno puede concluir que, con los años, el punto central de discusión se ha trasladado y ha sufrido un cambio, de forma que no está tanto en discutir si se puede aplicar o no el ACCP a variables nominales y ordinales como en saber si esta técnica es la que puede ofrecer los mejores resultados en la reproducción de la información original.

Los nuevos métodos de que se dispone desde 1990 denominados el Sistema Gifi, son técnicas que, por así decirlo, preparan los datos para que, al aplicarles un ACCP o un Análisis Canónico, proporcionen los mejores resultados posibles. La idea resulta muy interesante y atractiva: no tenemos datos cuantitativos numéricos pero podemos "numerizarlos" antes de efectuar los análisis.

Antes de pasar al desarrollo de las características de este sistema y de analizar su aplicabilidad al estudio del Comportamiento Electoral, conviene saber que no es el único planteamiento posible para el tratamiento de datos cualitativos y que, a veces, puede no ajustarse a determinadas preguntas planteadas por los politólogos.

Su difusión puede llegar a ser rápida dado que la aplicación SPSS la ha incorporado como uno de sus módulos, pero a nivel académico, al menos en nuestras universidades, todavía no se utiliza.

4.2.2 PLANTEAMIENTO GENERAL DEL SISTEMA GIFI

Las técnicas de análisis estadístico más populares para analizar las relaciones entre variables, en el terreno que nos ocupa, son las de regresión múltiple y componentes principales.

El Sistema Gifi constituye una propuesta diferente debido a que el tratamiento matemático es distinto del que se emplea en las dos técnicas mencionadas.

Habitualmente, la regresión se presenta en los manuales clásicos como un método de ajuste de un modelo llamado lineal. Este modelo supone que cada una de nuestras observaciones proviene de una muestra distribuida normalmente. Las distribuciones normales de las diferentes variables tienen varianza constante, pero diferentes medias y, estas medias están en un sub-espacio de "p" dimensiones de un espacio de "n" dimensiones. Estas hipótesis de partida proporcionan un desarrollo matemático elegante y que justifica el empleo de la regresión en muchos casos. Sin embargo, en la realidad, se ha abusado de la asunción de dichas hipótesis no siendo siempre correcta la aplicación de la técnica.

El sistema Gifi se relaciona mucho más con el análisis de componentes principales que con el de regresión múltiple. Esta técnica ha sido presentada como algo diferente al análisis de regresión: el hallazgo de combinaciones lineales de las variables originales que recogiesen el máximo de varianza de las mismas satisfaciendo una serie de criterios y estando muy relacionado con la interpretación geométrica de los resultados.

El análisis multivariante no lineal difiere de los anteriores precisamente por no buscar relaciones de tipo lineal entre las variables implicadas, basándose en el concepto de cuantificación óptima que va a estar presente a lo largo de toda esta exposición. La idea básica que subyace en el sistema Gifi es que las variables pueden agruparse en subconjuntos de diversas formas y ser cuantificadas empleando diversos tipos de restricciones. Combinando particiones de las variables con distintos tipos de restricciones en las mediciones de las mismas, se puede cubrir el espectro de la mayor parte de las técnicas multivariantes clásicas y, además, añadir unas cuantas más.

Así, haciendo referencia al contenido del módulo Categories del SPSS que permite la aplicación de estas técnicas, se pueden distinguir:

Diferentes generalizaciones del análisis de componentes principales: PRINCALS, HOMALS (análisis múltiple de correspondencias) y ANACOR (análisis de correspondencias).

Generalización del análisis canónico de correlaciones: CANALS.

Generalización de una forma de análisis canónico de correlaciones generalizado o múltiple: OVERALS.

De hecho, todas las aplicaciones descienden del OVERALS que es, por así decirlo, el generador de las mismas.

En definitiva, el planteamiento general del Sistema Gifi es el de una metodología de cuantificación óptima, es decir, de asignar a los valores de una variable, unos nuevos valores que sean óptimos para el propósito del análisis. Se siguen buscando relaciones entre las variables, pero no necesariamente de tipo lineal.

Las variables analizadas (aunque más que analizadas deberíamos decir transformadas, al menos en la primera parte del análisis), siempre son cualitativas[197], de forma que pueden ser codificadas mediante matrices indicador (filas con un 1 y resto de ceros), e imponer diversas restricciones de tipo ordinal y numérico en la cuantificación[198] de las categorías, como se verá más adelante.

Uno de los autores que más ha trabajado en la difusión de los conceptos teóricos del sistema Gifi es John Van der Geer. En su exposición destaca que una de las claves del método es mostrar que las relaciones entre variables pueden mejorar después de una cuantificación óptima de sus categorías.

Si tenemos en cuenta que la relación entre variables se puede medir a través de la correlación, entonces, resulta que, en definitiva se trata de incrementar la correlación que presentan las variables originales, si ello es posible, a base de cambiar los valores originales por unos nuevos bajo ciertas restricciones.

El sistema Gifi es una familia de técnicas con una serie de puntos en común y que difieren según las restricciones que se impongan en el análisis, el tratamiento de las variables y otros factores que seguidamente serán presentados. Por eso, antes de entrar en una técnica en particular, conviene familiarizarse con la terminología común de todas ellas y con los principales conceptos que intervienen en sus desarrollos.

4.2.3 LAS VARIABLES Y LOS "OBJETOS" EN EL SISTEMA GIFI

El centro de interés del sistema es el análisis de datos procedentes de variables categóricas o cualitativas (las más frecuentes en las encuestas). Estas variables clasifican objetos o individuos en un número limitado de grupos que habitualmente llamamos categorías o códigos.

En nuestro caso, la variable por excelencia sería la opción votada con los distintos partidos, la abstención, el voto en blanco y el nulo como categorías y los objetos, los electores entrevistados.

Cuando esta pregunta aparece en una encuesta electoral, lo único que sabemos es que los que han escogido el mismo partido u opción son electores "similares" y los que han escogido otras opciones son "disimilares"[199].

Una de las principales características del planteamiento de este sistema y su punto de partida es, precisamente, la forma en que el investigador percibe las variables.

El investigador trabaja con unos datos procedentes de encuestas de tal forma que, en la nomenclatura original del sistema, un elector o entrevistado es un "objeto" (object) acerca del cual se han obtenido diversas informaciones o "variables".

Se define una variable como el mecanismo que facilita la posibilidad de clasificar objetos en categorías diferentes y mutuamente excluyentes.

---

[197] Son cualitativas porque así lo requiere el planteamiento teórico del Sistema, pero a la práctica, en los análisis se emplean tanto variables cualitativas como combinaciones de cualitativas con cuantitativas.

[198] Se entiende por cuantificación la asignación de valores a las distintas categorías de una variable. Es un término que se usa constantemente en el sistema Gifi y con el que hay que familiarizarse.

[199] En estadística es habitual de hablar de sujetos similares y disimilares en lugar de similares y diferentes, de forma que emplearemos esta terminología.

Por ejemplo, si la variable es la edad, los objetos son las personas que han quedado clasificados en un grupo determinado de edad. Si la variable es el voto, los objetos son los votantes y quedan clasificados en las diferentes opciones que ofrecía la pregunta en la encuesta. Por tanto, las variables se definen mediante sus categorías.

De esta forma, un "objeto" en este contexto, es un término abstracto que define a cualquiera que haya quedado clasificado en las categorías de una variable.

4.2.4 TIPOS DE VARIABLES

En el sistema Gifi se distingue principalmente entre tres tipos de variables: nominales, ordinales y numéricas (o proporcionales).

Al igual que las clasificaciones clásicas, las variables nominales en este terreno serían aquellas en que no existe un orden definido a priori para las opciones y que no se puede medir la distancia entre una categoría y otra. Un ejemplo claro es el de la variable opción política votada: los códigos sólo sirven para distinguir entre unos partidos y otros o la abstención, el voto nulo y el voto en blanco.

La variable ordinal, en cambio, será aquella en que se tenga en cuenta que las categorías tienen fijado un orden a priori. Por ejemplo, una variable que recoja el grado de autosituación política en una escala de 1 (extrema izquierda) a 9 (extrema derecha). En este caso tampoco es posible medir en términos de distancia numérica la diferencia entre una categoría y la siguiente.

En el tercer tipo de variable, las categorías se cuantifican a priori con la intención de que las diferencias entre los números puedan interpretarse como diferencias entre los objetos.

En este contexto, las variables pertenecen a la categoría de nominales, ordinales o numéricas por decisión del investigador y no sólo por su naturaleza. Así, variables tradicionalmente identificadas como nominales, tales como la opción política votada, pueden pasar a ser ordinales si el investigador justifica que le interesa ordenar las categorías de izquierda a derecha o viceversa. Variables proporcionales como la edad, pueden transformarse en nominales cuando el interés se centra en aspectos tales como tener edad para votar o no, etc.

4.2.5 LOS CASOS DE NO RESPUESTA O "MISSINGS" EN EL SISTEMA GIFI

También es muy importante en este ámbito el explicitar que se entiende por "missing"[200] o caso de no respuesta".

Por comodidad, se mantiene la denominación inglesa de los casos de no respuesta "missings", ya que el término es más corto y se comprende de forma inmediata.

El tratamiento de los missings en el contexto del sistema Gifi puede responder a varios planteamientos:

[200] Un "missing" es un caso de no respuesta o casilla vacía en una base de datos. A lo largo de esta parte de la tesis se mantiene el término anglosajón por comodidad y por ser una expresión ampliamente aceptada en el ámbito estadístico.

a) Si hay individuos que tienen missing en todas las variables del estudio, lo más natural es descartarlos por completo del mismo, asumiendo el riesgo de sesgar la muestra.

b) Si hay individuos que tienen missings en algunas variables y en otras no, entonces:

b.1) Se puede introducir una nueva categoría para recoger los casos de no respuesta en las variables que lo necesiten. Esto da como resultado que los individuos que no tienen respuesta pasan a ser considerados como iguales, lo cual es discutible en la mayoría de las ocasiones. Esta opción se llama "active single".

Ejemplo:

VOTO: 1 3 4 6 3 1 1 1 etc.

VOTO 1 3 4 9 6 3 9 1 1 1 etc.

Se ha puesto el código 9 a todos los missings.

b.2) Se puede crear dentro de cada variable que tenga missings una nueva categoría para cada individuo que tenga un missing. Esto implica que sólo hay un individuo en esta categoría adicional. Este sistema se llama "active multiple" y su principal desventaja es que puede aumentar mucho las categorías de las variables correspondiendo a estas categorías una frecuencia marginal mínima.

Ejemplo:

VOTO: 1 3 4 6 3 1 1 1 etc.

VOTO 1 3 4 90 6 3 91 1 1 1 etc.

Se ha puesto el código 90 al primer missing, el 91 al segundo, etc.

b.3) Finalmente, se pueden dejar los missings que hay tal y como están. A este sistema se le llama "pasive".

La forma de tratar los missings en este contexto, depende del investigador. Por ejemplo, en preguntas que tienen como opciones de respuesta "si" y "no", puede considerarse que los que contestan "si" son iguales, mientras que pueden quedar dudas acerca de si los que contestan "no" también son iguales entre si. En un caso como este, el investigador puede ser que decida tratar las respuestas "no" como missings.

En preguntas que tienen la opción "no opina", el entrevistado puede escoger esta respuesta por muchos motivos diferentes: no entender la pregunta, no tener información para responder, no querer que se sepa lo que piensa, etc. En este caso, los individuos tampoco tienen porqué ser considerados como iguales y puede que el investigador decida tratarlos como missings.

También puede ser elección del investigador el usar diferentes tipos de missings en distintas variables.

Por otro lado, también se puede hacer extensivo el tratamiento activo y pasivo a las variables en lugar de limitarlo a los individuos. Una variable se trata de forma activa cuando se tienen en cuenta todos las observaciones de la misma para el análisis.

Una variable se trata de forma pasiva cuando es desechada para el análisis. Sin embargo, una vez llevado a cabo el análisis, no hay ningún inconveniente en analizar lo que sucede en las categorías de la variable pasiva. Por ejemplo, supongamos que se lleva a cabo un análisis de componentes principales en que se barajan las valoraciones de diez políticos y que se sabe el género a que pertenece cada entrevistado. El análisis puede llevarse a cabo dejando de lado la variable género. Esta variable no interviene en el cálculo de las valoraciones de los entrevistados, pero una vez finalizado el análisis, no hay nada que impida estudiar por separado las valoraciones de hombres y mujeres. Si la variable género hubiese intervenido, podría haber afectado a los resultados del análisis de alguna forma.

Por tanto, en este contexto, el investigador dispone de muchas opciones, lo cual no significa que esto de pie a que haya cierta arbitrariedad en el trabajo, sino que invita a los estudiosos a efectuar diversos planteamientos y a comparar los resultados, que en caso de resultar diferentes pueden dar lugar a nuevas preguntas.

4.2.6 LA CUANTIFICACIÓN

En el entorno del sistema Gifi se trabaja básicamente con variables cualitativas. Estas variables organizan a los individuos u objetos en un número de grupos limitado que llamamos categorías y que en principio representan una primera cuantificación de las variables.

Una de las cuestiones clave del sistema Gifi es mostrar que las relaciones entre las variables pueden mejorar después de una cuantificación óptima de sus categorías. Supongamos, por ejemplo, que la variable edad se relaciona con otras variables y que estas relaciones no son lineales. Si las relaciones fuesen lineales, una cuantificación óptima de las edades concordaría con los puntos medios de cada intervalo de edad. Si la relación fuese logarítmica, la cuantificación óptima sería una función logarítmica de las marcas de clase y, si la relación fuese cuadrática, la cuantificación óptima sería mayor para las edades medias y menor para los jóvenes y los mayores.

Obviamente, una cuantificación es óptima siempre y cuando una variable se relaciona con otra, ya que cuando la relación cambia, lo más probable es que la cuantificación óptima cambie también.

Para ilustrar este concepto, supongamos que tenemos una cuantificación hecha a priori para unos intervalos de edad. En un caso así, es posible confeccionar una gráfico que relacione la cuantificación a priori con la cuantificación óptima: en el eje de abcisas se coloca la cuantificación a priori y en el de ordenadas la óptima. La función resultante indica que la cuantificación óptima es alguna función matemática de la cuantificación a priori, pudiendo ser lineal, logarítmica, exponencial, cuadrática, etc.

