


UNIVERSIDAD CENTRAL DE VENEZUELA
FACULTAD DE CIENCIAS ECONOMICAS Y SOCIALES
COMISION DE ESTUDIOS DE POSTGRADO
AREA DE POSTGRADO EN ESTADISTICA Y ACTUARIADO
DOCTORADO EN ESTADÍSTICA


VEREDICTO

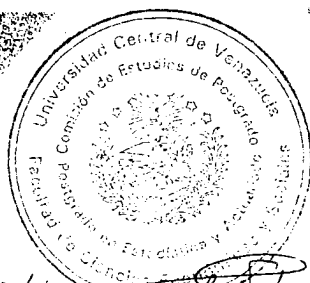
Quienes suscriben, miembros del jurado designado por el Consejo de la Facultad de Ciencias Económicas y Sociales y el Consejo Central de Postgrado de la Universidad Central de Venezuela para examinar la Tesis Doctoral titulada: ***“UNA PROPUESTA PARA EL TRATAMIENTO DE TABLAS DE CONTINGENCIAS MÚLTIPLES BAJO UNA PERSPECTIVA DE ADAPTACIÓN DE LA METODOLOGÍA STATIS”*** presentado por **RAIZA TORRES**, Cédula de Identidad N° 4.210.340, a los fines de cumplir con el requisito legal para optar al grado de **DOCTOR EN ESTADISTICA**, dejan constancia de lo siguiente:


- 1.- Leído como fue dicho trabajo por cada uno de los miembros del Jurado, éste fijó el día 15 de Enero de 2010 a las 10:00 a.m. para que la autora lo defendiera en forma pública, lo que ésta hizo en la Sala de Conferencias del Área de Postgrado en Estadística y Actuariado, mediante un resumen oral de su contenido, luego de lo cual respondió **SATISFACTORIAMENTE** a las preguntas que le fueron formuladas por el jurado; todo ello conforme a lo dispuesto en el Reglamento del Consejo de Estudios de Postgrado.
- 2.- Finalizada la defensa pública del trabajo, el jurado decidió **APROBARLO** por considerar, sin hacerse solidario de las ideas expuestas por la autora, que se ajusta a lo dispuesto y exigido en el Reglamento de Estudios de Postgrado. Para dar este veredicto el Jurado estimó que dicho trabajo constituye un aporte novedoso en el tratamiento de datos categóricos, particularmente evidenciado en la propuesta de una distancia entre estructuras de datos cualitativos función de las inter-asociaciones entre las variables. La contribución cobra tanto más valor ante la carencia de medidas similares en el ámbito de la estadística aplicada a datos categóricos.

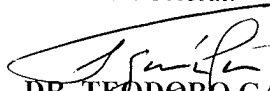
En fe de lo cual se levanta la presente Acta a los quince días del mes de Enero de dos mil diez, dejándose también constancia de que, conforme a lo dispuesto en la normativa jurídica vigente, actuó como Coordinadora del Jurado, la Profesora Maura Vásquez, tutora de la Tesis Doctoral.

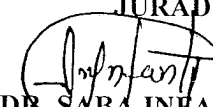

DR. GUILLERMO RAMIREZ
C.I. N° 3.609.750
JURADO


DRA. OLESIA CÁRDENAS
C.I. N° 3.967.050
JURADO




DRA. MAURA VÁSQUEZ
C.I. N° 3.100.452
TUTORA


DR. TEODORO GARCÍA
C.I. N° 7.182.927
JURADO


DR. SABA INFANTE
C.I. N° 8.795.671
JURADO



UNIVERSIDAD CENTRAL DE VENEZUELA
FACULTAD DE CIENCIAS ECONÓMICAS Y SOCIALES
COMISIÓN DE ESTUDIOS DE POSTGRADO
ÁREA DE POSTGRADO EN ESTADÍSTICA Y ACTUARIADO
PROGRAMA INTEGRADO DE POSTGRADO EN ESTADÍSTICA

**UNA PROPUESTA PARA EL TRATAMIENTO DE TABLAS
DE CONTINGENCIAS MÚLTIPLES BAJO UNA PERSPECTIVA DE ADAPTACIÓN
DE LA METODOLOGÍA STATIS**

Tesis Doctoral presentada ante la
Universidad Central de Venezuela
Como requisito parcial para optar al grado de
Doctora en Estadística

Autora: Raiza Priscila Torres Wills
Tutor: Maura Vásquez

Caracas, enero de 2010

REPUBLICA BOLIVARIANA DE VENEZUELA
UNIVERSIDAD CENTRAL DE VENEZUELA
FACULTAD DE CIENCIAS ECONOMICAS Y SOCIALES
AREA DE POSTGRADO EN ESTADÍSTICA Y ACTUARIADO
PROGRAMA INTEGRADO DE POSTGRADO EN ESTADÍSTICA

**UNA PROPUESTA PARA EL TRATAMIENTO DE TABLAS DE CONTINGENCIAS
MÚLTIPLES BAJO UNA PERSPECTIVA DE ADAPTACIÓN DE LA
METODOLOGÍA STATIS**

Autora: Raiza Torres
Tutora: Maura Vásquez

RESUMEN

En situaciones que surgen en diferentes ciencias, se miden T conjuntos de variables o un solo conjunto de variables categóricas observadas en T ocasiones, sobre un mismo grupo de individuos. Esta información permite conceptualizar la realidad, originando T estudios representados por objetos matemáticos que contienen información interpretable estadísticamente, arreglada en T tablas de datos, que precisa de evaluaciones comparativas, para explorar las estructuras de interasociaciones entre variables y/o interdistancias entre individuos, desde una perspectiva gráfica sustentada por índices.

En esta investigación se propone un método para analizar información con una estructura como la descrita, fundamentada en una adaptación del STATIS, metodología propuesta para analizar datos con estructuras similares, pero cuantitativas. A estos efectos, se formula una distancia en el sentido Hilbert-Schmidt, que mide las diferencias entre las estructuras de interdistancias utilizadas para comparar posiciones de individuos caracterizados en dos estudios diferentes y clasificados en Tablas de Contingencias Múltiples.

Esta distancia denominada d_{HS-C} , se define como función inversa de la interasociación global entre las variables categóricas que identifican las estructuras comparadas, con la propiedad de que acepta una desagregación para medir la contribución, en términos de asociación, de dos variables categóricas a la interasociación global, y a la distancia d_{HS-C} .

Palabras clave: Tablas de Contingencias Múltiples, Análisis de Correspondencias Binarias, Análisis de Correspondencias Compuesto, STATIS, distancia entre individuos.

AGRADECIMIENTO

A DIOS, que nos proveyó de paciencia a mí mi querida tutora y a mi persona.

A DIOS, que les dio fortaleza a mis hijas y mi familia, para soportar el tiempo que les robé, y el sufrimiento de saberme en carretera todas las semanas.

A los profesores de Postgrado:

Maura Vásquez, por su muy acertada orientación en la tutoría del trabajo, y la precisión en la identificación de uno de los conceptos más fundamentales de la tesis.

Olesia Cárdenas, por su generosa ayuda en proporcionarme material bibliográfico.

Alberto Camardiel, por su paciencia en mi pre-defensa.

Y al Prof. Guillermo Ramírez, que sé que estuvo presente de alguna manera en el trabajo de la Tesis.

A la Sra. Roxana Revette, solidaria y amiga.

ÍNDICE DE CONTENIDO

Pág.

Introducción.....	1
CAPÍTULO I: EL PROBLEMA.....	4
1.1 Planteamiento del problema.....	4
1.2 Preguntas de Investigación.....	6
1.3 Objetivos de la investigación.....	7
1.3.1 Objetivo General.....	7
1.3.2 Objetivos Específicos.....	7
1.4 Antecedentes.....	8
CAPÍTULO II MARCO TEÓRICO.....	13
2.1 Bases Teóricas.....	13
2.1.1 Estructura de Datos.....	13
2.2 Metodología STATIS.....	15
2.2.1 Escenarios de aplicación de la Metodología STATIS.....	16
2.2.2 Fundamentos Teóricos del STATIS.....	19
2.2.3 Arreglos de datos e información derivada en STATIS.....	19
2.2.3.1 Definición del Objeto Representativo de los T estudios E_t	19
2.2.3.2 Producto Escalar entre objetos representativos W_t	20
2.2.3.3 Imagen euclídea de los objetos W_t	23
2.2.3.4 Determinación del Compromiso.....	25
2.2.3.5 Trazado de trayectorias.....	28
2.3 Análisis de Correspondencia Compuesto.....	30
2.4 Definiciones y Términos Básicos.....	34
2.4.1 Definiciones del Análisis de Correspondencias Binarias (ACB).....	35
2.4.2 Tablas de Contingencias Múltiples (TCM).....	37
2.4.2.1 TCM producto del cruce de dos conjuntos de variables categóricas X e Y.....	38

2.4.2.2	TCM: Producto del cruce de un conjunto de variables X consigo mismo.....	39
2.5	Definiciones planteadas en el STATIS-C para el tratamiento de datos cualitativos.....	40
	CAPÍTULO III: PROPUESTA QUE ADAPTA EL STATIS A DATOS CATEGÓRICOS (STATIS-C).....	41
3.1	Fundamentos de la propuesta.....	41
3.1.1	Enfoque de ACB de las Tablas de Contingencias Múltiples.....	41
3.1.1.1	Enfoque ACB sobre la TCM F_{YX}	42
3.1.1.2	Enfoque ACB sobre la TCM F_{XX}	45
3.2	Estructura de Datos.....	50
3.3	Adaptación de la propuesta al caso de dos estudios E_X y E_Y	50
3.4	Modelo Teórico de la Adaptación del STATIS en el Análisis de Datos Categóricos.....	50
3.4.1	Adaptación de la Etapa de la Interestructura.....	52
3.4.1.1	Definición del Objeto Representativo del t -ésimo estudio E_t	52
3.4.1.2	Definición del Producto Escalar entre Objetos.....	53
3.4.1.3	Definición de la Distancia entre Objetos.....	55
3.4.1.3.1	Propiedades algebraicas del producto escalar HS entre objetos categóricos W_i	56
3.4.1.3.2	Propiedades estadísticas del producto escalar HS entre objetos categóricos W_i	57
3.4.1.4.3	Índice de Contribución de las TCM.....	58
3.4.1.4.4	Índice de contribución de las tablas de contingencias bidimensionales..	58
3.4.1.5	Construcción de la imagen euclídea de los objetos W_i	59
3.4.2	Construcción del Compromiso W	61
3.4.3	Estudio de la Intraestructura	62
	CAPÍTULO IV METODOLOGÍA USADA PARA LA APLICACIÓN DEL STATIS-C.....	65
4.1	Simulación de datos.....	65
4.2	Tipos de tablas de datos simulados a analizarse mediante STATIS-C.....	67

	CAPÍTULO V: APLICACIÓN DEL STATIS-C.....	70
5.1	Implementación del STATIS-C.....	70
5.1.1	Estudio de tres dimensiones con estructuras exactamente iguales.....	70
5.1.2	Estudio simultáneo de las dimensiones Z_1 , Z_2 y Z_3 con estructuras similares.....	74
5.1.3	Estudio de las dimensiones Z_1 , Z_2 , Z_3 de estructura similar, y la tabla Y con una estructura diferente.....	77
5.1.4	Análisis de las dimensiones X, Y y Z, con estructuras de correlaciones diferentes y comparación de resultados con el uso de objetos W_t y objetos normados.....	79
5.1.5	Análisis de los constructos X, Z_1 , Y, cuyas estructuras de correlaciones son diferentes, y diferentes números de variables, diferentes números de categorías, y diferentes tamaños de muestra	84
5.2	Análisis de Datos Reales.....	86
	CAPÍTULO VI. HALLAZGOS, CONCLUSIONES Y RECOMENDACIONES.....	95
	Referencias Bibliográficas.....	97
	Anexos.....	100

INTRODUCCIÓN

El análisis de tablas de contingencias múltiples consta de un conjunto de herramientas estadísticas que permiten estudiar fenómenos del mundo real con alta carga de complejidad, debido particularmente a que la información necesaria para su abordaje se obtiene a partir de grandes volúmenes de información generados al caracterizar numerosos individuos de acuerdo con uno o varios conjuntos de variables de naturaleza cualitativa, lo que agrega una mayor complicación a su tratamiento estadístico.

La información adecuada para el análisis de tablas de contingencias múltiples se obtiene a partir de situaciones prácticas en las que se miden T conjuntos de variables categóricas sobre un grupo de individuos, o se dispone de un conjunto de variables, igualmente categóricas, medidas en T ocasiones diferentes sobre los mismos individuos. Numerosas propuestas han sido desarrolladas en las últimas décadas para analizar simultáneamente tablas múltiples de datos cuantitativos, siguiendo cuatro direcciones básicas: a) la metodología STATIS; b) el Análisis Procrustes; c) el Análisis Factorial Múltiple; y d) el Análisis Canónico Generalizado. Siguiendo esta línea de investigación, numerosos autores han abordado la misma problemática, pero en el ámbito de datos categóricos, con la finalidad de estudiar las inter-asociaciones y la modelación de los datos que contienen las tablas de contingencias múltiples.

En este trabajo de investigación se construye una propuesta basada en una adaptación de la metodología STATIS para el tratamiento estadístico de datos categóricos provenientes de la caracterización de n unidades de análisis de acuerdo a T conjuntos de variables categóricas (diseño transversal) o de la observación del mismo conjunto de variables en T ocasiones diferentes (diseño longitudinal).

La presentación de todo el proceso y de los resultados de investigación se organizan de acuerdo a los siguientes capítulos:

En el Capítulo I describe fundamentalmente el problema de la investigación, en el cual se plantea la búsqueda de una distancia que sirva como criterio para comparar las tablas de

contingencias múltiples consideradas en el análisis es la idea fundamental; es decir, el problema puede resumirse en la pregunta de la investigación:

¿Es posible tratar esta problemática mediante la adaptación de alguna metodología que con un propósito similar haya sido desarrollada para el caso de que las variables sean cuantitativas?

En resumen, la idea es adaptar la metodología STATIS para el análisis de datos categóricos organizado en Tablas de Contingencia Múltiples.

En el Capítulo II se desarrollan los fundamentos teóricos; se presentan la Metodología STATIS diseñada para analizar simultáneamente T tablas de datos cuantitativos, el Análisis de Correspondencias Compuesto (ACc) planteado para analizar una Tabla de Contingencias Múltiple (TCM), la cual concatena tablas de contingencias bidimensionales definidas por el cruce de dos conjuntos de variables categóricas. Seguidamente, se plantean conceptos propios del Análisis de Correspondencias Binarias (ACB), y su respectiva generalización implícita en el ACc a las tablas TCM, para finalmente presentar las relaciones existentes entre las técnicas señaladas, y así proponer la adaptación del STATIS al análisis de datos categóricos, la cual será denominada **STATIS-C**.

En el Capítulo III se presenta el desarrollo teórico de la propuesta metodológica para el análisis de datos categóricos, cuyo aporte se centra en la definición de la distancia de Hilbert-Schmidt como criterio de comparación de las T tablas de datos objeto de análisis, distancia que ha sido denominada d_{HS-C} .

Seguidamente, se plantea en el Capítulo IV, la metodología usada para la aplicación de la propuesta de adaptación del STATIS para el análisis de datos categóricos, que hemos denominado STATIS-C, a datos simulados y datos reales. Esta implementación sobre datos simulados y reales se presenta en el Capítulo V, mostrando algunas conclusiones específicas que dan cuenta de la viabilidad del método en cada caso.

El Capítulo VI se presenta los hallazgos, conclusiones y recomendaciones obtenidas en la investigación realizada al adaptar la metodología STATIS para el tratamiento de datos categóricos.

CAPÍTULO I

EL PROBLEMA

1.1 PLANTEAMIENTO DEL PROBLEMA

En un sinnúmero de situaciones prácticas que se presentan en diversas disciplinas científicas, se observan T conjuntos de variables sobre un mismo grupo de individuos, o se dispone de un mismo conjunto de variables medidas en T ocasiones diferentes también sobre los mismos individuos. La información que se genera en cualquiera de estas dos situaciones, utilizada para construir aproximaciones conceptuales de la realidad que se analiza, conduce a T **estudios** diferentes que pueden ser identificados en términos de **objetos** descritos como arreglos que describen la estructura de **Interasociaciones** entre las variables o de **inter-distancias** entre individuos.

El tratamiento estadístico de este tipo de datos requiere, en primer lugar, que se efectúen comparaciones entre objetos contentivos de información que identifica a cada uno de los T estudios –interestructura- con el objeto de producir un primer diagnóstico de la heterogeneidad presente entre estudios. En segundo lugar, se trata de caracterizar la estructura de relaciones que se producen a lo interno de cada uno de los T estudios –intraestructura- con el propósito de profundizar en la explicación de las diferencias entre los objetos, si las hubiere. El abordaje de estas tareas desde una perspectiva exploratoria y/o confirmatoria, requiere definir estrategias de tratamiento de la información que permitan desarrollar un análisis comparativo entre los objetos, y otro que conduzca a reconocer las estructuras de información que ellos contienen internamente.

En las últimas décadas, se han desarrollado numerosas investigaciones en torno al análisis simultáneo de tablas múltiples de datos cuantitativos, siguiendo cuatro direcciones básicas: a) la metodología STATIS, introducida por Y. Escouffier et H. L'Hermier en 1976 y desarrollado por C. Lavit (1988) realiza el análisis en tres etapas: Análisis de la interestructura, en la que se compara globalmente las matrices originales sobre un mismo subespacio; búsqueda de una matriz consenso o compromiso, construida como una función de las matrices originales, a cada una de las cuales se le asigna ponderaciones convenientemente elegidas; Análisis de la

intraestructura, en la que se efectúa un estudio detallado de cada uno de los elementos, individuos y/o variables, que definen la estructura de las matrices originales, sobre un subespacio común generado en la etapa anterior. b) el Análisis Procrustes de Tucker en 1958 y Procrustes Generalizado propuesto por Gower en 1975, que compara diferentes configuraciones de datos mediante traslación, rotación y dilatación de éstas, a posiciones de mejor ajuste de las unas con respecto a las otras, para luego calcular una medida del grado de coincidencia entre configuraciones, c) el Análisis Factorial Múltiple, (AFM) desarrollado por Escofier y Pagès (1992), que se desarrolla en dos fases: en primer lugar, analiza cada sub-tabla separadamente, mediante un Análisis de Componentes Principales (ACP). En la segunda fase analiza una tabla definida por la yuxtaposición de las tablas originales ponderadas por el inverso del primer valor propio del ACP en la primera etapa, y d) el Análisis Canónico Generalizado (ACG) presentado por Carroll (1968) y Kettenring (1976), dirigido a la búsqueda de relaciones entre dos conjuntos de variables a partir de ejes canónicos.

Siguiendo esta línea de investigación, autores tales como, Goitisoló (2002), Bécue y Pagés (2004), Vivien y Sabatier (2004), Rodríguez, Galindo-Villardón, Vicente-Villardón (2001), Christensen (1990), Agresti (1984) entre otros, han abordado la misma problemática, pero en el ámbito de datos cualitativos, con la finalidad de estudiar las Interasociaciones y la modelación de tablas de contingencias múltiples resultantes de la observación de variables categóricas. En este orden de ideas, Kiers (1988) establece que los datos de una tabla múltiple, o datos multivía, poseen una configuración de la forma:

Individuos * Conjuntos de Variables * Ocasiones.

Ajustada a las situaciones descritas, la estructura de datos considerada en este trabajo es de tipo multivía, proveniente de la caracterización de las mismas unidades de análisis según T conjuntos de variables categóricas en un diseño de corte transversal, o de la observación del mismo conjunto de variables en T ocasiones diferentes. En ambos casos se dispone de T tablas, cada una de las cuales es obtenida al caracterizar n individuos de acuerdo a un conjunto de variables cualitativas, que a efectos operativos se describen sobre matrices disyuntivas completas yuxtapuestas.

En consideración a este planteamiento, la finalidad de esta investigación, consistió en la búsqueda de un método que permita explorar simultáneamente la información contenida en cada uno de los **T objetos** contruidos especialmente para describir la estructura de interdistancias entre individuos, caracterizados por variables categóricas , a los propósitos de evaluar cambios, o estabilidad, en las estructuras comparadas.

En el proceso de revisión del estado del arte de las técnicas disponibles para analizar simultáneamente varias tablas de datos cuantitativos, se evaluó la factibilidad de adaptar algunas de ellas al tratamiento de datos cualitativos. En este trabajo, se seleccionó la metodología STATIS, dada la analogía que se pudo establecer entre esta técnica y el Análisis de Correspondencias presentado por Benzecri (1973), fundamentalmente en todo lo relacionado con la formulación de las **métricas requeridas para la comparación de los objetos**. En consecuencia, la propuesta que se presenta en este trabajo de investigación en lo esencial, desarrolla un procedimiento que adapta el modelo teórico del STATIS para comparar simultáneamente los **objetos** en los que se registran estructuras de interdistancias entre individuos caracterizados por variables categóricas.

1.2 PREGUNTAS DE INVESTIGACIÓN

El problema de investigación en lo fundamental se plantea analizar la información que se genera al caracterizar simultáneamente un mismo grupo de individuos de acuerdo a T conjuntos de variables categóricas, con lo cual cabe preguntarse si:

¿Es posible abordar esta problemática mediante la adaptación de alguna metodología que con un propósito similar haya sido desarrollada para el caso de variables cuantitativas?

Una tarea que necesariamente debe efectuarse en la dirección de este problema es comparar las estructuras que describen las interdistancias entre individuos caracterizados por cada uno de los diferentes conjuntos de variables categóricas consideradas. En torno a estas ideas surge como pregunta central de investigación:

¿Es posible cuantificar en una medida de distancia las diferencias existentes entre las estructuras de interés?

A este respecto, y en un intento que generalizaría la pregunta que se hace un investigador cuando trata de analizar el valor de una distancia entre individuos en torno a la contribución de las variables, se plantea la siguiente interrogante:

¿Es posible que la medida de distancia propuesta para comparar dos objetos de interés, tome en cuenta las interasociaciones que se producen entre los respectivos conjuntos de variables categóricas que los identifican?

1.3 OBJETIVOS DE INVESTIGACIÓN

1.3.1 Objetivo General

Proponer un procedimiento estadístico basado en una adaptación de la metodología STATIS para efectuar un análisis comparativo de carácter exploratorio entre los **objetos** que identifican las estructuras de interdistancias correspondientes a T **estudios** descritos por diferentes conjuntos de variables categóricas, con el propósito de evaluar diferencias y/o semejanzas entre ellos.

1.3.2 Objetivos Específicos

1.3.2.1 Obtener los elementos básicos para el análisis de la interestructura mediante la construcción de:

- a) Un **objeto** representativo de la estructura de interdistancias entre los individuos a lo interno de un **estudio**.
- b) Un producto escalar entre los **objetos** representativos de dos **estudios** descritos por conjuntos diferentes de variables categóricas, que permita definir una medida de covariación entre los objetos en el sentido del coeficiente RV de Escoufier.
- c) Una distancia entre **objetos representativos** de diferentes **estudios**, en el sentido de la distancia de Hilbert-Schmidt.

1.3.2.2 Representar aproximadamente las semejanzas y/o diferencias entre los objetos sobre un espacio euclídeo construido con la información reportada por:

- a) Una matriz **S** de productos escalares entre los **objetos**.
- b) Los vectores que describen las direcciones principales de variabilidad entre los objetos obtenidos a partir de la descomposición espectral de la matriz **S**.