Esto es lo que permite redefinir los conceptos de variables nominales, oridnales y proporcionales, dado que en ellas siempre existe una cuantificación a priori.

Así, una variable se considera proporcional si se requiere que la transformación proporcione una línea recta, es decir, sólo se acepta una transformación lineal de la cuantificación a priori.

Una variable se considera ordinal si se permite a la transformación mostrar una función monotónica creciente o decreciente. La cuantificación de estas variables requiere que los objetos de la misma categoría tengan la misma cuantificación, pero

si la categoría "j" tiene una cuantificación a priori superior a la de la categoría "i", el requerimiento es sólo que la cuantificación óptima de "j" no sea inferior a la de "i". En otras palabras, no se permite que "j" e "i" obtengan la misma cuantificación óptima.

Una variable se trata como nominal si no hay restricción en el tipo de transformación, aparte de requerir que los objetos de la misma categoría tengan la misma cuantificación. De todas formas, esto no implica que sujetos de diferentes categorías deban obtener cuantificaciones distintas y está permitido que distintas categorías obtengan idéntica cuantificación (esto se comprende mejor viendo la aplicación práctica). De todo ello, puede resultar que el tratamiento nominal de una variable podría llevarse a cabo mediante una transformación que fuese monotónicamente creciente. Esto indica que se puede obtener el mismo resultado si la variable es tratada como ordinal en lugar de cómo nominal.

También hay que remarcar que si una variable tiene sólo dos categorías, como en el caso de las dicotómicas, entonces, sólo hay dos puntos en el gráfico de transformación, que siempre formarán una línea recta. De ello, se puede desprender que no hay diferencia si una variable binaria es tratada como proporcional, ordinal o nominal.

El análisis clásico multivariante asume que cada variable tiene una cuantificación a priori y que las variables deben ser tratadas como proporcionales, de forma que todos los gráficos de transformación deben mostrar líneas rectas. Por tanto, si se permite que los gráficos de transformación no tengan que ser líneas rectas, entonces, es cuando se puede hablar del modemo análisis multivariable no lineal (que no excluye del todo las rectas).

Para comprender mejor el funcionamiento de la cuantificación, supongamos que disponemos de una variable nominal como la provincia de origen del entrevistado. Imaginemos que no tenemos una cuantificación a priori. En ese caso, se puede establecer, de forma arbitraria, una cuantificación como, por ejemplo, sustituir las etiquetas de las categorías por números naturales:

Variable: Provincia Cuantificación a priori (arbitraria)

Barcelona	1
Barcelona	1
Tarragona	2
Girona	3
Tarragona	2
Lléida	4
Lléida	4

Si una vez hecha esta operación tratamos la variable como nominal, los valores numéricos de estas categorías no juegan ningún papel más que el de permitirnos distinguir entre ellas. Esto significa que la cuantificación óptima resultante, será igual que la que se obtendría escogiendo, a priori, otros números para las provincias (por ejemplo: Barcelona=1, Girona=2, Lléida=3 y Tarragona=4).

Esto significa que un gráfico de transformación basado en una numeración a priori arbitraria no tiene ningún significado. Sin embargo, las aplicaciones informáticas desarrolladas por el software del sistema Gifi, requieren que las categorías de las variables tengan valores numéricos y, por eso, el usuario es obligado de una forma u otra a escoger una cuantificación a priori y, mientras la variable sea tratada como nominal, no tiene importancia la numeración que se haya escogido.

De lo anterior se desprende que los gráficos de transformación no tendrán sentido o significado si la cuantificación es arbitraria, por lo que se aconseja al investigador que escoja una cuantificación que aparente tener algún tipo de significado. Por ejemplo, si la variable recoge preferencias por partidos políticos, la cuantificación a priori que mejor podría funcionar sería una en que los partidos estuviesen ordenados de izquierda a derecha o viceversa. Incluso cuando esta ordenación es una vaga hipótesis, el efecto es que los gráficos de transformación resultan mucho más interesantes que los de cuantificaciones arbitrarias y, al final, son más útiles para interpretar los correspondientes a la cuantificación óptima producida por el programa.

4.2.7 LOS NIVELES DE MEDICIÓN EN EL SISTEMA GIFI

Para finalizar la presentación de conceptos previos, necesarios para entender las aplicaciones que veremos a continuación, es importante hablar del nivel de medición de las variables cuando se va a trabajar en este contexto.

El análisis clásico de componentes principales sólo admite variables numéricas.

El análisis no lineal de componentes principales admite variables nominales y ordinales. El tratamiento que pueden tener las variables que se cuantifican en este contexto es el siguiente:

Numérico: asume que una variable observada ya tiene valores numéricos en sus categorías. En la cuantificación, estos valores numéricos se respetan en el sentido de que sólo se permite una transformación en otros valores numéricos, pero como en el proceso de cuantificación la nueva matriz debe tener los valores estandarizados, sólo hay una solución de cuantificación, lo cual indica que, si todas las variables del análisis son numéricas el sistema Gifi proporcionará una solución idéntica a la del análisis clásico de componentes principales.

Ordinal: las categorías de las variables una vez cuantificadas tienen el mismo orden que las originales. La solución que se obtiene con este tratamiento no está condicionada por el número de dimensiones que ofrece. Los resultados siempre mostrarán, invariablemente que se producen cuantificaciones en las cuales algunas categorías adyacentes quedan agrupadas.

Nominal simple (single nominal): los objetos que tienen la misma categoría en las variables observadas, tendrán la misma categoría en las variables cuantificadas. El efecto de este tratamiento en una o más variables es que la solución deja de estar vinculada al número de factores. Por ejemplo, si tomamos una solución PRINCALS de dos dimensiones, en ésta se maximiza la suma de los dos primeros valores propios, pero esto no implica que su primer valor propio sea el mismo que en una solución de una sola dimensión. De hecho, será menor, pero la pérdida se compensa en una ganancia en el segundo valor propio.

Se podría argumentar que el tratamiento nominal simple de una variable tiene poco sentido pero se aconseja siempre que el investigador no tiene una idea a priori de cómo deben ser cuantificadas todas o, incluso sólo algunas de las categorías. Con ello se espera conseguir que el tratamiento acabe asignando una cierta posición u orden interpretable por medio de una cuantificación óptima.

Nominal múltiple (multiple nominal): como se comentará inmediatamente, la matriz que transforma o cuantifica las variables originales, puede proporcionar diversas soluciones en el caso de usar este nivel de tratamiento de las variables. Por tanto,

en este caso, las cuantificaciones de las observaciones originales pueden diferir en cada una de las soluciones.

En resumen: la técnica PRINCALS nos permite tratar a cada variable como queramos, ya sea como proporcional, ordinal, nominal simple o nominal múltiple, pero una vez escogido el tratamiento ordinal o nominal simple para alguna o algunas de ellas, la solución no quedará condicionada y los resultados dependerán del número de dimensiones que le hayamos pedido que extraiga.

4.2.8 El Análisis de Componentes Principales no lineal

Para comprender el análisis multivariante no lineal es necesario estar familiarizado con los análisis clásicos de Componentes Principales y Canónico, así como con la generalización de éstos, llamada Análisis Canónico generalizado, supuesto del que se parte en esta parte de la tesis, dado que la utilización de dichos análisis está muy extendida entre los politólogos y los estadísticos y, existe amplia bibliografía y documentación acerca de los mismos que no vale la pena reproducir en este trabajo.

El Análisis de Componentes Principales juega un papel muy importante en el análisis no lineal, ya sea en forma directa o indirecta (cuando se usa como herramienta intermedia). En este apartado se trata el tema de la aplicación de este análisis a variables categóricas siguiendo el desarrollo de Van de Geer mediante un ejemplo que después será reproducido mediante análisis no lineal para comparar resultados.

En la aplicación informática SPSS existe un módulo llamado "Categories" en que se hallan las principales técnicas relacionadas con el sistema Gifi. Para comenzar, vamos a centrar la atención en la técnica llamada PRINCALS.

La expresión PRINCALS indica que se refiere al Análisis de Componentes Principales (Principal Components), pero que su algoritmo de cálculo se lleva a cabo mediante Mínimos Cuadrados Alternantes (Alternating Least Squares).

Se trata de un programa con diversas opciones entre las que se hallan: el análisis de componentes principales clásico (cuando todas las variables son tratadas como numéricas o proporcionales), la solución que proporciona la técnica HOMALS (cuando todas las variables son tratadas como nominales múltiples) y las suyas propias en que se tratan conjuntamente variables nominales y ordinales.

El primer elemento que se necesita para iniciar un PRINCALS es una matriz o base de datos.

Vamos a llamar *H* a dicha matriz y consideremos que tiene *n* filas, cada una de las cuales es un caso o individuo y *m* columnas, cada una de las cuales representa a una variable. Las variables organizan a los individuos en categorías.

El programa PRINCALS cuantifica la matriz de datos *H* asignando valores numéricos a las diferentes categorías de cada variable. Esta operación proporciona una matriz que llamaremos *Q* y que contendrá las cuantificaciones de los datos de *H* con *n* filas y *m* columnas.

La técnica PRINCALS proporciona cuantificaciones de los sujetos en forma de columna. Cada columna tiene *n* cuantificaciones que llamaremos *x* o "object scores". En este contexto, el programa proporcionará diferentes soluciones de *Q* y *x*

a las que llamaremos dimensiones. Por tanto, ponemos un subíndice a Q y x para indicar las diferentes soluciones: Q_s y x_s. La s puede variar de 1 a p que es el número total de dimensiones.

El criterio de cuantificación de H es que x_s debe mostrar unas correlaciones muy elevadas con cada una de las variables que hay en Q_s, de forma que una solución es "buena" cuando satisface este criterio.

El PRINCALS actúa de forma que las columnas de las matrices Q_s tienen valores estandarizados (con media cero y varianza uno). Admite cuatro formas de cuantificación de las variables:

Numérica: ya se ha comentado que conduce a la solución clásica, precisamente a causa de la estandarización de los datos de las Q_s.

Ordinal: sean h_i las categorías observadas y q_i las cuantificadas. En este caso, para dos objetos g e i, se cumple que:

Si $h_{gi} = h_{ij}$ entonces $q_{gj} = q_{ij}$

Si $h_{gj} > h_{ij}$ entonces $q_{gj} \geq q_{ij}$

Nominal simple (single nominal): Si $h_{gj} = h_{ij}$ entonces $q_{gj} = q_{ij}$

Estas tres primeras posibilidades tienen en común que h_j tiene la misma cuantificación en todas las dimensiones, es decir, la columna j de Q_s es la misma sea cual sea la s. Las tres posibilidades varían en la cantidad de restricciones impuestas en la cuantificación, siendo la más restrictiva la de las variables numéricas.

Nominal múltiple (multiple nominal): la columna h i-ésima de Q_1 no será la misma que la de Q_2, etc.

A partir de esto, hay que comentar que el programa PRINCALS tiene cinco opciones de tratamiento de variables. Las primeras cuatro opciones implican que todas las variables sean tratadas de la misma forma: todas como numéricas, todas como ordinales, todas como nominales simples o todas como nominales múltiples. La quinta opción constituye la mezcla de las anteriores y, entonces, el investigador debe especificar en el programa qué tratamiento va a dar a cada una de las variables.

El programa PRINCALS no incluye la posibilidad de tratar a las variables como ordinales múltiples (multiple ordinal). Esto significaría que las cuantificaciones de las hj deberían ser diferentes en cada dimensión pero, además, obedeciendo las restricciones de orden y no hay suficientes grados de libertad como para que lo pueda hacer. Algo parecido sucedería con un tratamiento numérico múltiple (multiple numerical), de forma que tampoco se contempla esta posibilidad en la aplicación.

Tras leer lo anterior, lo lógico es preguntarse: ¿cómo hay que organizarse para definir el tratamiento de las variables al iniciar un PRINCALS?

La respuesta es que se trata de una cuestión de práctica y experiencia, además de razonamiento y reflexión. Por ejemplo, la variable que organiza a los electores según la opción votada es aparentemente nominal. Pero el investigador puede estar interesado en la búsqueda de una dimensión izquierda-derecha y, sobre esta base,

podría ordenar los partidos políticos de izquierda a derecha y solicitar que la cuantificación de los partidos respetase ese orden imponiendo un tratamiento ordinal. A veces, disponemos de variables cuyas categorías son escalas de ordenación del tipo "en contra de algo", "completamente a favor" con la opción "no sabe/no responde". Si consideramos que esta opción no se puede tomar como intermedia de las anteriores, entonces, lo mejor es dar un tratamiento nominal a la variable. El mismo tipo de razonamiento se seguiría con cualquier variable que interviniese en el análisis.

Otros aspectos importantes que hay que tener en cuenta antes de realizar un análisis con PRINCALS es que las variables no deberían sobrepasar las 12 categorías y que en cada categoría debe tener una frecuencia de casos razonable. Por ejemplo, en el caso de tener una lista con demasiados partidos políticos, algunos de los cuales tienen frecuencias de voto muy bajas, se puede proceder a efectuar agrupaciones lógicas de los mismos. Si al investigador le preocupa el efecto que estas operaciones pueden tener respecto de la cuantificación, lo que puede considerar es efectuar el análisis sin agrupar categorías y agrupándolas y comparar las cuantificaciones de las categorías. Si estas se parecen, no hay problema y los resultados finales no se verán afectados.

4.2.9 ILUSTRACIÓN CON UNA APLICACIÓN PRÁCTICA

Supongamos que tenemos los datos de 7 entrevistados respecto de 5 variables que tienen 4, 4, 4, 2 y 3 categorías respectivamente. Los datos aparecen en la siguiente tabla, siendo las filas los sujetos y las columnas las variables y, representan a la matriz *H* a la que hemos hecho referencia anteriormente:

1	1	1	2	2
2	2	2	2	2
1	1	2	1	1
4	2	4	2	3
4	4	4	2	3
3	3	3	1	2
2	3	3	2	2

Las variables son: el voto emitido en tres elecciones distintas, el género y el posicionamiento político de la persona.

Las categorías de estas variables tienen el siguiente significado:

Para el voto en las elecciones (columnas 1-3): 1 (IC), 2 (PSOE), 3 (CIU), 4 (PP)
Para el género (columna 4): 1 (hombre) 2 (mujer)
Para el posicionamiento político (columna 5): 1 (izquierda), 2 (centro) y 3 (derecha)

Las cinco variables son cualitativas y, por tanto, en teoría no es posible aplicarles un análisis clásico de componentes principales. Sin embargo, como investigadores, nosotros deseamos resumir la información contenida en estas variables en una o dos dimensiones, tal y como lo haríamos en dicho análisis. Para ello, podemos aplicar la técnica PRINCALS y ver como, de datos cualitativos, conseguimos extraer unos resultados interpretables y mejores a base de un análisis de componentes principales no lineal.

Sin embargo, antes de ver y comentar resultados calculados automáticamente por la aplicación informática, vamos a presentar los diferentes elementos del análisis y algunas relaciones algebraicas importantes del proceso:

H = matriz original de datos. Contiene *n* individuos (filas) y *m* variables (columnas).
Q = matriz que resulta tras cuantificar los valores de *H*. Tiene *n* filas y *m* columnas.

x = object scores (puntuaciones o valores de los individuos en las nuevas dimensiones). Hay 7 valores para cada dimensión que extraiga el análisis. En un análisis clásico, equivalen a las puntuaciones obtenidas por los individuos en cada componente principal extraída, de forma que podemos usar estos valores para representar a los individuos sobre cada dimensión.

a = component loadings o correlaciones entre las columnas de object scores y las columnas correspondientes de la matriz *Q*.

λ = valor propio, o variación original reproducida por una componente principal o dimensión.

Por las explicaciones que acompañan a cada elemento, se puede ver fácilmente que existe una estrecha relación entre todas ellas y que salvo el proceso de cuantificación, el resto se puede expresar mediante fórmulas matemáticas sencillas. De todas formas, en ningún momento se pretende desarrollar todo el algoritmo de cálculo de la cuantificación y si alguien desea más información sobre ello, puede consultar el libro de Gifi (1981ª, pgs. 163-196). Nuestro objetivo principal es mostrar la utilidad de esta técnica como sustituta del análisis de componentes principales clásico tan utilizado en el terreno del Análisis del Comportamiento Electoral, para el tipo de variables que se usan en él con mayor frecuencia: las cualitativas.

En primer lugar, a partir de la matriz original *H* el programa inicia un proceso iterativo para llevar a cabo la cuantificación de los valores originales de las variables de manera que estos valores cumplan una serie de requisitos. Ya se ha comentado que se entiende por cuantificación: la asignación de unos nuevos números a las diferentes categorías de una variable. Las categorías originales no son más que códigos y, en cambio, las cuantificaciones son números en el sentido más estricto de la palabra. El principio en el que se basa la cuantificación es que las puntuaciones que obtengan los individuos (las *x*) tengan la máxima correlación posible con los valores de cada columna de la matriz cuantificada *Q*. Por otro lado, hay que decir que PRINCALS cuantifica las variables de forma que las columnas de la matriz *Q* queden estandarizadas (es decir, tengan media cero y varianza 1).

Antes de llevar a cabo la cuantificación, es preciso escoger un nivel de medición para las variables y aclarar nuestro criterio acerca de los missings.

En este caso, decidimos asignar un nivel de medición ordinal simple a todas las variables basándonos en que las fuerzas políticas están ordenadas de izquierda a derecha y el posicionamiento político también. El género, podemos dejarlo también en este mismo nivel para no complicar más la primera aproximación que efectuamos, pero podría haber sido nominal simple, por ejemplo.

Asimismo, vamos a trabajar con todos los datos no declarando missing a ningún individuo, ni a ninguna variable o categoría de las mismas y, finalmente, vamos a extraer dos componentes principales.

Una vez efectuada la cuantificación, la matriz Q que se obtiene en este caso es la siguiente:

-1.53	-1.55	-2.20	0.63	0.03
0.36	0.51	-0.30	0.63	0.03
-1.53	-1.55	-0.30	-1.58	-2.20
0.98	0.51	0.70	0.63	1.04
0.98	1.07	0.70	0.63	1.04
0.38	0.51	0.70	-1.58	0.03
0.36	0.51	0.70	0.63	0.03

Comparándola con la original, se puede apreciar que el nivel ordinal escogido ha proporcionado unas cuantificaciones que respetan el orden de las categorías. Así, para el valor más bajo de la primera variable original (1) la cuantificación es el valor más bajo de la columna correspondiente en Q, es decir (-1.53), para el valor 2 tenemos 0.36, para el valor 3 tenemos 0.38 y, para el valor 4, que es el más alto, tenemos 0.98. Lo mismo sucede con el resto de los valores de ambas matrices.

Posteriormente, otro resultado que aparece son las "Component Loadings" o correlaciones entre las puntuaciones de los individuos en cada dimensión (*x*) y las columnas de la matriz Q. Estos valores ejercen una función parecida a las correlaciones entre las componentes de un análisis clásico y sus variables originales y ayudan a interpretar y dotar de significado a los resultados. En un análisis PRINCALS se cumple que un valor propio de una componente principal:

$\lambda = \Sigma a_j^2/m$ donde *m* es el número de columnas de *H* o de *Q*.

PARTI1	-,987	,136
PARTI2	-,968	,169
PARTI3	-,733	,644
GENERO	-,486	-,826
POSICPOL	-,842	-,437

Al ver estas correlaciones, podemos asociar a la primera columna, que es la correspondiente a la primera componente principal o dimensión, con todas las variables de corte político y a la segunda con el género. Por tanto, se podría decir que:

Dimensión 1 = actuación, características políticas del individuo
Dimensión 2 = género del individuo

Con ello se puede comprobar que los resultados se van a parecer mucho a los de componentes principales clásico en cuanto a la forma de llevar a cabo la interpretación. El signo negativo de las correlaciones está indicando que las dimensiones se interpretarán a la inversa de lo que cabría esperar, es decir, el eje político irá de derecha a izquierda, en lugar de ir de izquierda a derecha y, en la segunda dimensión, los hombres tendrán valores positivos y las mujeres negativos, también contra lo que parecía que tendría que ser. Por ejemplo, la correlación entre lo votado en la primera elección y la dimensión 1 es negativa y muy alta. Eso significa que a valores altos de PARTI1 (derechas) le corresponden valores bajos de la dimensión 1 y a valores bajos de PARTI1 (izquierdas), valores altos de la dimensión 1. Por otro lado, la correlación entre el GÉNERO y la dimensión 2 es también negativa e intensa, de forma que al valor más alto (mujer=2) le corresponden valores bajos de esa dimensión y, al valor bajo (hombre=1) le corresponden valores altos de esa dimensión.

Para calcular la cantidad de variación recogida por cada componente, basta con aplicar la fórmula anterior a estos números:

$\lambda_1 = (-0.987^2)+(-0.968^2)+(-0.733^2)+(-0.486^2)+(-0.842^2)/5 = 0.6787$

$\lambda_2 = (0.136^2)+(0.169^2)+(0.644^2)+(-0.826^2)+(-0.437^2)/5 = 0.2670$

De esta forma tenemos los valores propios de cada una de las dos componentes principales y sabemos que la primera recoge el 67.87% de la información original (lo cual justifica la necesidad de una segunda componente al no alcanzar el 75% deseable en estos casos)y que, la segunda recoge el 26.7% de la misma. En total, un análisis como este reproduciría un 94.57% de la variación original, cifra que indica que se trata de una aplicación bastante acertada.

Las puntuaciones de los individuos en cada dimensión u "object scores" se calculan de la siguiente forma:

$$x = \Sigma q_j \cdot a_j / m \cdot \lambda$$

Por ejemplo, la primera puntuación del primer individuo sobre la primera componente principal sería:

$$x = (-1.53)(-0.987)+(-1.55)(-0.968)+(-2.20)(-0.733)+(0.63)(-0.486)+(0.03)(-0.842)/5(0.6787) = 1.27$$

En conjunto, el análisis proporciona los siguientes "object scores":

Dimensión 1	Dimensión 2
1.27	-1.81
-0.28	-0.45
1.72	1.20
-0.93	-0.23
-1.09	-0.16
-0.19	1.41
-0.50	0.04

De forma que el quinto individuo es el que tiene un comportamiento político más acentuado hacia la derecha, tanto en votaciones como en posicionamiento y hay que identificarlo con la parte inferior de la segunda dimensión, es decir, que tiene que ser una mujer con casi toda seguridad. En cambio, el tercer individuo es el que tiene un comportamiento más de izquierdas y un valor alto y positivo en la segunda dimensión, de forma que se trata casi con toda seguridad de un hombre. Los datos originales acerca de estos dos individuos así lo confirman. Para el quinto individuo teníamos PP, PP, PP, Mujer, Derechas y para el tercero IC, IC, PSOE, Hombre, Izquierdas.

Por tanto, insistimos en que la primera dimensión va de derecha a izquierda en cuanto a características y actuación política y, la segunda dimensión coloca al género masculino hacia los valores positivos de su escala y a las mujeres en los negativos.

En los resultados se ofrecen las llamadas "single category coordinates" o coordenadas simples que se calculan multiplicando las cuantificaciones por su

correspondiente component loading *a*. Es decir, que se hallan multiplicando cada valor de la primera columna de *Q* por el primer valor de la columna de las component loadings, cada valor de la segunda columna de *Q* por el segundo valor de la columna de las component loadings y así sucesivamente. Por ejemplo, para la primera dimensión:

(-1.53)(-0.987) = 1.51
(0.36)(-0.987) = -0.35
(-1.53)(-0.987) = 1.51 etc. etc.

1.51	1.50	1.61	-0.31	-0.02
-0.35	-0.48	0.22	-0.31	-0.02
1.51	1.50	0.22	0.77	1.85
-0.97	-0.48	-0.51	-0.31	-0.88
-0.97	-1.03	-0.51	-0.31	-0.88
-0.38	-0.49	-0.51	0.77	-0.02
-0.35	-0.49	-0.51	-0.31	-0.02

Cuando se dispone de estas coordenadas y de las de la segunda dimensión, es posible representar a los individuos en un par de ejes catersianos cada uno de los cuales sea una componente principal extraída. Se usan estas coordenadas si el nivel de medición de las variables ha sido simple (single) y las que presentamos a continuación si ha sido nominal múltiple (multiple nominal).

Finalmente, a partir de los "object scores" y del número de individuos que aparecen con el mismo valor en las variables originales, se calcula una matriz *Y* llamada de coordenadas múltiples (multiple category coordinates). Esta matriz tiene, al igual que *H* y que *Q*, *n* filas y *m* columnas. Sus elementos son medias de "object scores" en función del número de individuos u objetos que están en la misma categoría. Por ejemplo, en la variable PARTI1 hay 2 individuos en la categoría 1 (IC) que son el primero y el tercero. La multiple category coordinate que les corresponde es:

(1.27 + 1.72)/2 = 1.49

la segunda de la primera columna sería:

(-0.28)+(-0.50)/2 = -0.39

la tercera de la primera columna sería:

(-0.93)+(-1.09)/2 = -1.01

y así sucesivamente hasta completar toda la matriz que se ofrece a continuación, correspondiente a la primera dimensión:

1.49	1.49	1.27	-0.31	-0.32
-0.39	-0.61	0.72	-0.31	-0.32
1.49	1.49	0.72	0.77	1.85
-1.01	-0.61	-1.01	-0.31	-0.88
-1.01	-1.09	-1.01	-0.31	-0.88
-0.19	-0.34	-0.34	0.77	-0.32
-0.39	-0.34	-0.34	-0.31	-0.32

Una vez presentados los principales conceptos y las relaciones algebraicas más importantes del programa PRINCALS, vamos a dejar que la aplicación informática trabaje con los datos anteriores y a ver la forma en que los presenta y los podemos interpretar.

Por tanto, indicamos en el programa que todas las variables tienen nivel de medición ordinal simple y aplicamos un PRINCALS a los datos con el SPSS. Los resultados son los siguientes:

```
                P R I N C A L S - VERSION 0.6
                          BY
                DEPARTMENT OF DATA THEORY
            UNIVERSITY OF LEIDEN, THE NETHERLANDS

 The number of observations used in the analysis = 7

                      List of Variables

                                                    Number of    Measurement
 Variable  Variable Label                          Categories      Level
 --------  --------------                          ----------   -----------
 PARTI1    PARTIDO VOTADO EN ELECCIÓN 1                 4       Ordinal
 PARTI2    PARTIDO VOTADO EN ELECCIÓN 2                 4       Ordinal
 PARTI3    PARTIDO VOTADO EN ELECCIÓN 3                 4       Ordinal
 GENERO    GENERO DEL ELECTOR                           2       Ordinal
 POSICPOL  POSICIONAMIENTO POLÍTICO                     3       Ordinal
```

En primer lugar podemos apreciar el nombre, versión y procedencia del programa. En segundo lugar se nos informa acerca del número de observaciones utilizadas en el análisis (7 en este caso). Seguidamente disponemos de una lista de las variables con el número de categorías que tiene cada una y el nivel de medición que hemos decidido otorgarles (en este caso ordinal). Se trata de un ejemplo elaborado con los mismos datos ficticios que hemos usado para poder comentar cuestiones de cálculo del algoritmo. Más adelante se presenta un caso con datos reales una vez dominada la parte de notación e interpretación de resultados.

Por tanto, se supone que 7 electores nos han dicho lo que votaron en tres elecciones diferentes y que sabemos el género al que pertenecen y su posicionamiento político. Recordemos que el significado de las categorías de las variables es el siguiente:

Para el voto en las elecciones: 1 (IC), 2 (PSOE), 3 (CIU), 4 (PP)

Para el género 1 (hombre) 2 (mujer)

Para el posicionamiento político 1 (izquierda), 2 (centro) y 3 (derecha)

Seguidamente, el programa nos ofrece una tabla con las frecuencias marginales de cada una de las variables. Así, por ejemplo, ninguna variable tiene missings, 2 electores votaron a IC en la primera elección, 2 al PSOE, 1 a CIU y 2 al PP. Disponemos de los datos de 2 hombres y 5 mujeres y, había una persona de izquierdas, 4 de centro y 2 de derechas.

Marginal Frequencies

Variable	Missing	Categories 1	2	3	4
PARTI1	0	2	2	1	2
PARTI2	0	2	2	2	1
PARTI3	0	1	2	2	2
GENERO	0	2	5		
POSICPOL	0	1	4	2	

A partir de este punto, el programa inicia el proceso de búsqueda de la cuantificación óptima de las observaciones de la matriz *H*, proceso que es de tipo iterativo y que se detiene cuando se alcanza la convergencia. Las iteraciones dejan de sucederse cuando el ajuste total (Total Fit) entre las dos últimas iteraciones es inferior al valor de convergencia que tiene por defecto el programa. Este valor es 0.00001. En la columna de la derecha (Iteration Change) se puede seguir su evolución y ver que el último valor es 0.000006 < 0.00001.