1.3.2.3 Evaluar algunas propiedades del producto escalar propuesto, de la distancia asociada, y de la representación de los objetos sobre espacios euclídeos.

1.3.2.4 Utilizar el procedimiento estándar definido por la metodología STATIS para completar las etapas de análisis de la intraestructura en un espacio compromiso.

1.3.2.5 Diseñar herramientas automatizadas bajo el lenguaje de programación Matlab que permitan la implementación de la propuesta.

1.3.2.6 Ensayar la propuesta utilizando datos simulados con diferentes grados de similitud, diferentes esquemas de asociación: nula, moderada y fuerte; diferentes tamaños de muestra; y diferente números de variables en cada uno de los conjuntos que identifican los T estudios.

1.3.2.7 Realizar una aplicación sobre un problema real.

1.4 ANTECEDENTES

A continuación se hace una revisión de algunos desarrollos recientes del tratamiento estadístico de varias tablas de datos múltiples. Entre otros autores que han trabajado en la misma línea, se citan los siguientes autores:

Israëls (1987) reseña a los autores Keller y Wansbeek, quienes en 1983, fundamentados en un trabajo presentado de Leclerc (1974) sobre tratamiento de datos categóricos multivía, plantea el Análisis de Correspondencias Compuesto (ACc) como una técnica para cuantificar variables categóricas, utilizando un enfoque que generaliza el Análisis de Correspondencias Binarias (ACB) sobre una tabla de contingencias múltiple (TCM). La finalidad de esta metodología consiste en estudiar la asociación entre los dos conjuntos de variables categóricas

$X = \{X_j, j=1,2,...J\}$ e $Y = \{Y_i, i=1,2,...I\}$, con un enfoque que conduce a la cuantificación de las categorías de cada una de las variables en los dos conjuntos considerados. En su trabajo, Israëls (op cit) justifica la construcción de una medida de asociación global tipo chi-cuadrado para la TCM definida por el cruce de las variables X 's e Y 's, encontrando que es posible definirla como un promedio de los estadísticos chi-cuadrado correspondientes a las subtablas definidas por el cruce de cada una de las variables Y con cada una de las X . Este resultado suministra una medida que permite evaluar la importancia de la asociación en una sub-tabla respecto de la asociación global. Estos resultados serán desarrollados ampliamente en el cuerpo metodológico de la propuesta por constituir una base referencial muy importante para el logro de los objetivos planteados.

Greenacre (1993) desarrolla una generalización del Análisis de Correspondencias Binarias para explorar las asociaciones en tablas de tres-vías, organizando los datos de acuerdo con las diferentes estructuras que se describen a continuación:

En primer lugar, resume la información en la tabla de tres vías sobre un arreglo de dos vías, describiendo para cada categoría de una tercera variable la tabla de contingencias resultante del cruce de las dos primeras. En este sentido, propone una secuencia de gráficos que le permiten visualizar las asociaciones e interacciones entre las variables:

- a) Un primer gráfico en el que se representan los perfiles fila de la primera variable, definidos por las categorías de la segunda, para cada una de las categorías de la tercera variable.
- b) Un segundo gráfico obtenido a partir de la misma tabla anterior pero a la que se le han agregado como categorías suplementarias las filas de dos tablas bidimensionales colapsadas que son obtenidas al cruzar la primera variable con la segunda, y la tercera con la segunda. Los perfiles fila de estas tablas, que son esencialmente los centros de gravedad en el arreglo bidimensional en el que se organiza la información de las tres variables, son representados como ilustrativos.
- c) En un tercer gráfico se representan los perfiles columna de la segunda variable definidos por las categorías de la variable resultante de combinar la primera y la tercera variables.

- d) Este autor también desarrolla el Análisis de Correspondencia Conjunto (ACJ) para explorar simultáneamente las interasociaciones entre pares de variables. Los resultados los representa sobre un único gráfico de manera análoga a la representación de una matriz de varianzas y covarianzas en el Análisis de Componentes Principales.

Djahuri (1998) introduce dos proposiciones matemáticas para construir un método para seleccionar variables categóricas basadas en el coeficiente RV de Escoufier en el marco del ACB. Basado en estas proposiciones, la selección de variables consiste en la escogencia de un conjunto de operadores de Escoufier cuyo promedio maximiza el coeficiente RV.

Rodríguez, Galindo-Villardón, Vicente-Villardón (2001) proponen un método para la comparación de los resultados provenientes de dos análisis Biplot aplicados a las mismas variables en dos grupos de individuos mediante la integración de subespacios utilizando rotaciones y escalamientos, con la finalidad de identificar las semejanzas y/o diferencias entre los grupos. La propuesta consiste en determinar un consenso con el cual sea posible contrastar las configuraciones resultantes de los análisis en consideración, y se trata de responder a las siguientes preguntas: ¿Cuáles grupos muestran respuestas similares? ¿Cómo difieren los grupos? ¿Cómo difieren los individuos en sus respuestas en los diferentes grupos?

Bécue y otros (2003) utilizan una metodología basada en el Análisis Factorial Múltiple para tablas de contingencias múltiples cuyas siglas son AFMTC, que facilita la comparación de varias tablas. Esta metodología produce una descripción global de la tabla múltiple y una comparación sistemática de las subtablas (descripción parcial). En la siguiente Figura 1 se presenta la tabla analizada en la investigación en referencia, allí puede observarse la estructuración de tablas por bloques.

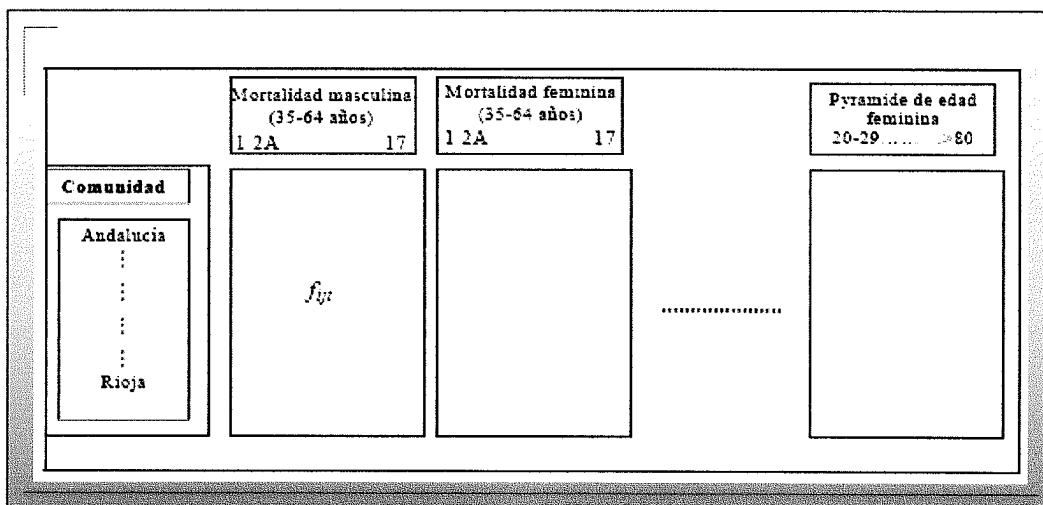


Figura 1: Gráfico del Esquema del arreglo de datos de Bécue-Pagès (2003)

Vivien y Sabatier (2004) proponen el método DO-ACT para determinar dimensiones comunes dentro de dos tablas referidas por $\{X_k\}$ ($k=1; \dots; K$) and $\{Y_l\}$ ($l=1; \dots; L$) y denominadas multibloques. Donde K y L son los tamaños, no son simultáneamente igual a 1. Cada matriz X_k y Y_l son de orden $n \times p_k$ y $n \times p_l$ respectivamente, definidas por variables cuantitativas medidas sobre los mismos individuos. Una tabla multibloque es constituida por un conjunto de matrices, no necesariamente del mismo tamaño. La propuesta de estos autores está relacionado con la metodología STATIS y el método inter-baterías de Tucker (1958). El procedimiento es realizado en tres pasos: el primero proporciona un resumen de las dos tablas multibloque; en el segundo, se construye una grafica de representaciones óptimas de las observaciones (una para cada tabla); y el tercero, una descripción global de cada tabla de las dos multibloque. El propósito del método es describir las Interasociaciones entre las dos tablas de diferente tamaño, en una dimensión común.

Bécue-Bertaut y Pagès (2004) presentan una nueva metodología para comparar la estructura de varias tablas de contingencias, desarrollada sobre diferentes muestras o poblaciones, en la que la variable que define las filas de las tablas es la misma, pudiéndose considerar diferentes variables para las columnas. Esta metodología combina algunos aspectos de los métodos de coordenadas principales, análisis de correlación canónica y análisis Procrustes.

Bécue-Bertaut y Pagès (2008) proponen un método para analizar y agrupar individuos descritos por un conjunto mixto de variables cuya información es arreglada en una tabla

múltiple que yuxtapone conjuntos de variables cuantitativas, variables indicatrices y frecuencias, como se ilustra en la siguiente figura 2.

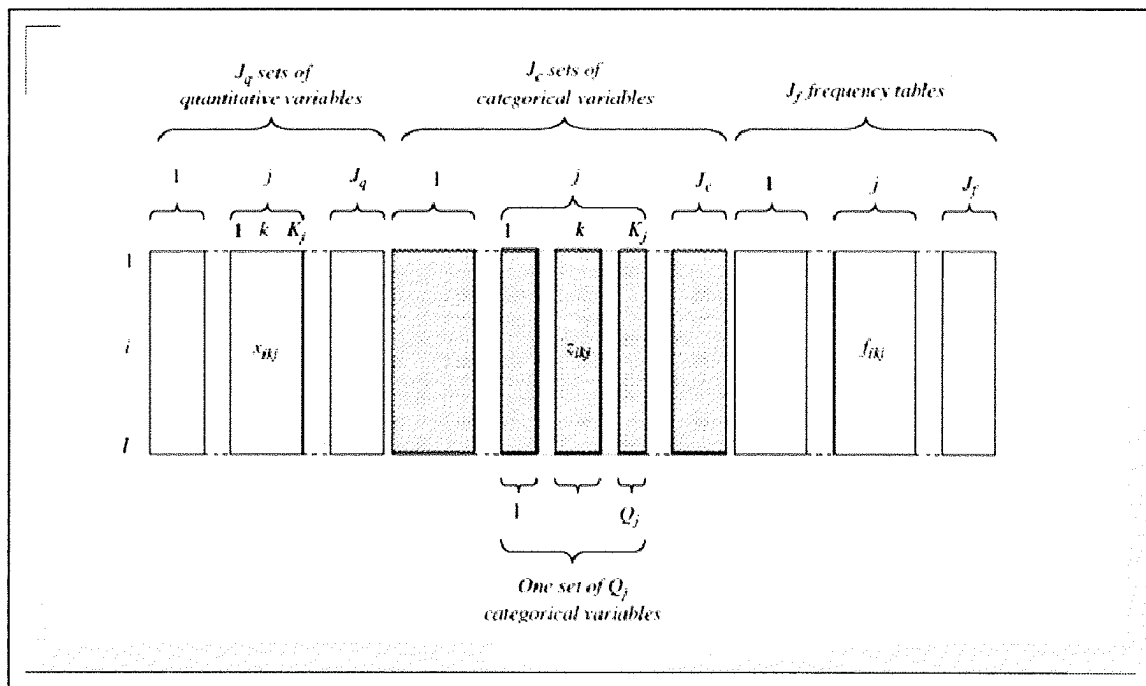


Figura 2: Esquema de los datos de Bécue-Bertaut y Pagès (2008)

Los autores consideran el problema del peso de los individuos debido a que este es fijado, usualmente uniforme, por el usuario, en los ACP y ACM, pero impuesto por las marginales en el ACB. El problema consiste en definir una distancia global mediante la combinación de las distancias, llamadas distancias separadas, obtenidas mediante el ACP, el ACM y el ACB, a partir de una extensión del AFM.

En las últimas décadas, el problema del análisis de tablas multivías basada en la comparación de las mismas como punto inicial del análisis, constituye una línea de investigación aun no acabada, donde la definición de una distancia generalmente es requerida como una medida de la variabilidad entre y dentro de este tipo de datos arreglados en tablas de diversas formas.

CAPÍTULO II

MARCO TEÓRICO

2.1 BASES TEÓRICAS

En este capítulo se desarrollan las bases teóricas de un conjunto de técnicas que en lo fundamental constituyen referencia para la propuesta que se hace en esta investigación.

2.1.1 Estructura de Datos

A los efectos de establecer la estructura de la información que sirve de referencia para el desarrollo del problema de investigación planteado, se especifica el mecanismo generador de la misma y la configuración de los datos. Dos escenarios se pueden presentar para ello; el primero corresponde a un diseño de corte transversal, en el que las unidades de análisis se caracterizan de acuerdo a T conjuntos de variables categóricas, cada variable con el mismo o diferente número de categorías. La tabla de datos que se describe seguidamente, es ilustrativa al respecto:

Tabla 2.1: T matrices de Datos Cualitativos obtenidas en un Diseño Transversal

	T tablas de datos				
	Dimensión 1 Conjunto de variables $\overbrace{X_1 \dots X_j \dots X_J}^x$...	Dimensión t Conjunto de variables $\overbrace{Y_1 \dots Y_i \dots Y_I}^y$...	Dimensión T Conjunto de variables $\overbrace{Z_1 \dots Z_i \dots Z_I}^z$
	Tabla T ₁ J matrices disyuntivas completas	...	Tabla T _t I matrices disyuntivas completas	...	Tabla T _T K matrices disyuntivas completas
n individuos {					

La t-ésima tabla correspondiente a la dimensión t, es una matriz de orden $n \times L_t$, donde $L_t = K_{1(t)} + \dots + K_{i(t)} + \dots + K_{I(t)}$ es el número total de categorías de las I variables del conjunto Y que la definen la t-ésima dimensión, y $K_{i(t)}$ es el número de categorías de la i-ésima variable Y_i . Este número total de categorías L_t define el número de variables indicatrices que conforma la tabla T_t , en la cual se juxtapone I matrices disyuntivas completas; $t=1,2,\dots,T$.

El segundo escenario queda descrito por la información correspondiente a un diseño longitudinal, en el cual las unidades de análisis son medidas en términos de un solo conjunto de variables categóricas, digamos X , en T ocasiones definidas temporal o espacialmente. Los datos en este caso presentan la configuración de la siguiente Tabla 2.2.

Tabla 2.2: T matrices de Datos Cualitativos obtenidas en un Diseño Longitudinal

Mismo grupo de individuos	T tablas de datos				
	Ocasión 1 Conjunto de variables $\overbrace{X_{1(1)} \cdots X_{j(1)} \cdots X_{J(1)}}^X$...	Ocasión t Conjunto de variables $\overbrace{X_{1(t)} \cdots X_{j(t)} \cdots X_{J(t)}}^X$...	Ocasión T Conjunto de variables $\overbrace{X_{1(T)} \cdots X_{j(T)} \cdots X_{J(T)}}^X$
	Tabla T_1 J matrices disyuntivas completas	...	Tabla T_t J matrices disyuntivas completas	...	Tabla T_T J matrices disyuntivas completas
n individuos {					

La t -ésima tabla correspondiente a la ocasión t , es una matriz de orden $n \times L_t$, donde $L_t = L_{1(t)} + \cdots + L_{j(t)} + \cdots + L_{J(t)}$ es el número total de categorías de las J variables del conjunto X . Este número total de categorías L_t definen el número de variables indicatrices que conforma la tabla T_t , en la cual se yuxtapone J matrices disyuntivas completas; $t=1,2,\dots,T$.

La información obtenida en un diseño transversal es generada por una muestra aleatoria, y la del longitudinal por T muestras aleatorias relacionadas. Es obvio que el tipo de análisis, que es una condición intrínsecamente relacionada con el diseño, requiere la exploración de la **interestructura** y de la **intraestructura** de los objetos que se construyan para representar las T tablas de datos cualitativos, ello coincide esencialmente con el objetivo de la **Metodología STATIS** para el caso de datos cuantitativos Lavit (1988). En consideración a que la estructura de datos del problema planteado en esta investigación es esencialmente similar a la del STATIS, excepto por la escala de medida, se ha considerado plausible definir un procedimiento para analizar las T tablas de datos cualitativos mediante una adaptación de la metodología STATIS.

La metodología STATIS presenta dos versiones según el interés de análisis; el STATIS y el STATIS DUAL, por lo tanto la contextualización del problema planteado, requiere que se especifique cual de los dos escenarios de aplicabilidad de la metodología STATIS es adecuada para su adaptación a datos cualitativos.

2. 2 Metodología STATIS

La metodología STATIS (Structuration des Tableaux a Trois Indices de la Statistique), que es introducida por Escouffier y L'Hermier (1976) y desarrollado por Lavit (1988), permite la exploración simultánea de varias tablas de datos provenientes de la medición de variables cuantitativas que se clasifican de acuerdo con las diferentes dimensiones que fundamentan diferentes aspectos de un determinado fenómeno bajo estudio. Esta metodología es aplicable a dos situaciones que difieren en cuanto al objetivo de investigación, y en consecuencia en cuanto al diseño: a) La primera de ellas, centra su objetivo de análisis en evaluar los cambios en las posiciones relativas de los individuos. b) La segunda da prioridad al análisis de los cambios en las relaciones entre las variables.

Los objetivos de análisis de la metodología STATIS se pueden resumir de la siguiente manera:

- 1) Comparar globalmente las estructuras de interés que identifican las diferentes tablas, utilizando para ello representaciones euclídeas, sobre las cuales es posible visualizar similitudes y/o diferencias entre las estructuras.
- 2) Determinar una estructura que resuma, a modo de promedio ponderado, las correspondientes estructuras de las diferentes tablas bajo consideración.
- 3) Identificar los individuos, o las variables responsables de tales similitudes o diferencias.

2.2.1 Escenarios de aplicación de la Metodología STATIS

Según los objetivos de la investigación, sea centrado en el análisis de la posición relativa de los individuos, o de las relaciones entre las variables, la metodología STATIS es aplicable en uno de los dos escenarios siguientes:

Escenario 1: STATIS

Este escenario corresponde a un diseño de corte transversal, la información se dispone sobre un arreglo particionado en T tablas, y el análisis se centra en los cambios que ocasionalmente se producen en las posiciones relativas de los individuos, al ser medidos por uno u otro conjunto de variables

n individuos	T Conjuntos de Variables Cuantitativas				
	Conjunto 1	...	Conjunto t	...	Conjunto T
	$\{X_{1(1)}...X_{j(1)}...X_{J(1)}\}$...	$\{Y_{1(t)}...Y_{i(t)}...Y_{I(t)}\}$...	$\{Z_{1(T)}...Z_{K(T)}...Z_{K(T)}\}$

Una variante del escenario anterior corresponde a un diseño de corte longitudinal, donde las mismas unidades estadísticas son caracterizadas de acuerdo a un sólo conjunto de variables cuantitativas, medidas en diferentes ocasiones. La información al igual que antes, se dispone sobre un arreglo particionado en T tablas, y el análisis se centra en los cambios que eventualmente se producen en las posiciones relativas de los individuos, al ser medidos en las diferentes ocasiones.

n individuos caracterizados de acuerdo a J variables cuantitativas medidas en T ocasiones	T Ocasiones				
	Ocasión 1	...	Ocasión t	...	Ocasión T
	$\{X_{1(1)}...X_{j(1)}...X_{J(1)}\}$...	$\{X_{1(t)}...X_{j(t)}...X_{J(t)}\}$...	$\{X_{1(T)}...X_{j(T)}...X_{J(T)}\}$

Escenario 2: STATIS DUAL

Este escenario corresponde a un diseño de corte transversal, la información se dispone sobre un arreglo particionado en T tablas, y el interés del análisis radica en la evaluación de los cambios que se producen en la estructura de correlaciones entre las variables, al pasar de un grupo de individuos a otro.

plano es posible visualizar la interestructura; es decir, si dos objetos aparecen muy cercanos, ello es indicativo de que las estructuras de interdistancias entre los individuos de las tablas correspondientes, son similares.

En particular, también en lo relacionado con el coeficiente de Escoufier entre dos objetos W_t y W_s será posible su descripción geoméricamente como un coseno de ángulo entre los vectores que describen las posiciones de esos objetos sobre el plano factorial en consideración:

$$RV(t,s) = \left\langle \frac{W_t}{\|W_t\|_{HS}} \middle| \frac{W_s}{\|W_s\|_{HS}} \right\rangle_{HS} = \frac{S_{ts}}{S_{tt}^{1/2} S_{ss}^{1/2}} = \cos(\theta_{A_t, \theta_{A_s}})$$

es decir, si el ángulo entre los dos objetos representados sobre el plano tiende a ser pequeño, ello indica que los estudios respectivos guardan entre sí una fuerte relación.

Por otra parte, debido a que S es una matriz simétrica con todos sus elementos positivos, por el Teorema de Frobenius, citado en Dazy y Le Barzic (1996), admite un primer autovector con todos coordenadas del mismo signo. Es decir, en general, en las representaciones gráficas sobre el primer plano factorial, la variabilidad de los estudios se distingue con mayor claridad sobre el eje 2, tal como se ilustra en la siguiente figura 2.1:

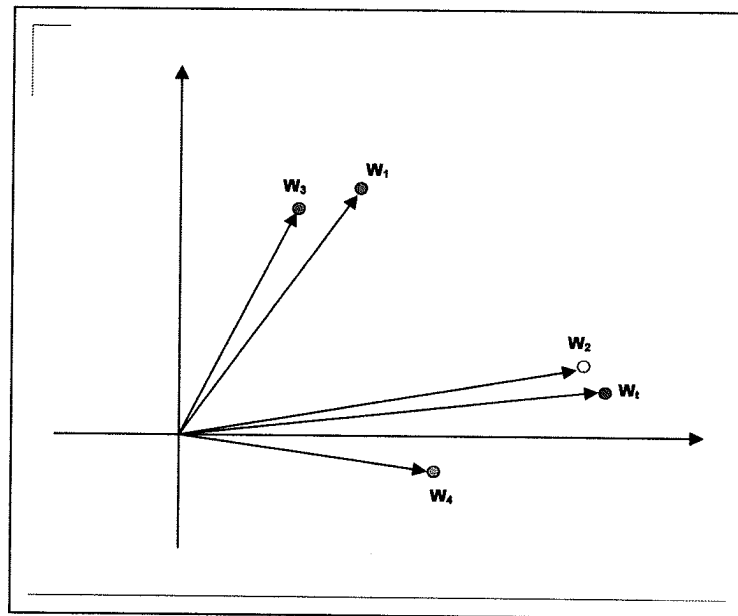


Figura 2.1: Ilustración de la Representación de T Tablas de datos

En esta imagen euclídea se muestra la existencia de una estructura común para los estudios representados por W_2 y W_t , y otra para los estudios representados por W_1 y W_3 ; que a su vez, aparecen diferenciadas de la estructura representada por el punto W_4 .

2.2.3.4 Determinación del Compromiso

La segunda etapa del STATIS tiene como objetivo resumir las estructuras de interdistancias entre los individuos bajo estudio, que son organizadas sobre los objetos representativos de las T tablas de datos, en una sola denominada compromiso W , arreglo es definido como un promedio ponderado de los objetos, o de los objetos normados, en la forma

$$W = \sum_{t=1}^p \alpha_t W_t \quad \text{o} \quad W = \sum_{t=1}^T \alpha_t \frac{W_t}{\|W_t\|_{HS}}, \text{ respectivamente.}$$

El vector sobre el que se organizan los coeficientes α de la combinación lineal, es seleccionado de tal manera que el arreglo compromiso W sea el objeto más correlacionado con los objetos W_t $t=1,2,\dots,T$. Es decir, bajo esta premisa debe ocurrir que la norma del compromiso W queda en la forma $\|W\|_{HS} = \sum_{t=1}^p \pi_t \|W_t\|_{HS}$ si los objetos no son normados.