*** The History of Iterations ***

Iteration	Total Fit	Total Loss	Multiple Loss	Single Loss	Iteration Change
1	,9278140	1,0721860	,8150777	,2571083	,0247510
2	,9348597	1,0651403	,8192426	,2458977	,0070457
3	,9362686	1,0637314	,8243141	,2394173	,0014089
4	,9370361	1,0629639	,8261222	,2368417	,0007675
5	,9376420	1,0623580	,8264488	,2359092	,0006058
6	,9381701	1,0618299	,8266743	,2351556	,0005282
7	,9386847	1,0613153	,8271820	,2341333	,0005145
8	,9391996	1,0608004	,8279896	,2328108	,0005149
9	,9397122	1,0602878	,8290007	,2312871	,0005126
10	,9402204	1,0597796	,8301226	,2296570	,0005082
11	,9407238	1,0592762	,8313006	,2279756	,0005035
12	,9412220	1,0587780	,8325116	,2262664	,0004981
13	,9417132	1,0582868	,8337484	,2245385	,0004912
14	,9421950	1,0578050	,8350068	,2227983	,0004818
15	,9426645	1,0573355	,8362810	,2210545	,0004695
16	,9431186	1,0568814	,8375624	,2193190	,0004541
17	,9435544	1,0564456	,8388403	,2176052	,0004358
18	,9439693	1,0560307	,8401034	,2159273	,0004148
19	,9443609	1,0556391	,8413407	,2142984	,0003916
20	,9446977	1,0553023	,8425421	,2127602	,0003368
21	,9448972	1,0551028	,8442762	,2108266	,0001995
22	,9450166	1,0549834	,8465688	,2084146	,0001194
23	,9450887	1,0549113	,8484978	,2064135	,0000721
24	,9451325	1,0548675	,8500470	,2048205	,0000438
25	,9451594	1,0548406	,8512709	,2035697	,0000269
26	,9451761	1,0548239	,8522339	,2025900	,0000167
27	,9451866	1,0548134	,8529924	,2018210	,0000105
28	,9451932	1,0548068	,8535918	,2012150	,0000066

The iterative process stops because the convergence test value is reached.

El siguiente resultado proporciona los valores propios (eigenvalues) de las dos dimensiones o componentes principales que hemos solicitado que extraiga el análisis. Su sentido es equivalente al de los valores propios de un Análisis Clásico de Componentes Principales y, por tanto, miden la varianza de la información original, reproducida por cada componente. A pesar de que no se nos da un informe en términos de porcentaje, la suma de los dos valores propios es 0.9452, lo cual representa un 94.52% de reproducción de la variación original, que es más que suficiente. Por tanto, no es necesario pedir más de dos componentes y con una

hubiésemos alcanzado el 67.84% de reproducción de la información original. Para calcular estos valores propios ya hemos visto que se necesitan elementos que salen en la parte final de los resultados, concretamente las llamadas "Component Loadings" que simbolizamos mediante *ai*. Cada valor propio se obtiene calculando:

$\lambda = \Sigma\ a_j^2 / m$

Así, $0.6784 = [(-0.987^2)+(-0.968^2)+(-0.733^2)+(-0.486^2)+(-0.842^2)]/ 5$

Como ya habíamos hecho anteriormente.

Dimension	Eigenvalue
1	,6784
2	,2668

Finalmente, se puede apreciar que la suma de los dos valores propios es igual al último valor que aparece en la columna Total Fit (Ajuste total) de la tabla que resume el proceso de las iteraciones y representa la cantidad de variación original reproducida por el análisis.

En un Análisis Clásico, el siguiente paso sería tratar de dotar de significado a estas dimensiones mediante el análisis de su correlación con las variables originales. En este tipo de análisis, el proceso de presentación es algo diferente, aunque el objetivo sea el mismo.

Los siguientes resultados proporcionan las frecuencias marginales de las categorías de cada variable y la cuantificación óptima de dichas categorías, es decir, los valores de la matriz Q, sólo que en lugar de estar reunidos en una tabla, se nos dan para cada variable por separado.

Variable: PARTI1 PARTIDO VOTADO EN ELECCIÓN 1

Type: Ordinal Missing: 0

Category:		Marginal Frequency	Quantification
1	IC	2	-1,53
2	PSC	2	,36
3	CIU	1	,38
4	PP	2	,98

Single Category Coordinates

Category	Dimension 1	Dimension 2
1	1,51	-,21
2	-,35	,05
3	-,38	,05
4	-,97	,13

Multiple Category Coordinates

Category	Dimension 1	Dimension 2
1	1,49	-,31
2	-,39	-,21
3	-,19	1,41
4	-1,01	-,19

Tomemos por ejemplo, la primera variable que es el partido votado en la elección 1. Según se puede apreciar se nos da el nombre y significado de la variable, se nos indica que no hay missings y que el nivel de medición es ordinal.

Seguidamente se puede ver la frecuencia marginal de la categoría 1 (IC) que son dos votantes y la cuantificación óptima de esta categoría que es –1.53. Si recopilamos todas las cuantificaciones óptimas de las categorías de las cinco variables, podemos construir la matriz *Q*, que ya hemos presentado anteriormente aunque a efectos de análisis no es necesario, ni el programa la proporciona en estos términos:

-1.53	-1.55	-2.20	0.63	0.03
0.36	0.51	-0.30	0.63	0.03
-1.53	-1.55	-0.30	-1.58	-2.20
0.98	0.51	0.70	0.63	1.04
0.98	1.07	0.70	0.63	1.04
0.38	0.51	0.70	-1.58	0.03
0.36	0.51	0.70	0.63	0.03

Por tanto, en los resultados del programa, cuando trabajamos con variables medidas a nivel proporcional, ordinal y nominal simple, los elementos de la matriz Q o cuantificaciones óptimas de las categorías de dichas variables, aparecen bajo la denominación "Quantifications". Estas cuantificaciones sirven para los cálculos posteriores de otros elementos que aparecen en los resultados y que son la base de la interpretación final.

Si seguimos con la segunda variable, que presenta un cuadro similar a las restantes, veremos las cifras agrupadas bajo el nombre de "Single Category Coordinates" y otras bajo el nombre de "Multiple Category Coordinates". Ya hemos explicado el significado de estas coordenadas y su forma de cálculo. Por tanto, ponemos las tablas de las cuatro variables que quedan y continuamos las explicaciones tras ellas.

```
Variable: PARTI2     PARTIDO VOTADO EN ELECCIÓN 2
---------
Type: Ordinal              Missing:       0

Category:                          Marginal Frequency          Quantification
---------                          ------------------          --------------
       1  IC                                        2                   -1,55
       2  PSC                                       2                     ,51
       3  CIU                                       2                     ,51
       4  PP                                        1                    1,07

Single Category Coordinates
---------------------------
Category          Dimension

                     1         2

       1        1,50      -,26
       2        -,49       ,09
       3        -,49       ,09
       4       -1,03       ,18

Multiple Category Coordinates
-----------------------------
Category          Dimension

                     1         2

       1        1,49      -,31
       2        -,61      -,34
       3        -,34       ,73
       4       -1,09      -,16
```

Variable: PARTI3 PARTIDO VOTADO EN ELECCIÓN 3

Type: Ordinal Missing: 0

Category:		Marginal Frequency	Quantification
1	IC	1	-2,20
2	PSC	2	-,30
3	CIU	2	,70
4	PP	2	,70

Single Category Coordinates

Category	Dimension 1	Dimension 2
1	1,61	-1,42
2	,22	-,20
3	-,51	,45
4	-,51	,45

Multiple Category Coordinates

Category	Dimension 1	Dimension 2
1	1,27	-1,81
2	,72	,37
3	-,34	,73
4	-1,01	-,19

Variable: GENERO GÉNERO DEL ELECTOR

Type: Ordinal Missing: 0

Category:		Marginal Frequency	Quantification
1	HOMBRE	2	-1,58
2	MUJER	5	,63

Single Category Coordinates

Category	Dimension 1	Dimension 2
1	,77	1,31
2	-,31	-,52

Multiple Category Coordinates

Category	Dimension 1	Dimension 2
1	,77	1,31
2	-,31	-,52

```
Variable: POSICPOL   POSICIONAMIENTO POLÍTICO
---------
Type: Ordinal              Missing:      0

Category:                          Marginal Frequency          Quantification
---------                          ------------------          --------------
       1  IZQUIERDA                                1                   -2,20
       2  CENTRO                                   4                     ,03
       3  DERECHA                                  2                    1,04

Single Category Coordinates
---------------------------

Category          Dimension

                      1         2

       1        1,85       ,96
       2        -,02      -,01
       3        -,88      -,46

Multiple Category Coordinates
-----------------------------

Category          Dimension

                      1         2

       1        1,72      1,20
       2         ,08      -,20
       3       -1,01      -,19
```

En la siguiente tabla resumen del análisis, las cifras que aparecen bajo el nombre de "Single Fit" son las "Component Loadings" al cuadrado, que aparecen más abajo. Así, por ejemplo:

$0.974 = (-0.987^2)$
$0.018 = (0.136^2)$ etc.

La suma de estos valores de las dos dimensiones proporciona las cifras que aparecen bajo el rótulo "Row Sums" y al final, se extraern y presentan las medias de las tres columnas. Cuando en el análisis se emplean niveles de medición múltiple se consulta el apartado Multiple Fit y, cuando se emplean niveles de medición simples (como en nuestro caso), se consultan las cifras del apartado Single Fit. La función de esta información es análoga a la de las comunalidades en un análisis clásico. Así, siempre mirando las columnas del "Single Fit", poedmos decir que la primera dimensión capta el 97,4% de las variable PARTI1, mientras que la segunda dimensión sólo capta un 1.8%. Entre ambas captan el 99.2% de su información o variación. Las medias indican que la dimensión uno capta, por término medio un 67.8% de la variación original y que la dimensión 2 capta un 26.7%. Entre ambas, alcanzan una media de 95.4%, cifras éstas últimas que coinciden con los valores propios y su interpretación. Las cifras del apartado Multiple Fit no tienen sentido en este caso y, por tanto, no afecta para nada el que salgan valores superiores a 1 en lo que se supone que son una especie de comunalidades. Sin embargo, se ha optado por dejar todos los resultados que ofrece el análisis al completo, precisamente para aclarar cuáles deben ser considerados y cuáles no en función de las decisiones que hayamos tomado al organizar los datos.

La conclusión es que la dimensión 1 se forma, sobre todo, a base de la información contenida en las variables PARTI1 y PARTI2 y que la dimensión 2 se forma, sobre todo sobre la variable GÉNERO, siendo estas cifras una medida de la importancia de las mismas en las nuevas dimensiones. Por tanto, es más consistente y clara la dimensión 1 que la 2.

Summary of Analysis

Multiple Fit

Variable	Row Sums	Dimension 1	2
PARTI1	1,316	,980	,336
PARTI2	1,162	,947	,215
PARTI3	1,376	,705	,671
GENERO	,918	,236	,682
POSICPOL	,960	,721	,239
Mean:	1,146	,718	,428

Single Fit

Variable	Row Sums	Dimension 1	2
PARTI1	,992	,974	,018
PARTI2	,965	,937	,028
PARTI3	,951	,537	,415
GENERO	,918	,236	,682
POSICPOL	,899	,708	,191
Mean:	,945	,678	,267

El resultado que ofrece la siguiente tabla son las llamadas "Component Loadings" que son las correlaciones entre los "Object Scores" (antes citados como *x*) y las columnas de la matriz *Q* de cuantificaciones. Se puede comprobar que si elevamos al cuadrado los coeficientes de correlación de cada columna, los sumamos y dividimos por el número de columnas de la matriz *Q* (5 en este caso), obtendremos, respectivamente, cada uno de los valores propios de las dimensiones, comprobación que no repetimos aquí por haberla realizado en la parte de presentación de relaciones algebraicas. Recordemos que la interpretación que se hizo en dicha representación es que la dimensión 1 está negativamente relacionada con todas las variables políticas, con coeficientes muy elevados, especialmente respecto de las dos primeras variables y que la dimensión dos está negativa y altamente correlacionada con el género, lo cual nos ha permitido dotarlas de significado.

Component Loadings

Variable	Dimension 1	2
PARTI1	-,987	,136
PARTI2	-,968	,169
PARTI3	-,733	,644
GENERO	-,486	-,826
POSICPOL	-,842	-,437

En la siguiente parte de los resultados aparece una matriz de correlaciones entre las variables una vez llevada a cabo su cuantificación óptima. La correlación más elevada se da entre PARTI1 y PARTI2, como era de esperar por su participación en la primera dimensión y la más baja, entre el GÉNERO y PARTI3, con signo negativo.

* Correlations between Optimally Scaled Variables *

	PARTI1	PARTI2	PARTI3	GENERO	POSICPOL
PARTI1	*				
PARTI2	,979	*			
PARTI3	,802	,794	*		
GENERO	,363	,330	-,126	*	
POSICPOL	,770	,722	,300	,686	*

Ya se ha explicado que los "Object Scores" (x) son una suma ponderada de las columnas de la matriz Q, con pesos igual a los valores proporcionados por las "Component Loadings" divididos por el número de columnas que viene multiplicado por el valor propio correspondiente a cada dimensión. Por tanto, si llamamos a_i a cada "Component Loading", podemos escribir:

$x = \Sigma\, q_j \cdot a_j / m \cdot \lambda$

Siendo λ un valor propio y m el número de columnas de la matriz Q.