En caso de que los objetos sean normados, la norma del compromiso W es igual a 1.

En la construcción del compromiso, los coeficientes α_t son tales que:

$$\alpha_t = \frac{1}{\sqrt{\lambda_1}} \left(\sum_{t=1}^T \pi_t \sqrt{S_{tt}} \right) \pi_t \gamma_{1(t)}, \text{ si el objeto representativo es } W_t$$

Siendo $\gamma_{1(t)}$ la t -ésima coordenada del vector $\gamma_1 = (\gamma_{1(1)} \cdots \gamma_{1(T)})'$ el autovector de $S\Delta$ asociado al autovalor λ_1 y $S_{tt} = \|W_t\|_{HS}^2$ es el t -ésimo elemento de la diagonal de la matriz S ; y por otra parte:

$$\alpha_t = \frac{1}{\sqrt{\lambda_1}} \pi_t \gamma_{1(t)} \quad \text{si el objeto es } \frac{W_t}{\|W_t\|_{HS}}$$

El compromiso W se incorpora como un arreglo ilustrativo en el primer plano factorial de la representación euclídea de los objetos, construido en la etapa de la interestructura,

básicamente con el propósito de determinar si este nuevo arreglo representa adecuadamente las T tablas comparadas, o también para establecer a cuales tablas puede representar en forma apropiada. Se interpreta, gráficamente, como una referencia para observar la dispersión de los T objetos representativos de los Estudios, y determinar en función de la distancia entre éstos y el compromiso, si los representa adecuadamente; es decir, mientras menor sea la distancia, mayor será la adecuación de la representación de los objetos representativos por compromiso. Analíticamente, los elementos del compromiso W son productos de los individuos escalares compromiso, definidos como una combinación lineal de los productos escalares entre individuos obtenidos en cada uno de los objetos representativos.

La siguiente figura ilustra una representación simultánea de los T objetos y de su compromiso W como información ilustrativa.

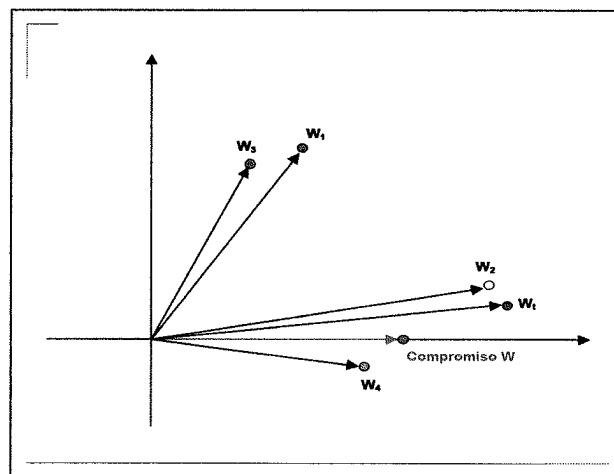


Figura 2.2: Ilustración de la representación del compromiso W en el plano de la interestructura

La figura 2.2. indica que las tablas de datos considerados no son representadas adecuadamente por el compromiso W . Se observan tres estructuras bien diferenciadas, por lo tanto en este caso se recomendaría analizar por separado, los grupos de tablas $\{T_1, T_3\}$, $\{T_2, T_t\}$ y la tabla T_4 , obviamente, el compromiso W no podría utilizarse para resumir las T tablas de datos.

El estudio de la intraestructura conduce a representar posiciones compromiso de los individuos. En esta imagen, un punto corresponde a la posición promedio que un individuo

tendría en los objetos que identifican las T tablas. La representación es proporcionada por la diagonalización del arreglo compromiso WD de orden $n \times n$.

De esta manera, si se consideran $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$, los autovectores de la matriz WD , asociados respectivamente a los autovalores μ_i ; $i=1, \dots, n$; y se denota la imagen euclídea de los individuos compromiso mediante los puntos B_1, \dots, B_n , cuyas coordenadas de proyección quedan definidas en la forma:

$$\sqrt{\mu_k} \varepsilon_k = \frac{1}{\sqrt{\mu_k}} WD \varepsilon_k$$

Un individuo compromiso es, entonces, un promedio ponderado del producto escalar entre individuos descritos en cada uno de los objetos representativos de los T estudios, los cuales definen una distancia entre los i, j -ésimos individuos compromiso.

En el espacio euclídeo definido por el compromiso, se incluyen como puntos suplementarios las interdistancias de individuos en las T tablas, lo que da lugar a representación en proyección de nT puntos. Sobre esta gráfica es posible determinar cuáles individuos son responsables de las diferencias que eventualmente existen entre dichas tablas.

Las coordenadas de proyección de los n individuos descritos de acuerdo a la configuración W_t sobre ε_k , el vector director del k -ésimo eje del espacio compromiso, que se denotan mediante $B_{1(t)k}, \dots, B_{n(t)k}$ para $t=1, \dots, T$, pueden obtenerse mediante la expresión:

$$\frac{1}{\sqrt{\mu_k}} W_t D \varepsilon_k, \text{ si los objetos representativos son } W_t$$

o mediante:

$$\frac{1}{\sqrt{\mu_k}} \frac{1}{\|W_t\|_{HS}} W_t D \varepsilon_k, \text{ si los objetos representativos son normados.}$$

A partir de la definición del arreglo compromiso, es fácil verificar que los individuos compromiso son el centro de gravedad de los puntos $B_{1(t)}, \dots, B_{n(t)}$; con $t=1, \dots, T$, ponderados por los coeficientes α_t .

2.2.3.5 Trazado de trayectorias

La representación de cada uno de los n individuos a partir de las configuraciones W_t , define T puntos en la gráfica compromiso, que permiten trazar una trayectoria con el propósito de evaluar los cambios que se producen en sus posiciones al pasar de uno a otro de los T estudios. Esto permite identificar aquellos individuos que son responsables de las diferencias entre las tablas.

Cuando se trata de tablas de datos producidas por un estudio transversal, las trayectorias se interpretan respecto de la posición compromiso de cada individuo; una trayectoria que es poco extensa alrededor de su posición compromiso formando un bucle pequeño, corresponde a un individuo cuya caracterización es similar en las T tablas. Por el contrario, una trayectoria de gran amplitud refleja diferentes caracterizaciones del individuo por las variables de una tabla a otra.

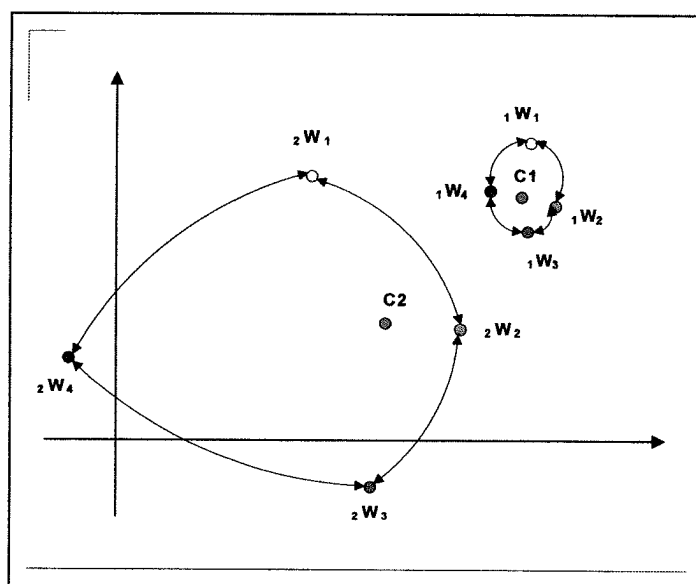


Figura 2.3: Ilustración del estudio de la intraestructura de dos individuos con diferentes interdistancias en los estudios de un diseño transversal

En la figura 2.3 el individuo 1 queda bien representado por su posición compromiso (C1); puede observarse que los $T = 4$ puntos (W_1, \dots, W_4) se encuentran a la misma distancia de la posición compromiso, ello indica que todas las dimensiones contribuyen en la misma medida a esa posición. El individuo 2, en los distintos estudios presenta caracterizaciones que difieren sensiblemente del compromiso, que parece estar mejor definida por el estudio 2. En este caso, cabe preguntarse, por las características que presenta este individuo en cada una de las tablas, con el propósito de indagar la razón por la que los distintos estudios contribuyen en diferente medida a su posición compromiso.

En caso de un estudio longitudinal, el dibujo de las trayectorias permite constatar la ocurrencia de cambios de un periodo de tiempo a otro. Las trayectorias se interpretan respecto a la evolución media, es decir, una trayectoria que es poco extensa, corresponde a un individuo cuya evolución sigue a la evolución media; es decir, que para cada conjunto de variables las diferencias con respecto al compromiso – individuo promedio– es similar de un año a otro. Por el contrario, una trayectoria de gran amplitud refleja cambios en las características del individuo con el curso de los años, diferentes de la evolución promedio. La siguiente figura 2.4, ilustra trayectorias hipotéticas en caso de que el estudio sea longitudinal.

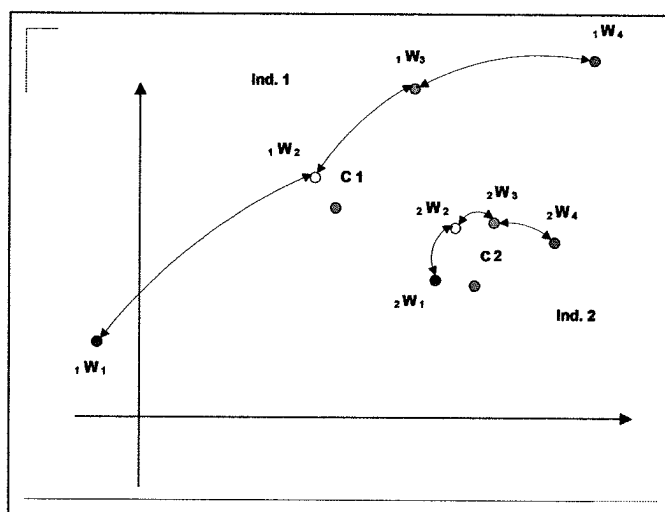


Figura 2.4: Ilustración del estudio de la intraestructura de dos individuos con diferentes interdistancias de un estudio longitudinal

2.3 Análisis de Correspondencias Compuesto (ACc)

Israëls (op cit) deduce a partir del Análisis de Correlación Canónica (ACC), entre otras, las técnicas ACB, ACM y el **Análisis de Correspondencias Compuesto (ACc)** como procedimientos para cuantificar variables categóricas, la cuales denominó técnicas de autovalores debido a que los modelos teóricos que las sustentan están basados en la descomposición en valores singulares.

La estructura de datos sobre la que es factible aplicar un ACc, es generada en un estudio transversal, en el que se observan dos conjuntos de variables categóricas sobre un mismo grupo de individuos, un primer conjunto de I variables categóricas $\mathbf{Y}=\{Y_1, \dots, Y_i, \dots, Y_I\}$, y un segundo conjunto de J variables categóricas $\mathbf{X}=\{X_1, \dots, X_j, \dots, X_J\}$.

donde:

$$Y_i = \left(y_i^{(1)}, \dots, y_i^{(k_i)}, \dots, y_i^{(K_i)} \right)$$

siendo K_i es el número de categorías de la i-ésima variable, Y_i , para $i=1, \dots, I$.

$$X_j = \left(X_j^{(1)}, \dots, X_j^{(l_j)}, \dots, X_j^{(L_j)} \right)$$

y L_j es el número de categorías de la j-ésima variable, X_j , para $j=1, \dots, J$.

Las mediciones en ambos conjuntos de variables son descritas mediante variables indicatrices, las cuales conforman dos tablas de datos; una que yuxtapone I matrices disyuntivas completas definidas por las variables Y's, y otra que yuxtapone J matrices disyuntivas completas definidas por las variables X's.

El cruce de estas dos tablas de datos reproduce una tabla que concatena $I \times J$ tablas de contingencias bidimensionales, de orden $\left(\sum_i K_i \right) \times \left(\sum_j L_j \right)$, y es denominada tabla concatenada o tabla compuesta.

El arreglo de información formulado por Israëls (op cit) es básicamente una tabla, que expresada en forma teórica, tiene la representación siguiente:

$$P = \frac{1}{I \times J} E(YX') = \frac{1}{I \times J} \begin{pmatrix} E(Y_1 X_1') & E(Y_1 X_2') & \dots & E(Y_1 X_J') \\ E(Y_2 X_1') & E(Y_2 X_2') & \dots & E(Y_2 X_J') \\ \vdots & \vdots & \ddots & \vdots \\ E(Y_I X_1') & E(Y_I X_2') & \dots & E(Y_I X_J') \end{pmatrix}$$

Donde la sub-matriz genérica $E(Y_i X_j')$ es una matriz tal que:

$$E(Y_i X_j') = \begin{pmatrix} P_{Y_i X_j}^{11} & P_{Y_i X_j}^{12} & \dots & P_{Y_i X_j}^{1L_j} \\ P_{Y_i X_j}^{k1} & P_{Y_i X_j}^{k2} & \dots & P_{Y_i X_j}^{kL_j} \\ \vdots & \vdots & \ddots & \vdots \\ P_{Y_i X_j}^{K_i 1} & P_{Y_i X_j}^{K_i 2} & \dots & P_{Y_i X_j}^{K_i L_j} \end{pmatrix}$$

Este elemento genérico $E(Y_i X_j')$, es una tabla descrita por los valores esperados de las frecuencias obtenidas al cruzar las variables i -ésima Y_i , y la j -ésima X_j , en consecuencia es una matriz de probabilidades, donde:

$P_{Y_i X_j}^{kl} = \text{Prob}(Y_i=k, X_j=l)$ es la probabilidad de que la variable Y_i tome el k -ésimo valor y la variable X_j el l -ésimo valor. En consecuencia:

$$\sum_{k=1}^{K_i} \sum_{l=1}^{L_j} P_{Y_i X_j}^{kl} = 1$$

Además, el vector que contiene las probabilidades marginales de las filas de la matriz P vienen dadas por:

$$\frac{1}{I \times J} \sum_{j=1}^J \sum_{l=1}^{L_j} P_{Y_i X_j}^{kl} = \frac{P_{Y_{i0}}^k}{I} = \frac{E(Y_i^k)}{I}$$

Análogamente el vector que contiene las probabilidades marginales de las columnas es:

$$\frac{1}{I \times J} \sum_{i=1}^I \sum_{k=1}^{K_i} P_{Y_i X_j}^{kl} = \frac{P_{X_{0j}}^l}{J} = \frac{E(X_j^l)}{J}$$

Entonces, es posible definir los arreglos:

$$D_Y = \begin{pmatrix} D_{Y_1} & & & & \theta \\ & \ddots & & & \\ & & D_{Y_i} & & \\ & & & \ddots & \\ \theta & & & & D_{Y_I} \end{pmatrix}$$

$$\text{siendo } D_{Y_i} = \text{diag} \left(\frac{P_{Y_{i0}}^1}{I}, \dots, \frac{P_{Y_{i0}}^k}{I}, \dots, \frac{P_{Y_{i0}}^{K_i}}{I} \right)$$

En forma similar se define el arreglo:

$$D_X = \begin{pmatrix} D_{X_1} & & & & \theta \\ & \ddots & & & \\ & & D_{X_j} & & \\ & & & \ddots & \\ \theta & & & & D_{X_J} \end{pmatrix}$$

$$\text{siendo } D_{X_j} = \text{diag} \left(\frac{P_{X_{0j}}^1}{J}, \dots, \frac{P_{X_{0j}}^l}{J}, \dots, \frac{P_{X_{0j}}^{L_j}}{J} \right)$$

El tratamiento que el Israëls (op cit) da a la tabla P de probabilidades, básicamente tiene por objeto la cuantificación de variables categóricas, formalizando su presentación desde la perspectiva de un Análisis de Correlaciones Canónicas aplicado sobre dos conjuntos de

vectores aleatorios, que respectivamente quedan definidos por variables aleatorias cualitativas. En términos de lograr la cuantificación, efectúa la descomposición en valores singulares de las matrices $[D_y^{-1} P D_x^{-1} P']$ y $[D_x^{-1} P' D_y^{-1} P]$, o trabajando con las autoecuaciones para hallar los autovectores asociados a sus respectivos autovalores λ al cuadrado:

$$[D_y^{-1} P D_x^{-1} P'] \times A = A \Lambda^2, \text{ con } A' D_y A = I$$

$$[D_x^{-1} P' D_y^{-1} P] \times B = B \Lambda^2, \text{ con } B' D_x B = I$$

Consiguientemente, $A' P B = \Lambda$

donde los autovectores referidos en A y B son las cuantificaciones buscadas por Israëls (op cit).

Finalmente, este planteamiento conduce a que proponga las siguientes relaciones:

Traza $((D_y^{-1/2} P D_x^{-1/2}) (D_x^{-1/2} P' D_y^{-1/2})) - 1 = \phi^2$; esto es:

$$\phi^2 = \frac{1}{I \times J} = \sum_{i=1}^I \sum_{j=1}^J \phi_{(i,j)}^2$$

donde $\phi_{(i,j)}^2$ es la medida de la asociación entre las variables Y_i y X_j que permiten la construcción de la tabla E ($Y_i X_j$).

Por consiguiente, la medida de asociación chi-cuadrado en una tabla concatenada obtenida a partir de la observación de los dos conjuntos de variables sobre N individuos, viene dada por:

$$\chi^2 = N I J \phi^2 = N \sum_{i=1}^I \sum_{j=1}^J \phi_{(i,j)}^2 = \sum_{i=1}^I \sum_{j=1}^J \chi_{(i,j)}^2$$

en la cual $\chi_{(i,j)}^2$ es el estadístico chi-cuadrado que corresponde a la (i,j)-ésima tabla de contingencias bidimensional.

2.4 DEFINICIONES Y TÉRMINOS BÁSICOS

Siguiendo la presentación que hacen Ramírez y Vásquez (2003), se describen a continuación los elementos fundamentales del Análisis de Correspondencias Binarias que son utilizados a los efectos de esta investigación.

2.4.1 Definiciones del Análisis de Correspondencias Binarias (ACB)

Definición 1

Dada la tabla de contingencias bidimensionales $F_{yx} = \frac{1}{n} Y'X = (f_{ij})$, las matrices diagonales de pesos de las categorías fila $D_Y = \frac{1}{n} Y'Y = \text{diag}(f_i)$ y de las categorías columna $D_X = \frac{1}{n} X'X = \text{diag}(f_j)$, la matriz de perfiles fila se define mediante el arreglo

$$D_Y^{-1}F_{YX} = \left(\frac{f_{ij}}{f_i} \right)$$

Definición 2

El centro de gravedad de la nube de perfiles fila se define como el promedio ponderado:

$$G = \sum_{i=1}^I f_i \left(\frac{f_{i1}}{f_i}, \dots, \frac{f_{ij}}{f_i}, \dots, \frac{f_{iJ}}{f_i} \right) = (f_{.1}, \dots, f_{.j}, \dots, f_{.J})$$

Definición 3

La distancia chi-cuadrado dotada de la métrica D_X^{-1} , entre los perfiles fila i -ésimo y s -ésimo, queda definida en la forma:

$$d_{\chi^2}^2 = \left(\frac{f_{i1}}{f_i} - \frac{f_{s1}}{f_s}, \dots, \frac{f_{ij}}{f_i} - \frac{f_{sj}}{f_s}, \dots, \frac{f_{iJ}}{f_i} - \frac{f_{sJ}}{f_s} \right) \begin{pmatrix} f_{i1}^{-1} & & & \mathbf{0} \\ & \ddots & & \\ & & f_{ij}^{-1} & \\ \mathbf{0} & & & f_{iJ}^{-1} \end{pmatrix} \begin{pmatrix} \frac{f_{i1}}{f_i} - \frac{f_{s1}}{f_s} \\ \vdots \\ \frac{f_{ij}}{f_i} - \frac{f_{sj}}{f_s} \\ \vdots \\ \frac{f_{iJ}}{f_i} - \frac{f_{sJ}}{f_s} \end{pmatrix}$$

$$= \sum_{j=1}^J \frac{1}{f_j} \left(\frac{f_{ij}}{f_i} - \frac{f_{sj}}{f_s} \right)^2$$

Lema 1

La variabilidad total de la nube de perfiles fila, descrita como un promedio ponderado de las distancias de esos perfiles a su centro de gravedad $\sum_{i=1}^I f_i d^2(i, G)$, es una medida de asociación global entre las variables Y y X.

En efecto:

$$\begin{aligned} \sum_{i=1}^I f_i d^2(i, G) &= \\ &= \sum_{i=1}^n \sum_{j=1}^p \frac{f_i}{f_j} \left(\frac{f_{ij}}{f_i} - f_j \right)^2 \\ &= \sum_{i=1}^n \sum_{j=1}^p \frac{1}{f_i f_j} (f_{ij} - f_i f_j)^2 \\ &= \frac{1}{N} \sum_{i=1}^n \sum_{j=1}^p \frac{(k_{ij} - e_{ij})^2}{e_{ij}} \\ &= \frac{1}{n} \chi^2 \end{aligned}$$

A continuación se señalan algunos resultados del modelo teórico del ACB.

a) La información que se genera al caracterizar n individuos de acuerdo a dos variables categóricas Y y X, puede describirse, a efectos operativos sobre matrices disyuntivas completas, cuyo producto da lugar a tablas de contingencias bidimensionales de la forma $F_{yx} = \frac{1}{n} Y'X$.

b) Una vez que el test estadístico chi-cuadrado (χ^2) indica significación de la asociación entre las dos variables Y y X, el siguiente paso en el análisis requiere necesariamente su explicación. A estos efectos, el Análisis de Correspondencias Binarias (Benzecri, 1973) surge como un potente procedimiento que permite explicar las direcciones principales de la asociación, entre las variables bajo estudio.

Lema 2

La traza de la matriz que define las principales direcciones de variabilidad del espacio de los perfiles fila (columna) coincide con la variabilidad total de la nube de perfiles fila (columna):

$$\text{traza} \left((D_Y^{-1/2} F_{YX} D_X^{-1/2}) (D_Y^{-1/2} F_{XY} D_Y^{-1/2}) \right) - 1 = \frac{\chi^2}{n}$$

Lema 3

La variabilidad total del espacio de los perfiles fila (columna), o medida de asociación $\frac{\chi^2}{n}$, se puede desagregar en porciones de asociación captada por cada uno de los autovalores descritos por las principales direcciones de variabilidad de la nube de perfiles fila (columna):

$$\frac{\chi^2}{n} = \sum_{\alpha=2}^{\min(I-1, J-1)} \lambda_{\alpha}$$

Es decir, el procedimiento que utiliza el ACB a los efectos de explicar la asociación entre las variables Y y X, consiste en descomponer la variabilidad total observada como un agregado de diferentes fuentes de asociación definidas básicamente por las direcciones de los subespacios que producen el mejor ajuste a la nube de perfiles fila (columna), Ramírez y Vásquez (op cit).