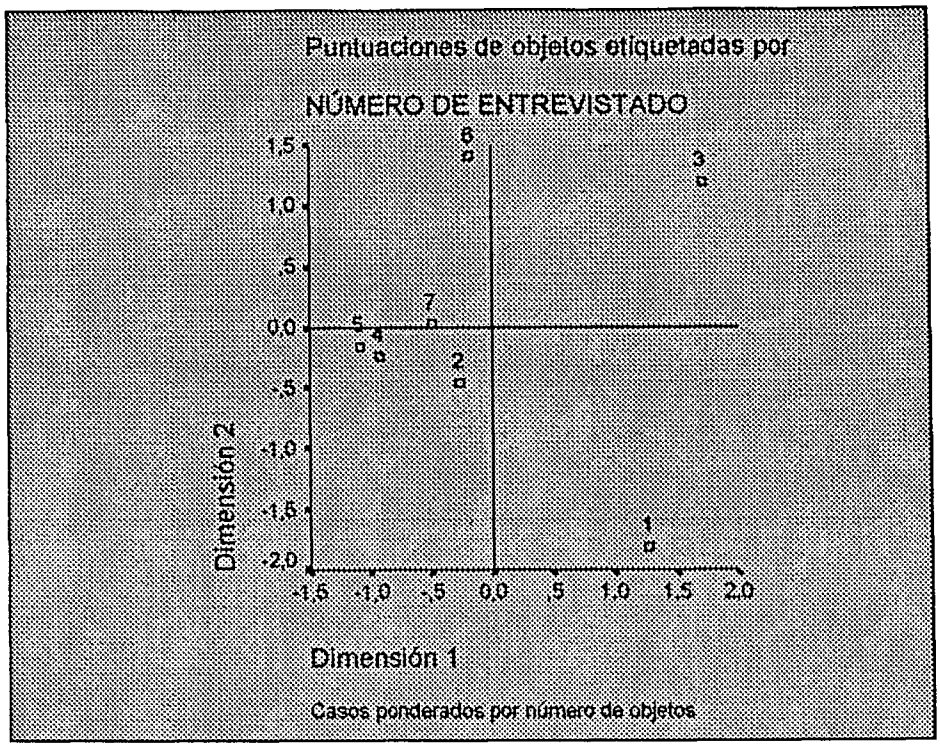
Así, por ejemplo, el valor 1.27 de la primera dimensión se obtiene haciendo:

(-1.53)(-0.987)+(-1.55)(-0.968)+(-2.20)(-0.733)+(0.63)(-0.486)+(0.03)(-0.842) / (5)(0.6784)

The Object Scores are:

Object	Dimension 1	Dimension 2
1	1,27	-1,81
2	-,28	-,45
3	1,72	1,20
4	-,93	-,23
5	-1,09	-,16
6	-,19	1,41
7	-,50	,04

Como ya se ha expuesto, estos valores equivalen a las puntuaciones que se obtienen en un análisis clásico y permiten situar a los individuos sobre la escala de las dimensiones o componentes halladas. Así, dando números del 1 al 7 a los entrevistados, se puede obtener la siguiente representación:



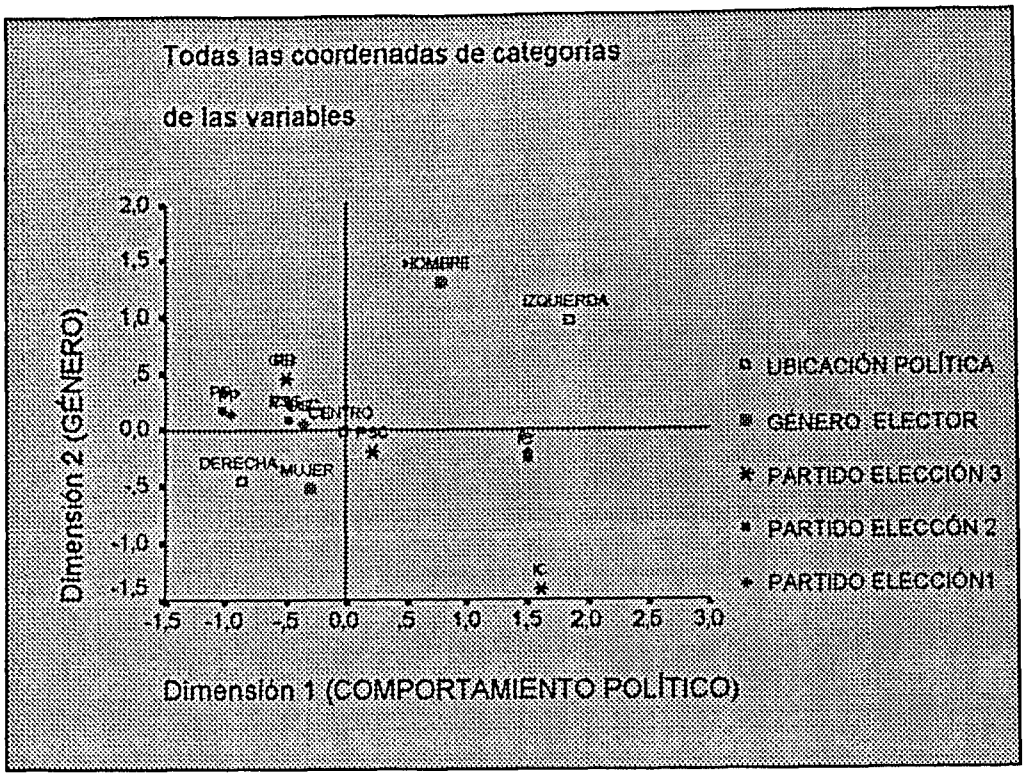
El sujeto número 1 está, en la parte de comportamiento político de izquierdas (recordemos que el eje va de derechas a izquierdas) y en la parte inferior de la dimensión 2, que es la correspondiente al género femenino. El sujeto 2 tiene un comportamiento más centrista y puede ser más bien de género femenino. El individuo 3 es un hombre de izquierdas, el 4 un hombre o mujer de derechas, el 5 casi es igual al 4, el 6 es un hombre centrista y el 7 podría ser un hombre centrista.

Si recordamos los datos originales:

1 IC	1 IC	1 IC	2 H	2 C
2 PSOE	2 PSOE	2 PSOE	2 H	2 C
1 IC	1 IC	2 PSOE	1 M	1 I
4 PP	2 PSOE	4 PP	2 H	3 D
4 PP	4 PP	4 PP	2 H	3 D
3 CIU	3 CIU	3 CIU	1 M	2 C
2 PSOE	3 CIU	3 CIU	2 H	2 C

Podremos ver que no está nada mal el retrato que nos ha proporcionado el análisis respecto a las variables políticas, siendo, en cambio más errático en lo referente al género. Todo ello era de esperar una vez interpretados todos los estadísticos que acompañan al resumen de toda esta información en sólo dos variables.

Otro gráfico que nos proporciona la solución del PRINCALS es el que representa a las categorías respecto de las dos dimensiones. Como se puede apreciar, se identifica claramente la izquierda con la derecha del eje horizontal, el centro con el centro del diagrama y la derecha con la izquierda del eje. Asimismo, se pueden ver las categorías del género con las mujeres en la parte inferior de la dimensión 2 y los hombres en la superior. Y, finalmente, se pueden ver los partidos votados en las tres elecciones, aunque menos claramente debido a solapamientos de comportamiento de los individuos.



Las coordenadas de los partidos políticos que quedan solapados en el gráfico anterior, las ofrece el programa en el siguiente cuadro:

```
Summary of multiple points in chart

   Dim1     Dim2     Actual label or name

   -,49      ,09     PSC
   -,49      ,09     CIU
   -,51      ,45     CIU
   -,51      ,45     PP
```

El programa, igual que en análisis clásico de componentes principales, añade, si lo deseamos, las puntuaciones de los individuos a la matriz original de datos para poder luego operar con ellos ya sea a nivel gráfico o para su empleo en otros tipos de análisis. Esta información y el nombre de las nuevas variables guardadas es la que aparece en el siguiente cuadro:

```
Object scores for 2 dimensions were saved with the rootname: PRI
Following object scores were added to the working file:

Name        Label

PRI1_1      Dimension 1
PRI2_1      Dimension 2
```

A partir de lo expuesto conviene volver atrás y efectuar una reflexión. En primer lugar, porqué se ha escogido la técnica PRINCALS como representante de las novedades dentro del terreno de la Estadística para el análisis de datos desagregados, ¿es que no hay otras?. La respuesta es que a pesar de que existen otras, a nivel práctico, así como los estadísticos siguen la dinámica de su parcela y aplican nuevas técnicas a todo tipo de datos, el politólogo, al menos el español, se

ha estancado un tanto en las técnicas multivariantes clásicas y no ha prestado suficiente atención a análisis como el de componentes principales no lineal.

Por tanto, se escoge precisamente esta técnica y no otras porque es la que representa el eslabón de continuidad con el punto en que se halla actualmente el politólogo de nuestro país. Se trata de advertir que ahora ya no hay excusa para seguir empleando técnicas multivariantes sin propiedad, puesto que aunque tienen ya unos diez años de existencia, las técnicas multivariantes para datos categóricos son muy desconocidas y, como constituyen la secuencia natural de las novedades que se van desarrollando, es imprescindible el presentarlas.

Por otro lado, ya se ha dicho en la introducción que este trabajo no pretende ser un compendio estadístico de técnicas y sus aplicaciones, sino que trata de situar una parcela de la Ciencia Política, tal y como se desarrolla en España, en la Ciencia en general y mostrar la relación que tiene con la Estadística en particular y el punto en que se encuentra dicha relación.

De esta forma, se justifica esta elección y se deja abierto el camino para explorar no sólo el análisis multivariante no lineal, sino muchas otras técnicas que se relacionan con análisis como el de correspondencias o el de regresión que cuentan con herramientas verdaderamente sofisticadas.

La exposición que se ha llevado a cabo es suficientemente asequible para cualquir analista político que haya trabajado con el análisis clásico de componentes principales y, con ella se espera animar y despertar la curiosidad de estos investigadores hacia nuevos caminos que pueden serles muy útiles a corto plazo. Por el momento, tras la amplia revisión de publicaciones, el foco de investigación en España en que se ha comenzado a aplicar alguno de estos tratamientos es el País Vasco, de forma que conviene una mayor generalización.

Asimismo, dado el carácter de la tesis es de rigor efectuar un comentario acerca de la cientificidad de estas técnicas y de su ubicación dentro de la investigación cualitativa y cuantitativa.

Acerca de este tema, la opinión que tenemos es que el marco de investigaciones en que se empleen estas técnicas es más cualitativo y descriptivo que cuantitativo. Las razones que avalan esta opinión son claras desde el punto de vista estadístico: el análisis de componentes principales no tiene carácter inferencial y, por tanto, aunque se trabaje con un algoritmo matemático, ello no implica que ofrezca resultados científicos en el sentido estricto de la palabra.

Lo que sí que se puede afirmar es que el investigador puede planificar y llevar a cabo una investigación con estas técnicas sigjiendo los pasos del método científico hasta donde sea posible y, si lo hace, su aportación será científica en el sentido ya defendido en el segundo capítulo, es decir como descripción científica.

No hay duda de que para aplicar estas técnicas en una investigación existirán diversas etapas previas que se rleacionarán directamente con el método científico: diseño de un cuestionario, diseño de una muestra representativa y compilación de los datos. Estos datos serán susceptibles de diversos tratamientos descriptivos para proporcionar los resultados básicos de la encuesta y, posteriormente, serán susceptibles de aplicaciones más sofisticadas, algunas inferenciales y otras no. En el caso de las componentes principales, se tratará de reducir la dimensionalidad y eliminar la información redundante de conjuntos de variables apropiados para ello. La novedad estriba en que ahora se pueden tratar las variables cuantitativas por un

lado y las cualitatvas por otro, e incluso efectuar combinaciones de ambas con rigor y propiedad.

Los resultados proporcionarán dimensiones que resuman el contenido de conjuntos de variables y que sean susceptbles de representaciones gráficas que permitan extraer conclusiones descriptivas acerca del comportamiento electoral de los entrevistados y de una población si ésta está bien representada. Así, es posible establecer perfiles de votantes, grupos de diversas tendencias, etc.

Por supuesto, otro camino que queda abierto a la especulación es el tratamiento de datos agregados con estas técnicas, que proporcionará, como en el caso del análisis clásico de componentes principales, representaciones o mapificaciones de parcelas territoriales en función de variables políticas o sociales.

Finalmente, se presenta el análisis de los datos empleados en este ejemplo, efectuado mediante la técnica clásica para comparar los resultados. La aplicación informática utilizada sigue siendo el SPSS con el algoritmo de cálculo que ya poseía desde los años ochenta y, por tanto anterior al que contiene el módulo donde desarrolla el PRINCALS.

En este caso, se ha pedido al programa una reducción de la dimensionalidad de las cinco variables originales, incluyendo el género del elector, que nosotros sabemos de antemano que son cualitativas y no cuantitativas.

Sin embargo, como podremos comprobar inmediatamente, aparte de estar aplicando una técnica inapropiadamente, veremos que, por lo demás, los resultados son muy parecidos a los obtenidos mediante el PRINCALS. Por tanto, no es de extrañar que se haya usado esta técnica de forma indiscriminada, puesto que, aparentemente, no proporciona unos malos resultados.

En primer lugar, podemos ver la tabla de comunalidades con dos columnas: la inicial y la final, una vez efectuada la extracción.

La comparación con los resultados equivalentes en PRINCALS indica que las comunalidades finales, es decir, la cantidad de variación reproducida por las componentes extraídas por este análisis es superior que en el caso anterior. Por tanto, una primera conclusión es que el análisis clásico tiende a sobrevalorar la reproducción de los valores originales de las variables, es decir, nos va a proporcionar un análisis "mejor" que el que habíamos obtenido. En este caso en concreto, la variable mejor reproducida es el genero del elector (96.9% de su variación) y la peor PARTI2 (74.1% de su variación), mientras que en el análisis no lineal, la variable mejor reproducida era PARTI1 (con un 99.2% de su variación) y la peor POSIPOL (con un 89.9% de su variación). Otra diferencia con el análisis no lineal es que las comunalidades vienen especificadas en dos partes: la que corresponde a la dimensión 1 y la que corresponde a la dimensión 2, lo cual es mucho más informativo. Así, habíamos visto que PARTI1 venía reproducida en un 97.4% en la primera dimensión y en un 1.8% en la segunda, POSIPOL en un 70.8% en la primera dimensión y en un 19.1% en la segunda y, el género, que interesa comparar por su importancia en el análisis clásico, venía reproducido en un 23.6% en la primera dimensión y en un 68.2% en la segunda, alcanzando un nivel total del 91.8%, inferior al 96.9% alcanzado en este análisis.

Por tanto, las variables tienen un peso diferente en el análisis clásico que en el no lineal, a pesar de que los resultados finales se parezcan.