2.4.2 Tablas de Contingencias Múltiples (TCM)

Dada la estructura de los datos descritas en las tablas 2.1 y 2.2, las cuales son una referencia muy importante para la problemática de que trata esta investigación, se definen dos tipos de Tablas de Contingencias Múltiples (TCM) de la siguiente manera.

2.4.2.1 TCM producto del cruce de dos conjuntos de variables categóricas Y y X

Sean:

$Y = \{Y_1, \dots, Y_i, \dots, Y_I\}$ un conjunto de I variables categóricas, cada una con K_i categorías;
 $i=1, \dots, I$, siendo $\sum_{i=1}^I K_i$ el número total de categorías en Y .

$X = \{X_1, \dots, X_j, \dots, X_J\}$ un conjunto de J variables categóricas, cada una con L_j categorías;
 $j=1, \dots, J$, siendo $\sum_{j=1}^J L_j$ es el número total de categorías en X .

Los datos obtenidos por la medición de los dos conjuntos de variables Y e X son arreglados en dos tablas de datos que yuxtaponen J e I matrices disyuntivas completas, respectivamente; el cruce de estas dos tablas genera una TCM como sigue:

Tabla 2.3: TCM DEL TIPO F_{YX}

$$\begin{pmatrix} F_{Y_1 X_1} & \cdots & F_{Y_1 X_j} & \cdots & F_{Y_1 X_J} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ F_{Y_i X_1} & & F_{Y_i X_j} & & F_{Y_i X_J} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ F_{Y_I X_1} & \cdots & F_{Y_I X_j} & \cdots & F_{Y_I X_J} \end{pmatrix}$$

El arreglo obtenido es de orden $\left(\sum_i K_i \right) \times \left(\sum_j L_j \right)$, su gran total es igual a $(I \times J) \times n$, y está conformado por $I \times J$ tablas de contingencias bidimensionales de la forma $F_{Y_i X_j}$, con $i=1, \dots, I$; $j=1, \dots, J$. Este arreglo es el análogo de la tabla teórica concatenada o compuesta objeto de análisis en el ACC, la cual es redefinida en esta investigación, de ahora en adelante, como TCM y **denotada por F_{YX}** .

2.4.2.2 TCM Producto del cruce de un conjunto de variables X consigo mismo

Sea $X = \{X_1, \dots, X_j, \dots, X_J\}$ un conjunto conformados por J variables categóricas, cada una de las variables con L_j categorías para $j=1, \dots, J$; donde $\sum_{j=1}^J L_j$ es el número total de categoría en el conjunto X .

Los datos provenientes de la medición de este conjunto de variables X son organizados en una tabla que yuxtapone J matrices disyuntivas completas. El cruce de esta tabla consigo misma genera una TCM como sigue:

Tabla 2.4: TCM DEL TIPO F_{XX}

$$\begin{pmatrix} F_{X_1 X_1} & \cdots & F_{X_1 X_j} & \cdots & F_{X_1 X_J} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ F_{X_j X_1} & & F_{X_j X_j} & & F_{X_j X_J} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ F_{X_J X_1} & \cdots & F_{X_J X_j} & \cdots & F_{X_J X_J} \end{pmatrix}$$

El arreglo obtenido es orden $\left(\sum_i K_i \right) \times \left(\sum_j L_j \right)$, su gran total es igual a $(I \times I) \times n$, está conformado por $J \times J$ tablas de contingencias bidimensionales, de las cuales, las definidas por $F_{X_j X_j}$, con $j=1, \dots, J$ son matrices diagonales. Este arreglo coincide con la tabla de Burt definida para el Análisis de Correspondencias Múltiple (ACM), en forma análoga al caso anterior, es redefinida en esta investigación como TCM y es **denotada por F_{XX}** .

2.5 Definiciones planteadas en el STATIS-C para el tratamiento de datos cualitativos

2.5.1 Planos Factoriales de la Etapa de la Interestructura

Como en general en todas las técnicas factoriales, las representaciones gráficas se realizan sobre planos factoriales, cuyos ejes directores quedan definidos por los autovectores asociados a los autovalores de la matriz que se diagonaliza. En la metodología propuesta, los planos factoriales, derivados de la diagonalización de la matriz S , sobre los cuales se representan los objetos representativos W_t de las T tablas de datos cualitativos, se denominan **Planos Interestructura**.

2.5.2 Planos Factoriales de la Etapa de la Intraestructura

Los planos factoriales, derivados de la diagonalización del Compromiso W , sobre los cuales se representan las posiciones relativas compromiso (promedio) de los individuos, y de manera suplementaria se proyecta la posición relativa de cada uno de los individuos descrita en las configuraciones definidas por T objetos representativos de las tablas de datos, se denominan **Planos Compromiso**.

CAPÍTULO III

UNA PROPUESTA PARA LA ADAPTACIÓN DEL STATIS AL ANÁLISIS DE DATOS CATEGÓRICOS

En este capítulo se presentan los aspectos teóricos principales subyacentes en el Análisis de Correspondencias Compuesto (ACc) según Israëls (op. cit), conceptos necesarios para proceder a escribir la propuesta de adaptación del STATIS para el análisis de datos categóricos, metodología originalmente desarrollada por Lavit (op. cit) en el ámbito de datos cuantitativos. También, se introducen resultados de carácter teórico, que resultan fundamentales para justificar la adaptación de los desarrollos en consideración a la propuesta de investigación, en un enfoque que presta especial atención a la definición del producto escalar que se requiere para definir la distancia que compara los objetos representativos de estructuras de datos resultantes de caracterizar n individuos mediante T conjuntos de variables categóricas, o mediante un único conjunto medido en T oportunidades diferentes.

Finalmente, se formula la propuesta cuyos fundamentos se apoyan en la integración de las técnicas anteriormente señaladas, con el objeto de desarrollar dispositivos gráficos multivariantes para visualizar las semejanzas y/o diferencias entre los objetos comparados.

3.1 FUNDAMENTOS DE LA PROPUESTA

3.1.1 Enfoque ACB sobre las Tablas de Contingencias Múltiples (TCM)

El enfoque de ACB descrito por Israëls (op cit) para el ACc, es aplicado sobre las TCM F_{YX} y F_{XX} descritas en el capítulo II, cuyos resultados son fundamentales para proponer la distancia entre las TCM subyacentes en la definición del Objeto Representativo de los estudios categóricos planteado para la adaptación del STATIS.

3.1.1.1 Enfoque ACB sobre la TCM F_{YX}

La TCM F_{YX} conformada por $I \times J$ tablas de contingencias bidimensionales $F_{Y_i X_j}$ es de la forma:

$$F_{YX} = \begin{pmatrix} \ddots & & \ddots \\ & F_{Y_i X_j} & \\ \ddots & & \ddots \end{pmatrix}$$

donde:

Y_i y X_j son la i -ésima y j -ésima variables de los conjuntos Y e X , respectivamente. K_i es el número de categorías de la variable Y_i ; L_j es el número de categorías de la variable X_j , con $i=1, \dots, I$; $j=1, \dots, J$.

La tabla de contingencias expresada en términos de frecuencias relativas respecto del total es tal de la forma:

$$F_{Y_i X_j} = \begin{pmatrix} \ddots & & \ddots \\ & f_{k_i l_j} = \frac{n_{k_i l_j}}{n} & \\ \ddots & & \ddots \end{pmatrix}$$

donde los vectores de información que se utilizan en el ACB generalizado y aplicado sobre una TCM F_{YX} , son los siguientes:

Vector columna de totales marginales por filas:

$$\mathbf{r}_Y = \left(\dots \frac{n_{k_{i1}}}{Jn} \dots \dots \frac{n_{k_{iL}}}{Jn} \dots \dots \frac{n_{k_{iJ}}}{Jn} \dots \right)' = (r'_{Y_1} \dots r'_{Y_i} \dots r'_{Y_I})$$

Vector Fila de totales marginales por columnas

$$\mathbf{c}'_x = \left(\dots \frac{n_{\cdot l_1}}{In} \dots \dots \frac{n_{\cdot l_j}}{In} \dots \dots \frac{n_{\cdot l_J}}{In} \dots \right) = (\mathbf{c}'_{x_1} \dots \mathbf{c}'_{x_j} \dots \mathbf{c}'_{x_J})$$

Los vectores \mathbf{r}_y y \mathbf{c}_x contienen respectivamente, los elementos de las matrices diagonales de pesos para las filas y para las columnas, de la tabla de contingencias múltiple F_{YX} , que se describen a continuación:

Matriz diagonal de pesos fila (inversa de la métrica en el espacio de las variables X 's)

$$D_y = \begin{pmatrix} D_{y_1} & & & \mathbf{0} \\ & \ddots & & \\ & & D_{y_i} & \\ & & & \ddots \\ \mathbf{0} & & & & D_{y_I} \end{pmatrix}$$

Para $i=1, \dots, I$

$$D_{Y_i} = \begin{pmatrix} f_{1i\cdot} & & & \mathbf{0} \\ & \ddots & & \\ & & f_{ki\cdot} & \\ & & & \ddots \\ \mathbf{0} & & & & f_{Ki\cdot} \end{pmatrix}$$

Matriz diagonal de pesos columna (inversa de la métrica en el espacio de las variables Y 's)

$$D_X = \begin{pmatrix} D_{x_1} & & & \mathbf{0} \\ & \ddots & & \\ & & D_{x_j} & \\ & & & \ddots \\ \mathbf{0} & & & & D_{x_J} \end{pmatrix}$$

Para $j=1, \dots, J$

$$D_{X_j} = \begin{pmatrix} f_{\cdot l(j)} & & & \mathbf{0} \\ & \ddots & & \\ & & f_{\cdot l_j} & \\ & & & \ddots \\ \mathbf{0} & & & & f_{\cdot l_j} \end{pmatrix}$$

Descripción del k_i -ésimo perfil fila ponderado y modificado de F_{YX}

$$R_{k_i}' = \left(\dots \frac{f_{k_i l_1}}{\sqrt{f_{k_i \cdot}} \sqrt{f_{\cdot l_1}}} \dots \frac{f_{k_i l_j}}{\sqrt{f_{k_i \cdot}} \sqrt{f_{\cdot l_j}}} \dots \frac{f_{k_i l_J}}{\sqrt{f_{k_i \cdot}} \sqrt{f_{\cdot l_J}}} \dots \right)$$

La matriz de perfiles fila ponderados y modificados es:

$$D_Y^{-1/2} F_{YX} D_X^{-1/2} = \begin{pmatrix} \ddots & & \\ & D_Y^{-1/2} F_{Y_i X_j} D_X^{-1/2} & \\ & & \ddots \end{pmatrix}$$

A partir de la Tabla de Contingencias Múltiple, F_{YX} , se define un arreglo con una estructura similar a la matriz que se diagonaliza en el ACB.

$$\left(\left(D_Y^{-1/2} F_{YX} D_X^{-1/2} \right) \left(D_Y^{-1/2} F_{YX} D_X^{-1/2} \right) \right) = \begin{pmatrix} \ddots & & \\ & \sum_{j=1}^J \left(\left(D_Y^{-1/2} F_{Y_i X_j} D_X^{-1/2} \right) \left(D_{X_j}^{-1/2} F_{X_j Y_i} D_{Y_i}^{-1/2} \right) \right) & \\ & & \ddots \end{pmatrix}$$

Esta matriz tiene como elementos en su diagonal principal, submatrices definidas como un agregado de J arreglos que describen las matrices que se diagonalizan en un ACB aplicado para explicar la asociación entre la i -ésima variable Y_i en \mathbf{Y} , con cada una de las J variables X_j en \mathbf{X} . En consecuencia, al tomar traza de la matriz antes descrita, se tiene:

$$\begin{aligned}
 \text{tr} \left(\left(D_Y^{-1/2} F_{YX} D_X^{-1/2} \right) \left(D_Y^{-1/2} F_{YX} D_X^{-1/2} \right)' \right) &= \sum_{i=1}^I \sum_{j=1}^J \left(\text{tr} \left(D_{Y_i}^{-1/2} F_{Y_i X_j} D_{X_j}^{-1/2} \right) \left(D_{X_j}^{-1/2} F_{X_j Y_i} D_{Y_i}^{-1/2} \right) \right) \\
 &= \sum_{i=1}^I \sum_{j=1}^J \left(\frac{\chi_{Y_i X_j}^2}{n} + 1 \right) \\
 &= \frac{\sum_{i=1}^I \sum_{j=1}^J \left(\frac{\chi_{Y_i X_j}^2}{n} \right) + (I \times J)}{(I \times J)} \\
 &= \frac{\sum_{i=1}^I \sum_{j=1}^J \chi_{Y_i X_j}^2}{(I \times J)n} + 1
 \end{aligned}$$

Es decir, la expresión que define la traza de la matriz en consideración, coincide con el promedio de los estadísticos chi-cuadrado que miden las interasociaciones entre las variables de los conjuntos Y 's y X 's, dos a dos, más la unidad.

3.1.1.2 Enfoque ACB sobre la TCM F_{XX}

La TCM F_{XX} conformada por $J \times J$ tablas submatrices descritas por tablas de contingencias bidimensionales $F_{X_j X_{j'}}$, $j, j' = 1, \dots, J$ de la forma:

$$F_{XX} = \begin{pmatrix} \ddots & & \ddots \\ & F_{X_j X_j} & \\ \ddots & & \ddots \end{pmatrix}$$

donde la submatriz correspondiente a la tabla de contingencias bidimensional que cruza las variables X_j y $X_{j'}$ expresada en términos relativa es de la forma:

$$F_{X_j X_{j'}} = \begin{pmatrix} \ddots & & \ddots \\ & f_{l_j l_{j'}} = \frac{n_{l_j l_{j'}}}{n} & \\ \ddots & & \ddots \end{pmatrix}$$

donde los vectores de información que se utilizan en el ACB generalizado aplicado sobre una TCM F_{xx} , son los siguientes:

Vector columna de totales marginales por fila:

$$\mathbf{r}_x = \left(\dots \frac{n_{l_j \cdot}}{Jn} \dots \vdots \dots \frac{n_{l_{j'} \cdot}}{Jn} \dots \vdots \dots \frac{n_{l_{j''} \cdot}}{Jn} \vdots \right)' = \left(\mathbf{r}'_{x_1} \dots \mathbf{r}'_{x_j} \dots \mathbf{r}'_{x_{j'}} \right)$$

Vector fila de totales marginales por columna:

$$\mathbf{c}_x' = \left(\dots \frac{n_{\cdot l_j}}{Jn} \dots \vdots \dots \frac{n_{\cdot l_{j'}}}{Jn} \dots \vdots \dots \frac{n_{\cdot l_{j''}}}{Jn} \vdots \right) = \left(\mathbf{c}'_{x_1} \dots \mathbf{c}'_{x_j} \dots \mathbf{c}'_{x_{j'}} \right)$$

Los vectores \mathbf{r}_x y \mathbf{c}_x son iguales, y contienen los elementos de la matriz diagonal de pesos para las filas (columnas), de la tabla de contingencias múltiple F_{xx} , que se describe a continuación:

Matriz diagonal de pesos fila (-inversa de la métrica en el espacio de las variables X'_s -)

Para $j=1, \dots, J$

$$D_x = \begin{pmatrix} D_{x_1} & & & \mathbf{0} \\ & \ddots & & \\ & & D_{x_j} & \\ \mathbf{0} & & & D_{x_J} \end{pmatrix}$$

Donde:

$$D_{x_j} = \begin{pmatrix} f_{\cdot l(j)} & & & \mathbf{0} \\ & \ddots & & \\ & & f_{\cdot l_j} & \\ \mathbf{0} & & & f_{\cdot l_J} \end{pmatrix}$$

Descripción del k-ésimo perfiles fila de F_{xx}

$$R_{l_i}' = \left(\dots \frac{f_{l_i l_i}}{\sqrt{f_{l_i \cdot}} \sqrt{f_{\cdot l_i}}} \vdots \dots \frac{f_{l_j l_j}}{\sqrt{f_{l_j \cdot}} \sqrt{f_{\cdot l_j}}} \vdots \dots \frac{f_{l_J l_J}}{\sqrt{f_{l_J \cdot}} \sqrt{f_{\cdot l_J}}} \dots \right)$$

Donde la matriz de perfiles filas ponderados y modificados es:

$$D_x^{-1/2} F_{xx} D_x^{-1/2} = \begin{pmatrix} \ddots & & \\ & D_x^{-1/2} F_{x_j x_j} D_x^{-1/2} & \\ & & \ddots \end{pmatrix}$$

A partir de la Tabla de Contingencias Múltiple, F_{xx} , se define un arreglo con una estructura similar a la matriz que se diagonaliza en el ACB.

$$\left(\left(D_x^{-1/2} F_{xx} D_x^{-1/2} \right) \left(D_x^{-1/2} F_{xx} D_x^{-1/2} \right) \right) = \begin{pmatrix} \ddots & & \\ & \sum_{j=1}^J \left(\left(D_{x_j}^{-1/2} F_{x_j x_j} D_{x_j}^{-1/2} \right) \left(D_{x_j}^{-1/2} F_{x_j y_j} D_{y_j}^{-1/2} \right) \right) & \\ & & \ddots \end{pmatrix}$$

Análogamente al caso anterior, esta matriz tiene como elementos en su diagonal principal, submatrices definidas por un agregado de J arreglos que describen las matrices que se diagonalizan en un ACB aplicado para explicar la asociación entre la j -ésima variable X_j en \mathbf{X} , con cada una de las J variables X_j 'también del mismo conjunto \mathbf{X} . En consecuencia, al tomar traza de la matriz antes descrita, se tiene que:

$$\begin{aligned} \text{tr} \left(\left(D_X^{-1/2} F_{XX} D_X^{-1/2} \right) \left(D_X^{-1/2} F_{XX} D_X^{-1/2} \right)' \right) &= \sum_{j=1}^J \sum_{j'=1}^J \left(\text{tr} \sum_{j=1}^J \left(\left(D_{X_j}^{-1/2} F_{X_j X_j} D_{X_j}^{-1/2} \right) \left(D_{X_j}^{-1/2} F_{X_j X_j} D_{X_j}^{-1/2} \right)' \right) \right) \\ &= \sum_{j=1}^J \sum_{j=1}^J \left(\frac{\chi_{X_j X_j}^2}{n} \right) + (J \times J) \\ &= \frac{\sum_{j=1}^J \sum_{j=1}^J \chi_{X_j X_j}^2}{(J \times J)n} + 1 \end{aligned}$$

Es importante señalar que cuando $j'=j=1, \dots, J$

$$\text{tr} \left(\left(D_{X_j}^{-1/2} F_{X_j X_j} D_{X_j}^{-1/2} \right) \left(D_{X_j}^{-1/2} F_{X_j X_j} D_{X_j}^{-1/2} \right)' \right) = L_j$$

que es el número de categorías de la j -ésima variable X .

En el siguiente apartado se presentan, a manera de lemas, los resultados teóricos que sirven de fundamento para la adaptación propuesta del STATIS.

Lema 4

Una medida de asociación en una TCM F_{YX} con la estructura de una tabla concatenada, propia de un ACC, es un promedio de la asociación de las tablas de contingencias bidimensionales que la conforman más la unidad:

$$\chi_{F_{YX}}^2 = \frac{\sum_{i=1}^I \sum_{j=1}^J \chi_{Y_i X_j}^2}{(I \times J)n} + 1$$

Lema 5

La medida de asociación de una Tabla de Contingencias Múltiple con la estructura de una Tabla de Burt $D_{XX}^{-1} F_{XX}$, definida por un conjunto $X = \{X_1, \dots, X_J, \dots, X_J\}$ de variables categóricas, es un promedio de la medida de asociación global tipo chi-cuadrado de las tablas de contingencias bidimensionales que la conforman más la unidad:

$$\chi_{F_{XX}}^2 = \frac{\sum_{j=1}^J \sum_{j=1}^J \chi_{(j,j)}^2}{(J \times J)n} + 1$$

Corolario 1

La medida de asociación en una tabla de contingencias formada por el cruce de una variable consigo misma, es el número de categorías de dicha variable menos la unidad.

En este orden de ideas, se relacionan algunas definiciones y lemas que constituyen la base para definición de la **distancia de Hilbert-Schmidt** entre objetos representativos de tablas de datos categóricos, siguiendo para ello la presentación de Ramírez y Vásquez (2003) explicada en la sección 2.2 del capítulo II, y los resultados presentados en este apartado.

La integración de los resultados anteriores constituye la base para definir la distancia de Hilbert-Schmidt resultado clave del aporte de esta investigación, lo que se fundamenta con los planteamientos presentados a continuación:

- a) El Análisis de Correspondencias compuesto de Keller y Wansbeek es un Análisis de Correspondencias Binarias generalizado aplicado sobre la Tabla de Contingencias

Múltiple (TCM), resultante de cruzar la información obtenida de la medición de los conjuntos de variables categóricas $Y = \{Y_1, \dots, Y_i, \dots, Y_I\}$, y $X = \{X_1, \dots, X_j, \dots, X_J\}$, denominada también, tabla concatenada.

- b) La adaptación del STATIS al tratamiento de datos categóricos requiere como tarea central, la utilización del producto escalar de Hilbert-Schmidt para definir una distancia entre las estructuras de los datos contenidas en los objetos representativos de los estudios identificados por T conjuntos de variables.

3.2 Estructura de Datos

La estructura de la Tabla 4.1, puede utilizarse para ilustrar el tipo de información que sirve de referencia para el desarrollo del problema de investigación planteado, ya sea que se trate de un diseño transversal o de uno longitudinal:

Tabla 4.1: T matrices de datos descriptivas de T dimensiones

	T tablas de datos categóricos				
	Dimensión 1		Dimensión t		Dimensión T
	$\overbrace{X_1 \dots X_j \dots X_J}$...	$\overbrace{Y_1 \dots Y_i \dots Y_I}$...	$\overbrace{Z_1 \dots Z_k \dots Z_K}$
	Tabla T_1	...	Tabla T_t	...	Tabla T_T
n individuos {					

3.3 Adaptación de la propuesta al caso de dos estudios E_X y E_Y

La adaptación de la metodología STATIS al tratamiento de datos categóricos, sin pérdida de generalidad, será desarrollada considerando que n individuos son caracterizados por dos conjuntos de variables categóricas $Y = \{Y_1, \dots, Y_i, \dots, Y_I\}$ y $X = \{X_1, \dots, X_j, \dots, X_J\}$.

Cada conjunto de variables identifica un estudio conforme a lo planteado en Djauhari (1998), en su propuesta para seleccionar variables categóricas basada en el coeficiente de Escoufier.

Estudio:

En general, la terna de arreglos que identifica el **Estudio** correspondiente al conjunto de variables categóricas $\mathbf{X} = \{X_1, \dots, X_j, \dots, X_J\}$, queda definido por:

$$\mathbf{E}_x = (\mathbf{X}, \mathbf{M}_x, \mathbf{D}_x)$$

donde:

X: Es una matriz disyuntiva completa $\mathbf{X} = [X_1 \dots X_j \dots X_J]$ de orden $(n \times L_t)$ conformada por J matrices disyuntivas completas, siendo $L_t = L_1 + \dots + L_J$ el número total de variables indicatrices generadas por las categorías de las variables X 's, obtenida al caracterizar n individuos. Siendo en particular L_{X_j} el número de categorías de la variable X_j .

D_x: Es una matriz diagonal de pesos que se asignan a los individuos de la forma:

$$D_x = \frac{1}{(J \times J)_n} I_{(n)}$$

donde $I_{(n)}$ es la matriz identidad de orden $n \times n$, y $(J \times J)_n$ es el gran total de la TCM que se produce con los datos de la t -ésima-dimensión.