Comunalidades

	Inicial	Extracción
PARTIDO VOTADO EN ELECCIÓN 1	1,000	,942
PARTIDO VOTADO EN ELECCIÓN 2	1,000	,741
PARTIDO VOTADO EN ELECCIÓN 3	1,000	,921
POSICIONAMIENTO POLÍTICO	1,000	,941
GÉNERO DEL ELECTOR	1,000	,969

Método de extracción: Análisis de Componentes principales.

Seguidamente, se pueden ver los resultados que indican que han sido extraídas dos componentes principales o dimensiones, lo cual también sucede en el análisis no lineal. La primera dimensión tiene aquí un valor propio de 3.435 y representa un 68.706% de la variación original total. En el análisis no lineal, el valor propio de la primera componente es 0.6784 y representa un 67.84% de la variación original total. Por tanto, el análisis clásico proporciona un resultado ligeramente más favorable que el no lineal.

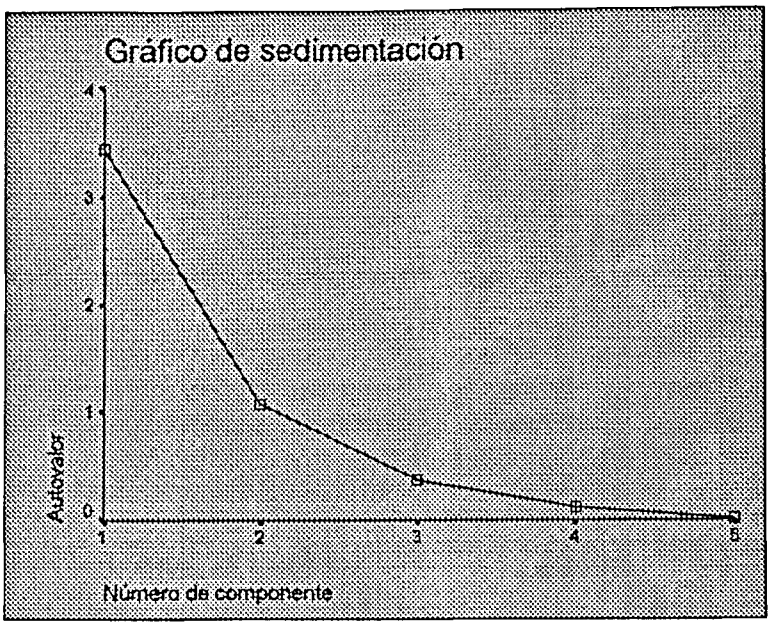
En cuanto al segundo valor propio, en la técnica clásica vale 1.080 y representa un 21.6% de la variación total y, en la técnica no lineal vale 0.2668 y representa un 26.68% de la variación total. Por tanto, en este caso, la segunda componente queda mejor valorada en la técnica no lineal.

Entre las dos componentes, en el análisis clásico reproducen un 90.306% de la información original y, en el análisis no lineal alcanzan el 94.52%, cifra superior a la anterior y en principio indicativa de un mejor aprovechamiento de los datos.

Varianza total explicada

	Autovalores iniciales			Sumas de las saturaciones al cuadrado de la extracción		
Componente	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
1	3,435	68,706	68,706	3,435	68,706	68,706
2	1,080	21,600	90,306	1,080	21,600	90,306
3	,361	7,224	97,530			
4	,114	2,282	99,812			
5	9,395E-03	,188	100,000			

Método de extracción: Análisis de Componentes principales.



El gráfico de sedimentación representa el número de componentes extraídas en función de los valores propios y no tiene mayor interés. En cambio, lo que sí que resulta intereante es la tabla de corelaciones entre las componentes extraídas y las variables originales, que nos ayudarán a interpretar el significado de estas nuevas variables.

Por las cifras, se puede ver rápidamente que la primera componente se relaciona con todas las variables políticas y en positivo y la segunda con el género y también en positivo. Por tanto, en apariencia tenemos lo mismo que en el análisis lineal, sólo que a la inversa en cuanto a representación: un eje izquierda-derecha y un eje vertical hombre-mujer. Sin embargo, no hay que despreciar los valores de los coeficientes de correlación, ya que en éste análisis, a diferencia del no lineal, parece estar sobrevalorada la claridad del género, es decir, se puede distinguir mucho mejor qué sujetos son hombre y mujeres, lo cual, indicaría que tienen un comportamiento político más diferenciado de lo que el análisis no lineal nos ha dado a entender.

En otras palabras, del análisis clásico, el investigador podría extraer la conclusión de que el género en esta población es un factor muy determinante del comportamiento político, mientras que el análisis no lineal (más adecuado para este tipo de datos) indica que no tiene tanta significación.

Matriz de componentes[a]

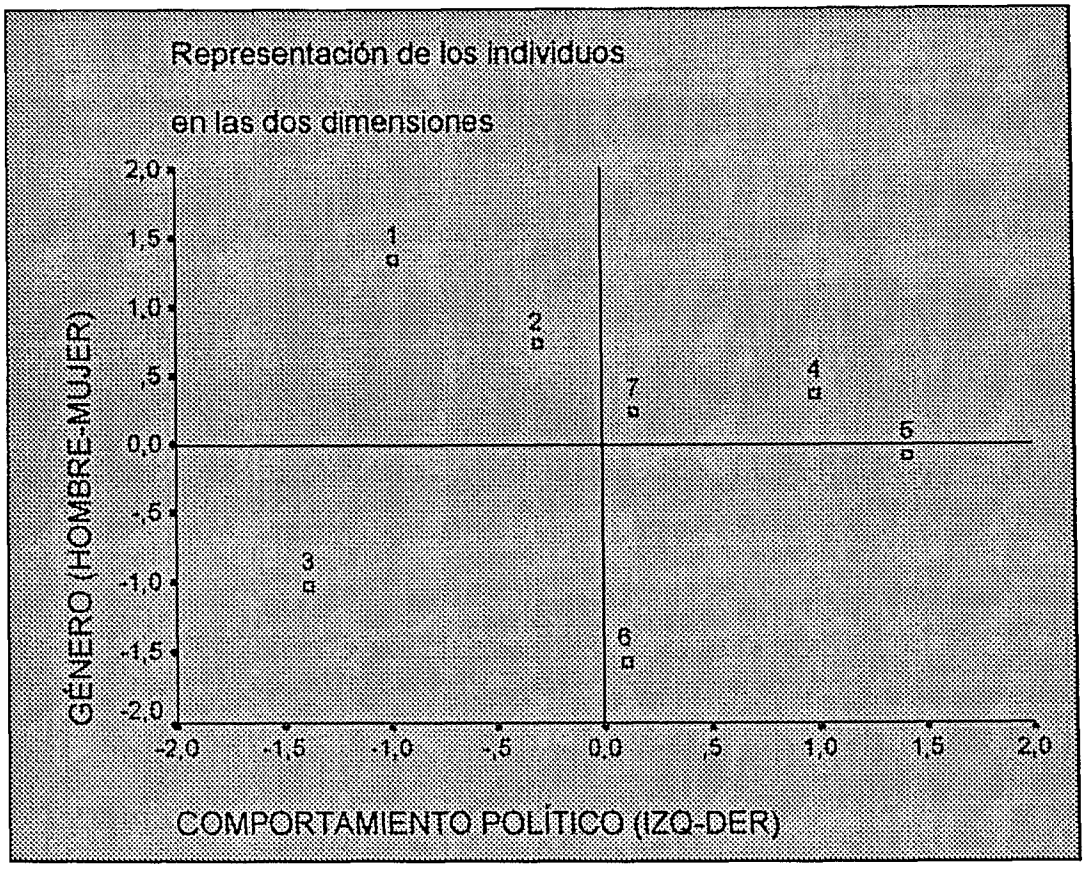
	Componente	
	1	2
PARTIDO VOTADO EN ELECCIÓN 1	,955	-,173
PARTIDO VOTADO EN ELECCIÓN 2	,816	-,273
PARTIDO VOTADO EN ELECCIÓN 3	,907	-,315
POSICIONAMIENTO POLITICO	,918	,314
GÉNERO DEL ELECTOR	,438	,882

Método de extracción: Análisis de componentes principales.

a. 2 componentes extraídos

Finalmente, al análisis clásico permite efectuar una representación gráfica de los individuos respecto de las componentes extraídas. Al igual que el análisis lineal, las coordenadas de los individuos (object scores) se guardan en la base de datos para su posterior utilización. La representación indicaría lo siguiente:

El individuo 1 es de izquierdas y mujer
El individuo 2 es de centro y mujer
El individuo 3 es de izquierdas y hombre
El individuo 4 es de derechas y mujer
El individuo 5 es de derechas y mujer
El individuo 6 es de centro y hombre
El individuo 7 es de centro y mujer



Las conclusiones a que habíamos llegado con el análisis no lineal eran:

El individuo 1 es de izquierdas y mujer (coincide)
El individuo 2 es de centro y probablemente mujer (coincide)
El individuo 3 es de izquierdas y hombre (coincide)
El individuo 4 es de derechas y no queda muy claro el género (coincide en lo político)
El individuo 5 es de derechas y no queda muy claro el género (coincide en lo político)
El individuo 6 es de centro y hombre (coincide)
El individuo 7 es de centro y probablemente hombre (coincide en lo político)

Si recordamos de nuevo los datos originales:

1 IC	1 IC	1 IC	2 H	2 C
2 PSOE	2 PSOE	2 PSOE	2 H	2 C
1 IC	1 IC	2 PSOE	1 M	1 I
4 PP	2 PSOE	4 PP	2 H	3 D
4 PP	4 PP	4 PP	2 H	3 D
3 CIU	3 CIU	3 CIU	1 M	2 C
2 PSOE	3 CIU	3 CIU	2 H	2 C

El individuo 1 se declara de centro pero vota a la izquierda y es un hombre, por tanto, ha sido bien captado por los dos tipos de análisis.
El individuo 2 se declara de centro votando al centro izquierda y es un hombre. Por tanto, se capta bien la parte política pero no el género.
El individuo 3 se declara de izquierdas y, en general lo es, siendo una mujer. Los análisis captan bien la parte política pero no el género.
El individuo 4 es de derechas, aunque en una ocasión vota PSOE y es un hombre. Los análisis captan bien la parte política y el análisis clásico se equivoca en el género.
El individuo 5 es de derechas y hombre. Los análisis captan bien la parte política y al análisis clásico se equivoca en el género.
El individuo 6 es de centro y mujer. Los análisis captan bien la parte política y mal el género.
El individuo 7 es de centro y hombre. Los análisis captan bien la parte política y el análisis clásico se equivoca en el género.

Por tanto, hay que concluir que, a pesar de que ambas técnicas ofrecen resultados muy parecidos, el balance de acierto, aunque sea por poco en este caso, es sin duda favorable al análisis no lineal. Este análisis ha puesto mejor que el otro de manifiesto que, el género no es un factor que tenga una actuación tan clara en la discriminación de los individuos de esta población respecto de su comportamiento político.

Así, de nuevo, se recomienda a los investigadores que tengan presente esta técnica y las que la acompañan para trabajar con variables cualitativas, especialmente, en el tratamiento de datos desagregados procedentes de encuestas y que sigan encarecidamente todo lo que se ha explicado acerca del método científico y de las pautas que pueden dar validez a un trabajo, ya sea descriptivo o inferencial.

En este caso se ha sido especialmente meticuloso para presentar todos los conceptos, pero, al tratarse de datos ficticios y escasos, se hace necesario presentar al menos una aplicación a una encuesta real para ver una aproximación más realista. Dicha aplicación se desarrolla a continuación, entrando sólo en la parte de interpretación y dando por comprendidos los conceptos y relaciones algebraicas del programa.

4.2.10 APLICACIÓN PRÁCTICA A UN CONJUNTO DE DATOS PROCEDENTES DE UNA ENCUESTA REAL

Los datos que se van a emplear en esta aplicación proceden de una encuesta realizada por el CIS con motivo de la convocatoria de elecciones generales en 1993.

Se trata de una encuesta post electoral muy amplia y compleja, que se presta a muchos tratamientos estadísticos. De todo el potencial que ofrece, nos vamos a

centrar en un apartado, que es el que se refiere al interés con que se ha seguido la campaña electoral en los medios de comunicación y la utilidad que ha supuesto para el elector la información proporcionada por dicha campaña.

Cuando iniciamos la investigación, sabemos que disponemos de 5001 entrevistas y que las preguntas que nos interesan son:

Como Vd. Recordará, el pasado domingo 6 de junio, se celebraron elecciones generales. Para empezar, me gustaría que me dijera con qué interés ha seguido Vd. la campaña electoral.

Con mucho interés	1
Con bastante interés	2
Ni con mucho ni con poco interés	3
Con poco interés	4
Con ningún interés	5
N.S.	8
N.C.	9

¿Me podría decir si, de una manera general, lo que ha visto u oido durante la campaña electoral le ha servido a Vd. mucho, bastante, poco o nada para...?

Conocer mejor a los líderes políticos
Informarse sobre qué soluciones propone cada partido (es decir, sus programas políticos)
Ver las diferencias que existen entre unos partidos y otros
Decidir su voto

La escala de valoración de cada uno de estos ítems es:

Mucho	1
Bastante	2
Poco	3
Nada	4
N.S.	8
N.C.	9

Supongamos que nuestra pretensión es reducir el anterior conjunto de variables a dos dimensiones que contengan el máximo de información que las variables originales y que una sea el grado de interés y la otra la utilidad general de la campaña sin diferenciar dicha utilidad en apartados concretos.

Para ello, observamos las variables y al ver que son de tipo ordinal, decidimos aplicar un análisis de componentes principales no lineal a sus datos.

Como se ha podido apreciar en el apartado anterior, esta técnica conlleva una serie de decisiones a tomar por parte del investigador y requiere cierta dosis de imaginación que la hace más atractiva que la clásica para este tipo de variables.

Si en el equipo de investigación hay un técnico estadístico, lo primero que hará es llamar la atención acerca de las escalas de valoración empleadas en las preguntas. Por un lado adevertirá que en el seguimiento de la campaña, la escala va de 1 (mucho) a 5 (nada), cuando lo apropiado sería lo inverso, es decir, de 1 (nada) a 5

(mucho). El seguir el primer modelo complica innecesariamente la interpretación de resultados, ya que conduce a pensar a la inversa.