M_x: Es la métrica que se utiliza para definir el producto escalar que sirve de base para obtener las distancias entre individuos, es una matriz de orden $L_t \times L_t$, simétrica, y definida positiva, en la forma:

$$\mathbf{M}_x = [\text{diag}(\mathbf{X}' \mathbf{D}_x \mathbf{X})]^{-1} = \left[\text{diag} \left(\frac{1}{(J \times J)_n} (\mathbf{X}' \mathbf{X}) \right) \right]^{-1}$$

$$= \mathbf{D}_X^{-1} = \begin{pmatrix} \mathbf{D}_{X_1}^{-1} & & & & 0 \\ & \ddots & & & \\ & & \mathbf{D}_{X_j}^{-1} & & \\ 0 & & & \ddots & \\ & & & & \mathbf{D}_{X_J}^{-1} \end{pmatrix}$$

$\mathbf{X}'\mathbf{X}$ es una TCM de la forma:

$$\begin{pmatrix} X_1'X_1 & \cdots & X_1'X_j & \cdots & X_1'X_J \\ & \ddots & \vdots & \ddots & \\ X_j'X_1 & & X_j'X_j & & X_j'X_J \\ & \ddots & \vdots & \ddots & \\ X_J'X_1 & \cdots & X_J'X_j & \cdots & X_J'X_J \end{pmatrix}$$

cuyo gran total es $(J \times J)n$, lo que implica que por fila y por columna cada individuo es contabilizado J veces.

El estudio correspondiente al conjunto de variables \mathbf{Y} , $\mathbf{E}_y = (\mathbf{Y}, \mathbf{M}_y, \mathbf{D}_y)$, se describe en forma análoga, y sus elementos \mathbf{Y} , \mathbf{M}_y y \mathbf{D}_y tienen la misma estructura que en el estudio \mathbf{E}_x .

3.4 Modelo Teórico de la Adaptación del STATIS en el Análisis de Datos Categóricos

3.4.1 Adaptación de la Etapa de la Interestructura

La adaptación de esta etapa requiere la definición de los siguientes elementos:

- Un objeto representativo W_i para cada uno de los T estudios.
- El producto escalar que provee la distancia que surge como criterio de comparación entre estudios.
- La matriz S de productos escalares entre objetos W_i , la cual se diagonaliza para obtener la representación gráfica de los estudios.

3.4.1.1 Definición del Objeto Representativo del t-ésimo estudio E_t .

En el STATIS el objeto representativo del t-ésimo estudio es definido en términos de productos escalares entre individuos, y viene dado por $W_t = X_t M_t X_t'$, específicamente, el objeto representativo para el estudio $E_x = (X, M_x, D_x)$ es una matriz W_x de orden $n \times n$ de la forma:

$$W_x = X M_x X' = [X_1 \cdots X_J \cdots X_J] \begin{bmatrix} D_{X_1}^{-1} & & & \theta \\ & \ddots & & \\ & & D_{X_j}^{-1} & \\ \theta & & & D_{X_J}^{-1} \end{bmatrix} \begin{bmatrix} X_1' \\ \vdots \\ X_j' \\ \vdots \\ X_J' \end{bmatrix} = \sum_{j=1}^J X_j D_{X_j}^{-1} X_j'$$

En forma análoga el objeto representativo para el estudio $E_y = (X, M_y, D_y)$ es la matriz W_y de orden $n \times n$:

$$W_y = Y M_y Y' = [Y_1 \cdots Y_I \cdots Y_I] \begin{bmatrix} D_{Y_1}^{-1} & & \\ & \ddots & \\ & & D_{Y_i}^{-1} \\ & & & \ddots \\ & & & & D_{Y_I}^{-1} \end{bmatrix} \begin{bmatrix} Y_1' \\ \vdots \\ Y_i' \\ \vdots \\ Y_I' \end{bmatrix} = \sum_{i=1}^I Y_i D_{Y_i}^{-1} Y_i'$$

Cabe resaltar que en la definición de los objetos W_t , subyace una descomposición de la información en sumandos proporcionados por cada una de las variables que definen el estudio.

3.4.1.2 Definición del Producto Escalar entre Objetos

El estudio de la interestructura en el STATIS queda básicamente descrito por el producto escalar de Hilbert-Schmidt entre los objetos representativos W_t .

El producto escalar HS entre dos objetos W_x y W_y definidos para las tablas de datos categóricos X e Y , queda definido en la forma:

$$\begin{aligned}
\langle W_x | W_y \rangle_{HS} &= \text{traza} (D_x W_x D_y W_y) \\
&= \text{tr}(D_x W_x D_y W_y) \\
&= \text{tr} \left[\frac{1}{(J \times J)_n} I \left(X D_x^{-1} X' \right) \frac{1}{(I \times I)_n} I \left(Y D_y^{-1} Y' \right) \right] \\
&= \text{tr} \left[\left(D_Y^{-1/2} \frac{1}{(I \times J)_n} F_{YX} D_X^{-1/2} \right) \left(D_X^{-1/2} \frac{1}{(I \times J)_n} F_{YX}' D_Y^{-1/2} \right) \right]
\end{aligned}$$

donde la matriz $\left(\left(\frac{1}{(I \times J)_n} I \right) (Y' X) \right)$ es una TCM F_{YX} (tabla concatenada del ACc), y por el lema 5 el producto escalar de HS entre dos objetos W_x y W_y es:

$$\langle W_x | W_y \rangle_{HS} = \text{tr} \left[\left(D_Y^{-1/2} F_{YX} D_X^{-1/2} \right) \left(D_X^{-1/2} F_{YX}' D_Y^{-1/2} \right) \right]$$

Por el lema 4, esta expresión es:

$$\langle W_x | W_y \rangle_{HS} = \frac{\sum_{i=1}^I \sum_{j=1}^J \chi_{(ij)}^2}{(I \times J)} + 1$$

Este resultado permite concluir que el producto escalar HS entre las estructuras de interdistancias de individuos definidas respectivamente por los objetos W_x y W_y , será tanto más grande cuanto mayor sea la interasociación global entre los dos conjuntos de variables, y viceversa para el caso de que las asociaciones sean débiles.

Nota 1:

La metodología STATIS define una única matriz de pesos D para todos los T estudios a ser analizados. En la adaptación de la fase del estudio de la Interestructura a datos categóricos, cada estudio posee su matriz de pesos D_t , lo cual podría estar en discordancia con la teoría. No

obstante, si se observa que las métricas $M_x = D_x^{-1} = \left[\text{diag} \left(X' \frac{1}{(J \times J)n} X \right) \right]^{-1}$ y

$M_y = D_y^{-1} = \left[\text{diag} \left(Y' \frac{1}{(J \times J)n} Y \right) \right]^{-1}$ poseen el término $(J \times J)n$ y $(I \times I)$ en el numerador, y la

matriz de pesos D_x y D_y en el denominador, entonces es perfectamente válido redefinir a la matriz de pesos, en general como $D = I_{(n)} = D$, común a los T estudios.

Análogamente a la metodología STATIS original, el producto escalar HS de un objeto consigo mismo $\langle W_x | W_x \rangle_{HS}$, donde la matriz de pesos de los individuos D se puede especificar como $D = I$, queda definido en la forma:

$$\begin{aligned} \langle W_x | W_x \rangle_{HS} &= \text{traza} (D_x W_x D_x W_x) \\ &= \text{tr} \left[D_x \left(X D_x^{-1} X' \right) D_x \left(X D_x^{-1} X' \right) \right] \\ &= \text{tr} \left(\left(D_x^{-1/2} \left(\left(\frac{1}{(J \times J)n} \right) (X' X) \right) D_x^{-1/2} \right) \left(D_x^{-1/2} \left(\left(\frac{1}{(J \times J)n} \right) (X' X) \right) D_x^{-1/2} \right) \right) \end{aligned}$$

donde la matriz $\left(\left(\frac{1}{(J \times J)n} \right) I \right) (X' X)$ es una TCM F_{XX} (tabla de Burt), y por el lema 5;

$$\langle W_x | W_x \rangle_{HS} = \frac{\sum_{j=1}^J \sum_{j=1}^J \chi_{(j,j)}^2}{(J \times J)n} + 1$$

Este resultado permite concluir que el producto escalar de un objeto consigo mismo, será tanto mayor cuanto más fuerte sea la interasociación global entre las variables categóricas que lo identifican, y viceversa para el caso en que las asociaciones sean débiles.

3.4.1.3 Definición de la Distancia entre Objetos

La distancia entre dos objetos W_x y W_y inducida el producto escalar HS queda determinada por la siguiente expresión:

$$\begin{aligned}
 d_{HS}^2(W_X, W_Y) &= \|W_X - W_Y\|_{HS}^2 = \|W_X\|_{HS}^2 + \|W_Y\|_{HS}^2 - 2\langle W_X | W_Y \rangle_{HS} \\
 &= \left(\frac{\sum_{j'=1}^J \sum_{j=1}^J \chi_{X_{j'} X_j}^2}{(J \times J)n} + 1 \right) + \left(\frac{\sum_{i=1}^I \sum_{j=1}^J \chi_{Y_i Y_j}^2}{(I \times I)n} + 1 \right) - 2 \left(\frac{\sum_{i=1}^I \sum_{j=1}^J \chi_{Y_i X_j}^2}{(I \times J)n} + 1 \right) \\
 &= \chi_{F_{XX}}^2 + \chi_{F_{YY}}^2 - 2\chi_{F_{YX}}^2
 \end{aligned}$$

Resultado 1

La distancia entre las estructuras de interdistancias entre individuos definidas respectivamente por los objetos W_x y W_y , será tanto más pequeña cuanto mayor sea la interasociación global entre los dos conjuntos de variables, y viceversa para el caso de que las asociaciones sean débiles.

El elemento en la posición (t, t') de la matriz S que contiene los productos escalares HS entre los objetos representativos de los T estudios, queda definido

en función de la interasociación global que existe entre los dos conjuntos de variables categóricas que definen a los estudios E_t y $E_{t'}$.

3.4.1.3.1 Propiedades algebraicas del producto escalar HS entre objetos categóricos W_t

Sea el conjunto $M_n(\mathbb{R})$ de matrices cuadradas de orden $n \times n$, y sea además la función definida $\langle \cdot | \cdot \rangle_{HS}: M_n(\mathbb{R}) \times M_n(\mathbb{R}) \rightarrow \mathbb{R}$ que asigna a cada par ordenado (W_x, W_y) de $M_n(\mathbb{R}) \times M_n(\mathbb{R})$ un número real $\langle W_x | W_y \rangle_{HS} = \text{tr}(DW_x DW_y)$ con $D=I$, entonces la función cumple las siguientes propiedades:

a) Distributividad

$$\begin{aligned} \langle W_X + W_Y, W_Z \rangle &= \langle W_X | W_Z \rangle + \langle W_Y | W_Z \rangle \text{ por propiedad del algebra matricial,} \\ &= \text{tr}(DW_X DW_Z) + \text{tr}(DW_Y DW_Z) \text{ por la definición del producto escalar de Hilbert-Schmidt,} \end{aligned}$$

$$= \chi_{F_{XZ}}^2 + \chi_{F_{YZ}}^2 + 2 \text{ por el lema 4:}$$

$$b) \langle \alpha W_X | W_Y \rangle = \text{tr}[D(\alpha W_X) DW_Y] = \alpha \text{tr}(DW_X DW_Y) = \alpha (\chi_{F_{YX}}^2 + 1) \text{ con } \alpha \in \mathbb{R}$$

c) Conmutatividad

$$\langle W_X | W_Y \rangle = \langle W_Y | W_X \rangle = \text{tr}(DW_Y DW_X) = \text{tr}(DW_X DW_Y) = \chi_{F_{YX}}^2 + 1 = \chi_{F_{XY}}^2 + 1$$

d) Positividad

$$\langle W_X | W_X \rangle_{HS} = \text{tr}(DW_X DW_X) = (\chi_{F_{XX}}^2 + 1) > 0 \text{ Para todo } W_X \in M_n \mathbb{R} \text{ con } W_X \neq 0$$

Resultado 2

Dados los resultados anteriormente descritos, el elemento genérico de la matriz $S = \langle W_t | W_{t'} \rangle_{HS}$ de productos escalares entre los objetos, queda descrito como función de una medida global de la interasociación entre las variables categóricas que identifican los estudios E_t y $E_{t'}$.

$$S = \begin{pmatrix} \ddots & & \\ & \langle W_t | W_{t'} \rangle_{HS} & \\ & & \ddots \end{pmatrix} = \begin{pmatrix} \ddots & & \\ & \chi^2_{Y_i X_j} + 1 & \\ & & \ddots \end{pmatrix}$$

A partir de los productos escalares HS entre los objetos que identifican T estudios diferentes descritos por variables categóricas, es posible construir la matriz S que se diagonaliza para representar gráficamente las diferencias o similitudes entre los objetos que identifican a esos estudios. Por lo tanto, el elemento genérico de la matriz $S = (\langle W_X, W_Y \rangle)$ queda descrito como función de una medida global de la interasociación entre las variables categóricas que identifican los estudios E_X y E_Y . Finalmente, la medida traza (S) queda definida como un agregado de las interasociaciones entre las variables que identifican cada una de las T tablas.

En caso de utilizar los objetos normados, la matriz \tilde{S} viene dada por:

$$\tilde{S} = \begin{pmatrix} \ddots & & \\ & \frac{\langle W_t | W_{t'} \rangle_{HS}}{\|W_t\|_{HS} \|W_{t'}\|_{HS}} & \\ & & \ddots \end{pmatrix} = \begin{pmatrix} \ddots & & \\ & \frac{\chi^2_{F_{tt'}} + 1}{\sqrt{\chi^2_{F_{tt}} + 1} \sqrt{\chi^2_{F_{t't'}} + 1}} & \\ & & \ddots \end{pmatrix} \text{ Para } t, t' = 1, \dots, T$$

3.4.1.3.2 Propiedades estadísticas del producto escalar HS entre objetos categóricos W_t

En el Capítulo II que contiene los fundamentos teóricos de la adaptación del STATIS, se obtuvo una medida sintética de la asociación global subyacente en una Tabla de Contingencias Múltiples descrita como un promedio de la asociación atribuible a cada uno de los pares de variables que definen las tablas de contingencias bidimensionales que conforman la TCM. Esta medida por construcción, puede descomponerse, permitiendo evaluar el aporte de cada uno de los pares de variables a la asociación global. Adicionalmente, ello hace posible determinar la contribución de cada uno de los estudios en la construcción del **subespacio interestructura**.

3.4.1.4.3 Índice de Contribución de las TCM

$$I_{cF_{it'}} = \frac{\chi_{F_{it'}}^2 + 1}{\sum_{t=1}^T \sum_{t'=1}^T (\chi_{E_{it'}}^2 + 1)}, \quad t, t' = 1, \dots, T$$

donde:

$\chi_{F_{it'}}^2 + 1$ es la medida de asociación global entre las variables categóricas que identifican los estudios E_t y $E_{t'}$ más la unidad.

$\sum_{t=1}^T \sum_{t'=1}^T (\chi_{F_{it'}}^2 + 1) = \sum_{t=1}^T \sum_{t'=1}^T (s_{tt'})$ es un agregado de las medidas de asociación global entre los T estudios considerados más la cantidad $(T \times T)$.

3.4.1.4.4 Índice de Contribución de las Tablas de Contingencias

$$I_{cF_{y_i x_j}} = \frac{\chi_{F_{y_i x_j}}^2}{\sum_{t=1}^T \sum_{t'=1}^T (\chi_{E_{it'}}^2)}, \quad i=1, \dots, I; j=1, \dots, J; t, t'=1, \dots, T$$

donde:

$\chi_{F_{y_i x_j}}^2$ Es la medida de asociación entre las variables Y_i y X_j pertenecientes a los conjuntos de variables categóricas Y 's y X 's que identifican los estudios E_Y y E_X .

$\sum_{t=1}^T \sum_{t'=1}^T \chi_{F_{it'}}^2 = \sum_{t=1}^T \sum_{t'=1}^T (s_{tt'} - 1)$ es un agregado de las medidas de asociación global entre los T estudios considerados.

3.4.1.5 Construcción de la imagen euclídea de los objetos W_t

En caso de que el investigador precise diferenciar los T estudios identificados por variables categóricas, de acuerdo a su importancia, se asignará un peso a cada uno, el cual depende del contexto teórico de la investigación propiamente dicha. En la presente adaptación del STATIS, se consideran que todas las tablas tienen igual importancia, por lo que la matriz de ponderación es:

$$\Delta = \begin{pmatrix} \pi_1 & & \theta \\ & \ddots & \\ \theta & & \pi_t \end{pmatrix} \quad \text{Con } \pi_t=1, \text{ para } t=1, \dots, T.$$

dado que la adaptación del STATIS a datos categóricos no es más que una ACP sobre la matriz $S\Delta$ o \tilde{S} , los dos primeros autovectores (ejes factoriales) asociados a sus respectivos autovalores, son utilizados como ejes para obtener la imagen euclídea plana aproximada de los T estudios. Sobre este primer plano factorial se realiza la gráfica de los puntos $B_1, \dots, B_t, \dots, B_T$ correspondientes a los T estudios representados por los objetos W_t , cuyas coordenadas de son:

$$\Psi_{ta} = \sqrt{\lambda_a} \gamma_{ta} \quad \text{para } t=1,2, \alpha=1, \dots, T$$

Ψ_{ta} es la coordenada de proyección de W_t sobre el eje γ^α , donde γ_{ta} es la t -ésima componente del α -ésimo autovector de ΔS asociado al α -ésimo autovalor λ_a .

Esta representación en el plano de los T estudios es tal que, la distancia entre dos puntos B_r y B_l es la mejor aproximación posible de la distancia HS entre los objetos representativos de las tablas r y l .

En este sentido, se propone un índice que mide la calidad con que la distancia euclídea entre dos objetos sobre el primer plano factorial aproxima a la distancia HS entre dos objetos en el espacio original.

A continuación se justifica la propuesta:

1.- En primer lugar, se considera la descomposición espectral de la matriz $S_{T \times T} = \gamma D_\lambda \gamma'$, donde $\gamma = (\gamma^1 \cdots \gamma^\alpha \cdots \gamma^T)$, es la matriz de autovectores de S y $\gamma^\alpha = (\gamma_{1\alpha} \cdots \gamma_{t\alpha} \cdots \gamma_{T\alpha})'$. Entonces, el elemento genérico de la matriz S en la posición (t,t') toma la forma:

$$\begin{aligned} S_{tt'} &= \langle W_t | W_{t'} \rangle_{HS} = \sum_{\alpha=1}^T \lambda_\alpha \gamma_{t\alpha} \gamma_{t'\alpha} \\ &= \sum_{\alpha=1}^T \left(\sqrt{\lambda_\alpha} \gamma_{t\alpha} \right) \left(\sqrt{\lambda_\alpha} \gamma_{t'\alpha} \right) \end{aligned}$$

2.- La proyección ortogonal de un objeto $W_{t'}$ sobre todos los ejes es de la forma:

$$\left(\sqrt{\lambda_1} \gamma_{1\alpha}, \cdots, \sqrt{\lambda_\alpha} \gamma_{t\alpha}, \cdots, \sqrt{\lambda_T} \gamma_{T\alpha} \right)$$

3.- La distancia euclídea entre las representaciones de los objetos W_t y $W_{t'}$ sobre todos los ejes factoriales es de la forma:

$$\begin{aligned} d_{e(W_t, W_{t'})}^2 &= \sum_{\alpha=1}^T \left(\sqrt{\lambda_\alpha} \gamma_{t\alpha} - \sqrt{\lambda_\alpha} \gamma_{t'\alpha} \right)^2 \\ &= \sum_{\alpha=1}^T \left(\sqrt{\lambda_\alpha} \gamma_{t\alpha} \right)^2 + \sum_{\alpha=1}^T \left(\sqrt{\lambda_\alpha} \gamma_{t'\alpha} \right)^2 - 2 \sum_{\alpha=1}^T \lambda_\alpha \gamma_{t\alpha} \gamma_{t'\alpha} \\ &= \langle W_t | W_t \rangle + \langle W_{t'} | W_{t'} \rangle - 2 \langle W_t | W_{t'} \rangle = d_{HS}(W_t, W_{t'}) \end{aligned}$$

En consecuencia de lo anterior, se plantea la construcción de un índice de calidad de la representación de las distancias entre dos objetos sobre el primer plano factorial, denotado por I_{cd} .

$$I_{Cd} = \frac{\sum_{\alpha=1}^2 \left(\sqrt{\lambda_{\alpha}} \gamma_{t\alpha} - \sqrt{\lambda_{\alpha}} \gamma_{t'\alpha} \right)^2}{d_{HS}(W_t, W_{t'})} \times 100\%$$

3.4.2 Construcción del Compromiso

La segunda etapa del STATIS tiene como objetivo resumir las estructuras de interdistancias obtenidas para los individuos en los T estudios, en una sola denominada compromiso W, que es definida como un promedio ponderado de los objetos W_t , o de los objetos normados, en la forma:

$$W = \sum_{t=1}^T \alpha_t W_t = \sum_{t=1}^T \left[\frac{1}{\sqrt{\lambda_1}} \left(\sum_{i=1}^T \pi_i \sqrt{S_{it}} \right) \pi_i \gamma_{ti} \right] W_t$$

Como en la adaptación del STATIS a datos categóricos, se tiene que:

$$S_{it} = \|W_t\|_{HS}^2 = \langle W_t | W_t \rangle_{HS} = \chi_{F_{it}}^2 + 1$$

$\pi_t = 1$ es la ponderación de la t-ésima tabla de datos

Entonces:

$$\alpha_t = \frac{1}{\sqrt{\lambda_1}} \left(\sum_{i=1}^T \sqrt{\sum_{j=1}^J \chi_{F_{it}}^2 + 1} \right) \gamma_{ti}$$

de esta manera, el compromiso W es de la misma naturaleza de los objetos W_t , su norma es el

promedio de las normas de los objetos W_t , $t=1, \dots, T$; $\|W\|_{HS} = \frac{\sum_{t=1}^T \|W_t\|_{HS}}{T}$, su representación gráfica en el primer eje factorial W estará a una distancia $\|W\|_{HS} = \frac{\sum_{t=1}^T \|W_t\|_{HS}}{T}$ a partir del origen (0,0), y el punto que lo representa es la proyección de un objeto ilustrativo en dicho plano.

3.4.3 Estudio de la Intraestructura

En esta etapa se procede igual que en el caso de la metodología STATIS aplicado sobre variables continuas, por lo tanto se representan posiciones compromiso como la posición promedio que él tiene considerada en términos de sus posiciones en los objetos que identifican las T tablas. La representación es proporcionada por la diagonalización del arreglo compromiso WD de orden $n \times n$.