Por otro lado, las escalas del resto de preguntas van de 1 a 4, con lo cual no son simétricas y no tienen punto medio. Además, el estadístico se pregunta ¿por qué en unas preguntas se emplea una escala y en otras una diferente y más confusa?. Finalmente, estas escalas de 1 a 4 cometen la misma imprudencia que la de 1 a 5, ya que asignan al valor 1 el mucho y al 4 el nada, siendo lo lógico el planteamiento inverso.

Así, se parte de unos datos que ya vienen con una estructura y que no se pueden cambiar. Lo máximo que se podría hacer es recodificar las escalas y ponerlas en su orden lógico y natural, pero lo que ya no se puede hacer es añadir un punto medio a las que van de 1 a 4.

Los investigadores tienen que tomar decisiones y, se decantan por lo siguiente:

Dejar las variables como están pero siendo conscientes del orden de las escalas en el momento de interpretar resultados.

Dejar como missings todos los casos de no sabe o no responde, por considerarlos fuera del objeto del estudio: aquí se pretende resumir el grado de interés de la campaña y su utilidad en general, de forma que sólo nos interesan individuos para los cuales haya sido de algún grado de utilidad (aunque se trate de un grado nulo). Las otras opciones serían susceptibles de otros análisis que determinasen, por ejemplo, el perfil de las personas que las escogieron.

Asignar un nivel de medición ordinal simple a todas las variables y aplicar un análisis de componentes principales no lineal para ver si logramos el propósito expuesto.

Los resultados son los siguientes:

En primer lugar tenemos la tabla de presentación con la lista de variables, su número de categorías y el nivel de medición.

```
PRINCALS - VERSION 0.6
                                BY
                    DEPARTMENT OF DATA THEORY
                UNIVERSITY OF LEIDEN, THE NETHERLANDS

 The number of observations used in the analysis = 5001

                          List of Variables

                                                    Number of    Measurement
Variable  Variable Label                            Categories      Level
--------  --------------                            ----------   -----------
INTCAM    INT. SEGUIMIENTO CAMPAÑA                         5     Ordinal
CONLID    INF EN CONOCER MEJOR LIDERES                     4     Ordinal
PROG      INF EN CONOCER PROGRAMAS                         4     Ordinal
DIFPAR    INF EN VER DIF ENTRE PARTIDOS                    4     Ordinal
VOTAR     INF EN DECIDIR SU VOTO                           4     Ordinal
```

En segundo lugar tenemos la tabla de frecuencias marginales de cada variable y sus categorías. En ella se puede apreciar que hay 25 missings en el interés por la campaña, 168 en su utilidad para conocer a los líderes políticos, etc., todo ello consecuencia de nuestra decisión de eliminar los casos de no respuesta.

Marginal Frequencies

Variable	Missing	Categories 1	2	3	4	5
INTCAM	25	619	1697	737	1228	695
CONLID	168	235	1427	1937	1234	
PROG	175	231	1486	1841	1268	
DIFPAR	193	266	1714	1646	1182	
VOTAR	210	198	936	1457	2200	

Seguidamente se dispone de la tabla que proporciona el historial de las iteraciones en busca de la cuantificación óptima. Tras 26 iteraciones, el ajusto conseguido, o el total de variación original reproducido, es del 86.34%, cantidad que está bastante bien.

*** The History of Iterations ***

Iteration	Total Fit	Total Loss	Multiple Loss	Single Loss	Iteration Change
1	,8360085	1,1639915	1,1576683	,0063232	,0030145
2	,8410634	1,1589366	1,1515694	,0073672	,0050549
3	,8454367	1,1545633	1,1457098	,0088535	,0043733
4	,8491848	1,1508152	1,1401152	,0107001	,0037481
5	,8523207	1,1476793	1,1349146	,0127647	,0031360
6	,8548803	1,1451197	1,1302219	,0148977	,0025596
7	,8569232	1,1430768	1,1260973	,0169795	,0020429
8	,8585232	1,1414768	1,1225459	,0189309	,0015999
9	,8597574	1,1402426	1,1195326	,0207099	,0012342
10	,8606987	1,1393013	1,1169996	,0223017	,0009413
11	,8614106	1,1385894	1,1148806	,0237088	,0007119
12	,8619460	1,1380540	1,1131103	,0249438	,0005354
13	,8623472	1,1376528	1,1116295	,0260233	,0004012
14	,8626474	1,1373526	1,1103871	,0269655	,0003002
15	,8628719	1,1371281	1,1093402	,0277879	,0002245
16	,8630399	1,1369601	1,1084536	,0285065	,0001680
17	,8631658	1,1368342	1,1076986	,0291356	,0001259
18	,8632604	1,1367396	1,1070521	,0296876	,0000946
19	,8633312	1,1366688	1,1064951	,0301737	,0000708
20	,8633775	1,1366225	1,1060439	,0305785	,0000464
21	,8634108	1,1365892	1,1057378	,0308514	,0000333
22	,8634365	1,1365635	1,1054640	,0310995	,0000257
23	,8634565	1,1365435	1,1052174	,0313261	,0000200
24	,8634721	1,1365279	1,1049947	,0315331	,0000156
25	,8634844	1,1365156	1,1047931	,0317225	,0000123
26	,8634941	1,1365059	1,1046100	,0318959	,0000097

The iterative process stops because the convergence test value is reached.

Los conceptos llamados "loss" y "fit" son medidas de lo buena o mala que es una solución. Un valor alto de "loss" implica que la solución es mala y un valor alto de "fit" lo contrario. La "Total loss" es la suma de la "Multiple loss" y la "Single loss". Se trata de medidas complementarias. La "Single fit" se puede interpretar como la variación explicada de cada elemento de la matriz de cuantificaciones, de forma que "Single loss" sería la variación no explicada contenida en dichos elementos. De todas formas, lo más interesante de la tabla anterior, a nivel práctico es saber que

para todas las variables y dimensiones consideradas conjuntamente, la bondad de la solución se indica en la columna "Total fit" y que su resultado final es igual a la suma de los valores propios de las componentes extraídas[201].

Seguidamente, los valores propios de las dos dimensiones extraídas nos proporcionan esa información pero dividida para cada una de ellas. Así, la primera dimensión reproduce el 64.45% de la información y la segunda el 21.9%. En principio basta con dos dimensiones para intentar lograr nuestro propósito.

Dimension	Eigenvalue
1	,6445
2	,2190

A continuación tenemos los resultados para cada variable y, al haber escogido un nivel de medición ordinal simple, sólo debemos observar las single category coordinates. La cuantificación de esta primera variable otorga el valor más alto al 5, como era de esperar, pero hay que recordar que, en este caso, 5 significa con ningún interés.

Variable: INTCAM INT. SEGUIMIENTO CAMPAÑA

Type: Ordinal Missing: 25

Category:		Marginal Frequency	Quantification
1	CON MUCHO INTERES	619	-,36
2	CON BASTANTE INTERES	1697	-,36
3	NI CON MUCHO NI CON	737	-,36
4	CON POCO INTERES	1228	-,31
5	CON NINGUN INTERES	695	2,54

Single Category Coordinates

Category	Dimension 1	Dimension 2
1	,18	,36
2	,18	,36
3	,18	,36
4	,16	,30
5	-1,30	-2,54

Multiple Category Coordinates

Category	Dimension 1	Dimension 2
1	,48	,20
2	,34	,28
3	,05	,44
4	-,06	,42
5	-1,28	-2,55

[201] Para más información y detalles, se puede consultar el libro PRINCALS de Albert Gifi, Department of Data Theory, Leiden, 1985

Tras estas tablas vienen otras iguales para cada una de las variables con escala de 1 a 4 y con el mismo tipo de características, de forma que no hay más comentarios que efectuar.

```
Variable: CONLID      INF EN CONOCER MEJOR LIDERES
---------
Type: Ordinal              Missing:     168

Category:                        Marginal Frequency          Quantification
---------                        ------------------          --------------
       1  MUCHO                                235                   -2,91
       2  BASTANTE                            1427                    -,67
       3  POCO                                1937                    -,04
       4  NADA                                1234                    1,39

Single Category Coordinates
---------------------------
Category          Dimension

                      1         2
       1          2,57      -,36
       2           ,59      -,08
       3           ,04      -,01
       4         -1,23       ,17

Multiple Category Coordinates
-----------------------------
Category          Dimension

                      1         2
       1          2,49      -,97
       2           ,61       ,03
       3           ,06       ,14
       4         -1,25       ,00
```

```
Variable: PROG        INF EN CONOCER PROGRAMAS
---------
Type: Ordinal              Missing:     175

Category:                        Marginal Frequency          Quantification
---------                        ------------------          --------------
       1  MUCHO                                231                   -2,97
       2  BASTANTE                            1486                    -,64
       3  POCO                                1841                    -,06
       4  NADA                                1268                    1,36

Single Category Coordinates
---------------------------
Category          Dimension

                      1         2
       1          2,68      -,43
       2           ,57      -,09
       3           ,05      -,01
       4         -1,23       ,20

Multiple Category Coordinates
-----------------------------
Category          Dimension

                      1         2
       1          2,59      -,99
       2           ,59       ,01
       3           ,07       ,14
       4         -1,26       ,04
```

Variable: DIFPAR INF EN VER DIF ENTRE PARTIDOS

Type: Ordinal Missing: 193

Category:		Marginal Frequency	Quantification
1	MUCHO	266	-2,72
2	BASTANTE	1714	-,56
3	POCO	1646	-,03
4	NADA	1182	1,45

Single Category Coordinates

Category	Dimension 1	Dimension 2
1	2,43	-,30
2	,50	-,06
3	,03	,00
4	-1,30	,16

Multiple Category Coordinates

Category	Dimension 1	Dimension 2
1	2,36	-,86
2	,51	,06
3	,05	,13
4	-1,32	,00

Variable: VOTAR INF EN DECIDIR SU VOTO

Type: Ordinal Missing: 210

Category:		Marginal Frequency	Quantification
1	MUCHO	198	-3,52
2	BASTANTE	936	-,83
3	POCO	1457	-,32
4	NADA	2200	,89

Single Category Coordinates

Category	Dimension 1	Dimension 2
1	2,63	-,80
2	,62	-,19
3	,24	-,07
4	-,67	,20

Multiple Category Coordinates

Category	Dimension 1	Dimension 2
1	2,52	-1,16
2	,66	-,07
3	,27	,03
4	-,69	,15

Summary of Analysis

Multiple Fit

Variable	Row Sums	Dimension 1	2
INTCAM	1,299	,296	1,004
CONLID	,836	,784	,052
PROG	,870	,818	,052
DIFPAR	,846	,800	,046
VOTAR	,625	,561	,064
Mean:	,895	,652	,244

El resumen del análisis, que equivale a las comunalidades que hay que mirar es el siguiente, puesto que el anterior sería adecuado si hubeise variables con medición múltiple. Así, el siguiente cuadro de resulktados indica que por término medio se ha aprovechado un 86.3% de las información de las variables originales entre las dos componentes y que en la primera el aprovechamiento medio ha sido del 64.5% y en la segunda del 21.9%. Ya se puede apreciar que la primera dimensión capta casi completamente la variable del interés por la campaña, tal y como era nuestra intención y que la segunda capta a las restantes, siendo la peor la influencia en el voto. Por tanto, la utilidad de la campaña se ha manifestado más en otros aspectos que en la influencia para que el elector decidiera su voto.

Single Fit

Variable	Row Sums	Dimension 1	2
INTCAM	1,257	,263	,994
CONLID	,799	,784	,016
PROG	,838	,817	,021
DIFPAR	,812	,799	,013
VOTAR	,612	,560	,052
Mean:	,863	,645	,219

El poder ver los resultados para un nivel múltiple, sirve para que los investigadores se planteen o no la conveniencia de cambiar el nivel de medición ordinal simple por un nominal múltiple en alguna o en todas las variables. Se trata, por tanto, de una información adicional valiosa para efectuar otras pruebas antes de quedarse con una solución definitva de la experiencia que estemos llevando a cabo.

Seguidamente, la matriz de correlaciones indica que la primera dimensión está muy correlacionada con todas las variables que deseábamos y sigue poniendo de manifiesto que la influencia en el voto ha sido la más baja. La segunda dimensión capta el interés por la campaña y por tanto, a raíz de estos resultados podríamos efectivamente poner título a las dimensiones:

Dimensión 1: utilidad de la campaña, al tener correlaciones negativas y al estar la escala al revés, su eje va de menor utilidad a la izquierda a mayor utilidad a la derecha.

Dimensión 2: grado de interés en el seguimiento de la campaña, al tener correlación negativa y estar la escla original a la inversa de lo lógico, su eje se interpreta de alto grado de interés en los valores más altos y poco grado de interés en los valores bajos (arriba mucho interés, abajo, poco interés).

Component Loadings

Variable	Dimension 1	2
INTCAM	-,513	-,997
CONLID	-,885	,125
PROG	-,904	,144
DIFPAR	-,894	,112
VOTAR	-,748	,228

Seguidamente, el programa nos recuerda el ajuste logrado, que es el Total Fit obtenido en la iteración 26: 86.35% de reproducción de la información original.