Se consideran a estos efectos $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ los autovectores de la matriz WD , asociados respectivamente con los autovalores μ_i ; $i=1, \dots, n$; y se denota la imagen euclídea de los individuos compromiso mediante los puntos B_1, \dots, B_n , las coordenadas de proyección coordenadas quedan definidas en la forma:

$$\frac{1}{\sqrt{\mu_k}} WD \varepsilon_k = \sqrt{\mu_k} \varepsilon_k$$

En este punto, la construcción de las trayectorias se hace conjuntamente con la imagen euclídea de los individuos compromiso, esto es; en el primer plano factorial se grafican nT individuos B_1^t, \dots, B_n^t como puntos suplementarios para cada $t=1, \dots, T$, cuyas coordenadas son como sigue:

Si las coordenadas de proyección de los n individuos compromiso vienen dadas por $\frac{1}{\sqrt{\mu_k}} WD \varepsilon_k = \sqrt{\mu_k} \varepsilon_k$, entonces las coordenadas de los n individuos definidas en los T objetos representativos W_t son:

$$\sqrt{\mu_k} \varepsilon_k = \frac{1}{\sqrt{\mu_k}} \left(\sum_{t=1}^T a_t W_t \right) \varepsilon_k = \sum_{t=1}^T \frac{1}{\sqrt{\mu_k}} a_t [W_t \varepsilon_k] = \sum_{t=1}^T a_t \left(\frac{1}{\sqrt{\mu_k}} W_t \varepsilon_k \right)$$

Por lo tanto, como las coordenadas de proyección de los n individuos sobre el espacio compromiso quedan claramente definidas como una combinación lineal de las proyecciones de las posiciones de esos individuos en cada uno de los T estudios, ello justifica la

representación de las proyecciones $\frac{1}{\sqrt{\mu_k}} w_{t \varepsilon_k}$ ($t=1,2,\dots,T$) como puntos suplementarios sobre el espacio de representación determinado por el arreglo compromiso W .

CAPÍTULO IV

METODOLOGIA USADA PARA LA APLICACIÓN DEL STATIS-C

La metodología diseñada permitió demostrar, mediante razonamientos lógicos deducidos del contexto teórico del problema, y sostener, mediante pruebas realizadas con datos simulados y reales, la tesis:

La metodología STATIS se puede adaptar al tratamiento de datos provenientes de la medición de variables categóricas

El carácter lógico-deductivo se fundamentó en los siguientes aspectos:

El conocimiento que se dispone para el análisis simultáneo de diversas tablas de datos cuantitativos.

El desarrollo de técnicas de análisis estadístico para datos cualitativos.

El modelo teórico subyacente en las técnicas factoriales, es decir en la diagonalización de una matriz contentiva de información que es interpretable estadísticamente.

En relación a la sustentación práctica de la tesis, a continuación se detallan los aspectos referidos a la aplicación de la propuesta metodológica de esta investigación.

4.1 Simulación de datos

En el proceso de simulación de datos se siguieron las siguientes etapas:

Etapas 1

En esta etapa se simuló $N = 300$ individuos caracterizados según la medición de un conjunto de 15 variables particionado en tres subconjuntos de variables $\mathbf{X} = \{X_1, \dots, X_5\}$, $\mathbf{Y} = \{Y_6, \dots, Y_{10}\}$, y $\mathbf{Z} = \{Z_{11}, \dots, Z_{15}\}$, de una población normal multivariante $N(\boldsymbol{\mu}_{15 \times 1}, \boldsymbol{\Sigma}_{15 \times 15})$. La estructura de correlaciones entre las variables se define de acuerdo a los siguientes esquemas:

Correlación nula entre las variables que conforman el conjunto **X**.

Correlación de baja a moderada entre las variables del conjunto **Y**

Correlación fuerte entre las variables del conjunto **Z**.

La estructura de correlaciones entre las variables del conjunto **X** y la variable Y_1 del conjunto de variables **Y** es moderada, y nula con las demás variables en **Y**.

Las correlaciones entre las variables Y 's y las variables Z 's en el conjunto **Z** es moderada.

La correlación entre la variable Y_1 y las variables en el conjunto **Z** presentan una correlación muy fuerte, las demás variables Y 's están correlacionadas en forma moderada con las variables Z 's. La matriz de correlaciones se muestra en el Anexo 5.1.

La simulación en referencia se realizó con una sintaxis propia del Matlab (Versión 7.2).
Anexo 5.2

Etapas 2

La distribución correspondiente a cada una de las quince variables generadas en la etapa anterior, fue sometida a una categorización en cinco clases, que utilizó como puntos de corte los percentiles 20%, 40%, 60% y 80%, en cada caso. Por consiguiente, se obtuvieron datos correspondientes a 300 unidades estadísticas caracterizadas según quince variables categóricas de cinco categorías cada una.

Etapas 3

Finalmente, se seleccionó una muestra aleatoria de aproximadamente el 50% de las observaciones generadas, utilizando el menú del SPSS (versión 13.0)

Etapas 4

Los datos correspondientes a cada una de las variables categóricas se organizaron sobre matrices disyuntivas completas.

4.2 Tipos de tablas de datos simulados a analizarse mediante STATIS-Categórico

A partir de los datos categóricos generados en el proceso de la simulación, organizados sobre tablas disyuntivas completas, se obtuvieron las diferentes Tablas que se describen a continuación.

4.2.1 Tablas de datos con estructuras idénticas

A los efectos de obtener datos simulados correspondientes a tres dimensiones con idéntica estructura, se seleccionó aleatoriamente una muestra de las unidades caracterizadas por las variables del conjunto **X**, la información es utilizada para construir tres tablas de datos idénticas, con la finalidad de mostrar gráficamente la igualdad de las tres tablas comparadas en un solo punto como es lo esperado, así como mostrar que las distancias entre los objetos representativos son iguales a cero.

4.2.2 Tablas de datos con estructuras similares

Con el objeto de obtener tres tablas de datos con estructuras similares, que identifiquen tres grupos de variables medidas sobre los mismos individuos, se procedió a:

- a) Seleccionar una muestra aleatoria del conjunto de las variables que presentan entre sí correlaciones fuertes, que son las variables del conjunto **Z**.
- b) Perturbar un 10% de las unidades de la muestra obtenida en primer término
- c) Efectuar una perturbación sobre el 15% de las unidades de la muestra obtenida en primer término.

El proceso de muestreo se efectuó utilizando el programa computacional SPSS (Versión 13.0).

Dado que se dispone de información correspondiente a cinco variables cualitativas descritas en términos de las variables indicatrices asociadas, cada una con cinco categorías, las perturbaciones se efectuaron de la siguiente manera:

Para cada variable se describen las permutaciones con repetición de 5 valores, de los cuales hay $n_1=1$ valor igual al número 1 y $n_2=4$ valores iguales al número 0. Estas permutaciones son las siguientes:

Tabla 4.1: Permutaciones en las variables indicatrices

1	0	0	0	0
0	1	0	0	0
0	0	1	0	0
0	0	0	1	0
0	0	0	0	1

Se hacen sucesivas selecciones al azar de la matriz de permutaciones y se sustituyen por la información de cada una de las unidades que han sido seleccionadas para ser perturbadas. se dispone finalmente de tres Tablas de datos, que se asumen identificadas por variables cuyos comportamientos dan lugar a estructuras de datos similares. Estas tres tablas de datos que simulan estructuras similares, son utilizadas para determinar cómo el STATIS-C captura las diferencias y semejanzas entre las tablas.

4.2.3 Tablas de datos con estructuras similares y una diferente

Se utilizaron las tablas de datos obtenidas en la sección 4.2.2 y la tabla definida por el conjunto de variables con correlaciones moderadas **Y**. La finalidad de esta ilustración es para observar el comportamiento de las tablas con estructuras similares cuando se agrega una cuarta con estructura diferenciada.

4.2.4 Tablas de datos con sucesivos tamaños diferentes

Para ilustrar esta situación, se consideraron los datos simulados correspondientes a los grupos de variables **X**, **Y** y **Z**, después de haber sido categorizadas. Con este juego de datos categorizados, se procedió a evaluar el efecto que tiene el tamaño de muestra en la adaptación del STATIS. Para ello, de la población de 300 individuos, se seleccionaron tres muestras aleatorias de aproximadamente 50%, 25% y 15% de los datos originalmente simulados.

4.3 Implementación de la adaptación del STATIS sobre datos reales

Los datos reales utilizados para la implementación pertenecen han sido recabados por la Doctora Magaly Torres de Cárdenas y reportados en Torres (2005), los cuales son descritos en el capítulo V de la implementación de la propuesta metodológica.

4.4 Programas para la aplicación del STATIS-C

Los programas para implementar la propuesta de esta investigación, fueron realizados en Matlab, Versión 7.2.

Para corroborar algunos de los resultados concernientes al uso de los resultados obtenidos en la utilización del enfoque de ACB sobre las TCM F_{YX} y F_{XX} en la adaptación del STATIS, se utilizó el SPSS versión 13 y el SPAD versión 4.5

CAPÍTULO V

APLICACIÓN DEL STATIS-C

En este capítulo se presenta la implementación de la propuesta que, análogamente a la metodología STATIS, efectúa el tratamiento de datos categóricos para comparar las diversas tablas de datos de que se dispone, y explorar simultáneamente el comportamiento de los individuos en dichas tablas. La aplicación de la propuesta metodológica se realiza en los contextos de datos simulados, y datos reales.

En primer lugar, se simularon datos que caracterizan a 257 unidades de análisis, descritos por tres conjuntos de variables categóricas **X**, **Y** y **Z** (diseño transversal). La información correspondiente se organiza sobre una tabla de datos que yuxtapone tantas matrices disyuntivas completas, como variables conformen el(los) conjunto(s).

En segundo lugar, la implementación de la propuesta metodológica se realiza con datos reales facilitados por la Dra. Magaly Torres, quien desarrolla un estudio de corte transversal cuyos propósitos fueron medir el impacto de la crisis venezolana en los patrones alimentarios de la población infantil, (Torres M., 2005).

5.1.- Implementación del STATIS-C

Para efectuar la validación del STATIS-C se simularon diferentes escenarios para las variables que identifican las *T* tablas de datos categóricos, las cuales dieron lugar a la obtención de tablas con estructuras: idénticas, similares, similares y una diferenciada de las primeras, y todas diferenciadas. Además, se evaluó el comportamiento de la propuesta metodológica en el caso de diferentes tamaños de muestra.

5.1.1 Estudio de tres dimensiones X_1 , X_2 y X_3 con estructuras exactamente iguales

Se aplica el STATIS-Categórico sobre datos simulados obtenidos al considerar el conjunto de variables $\mathbf{X}=[X_1...X_j...X_5]$ medido sobre 257 unidades de análisis. (Variables X 's categorizadas en cinco modalidades).

La información contenida en la tabla de datos identificada por las X's se repite para obtener tres tablas con estructuras similares, cuya información es organizada en una tabla de orden 257×25 , obtenida al yuxtaponer cada una de las cinco matrices disyuntivas completas de orden 257×5 correspondientes a su vez a cada una de las tablas.

Se dispone entonces de datos simulados correspondientes a tres dimensiones, representadas por los objetos $W_1 = W_2 = W_3 = W_x$, todos con estructura de interdistancias entre individuos idénticas.

Como era de esperarse, la implementación del STATIS-C arrojó los resultados siguientes:

- a) Un solo punto en la gráfica de la interestructura en la que se superponen los tres objetos y el Compromiso W.*
- b) Tanto las distancias d_{HS-C} entre los objetos W_x , como las distancias euclídea obtenidas en la representación sobre el primer plano factorial definido por las direcciones principales de la matriz S, son iguales a cero.*
- c) La norma del compromiso W es igual al promedio de la norma de los $T=3$ objetos W_x .*
- d) La posición relativa de los individuos compromiso y de los individuos en los tres objetos que describen las interdistancias entre individuos para las tres tablas de datos es la misma, por lo tanto, las cuatro nubes de puntos aparecen superpuestas en el plano compromiso.*

Resultados

A.- Estudio de la interestructura y el compromiso

Las variables simuladas del conjunto X, originalmente tienen correlaciones nulas; por consiguiente, también debe ocurrir que entre las variables categorizadas no se presenten asociaciones significativas.

La medida de asociación global promedio en la Tabla de Contingencias Múltiple, F_{XX} , es $\chi^2_{F_{XX}} = 0,8771$, lo que indica que las variables categóricas que conforman al conjunto X's no

están asociadas, si se compara con el valor teórico que correspondería a una tabla $F_{Xj \times k}$ de orden 5×5 , fijado un nivel de significación de $\alpha = 0.05$, cual es $\chi^2_{16,0.95} = 26.30$.

Por su parte, en la Tabla 5.5., se describe la matriz de productos escalares, cuyo término genérico es de la forma:

$$\langle W_X | W_X \rangle = \chi^2_{F_{XX}} + 1 = 1.8771$$

Tabla 5.5: Matriz S obtenida mediante el STATIS

$\langle W_t W_t \rangle$	W_1	W_2	W_3
W_1	1.8771	1.8771	1.8771
W_2	1.8771	1.8771	1.8771
W_3	1.8771	1.8771	1.8771

En la figura 5.1 puede observarse que, la proyección ortogonal de las matrices de interdistancias, W_x , sobre el primer plano factorial de la Interestructura son coincidentes. Los cuatro puntos, correspondientes a los tres objetos y al compromiso, se superponen. En la Tabla 5.6, se muestra que, en efecto las coordenadas de proyección son idénticas para las tablas en consideración.

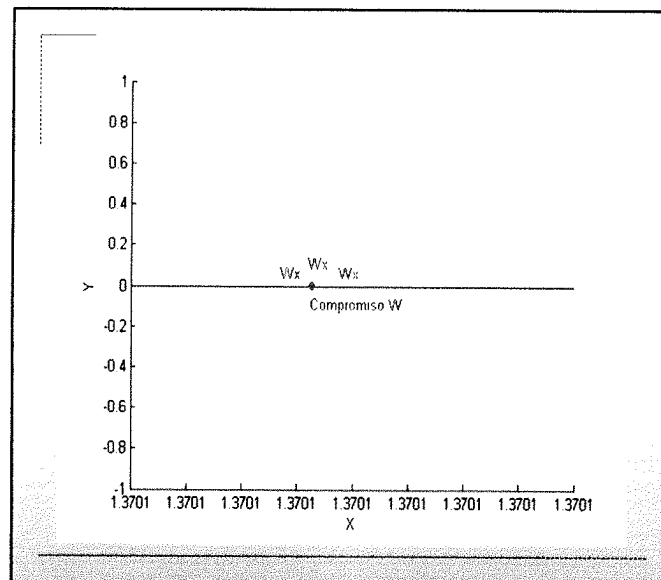


Figura 5.1: Gráfico de la Interestructura y Compromiso

Tabla 5.6: Coordenadas de los puntos representativos de W_1 , W_2 , W_3 y W

Objeto	X	Y
W_1	1.3701	0
W_2	1.3701	0
W_3	1.3701	0
Compromiso W	1.3701	0

Por su parte, como era de esperarse, al tratarse de arreglos idénticos, tanto las distancias euclídeas calculadas sobre las proyecciones del primer plano factorial de la matriz S, como las d_{HS-C} , son iguales a cero.

Tabla 5.7: Distancias entre los objetos

Distancias	Distancias Euclídea	d_{HS-C}
Dw_1w_2	0	0
Dw_1w_3	0	0
Dw_2w_3	0	0

B.- Análisis de la Intraestructura

Los individuos cuyas interdistancias quedan recogidas sobre los tres objetos W_x , al proyectarse de manera suplementaria sobre el plano compromiso concuerdan exactamente con la proyección de los individuos denominados compromiso, y en consecuencia aparecen superpuestos. En el figura 5.2, se percibe una única nube dibujada en color azul, que corresponde a la superposición de las cuatro nubes de individuos.

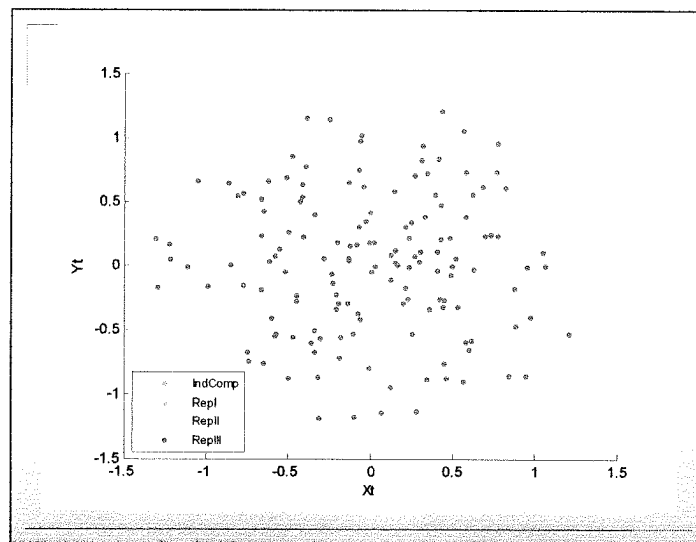


Figura 5.2: Gráfico de los individuos compromiso e individuos de los estudios E_x

Las variables del estudio E_x , que ha sido replicado tres veces, por construcción no presentan asociación entre sí, por lo tanto la nube de puntos que describe interdistancias compromiso entre individuos, queda representada de acuerdo a un patrón aleatorio.

5.1.2 Estudio simultáneo de las dimensiones Z_1 , Z_2 y Z_3 con estructuras similares

Las variables simuladas que originalmente permitieron definir el conjunto \mathbf{Z} conformado por cinco variables, presentan entre sí fuertes correlaciones. Los datos utilizados en esta sección han sido simulados a partir del conjunto de variables \mathbf{Z} , previa la categorización de cada una de las cinco variables en cinco categorías.

Se construyen entonces tres tablas de datos, la primera de ellas denotada por Z_1 , que coincide exactamente con la tabla original \mathbf{Z} ; las tablas Z_2 y Z_3 son obtenidas efectuando perturbaciones aleatorias de aproximadamente el 10% y el 15% de los registros, respectivamente, sobre la matriz original \mathbf{Z} .

Los resultados obtenidos en la etapa de la Interestructura correspondientes a productos escalares entre estudios quedan descritos sobre la matriz \mathbf{S} en la Tabla 5.8, a partir de cuya diagonalización se construye el espacio de representación de los objetos.

Tabla 5.8: Matriz \mathbf{S} obtenida mediante el STATIS

$\langle W_t W_{t'} \rangle$	W_{Z_1}	W_{Z_2}	W_{Z_3}
W_{Z_1}	4.3601	3.7098	3.6144
W_{Z_2}	3.7098	4.1297	3.0483
W_{Z_3}	3.6144	3.0483	4.0483

Tabla 5.9: Coordenadas de W_{Z_1} , W_{Z_2} , W_{Z_3} y W

Objeto	Coordenadas	
	X	Y
W_{Z_1}	2.0290	0.031015
W_{Z_2}	1.8872	0.697430
W_{Z_3}	1.8528	-0.74431
Compromiso W	2.0441	0

En cuanto a la estructura interna de interdistancias en los tres estudios E_{Z_1} , E_{Z_2} y E_{Z_3} información contenida en la tabla 5.10 muestra que el primer estudio muestra similitud con el segundo y el tercero, los cuales a su vez a parecen bien diferenciados. las coordenadas de proyección en la tabla 5.9, evidencian la semejanza entre los objetos W_{Z_1} , W_{Z_2} y W_{Z_3} , en la primera dirección principal definida por \mathbf{S} , observándose que el segundo eje captura las diferencia entre estos estudios, lo que resulta evidente sobre el espacio de representación de los objetos que se muestra en el figura 5.3. La calidad de la de representación de los objetos

sobre este plano se presenta en la tabla 5.11, por el I_{Cd} , indicando que las distancias entre los objetos W_{z2} y W_{z3} es muy bien aproximada.

Tabla 5.10: Distancias entre los objetos

	Distancia Euclídea	d_{HS-C}
$d_{W_{z1}W_{z2}}$	0.68134	1.0345
$d_{W_{z1}W_{z3}}$	0.79509	1.0861
$d_{W_{z2}W_{z3}}$	1.4421	1.4427

**Tabla 5.11: Índice de calidad de la
distancia Euclídea**

Distancias	I_{Cd}
$D_{W_{z1}W_{z2}}$	65.861%
$D_{W_{z1}W_{z3}}$	73.205%
$D_{W_{z2}W_{z3}}$	99.964%

A.- Interestructura y Compromiso

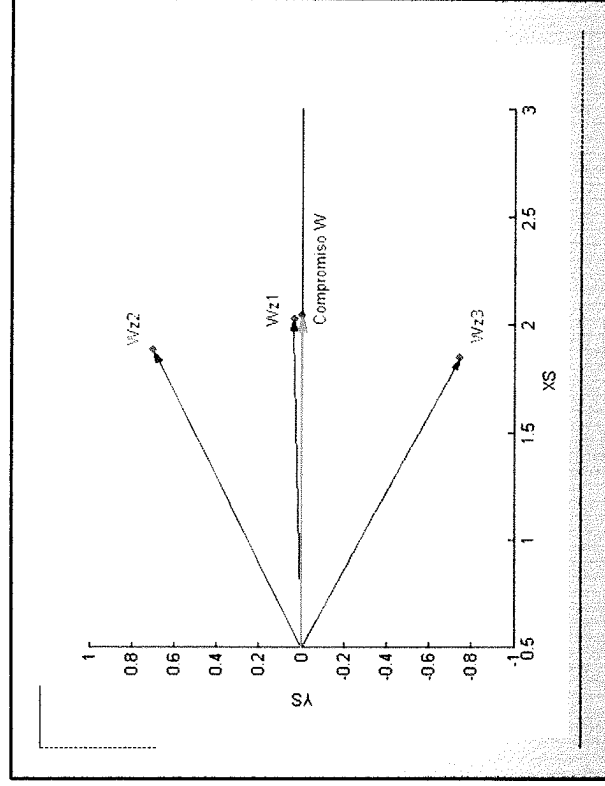


Figura 5.3a: Gráfico de tres estudios similares E_{Z1} , E_{Z2} y E_{Z3}

B.- Intraestructura

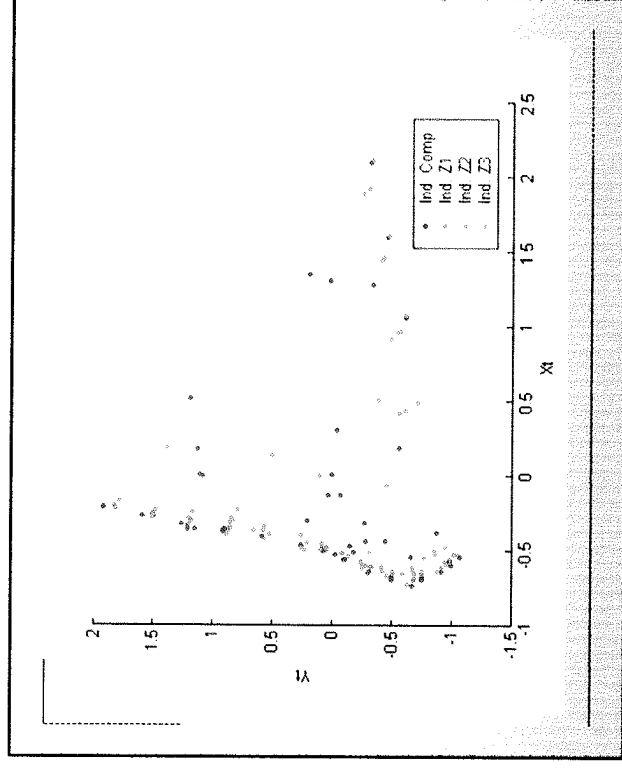


Figura 5.3b: Gráfico de individuos compromiso y descriptos por los objetos W_t

En la figura 5.3b, se observan las coincidencias entre la representación de los individuos compromiso (puntos azules), y el patrón de los individuos caracterizados en los objetos W_{Z1} , W_{Z2} y W_{Z3} , representados como puntos suplementarios, lo cual es evidencia de la existencia de una estructura de interdistancias común. Esto permite concluir que, en general, el compromiso representa adecuadamente a los tres objetos.