Iteration Number	Total Fit	Total Loss	Multiple Loss	Single Loss
26	,8635	1,1365	1,1046	,0319

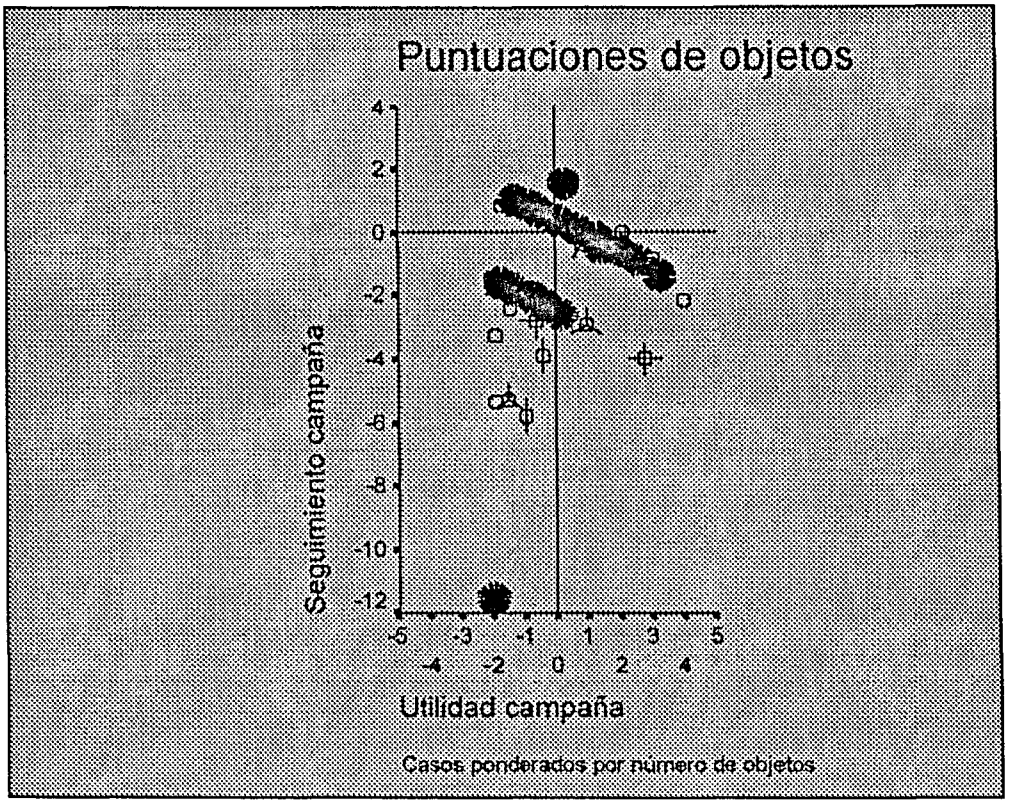
A continuación, si se desea, el programa proporciona las coordenadas de cada individuo respecto de las dos dimensiones, para su posterior representación gráfica. Evidentemente, reproducir aquí las aproximadamente 4000 que debe haber no conduce a nada, de forma que se incluyen unas cuantas a modo de ejemplo:

The Object Scores are:

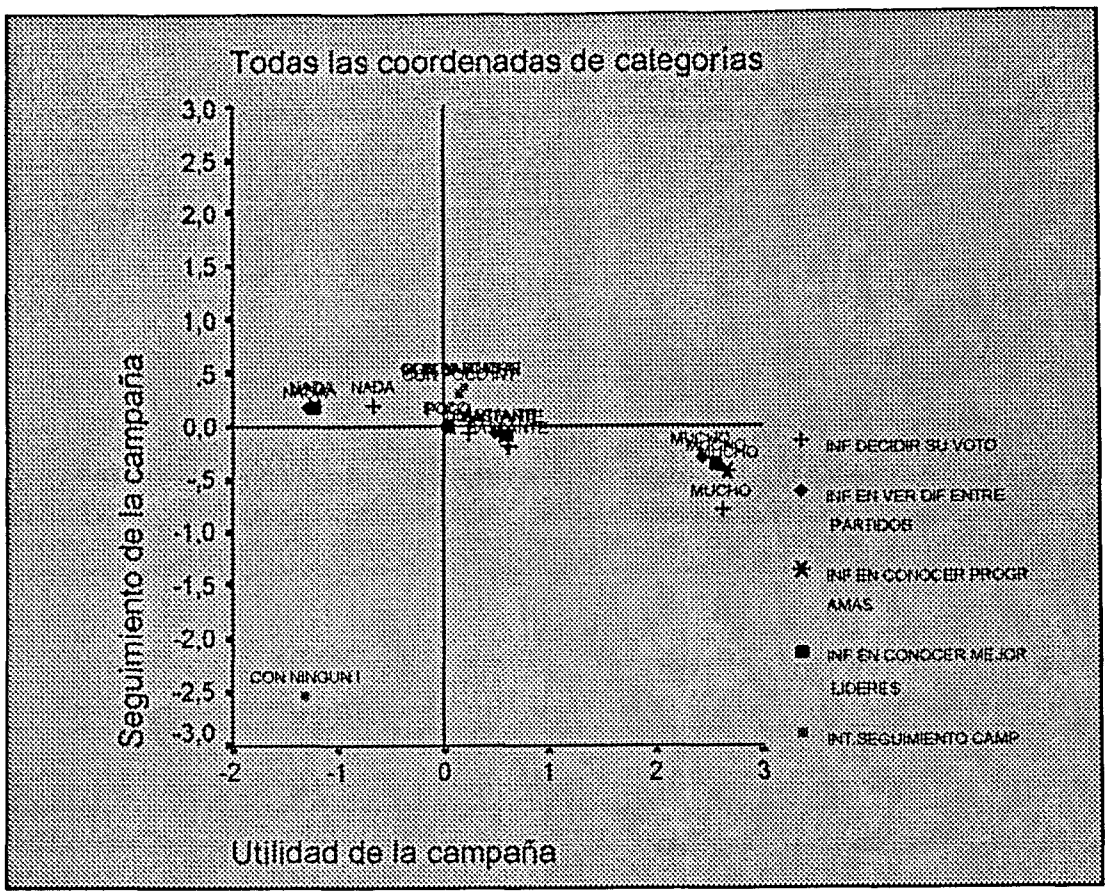
Object *	Dimension 1	2
1 *	,20	-2,55
2 *	,65	,04
etc.etc.		
4984 *	3,26	-1,35
4985 *	3,26	-1,35
4986 *	,25	1,42
4987 *	3,25	-1,40

Una vez representados los individuos, podemos ver que la gran cantidad de los mismos, proporciona un gráfico con muchos solapamientos de casos. Sin embargo, es posible ver cómo se distribuye la población. Por un lado, abajo a la izquierda, tenemos un bloqe importante de individuos que han seguido con muy poco interés la campaña (valores bajos de la dimensión dos) y para los cuales ha sido poco útil en la mayoría de los aspectos (valores negativos de la dimensión 2). Otra concentración pequeña dentro de las más destacadas, se da con valores positivos (cerca del 2) del seguimiento de campaña, es decir, que se trata de un colectivo

que la ha seguido con interés y, para los cuales, ha tenido utilidad, ya que se hallan en la parte derecha de la dimensión 1. En el centro del gráfico tenemos, el grueso de la población, para el cual hay un comportamiento variable que va desde un intenso seguimiento de la campaña con mucha utilidad, a un seguimiento menos profundo de la misma con menos utilidad. Es decir, que parece que lo más habitual es seguir con cierto detalle la campaña y que lo que se capta de ella influye de alguna forma en los aspectos mencionados: mejor conocimiento de líderes, programas electoales, etc., siendo el voto lo menos afectado por ella. Finalmente, la otra nube de puntos importante es similar a la mayor pero se trata de personas que han tenido un seguimiento de la campaña algo menos notable y cuya utilidad ha ido de bastante a menos pero sin llegar al extremo del nada.



El siguiente gráfico sirve para explicamos donde se sitúan las categorías de las variables y, por tanto, para ayudarnos en la confección de las explicaciones que acabamos de dar. Sin embargo, puede llegar a ser confuso y en el caso de que se solapen algunas categorías, el programa facilita una tabla que hay tras el gráfico con las coordenadas de las etiquetas que no se aprecian claramente. Abajo, a la izquierda, se puede ver por dónde está la categoría seguimiento de la campaña con ningún interés que se corresponde con el valor 5 original, lo cual confirma que los valores bajos de esta dimensión indican menor seguimiento y los altos mayor seguimiento de la campaña. A la derecha se puede ver claramente la etiqueta de mucha influencia en decidir el voto y a la izquierda la del nada, lo cual indica que el eje de utilidad va de izquierda a derecha de forma lógica, es decir, poca utilidad a la izquierda y mucha a la derecha. El resto de categorías está muy solapado y cuesta de distinguir, pero con la información que tenemos es suficiente para extraer una buena descripción.




```
Summary of multiple points in chart #2

   Dim1     Dim2     Actual label or name

    ,18      ,36     CON MUCHO IN
    ,18      ,36     CON BASTANTE
    ,18      ,36     NI CON MUCHO
```

Finalmente, el programa nos indica que ha añadido las variables que contienen estas puntuaciones de los individuos a nuestra base de datos. A partir de ellas, se podría, por ejemplo, efectuar una regresión o un análisis de correspondencias y ver hasta qué punto el interés por la campaña explica su utilidad. En ese caso, pasaríamos al terreno inferencial y contrastaríamos una hipótesis, empalmando con aspectos científicos y técnicas cuantitativas que irían más lejos de la descripción, para la población analizada.

```
Object scores for 2 dimensions were saved with the rootname: PRI
Following object scores were added to the working file:

Name        Label

PRI1_1      Dimension 1
PRI2_1      Dimension 2
```

4.2.11 CONCLUSIONES

En este capítulo se ha pasado revista a la situación del estudio del Comportamiento Electoral de los españoles a través de datos desagregados en la actualidad.

De todo lo expuesto se pueden extraer una serie de conclusiones:

En primer lugar, que el desarrollo de esta forma de estudio ha sido algo posterior al que se venía realizando con datos agregados.

En segundo lugar, que los medios de comunicación han difundido resultados simples desde el punto de vista estadístico, lo cual ha propiciado, a nuestro modo de ver, un cierto estancamiento en cuanto al conocimiento de las nuevas técnicas que se pueden aplicar a los datos procedentes de encuestas. El hecho de que se necesiten opiniones casi inmediatas antes y después de los eventos electorales implica que las tablas y gráficos simples y los cruzamientos hayan sido, por decirlo de algún modo, las aplicaciones dominantes en este contexto.

En tercer lugar, que los modelos de las encuestas adolecen de fallos técnicos, como los comentados en referencia a las escalas de valoración y de falta de renovación de preguntas que reflejen de manera más realista los condicionantes del voto, los factores que conducen al elector a tomar una decisión final. Esta parte se podría mejorar notablemente si se pusiera en práctica con mayor frecuencia el trabajo interdiciplinar, es decir, si colaborasen en mayor medida los politólogos y sociólogos con los técnicos estadísticos, teniendo ambos colectivos la visión más generalista posible del entorno en que se mueven. De hecho, se supone que los futuros licenciados y diplomados de todas estas disciplinas tendrán una mejor y mayor preparación en este sentido.

En cuarto lugar, que las encuestas de este tipo se llevan, mayoritariamente a cabo por parte de grandes organismos vinculados a las administraciones o, por empresas que tienen la suficiente infraestructura para ello, quedando un tanto al margen el ambiente académico o universitario que ha proporcionado mayores logros en la investigación con datos agregados y que tendría mucho que aportar a la de datos desagregados.

En quinto lugar, que el punto de conexión de lo que se viene haciendo con lo que se puede desarrollar a partir de ahora son las técnicas multivariantes no lineales que permiten tratar datos cualitativos y que, por ello, se ha escogido esa parcela de la estadística para establecer una línea de apertura en esta tesis: no se puede avanzar más si no se conoce primero lo que sigue a lo ya aplicado hasta el momento.

Aunque sólo se ha desarrollado el análisis de componentes principales no lineal con variables de nivel ordinal simple, los investigadores deben tener presente que ésta es sólo una de las múltiples posibilidades que ofrece esta área de la Estadística.

Así, si todas las variables son tratadas como numéricas, la solución que proporciona el análisis es única y coincidente con el análisis clásico. Si todas las variables son tratadas como nominales múltiples, estaremos desarrollando la técnica llamada HOMALS que proporciona diversas posibilidades de solución anidadas, de forma que en la primera solución de, por ejemplo 3 dimensiones, la primera dimensión es igual a la que tendríamos si sólo hubiésemos pedido extraer una componente, las dos primeras dimensiones son iguales que las que tendríamos si hubiésemos pedido una extracción de dos componentes y, en general, las primeras (p-i) dimensiones de una solución de p dimensiones, son iguales que las

dimensiones de una solución de (p-i) dimensiones. En cambio, las soluciones cuando se trabaja con variables medidas a nivel nominal u ordinal simple, no son de tipo anidado. Es decir, por ejemplo, las dos primeras dimensiones de una solución con p dimensiones no son iguales a las que proporcionaría una solución en que hubiésemos pedido la extracción de dos componentes.

Por otro lado, es importante comentar que las técnicas no lineales tienen muchos puntos en común con el análisis clásico. Así, por ejemplo, si nos preguntamos cuántas dimensiones o componentes hay que extraer, la respuesta en el caso de variables numéricas es la misma que en un análisis clásico: retener las primeras componentes para las cuales el valor propio sea superior a 1 dividido por el número de variables originales. Si una componente tiene un valor propio inferior a esa cantudad, entonces, reproduce menos variación que una variable individual y por tanto, no aporta nada al análisis.

En el caso de variables ordinales o nominales simples, se recomienda comenzar con la extracción de tres dimensiones. Si el valor propio de la tercera dimensión es inferior a 1 dividido por el número de variables, entonces se puede pedir la extracción de dos componentes. Si al elegir un número de dimensiones, la suma de los valores propios está muy próxima a (m-1)/m, siendo m el número de variables originales, entonces no es necesario aumentar el número de componentes a extraer. Si, por el contrario la suma de valores propios es inferior a (m-1)/m, entonces, hay que considerar el aumentar el número de componentes a extraer mientras el valor propio más pequeño siga estando lejos de 1/m.

Si trabajamos con variables nominales múltiples, las reglas acerca del número de componentes a extraer son más difíciles de precisar porque este tratamiento proporciona el mayor número de grados de libertad posible para la cuantificación óptima, de forma que, incluso con un conjunto de variables aleatorias incorrelacionadas, se podría obtener un número de valores propio superior al recíproco del número máximo. Por tanto, en general, en este terreno se espera que el investigador tenga buen criterio y pueda llevar a cabo el análisis con el mínimo número de dimensiones posible que hagan posible una buena interpretación.

En cuanto a la rotación, tan común en el análisis clásico, en el contexto no lineal no se realiza ni se aconseja porque la cuantificación óptima ya proporciona resultados suficientemente interpretables si el análisis es bueno, de manera que sería un paso supérfluo.

Finalmente, para terminar este capítulo, deseamos remarcar la parte de diseño y creatividad que puede conllevar el emplo de estas técnicas porque el investigador no sólo tiene que decidir acerca del nivel de medición de las variables, sino también acerca de cuáles considerar activas o pasivas y lo mismo respecto de los individuos entrevistados, con lo que todo ello puede implicar a nivel de representatividad de la muestra y otros aspectos importantes.

Por tanto, es un campo de inmensas posibilidades y que puede ayudar no sólo a describir situaciones y colectivos, sino a formular nuevas preguntas y a observar temas particulares dentro de otros más amplios.