5.1.3. Estudio de las dimensiones Z_1 , Z_2 , Z_3 de estructura similar, y la tabla Y con una estructura diferente

En este caso, se agregan a los estudios de la situación anterior, que se corresponde con tres estructuras similares, una cuarta dimensión Y con una estructura bien diferenciada respecto de las anteriores. El número de variables en Y es diferente al número de variables en las tablas Z_t $t=1,2,3$, y la estructura de correlaciones entre las variables del estudio Y , es de baja a moderada. Los resultados obtenidos al aplicar el STATIS-Catégorico se muestran en las figuras 5.4a y 5.4b:

A.- Interestructura y Compromiso

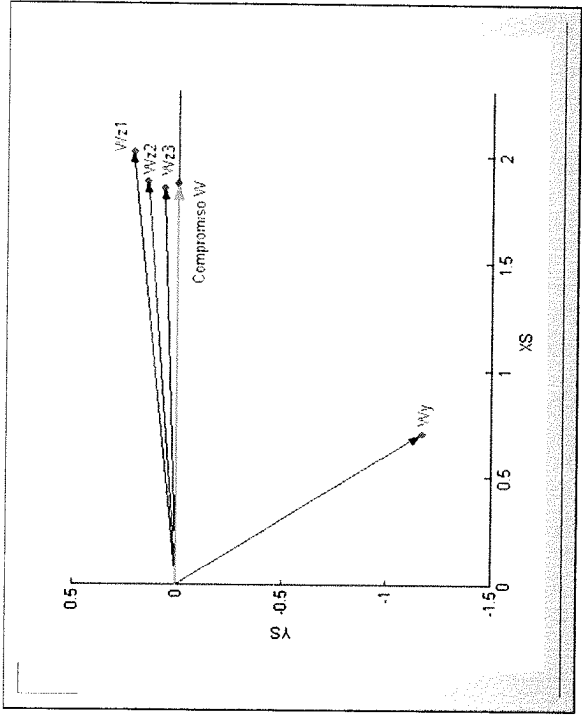


Figura 5.4a: Gráfico de tres estudios similares y uno diferente

B.- Intraestructura

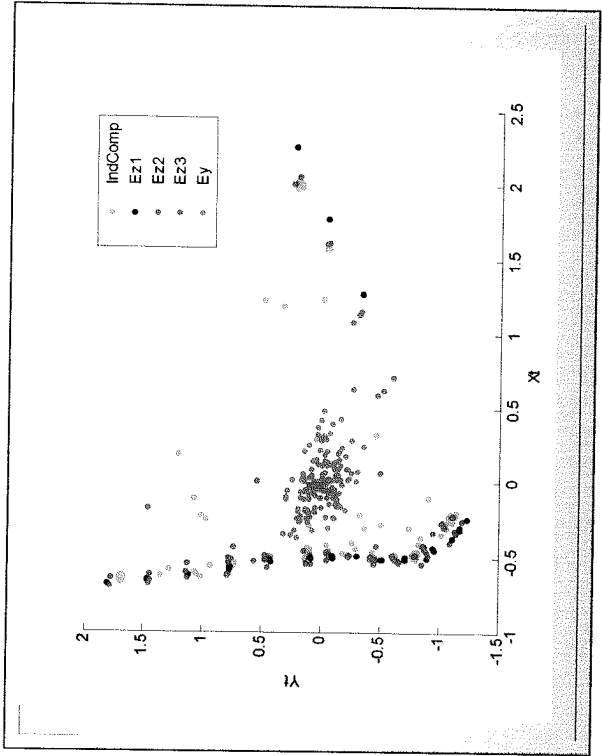


Figura 5.4b: Gráfico de los individuos compromiso y descritos en los Ez_1 , Ez_2 , Ez_3 y Ey

En las gráficas anteriores queda claramente comprobado que el cuarto estudio E_y es altamente diferente de los estudios E_{z1} , E_{z2} , E_{z3} , destacando además visualmente la notoria similitud entre éstos.

Por otra parte, en el plano compromiso, figura 5.4b, se muestran a los individuos caracterizados a partir de los estudios W_{z1} , W_{z2} , W_{z3} con un patrón más claro que en la figura 5.3b. por su parte, la nube de puntos rosados representa las interdistancias entre los individuos del estudio E_y , tienen una estructura diferente a la observada en los estudios E_{z1} , E_{z2} y E_{z3} . En este caso, también se observa que el compromiso representa adecuadamente a las tablas Z_1 , Z_2 y Z_3 .

Tabla 5.12: Matriz de interasociaciones globales S

$\langle W_t W_t \rangle$	W_{z1}	W_{z2}	W_{z3}	W_y
W_{z1}	4.3601	3.7098	3.6144	1.1837
W_{z2}	3.7098	4.1297	3.0483	1.1691
W_{z3}	3.6144	3.0483	4.0483	1.1748
W_y	1.1837	1.1691	1.1748	1.8771

Tabla 5.13: Índice de calidad I_{Cd}

Distancias	I_{Cd}
$D_{W_{z1}W_{z2}}$	14.652%
$D_{W_{z1}W_{z3}}$	20.88%
$D_{W_{z2}W_{z3}}$	6.3613%
$D_{W_{z1}W_y}$	97.384%
$D_{W_{z2}W_y}$	92.682%
$D_{W_{z3}W_y}$	89.364%

Como era de esperarse, y corroborando lo evidenciado en la figura 5.4a del primer plano de la Interestructura, las medidas de interasociación global a lo interno del estudio E_y y entre éste y los restantes estudios E_{z1} , E_{z2} , E_{z3} es bastante baja. Por su parte, el índice de calidad I_{Cd} muestra que las distancias entre el objeto W_y y los objetos W_{z1} , W_{z2} , W_{z3} del objeto W_y , quedan reflejadas sobre el plano con una calidad superior al 90%.

5.1.4 Análisis de las dimensiones X, Y y Z, con estructuras de correlaciones diferentes, y comparación de resultados con el uso de objetos Wt y objetos normados $\frac{W_t}{\|W_t\|}$

La simulación de datos descrita en la sección 4.2 en el Capítulo IV, dio lugar que los conjuntos de variables X's, Y's y Z's tuviesen internamente inter-correlaciones nulas, de baja a moderada y fuertes, respectivamente. En este caso, el STATIS-C se emplea para comparar los resultados de su aplicación sobre objetos no normados $W_t=XM_tX'$, con el caso en que se utilice objetos normados

Análisis comparativo de la interestructura y el compromiso con objetos no normados y objetos normados.

Tabla 5.14a: Matriz S con objetos Wt

$\langle W_t W_t \rangle$	Wx	Wy	Wz
Wx	1.8771	1.1125	1.1837
Wy	1.1125	1.9001	1.811
Wz	1.1837	1.811	4.3601

Tabla 5.14b: Matriz S con objetos normados

$\left\langle \frac{W_t}{\ W_t\ } \middle \frac{W_t}{\ W_t\ } \right\rangle$	Wx	Wy	Wz
Wx	1	0.58905	0.41375
Wy	0.58905	1	0.62918
Wz	0.41375	0.62918	1

En primer lugar, conviene hacer referencia a las medidas de asociación global entre los estudios dos a dos, que se presentan en orden creciente $\chi^2_{F_{xy}}=0.1125$, $\chi^2_{F_{xz}}=0.1837$, $\chi^2_{F_{yz}}=0.811$. Es decir, el mayor grado de interasociación global se produce entre los estudios E_Y y E_Z, orden que se modifica cuando se utilizan los objetos normados; esto es, $\chi^2_{F_{xz}}=0.41375$, $\chi^2_{F_{xy}}=0.58905$, $\chi^2_{F_{yz}}=0.62918$, que además describe asociaciones de la misma magnitud entre las variables F_{YX} y F_{ZX}.

Comparación de las Coordenadas y distancias con Objetos W_t y Objetos Normados.

Tabla 5.15a: Coordenadas de W_x , W_y , W_z y W

W_t	X	Y
W_x	0.88455	0.95375
W_y	1.1286	0.42176
W_z	1.9683	-0.67044
Compromiso W	1.6122	0

Tabla 5.15b: Coordenadas de W_x , W_y , W_z y W normados

Objeto Normado	X	Y
W_x	0.79131	0.57313
W_y	0.89572	-0.044571
W_z	0.81517	-0.50739
Compromiso W	1	0

Tabla 5.16a: Distancias entre los objetos W_x , W_y y W_z

Distancias	Distancia Euclídea	d_{HS-C}
dw_xw_y	0.5853	1.2459
dw_xw_z	1.9526	1.9672
dw_yw_z	1.3777	1.6243

Tabla 5.16b: Distancias entre los objetos normados

Distancias	Distancia Euclídea	d_{HS-C}
dw_xw_y	0.62647	0.90659
dw_xw_z	0.62647	1.0828
dw_yw_z	0.46977	0.86118

En las tablas 5.15a,b, y 5.16a,b se observa que tanto las coordenadas como las distancias son modificadas , en algunos casos sustancialmente.

Comparación del índice de calidad de aproximación de la distancia con objetos W_t y objetos Normados.

Tabla 5.17a: Índice de calidad con objetos W_t

Distancias	I_{Cd}
dw_xw_y	46.978%
dw_xw_z	99.256%
dw_yw_z	84.819%

Tabla 5.17b: Índice de calidad con objetos normados

Distancias	I_{Cd}
dw_xw_y	69.102
dw_xw_z	99.812
dw_yw_z	54.55

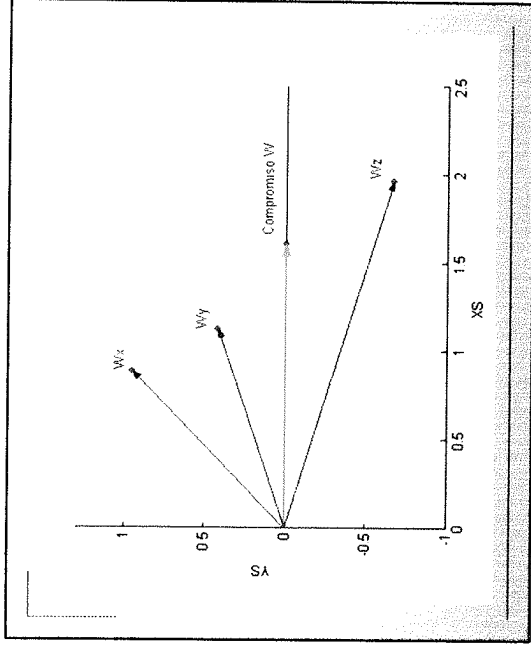


Figura 5.5a: Gráfico de tres estudios diferentes representados por objetos W_t

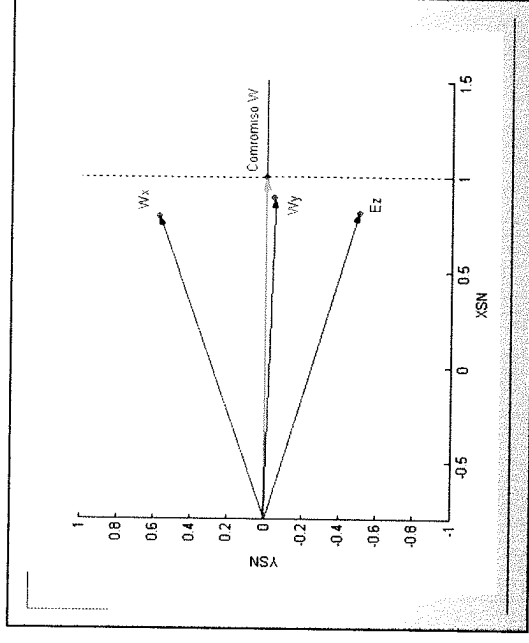


Figura 5.5b: Gráfico de tres estudios diferentes representados por objetos normados

El primer plano Interestructura, figura 5.5a, muestra que el objeto W_z incide en la formación del compromiso, y los objetos W_t son diferenciados en el plano por ambos ejes. En caso de los objetos normados, éstos favorecen la construcción del compromiso, pero la diferencia existente entre los estudios solo es mostrada en el segundo eje factorial

Las siguientes figuras 5.6a y 5.6b muestran que otra ventaja de los objetos normados está en que se detalla más la dispersión de los individuos compromiso y de las posiciones de los individuos en los tres estudios. Indistintamente del objeto utilizado, se conservan la estructura general de los estudios E_x , E_y , E_z .

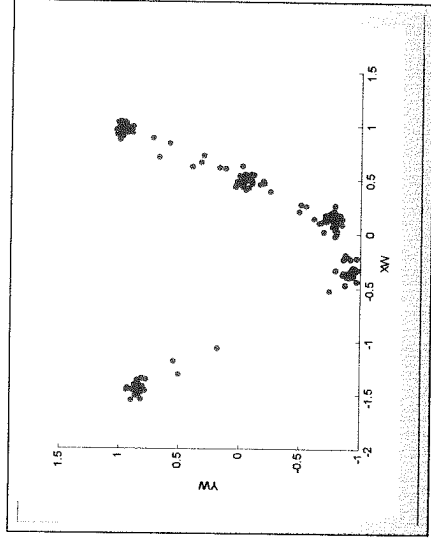


Figura 5.6a: Gráfico de individuos compromiso con objetos Wt

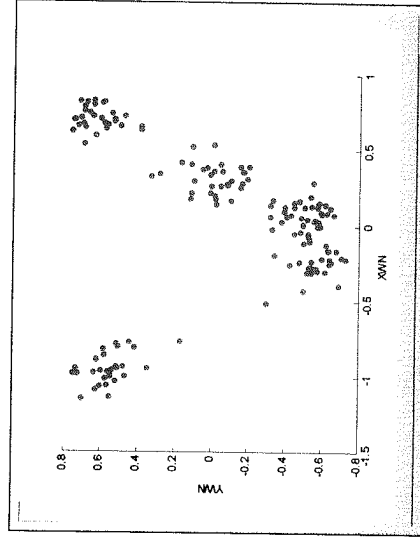


Figura 5.6b: Gráfico de individuos compromiso con objetos normados

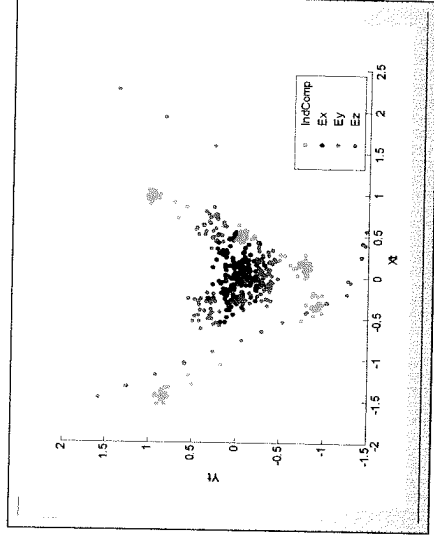


Figura 5.7a: Gráfico del primer plano intraestructura con objetos Wt

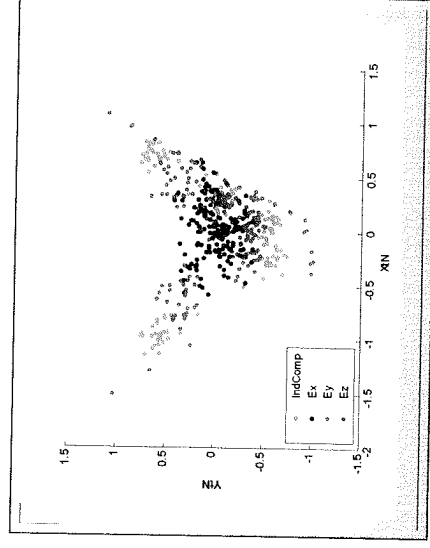


Figura 5.7b: Gráfico del primer plano intraestructura con objetos normados

Resultados de la comparación del uso de los objetos W_t y los objetos Normados

La adaptación de la etapa de la interestructura es modificada desfavorablemente con el uso de objetos normados; tanto la matriz a diagonalizar S , como las distancias entre los objetos representativos, muestran diferencias importantes, éstas son evidenciadas en el primer plano interestructura en el cual se exponen similitudes entre los objetos en el primer eje factorial. A este respecto, las figuras 5.5a y 5.5b ilustran esta desventaja del uso de objetos normados.

No obstante, la construcción del compromiso W y el primer plano compromiso son favorecidos con el uso de objetos normados; la norma del compromiso presenta una mejor aproximación al promedio de las normas ponderadas de los objetos W_t , lo cual es una propiedad del compromiso. Por otra parte, el primer plano compromiso, con el uso de los objetos normados, muestra una mejor visión de las posiciones relativas de los individuos. Las figuras 5.6a, 5.6b, 5.7a y 5.7b evidencian la ventaja del uso de los objetos normados en las etapas del compromiso e intraestructura: la mirada de las posiciones relativas compromiso y las descritas en cada uno de los objetos representativos de los estudios es una visión ampliada en la cual se pueden identificar mejor los individuos causantes de las similitudes y diferencias.

5.1.5 Análisis de los constructos X, Z1, Y, cuyas estructuras de correlaciones son diferentes, y diferentes números de variables, diferentes números de categorías, y diferentes tamaños de muestra.

Tabla 4.18a: Matriz S con n=157

Productos Escalares	$\langle W_x \rangle$	$\langle W_z \rangle$	$\langle W_y \rangle$
$\langle W_x \rangle$	1.8329	1.1653	1.1093
$\langle W_z \rangle$	1.1653	3.5866	1.6872
$\langle W_y \rangle$	1.1093	1.6872	1.9001

Tabla 4.18b: Matriz con n=79

Productos Escalares	$\langle W_x \rangle$	$\langle W_z \rangle$	$\langle W_y \rangle$
$\langle W_x \rangle$	1.9114	1.2097	1.2164
$\langle W_z \rangle$	1.2097	3.5533	1.741
$\langle W_y \rangle$	1.2164	1.741	1.9928

Tabla 4.18c: Matriz S con n=42

Productos Escalares	$\langle W_x \rangle$	$\langle W_z \rangle$	$\langle W_y \rangle$
$\langle W_x \rangle$	2.0003	1.3691	1.36
$\langle W_z \rangle$	1.3691	3.731	1.9501
$\langle W_y \rangle$	1.36	1.9501	2.1786

Tabla 4.19a: Índice de calidad de distancia. n=157

Distancias	I_{Cd}
$D_{W_x W_z}$	99.418%
$D_{W_y W_z}$	79.32%
$D_{W_x W_y}$	48.757%

Tabla 4.19b: Índice de calidad de distancia. n=79

Distancias	I_{Cd}
$D_{W_x W_z}$	99.344%
$D_{W_z W_y}$	79.982%
$D_{W_x W_y}$	48.332%

Tabla 4.19c: Índice de calidad de distancia. n=42

Distancias	I_{Cd}
$D_{W_x W_z}$	99.348%
$D_{W_z W_y}$	79.662%
$D_{W_x W_y}$	48.82%

Las tablas 4.18 a, b, c muestran que con la disminución del tamaño de la muestra, en general, ocurre un ligero incremento de las medidas de asociación global en las TCM involucradas en el análisis. Por otra parte, el tamaño de la muestra no afecta el índice de calidad de aproximación de la distancia entre objetos en el primer plano Interestructura, ni a la distancia d_{HS-C} .

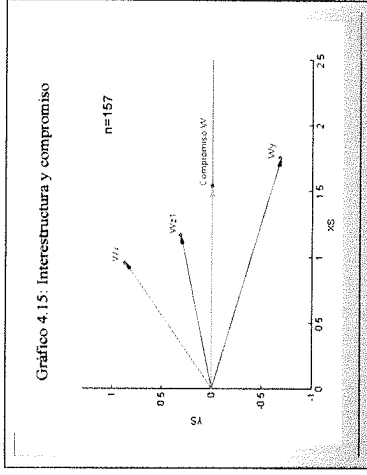


Figura 4.13a: Gráfico de interestructura. n=157

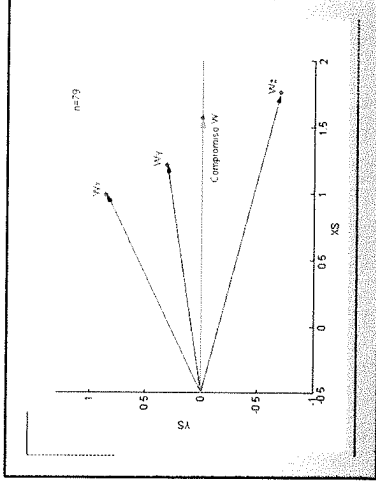


Figura 4.13b: Gráfico de interestructura. n=79

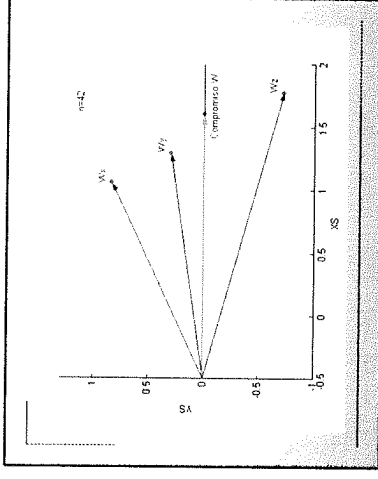


Figura 4.13c: Gráfico de interestructura. n=42

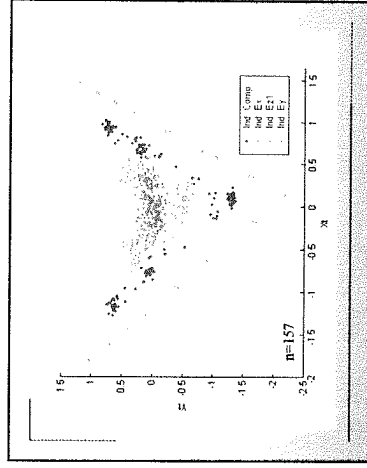


Figura 4.14a: Gráfico de plano compromiso. n=157

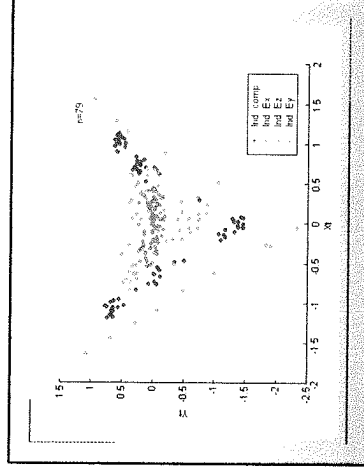


Figura 4.14b: Gráfico de plano compromiso. n=79

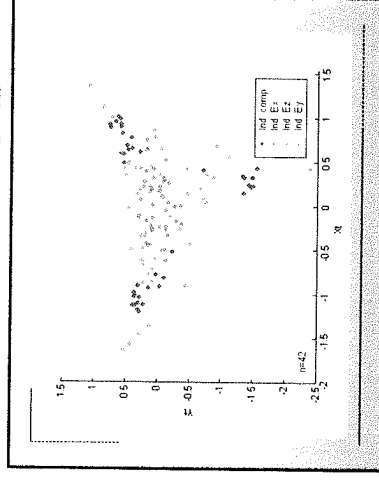


Figura 4.14c: Gráfico de plano compromiso. n=42

Los gráficos 4.13 a, b, c y 4.14 a, b, c, correspondientes al primer plano Interestructura, revelan que el tamaño de muestra no afecta las estructuras existentes entre las tablas de datos. Modifica ligeramente las coordenadas, pero en lo fundamental, se conservan las estructuras de interdistancias.

5.2 Análisis de Datos Reales mediante STATIS-Categórico

En esta ilustración se hace uso de los datos obtenidos en una investigación realizada por la Doctora Magaly Torres de la Facultad de Medicina de la Universidad Central de Venezuela, y reportada en Torres M. (2005), cuyo propósito fundamental es evaluar el impacto de la crisis venezolana en los patrones alimentarios de 256 niños de ambos sexos residentes en el Área Metropolitana de Caracas, con edades comprendidas entre 1 y 3 años, y pertenecientes a los estratos III, IV y V según Graffar. La autora enfoca esta problemática considerando cuatro dimensiones, operacionalizadas en términos de variables categóricas y variables cuantitativas, éstas últimas categorizadas basándose en conceptos teóricos sustantivos del ámbito biomédico. Las dimensiones en consideración, las variables categóricas utilizadas en esta ilustración, y las correspondientes categorías se presentan en la Tabla 5.20 a, b, c, d.

Tabla 5.20: Variables de Dimensión Nivel Socio-Económico

Dimensiones	Variables Categóricas	Categorías	Nº Total Cat.
Dim. 1 Nivel Socio-Económico	Tipo de familia	1= Nuclear 2=Extendida	15
	Ingreso familiar	1=<189,00 Bs. 2=190-220 3=221-300 4=301-500 5=>501	
	Gasto alimentación/semana	1=0-100 Bs. 2=100-200 3=200-300 4=300-400	
	Tamaño Núcleo Fliar.	1=2-4 Personas 2=5-8 3=9-12 4=13-16	

Tabla 5.21: Variables de Dimensión Antecedentes y Datos Antropométricos

Dim. 2 Antecedentes y Datos Antropométricos	Edad Madre	1=17-21 años 2=21-25 3=25-30 4=30-54	Nº Total Cat.
	Edo. Civil Madre	1=Soltera 2=Casada 3=Viuda 4=Concubina 5=Divorciada	21
	Nivel de Instrucción M.	1=ICOM 2=STSU 3=SITM 4=Primaria 5=Analfabeta	
	Peso del niño al nacer	1=Muy Bajo 2=Bajo 3=Normal 4=Normal Alto 5=Alto	
	Talla del niño al nacer	1=Pequeño 2=Normal 3=Alto	

Tabla 5.21: Variables de Dimensión Antecedentes Ingesta

Dim. 3 Ingesta	Consumo Real. de proteínas	1=0-20 2=20-40 3=40-60 4=60-80	Nº Total Cat.
	Adec. de proteínas	1=Prot. Insuficientes 2=Prot. Adecuadas 3=Prot. Exceso	24
	Adec. de Calorías	1=Calo. Insuficientes 2= Calo. Adecuadas 3= Calo. Exceso	
	Lactante o no	1=Si 2=No	
	Adec. de Vitaminas	1=VitA. Insuficiente 2= VitA. Adecuadas 3= VitA. Exceso	
	Adec. de hierro	1=Hie. Insuficiente 2= Hie. Adecuadas 3= Hie. Exceso	
	Adecuación de calcio	1=Cal. Insuficientes 2= Cal. Adecuadas 3= Cal. Exceso	
	Adec. de zinc	1=Zinc. Insuficientes 2= Zinc. Adecuadas 3= Zinc. Exceso	

Tabla 5.21: Variables de Dimensión Estado Nutricional Actual

Dim. 4 Edo. Nutricional Actual	Peso para la Talla	1=TP sobre la Norma 2= TP Normal 3= TP Deficiente	N° Total Cat. 12
	Talla para la Edad	1= Talla sobre la Norma 2= Talla Normal 3= Talla Deficiente	
	Peso para la Edad	1= Peso sobre la Norma 2= Peso Normal 3= Peso Deficiente	
	Diag. Edo. Nut. Combinado	1= Sobre la Norma 2=Normal 3=Desnutridos	

Dimensión 1: Nivel Socio-Económico

Las variables contempladas en esta dimensión son: Tipo de familia, Ingreso familiar, Gasto alimentación/semana y Tamaño del núcleo familiar, por considerar que ellas están directamente relacionadas con los patrones alimentarios del grupo familiar.

Dimensión 2: Antecedentes y Datos Antropométricos

Las variables que integran esta dimensión son: En primer lugar Estado Civil de la Madre y Nivel de Instrucción Madre, variables que dan cuenta del apoyo y de las potencialidades y/o limitaciones de la madre durante el embarazo. Se incorporan también variables antropométricas, que describen el estado nutricional del niño al nacer como son : Peso del niño al nacer, y Talla del niño al nacer.

Dimensión 3: Ingesta

Se consideraron todas las variables definidas técnicamente por Torres (op. cit.), que quedan descritas en la Tabla 5.20

Dimensión 4: Estado Nutricional Actual

También en esta dimensión se consideraron todas las variables definidas técnicamente por Torres (op. cit), que quedan descritas en la Tabla 5.20

A.- Estudio de la Interestructura y determinación del Compromiso W

A continuación se muestran los resultados obtenidos al aplicar el STATIS-Categórico utilizando un programa diseñado especialmente a estos efectos.

Sobre el Gráfico 5.15 se describen las diferencias y/o similitudes entre los objetos que describen cada una de las cuatro dimensiones de información consideradas en esta ilustración.

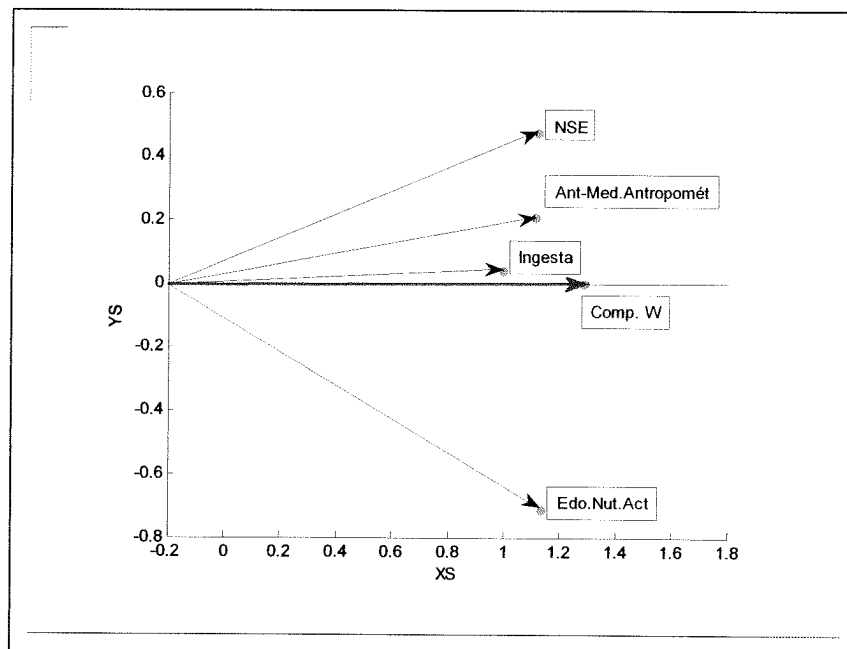


Figura 5.15: Gráfico de la representación euclídea de los estudios y compromiso

Los resultados básicos de la Interestructura reflejan que el primer eje define una dirección que no distingue entre los cuatro estudios bajo consideración. Sin embargo, el segundo eje factorial detecta ligeras diferencias entre las dimensiones Nivel Socio-Económico (NSE), Antecedentes y Datos Antropométricos (Ant-Med. Antrop), e Ingesta (Ingesta), que a su vez, presentan una marcada diferencia con respecto al objeto que describe interdistancias entre los 256 niños de acuerdo a Estado Nutricional Actual (E. Nut. Act.).

En consecuencia de lo anterior, puede señalarse que:

- a) Las diferencias medidas entre dos niños cualesquiera, en términos de su ubicación en algún estrato socio-económico, son concordantes con las que en

igual medida se observan entre sus perfiles definidos por los antecedentes de la madre y del niño al nacimiento, y también por el correspondientes al tipo de ingesta.

- b) La estructura de interdistancias que describe las diferencias entre los niños bajo estudio en lo relativo a su Estado de Nutricional, es marcadamente diferente de las que se perciben en las otras tres dimensiones.

B.- Análisis de la Intraestructura

En esta fase del análisis, se trata de indagar acerca de las razones que explicarían las diferencias entre los objetos observadas en la fase anterior. Para ello, en primer lugar se representan los niños sobre el espacio compromiso que se muestra en el Gráfico 5.16.

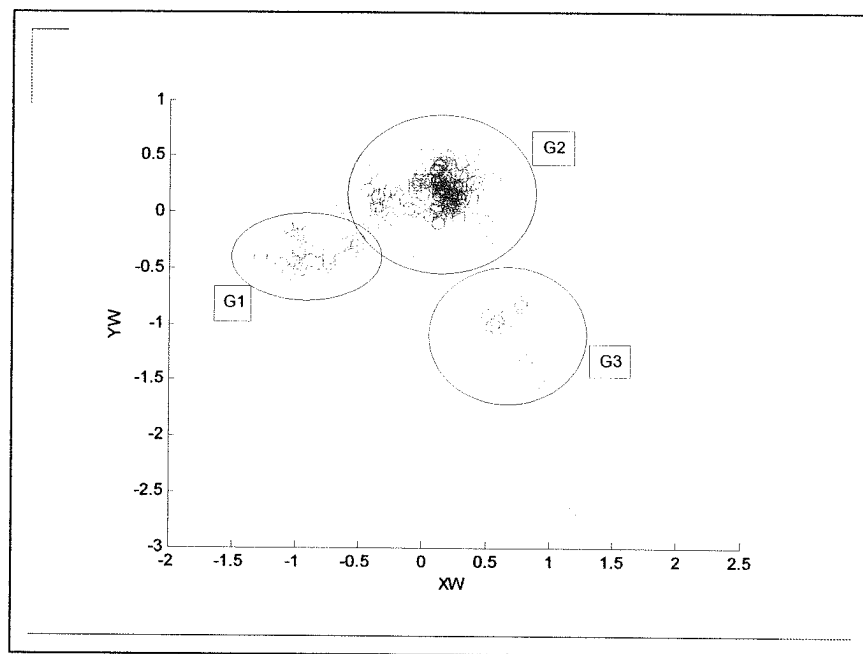


Figura 5.16: Gráfico de la representación de los individuos compromiso

La representación sobre el espacio compromiso sugiere la existencia de tres grupos de niños claramente diferenciados, con la presencia un niño atípico ubicado hacia la zona inferior derecha del gráfico.

En el Gráfico 5.17 se representan sobre las primeras dos dimensiones del espacio compromiso, a título ilustrativo interdistancias entre los niños, en las cuatro dimensiones bajo estudio.

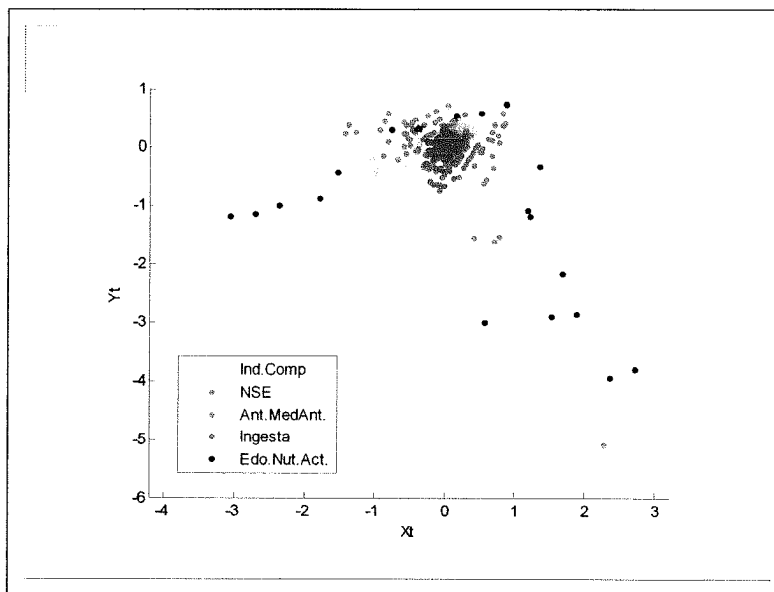


Figura 5.17:
Gráfico de la representación simultánea de interdistancias en el
compromiso y en las cuatro dimensiones bajo estudio

En el gráfico anterior se corrobora la diferencia observada de la dimensión Estado Nutricional –puntos negros- con respecto a las otras dimensiones, ya obtenida en la Interestructura. Se puede observar que, las unidades de análisis en la dimensión ingesta – puntos azules- se presentan poco distanciadas entre sí. En lo tocante a las otras dimensiones, parecieran reflejarse diferencias mayores entre los niños investigados, que particularmente se acrecientan en lo relacionado con el Estado Nutricional –puntos negros-.

Los resultados anteriores permiten concluir que en general, los niños participantes en el estudio presentan:

- a) Patrones en su ingesta alimentaria relativamente homogéneos.
- b) Diferencias marcadas en lo relacionado con su Estado Nutricional,

Las posiciones relativas de los niños en el espacio compromiso, descritas por las interdistancias definidas por el Estado Nutricional Actual, sugieren que esta dimensión es la determinante de la formación de los tres agrupaciones inicialmente visualizadas en el espacio compromiso.

Al indagar a nivel individual sobre las características de algunos de los niños ubicados en los grupos que se reflejan en la Interestructura, se obtiene lo siguiente:

Historia 221: correspondiente a un niño cuyo perfil en las variables que conforman las cuatro dimensiones se describe en la Tabla 5.21.

Tabla 5.21: Caracterización del niño en las dimensiones

Dimensiones	Variables Categóricas	Categoría
Dim. 1 Nivel Socio-Económico	Tipo de familia Ingreso familiar Gasto alimentación/semana Tamaño Núcleo Fliar.	Nuclear >501 Bs 200-300 13-16 personas
Dim. 2 Antecedentes y Datos Antropométricos	Edad Madre Edo. Civil Madre Nivel de Instrucción M. Peso del niño al nacer Talla del niño al nacer	>30 años Soltera Primaria Normal alto=3.95 Kg Normal
Dim. 3 Ingesta	Consumo Real. de proteínas Adec. de proteínas Adec. de Calorías Lactante o no Adec. de Vitaminas Adec. de hierro Adecuación de calcio Adec. de zinc	20-40 gr/kg Insuficiente Insuficiente No Vit A en exceso Adecuado Adecuado +baja en la Cat. Insuficiente
Dim. 4 Edo. Nutricional Actual	Peso para la Talla Talla para la Edad Peso para la Edad Diag. Edo. Nut. Combinado	Normal Normal Sobre la norma Sobre la norma

Este es un niño con características contradictorias, ya que si pertenece a una familia de tipo **Nuclear**, su madre no puede ser soltera, posiblemente es un error de transcripción. Por otra parte, tanto la adecuación del consumo de calorías y como el de proteínas de este niño es insuficiente, a lo que se agrega también insuficiencia en el consumo de zinc, sin embargo su

estado nutricional es clasificado como normal, y particularmente en cuanto a peso para la edad es diagnosticado sobre la norma.

Las características discordantes de las variables en el niño 221 parecen ser la razón de que éste aparezca en los gráficos de la Intraestructura, como un punto atípico.

Trayectorias en un Grupo de niños

El análisis de la Intraestructura culmina con el trazado de las trayectorias que describen los cambios en las posiciones relativas de los individuos de una dimensión a otra. Para efectos ilustrativos se construyeron las trayectorias de un grupo de tres niños cuya selección utilizó la variable Diagnóstico Combinado, que define entre otras, una categoría de niños calificados con estado nutricional Sobre la Norma y para los cuales se sugiere investigar la razón de Talla Baja (STIB), sus coordenadas compromiso son las siguientes:

120=	-0.22697537	-0.560681215	-0.251775809	0.0863521	-0.030854921	-0.001828036	0.15733625	0.568955293	-0.104265916	0.029855296
205=	0.19751633	-0.498883193	-0.025230358	-0.0865455	0.075035252	0.042931589	0.51573382	0.603522446	0.230109673	0.019326546
239=	-0.02115911	-0.115555445	-0.025230358	-0.0865455	-0.241327727	-0.035782409	-3.06748777	-1.173552547	-1.017192863	-0.429516775

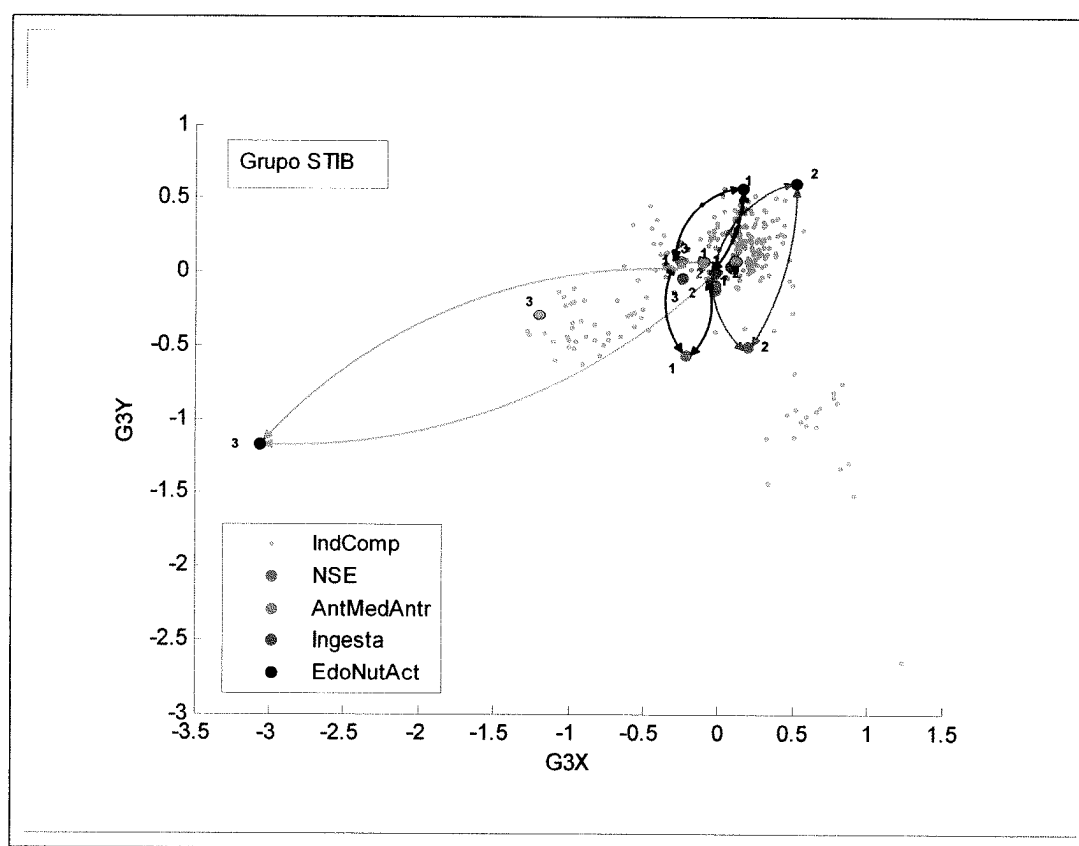


Figura 5.20: Grupo con diagnóstico combinado STIB; n=3

Tabla 5.21: Perfil de los niños STIB en las cuatro dimensiones bajo estudio

Dimensión	Variables Categóricas	Tabla de Datos Original		
		Niño 1	Niño 2	Niño 3
Dim. 1 Nivel Socio-Económico	Tipo de familia Ingreso familiar Gasto alimentación/semana Tamaño Núcleo Fliar.	Nuclear 301-500 Bs 0-100 Bs 2-4 personas	Nuclear <189 Bs 0-100 Bs 2-4 personas	Extendida 301-500 Bs 0-100 Bs 5-8 personas
Dim. 2 Antecedentes y Datos Antropomé- tricos	Edad Madre Edo. Civil Madre Nivel de Instrucción M. Peso del niño al nacer Talla del niño al nacer	22 - 25 años Casada Sec. Inc. o Tec. M. Alto Normal	> 30 años Concubina Sec Com o Tec S.Comp Normal Alto Normal	> 30 años Soltera Sec. Inc. o Tec. M. Normal Pequeño
Dim. 3 Ingesta	Consumo Real. de proteínas Adec. de proteínas Adec. de Calorías Lactante o no Adec. de Vitaminas Adec. de hierro Adecuación de calcio Adec. de zinc	20-40 Prot. Insuficientes Calorías Insuf. Lactante Vit A Exceso Hierro Insuficiente Calcio en exceso Niv. bajo Cat Isuf	40-60 Prot. Insuficientes Calorías Insuf. Lactante Vit A Insuficiente Hierro Adecuado Calcio en exceso Niv. bajo Cat Isuf	0-20 Prot. en exceso Calorías Insuf. No Lactante Vit A Exceso Hierro Insuficiente Calcio Insuficiente Niv. bajo Cat Isuf
Dim. 4 Edo. Nutricional Actual	Peso para la Talla Talla para la Edad Peso para la Edad Diag. Edo. Nut. Combinado	Sobre la norma Talla Deficiente Peso Normal Normal	Sobre la norma Talla Deficiente Peso Normal Normal	Sobre la norma Talla Deficiente Peso Normal Desnutrido

Puede observarse que en general, el perfil del Niño No. 3 es concordante con el de los otros dos en cuanto a su posición compromiso y a las posiciones relativas definidas por la Ingesta, los Antecedentes de la madre y del niño al nacimiento, y del Nivel Socio-económico, presentando sin embargo, un comportamiento notablemente diferenciado en lo tocante a su Nutricional Actual, lo cual es explicable en términos de que su diagnóstico nutricional es de déficit – desnutrido-, mientras que los otros dos niños, son diagnosticados como normales.

CAPITULO VI

HALLAZGOS, CONCLUSIONES Y RECOMENDACIONES

La problemática de investigación esencialmente se ha centrado en la búsqueda de un procedimiento para el tratamiento de la información que se genera en un diseño de corte transversal al caracterizar simultáneamente un mismo grupo de individuos de acuerdo a T conjuntos de variables cualitativas, o en uno longitudinal al caracterizar un mismo grupo de individuos de acuerdo a un único conjunto de variables cualitativas medidas a lo largo de T ocasiones.

En la dirección del objetivo general planteado, en esta investigación se ha logrado:

Abordar el problema de investigación mediante la adaptación de la metodología STATIS, originalmente desarrollada con propósitos similares para el caso de que las variables sean cuantitativas.

En la dirección de los objetivos específicos, la tarea de mayor envergadura que debió enfrentarse consistió en definir una medida para comparar las estructuras que describen las interdistancias entre individuos caracterizados por los diferentes conjuntos de variables cualitativas en consideración. En torno a este objetivo, se ha encontrado que:

Es posible cuantificar en una medida de distancias d_{HS-C} tipo Hilbert-Schmidt, las diferencias existentes entre las estructuras de interdistancias que se producen al comparar n individuos de acuerdo a dos conjuntos diferentes de variables cualitativas. La medida de distancia propuesta para comparar los objetos de interés, toma en cuenta las interasociaciones que se producen entre los respectivos conjuntos de variables cualitativas que respectivamente los identifican, y admite una desagregación que permite obtener la medida en que la asociación entre cualesquiera dos pares de variables cualitativas de los conjuntos

en consideración, contribuye a la interasociación global, y también en lo fundamental a la distancia d_{HS-C} propuesta.

La distancia d_{HS-C} es definida como sigue:

$$\begin{aligned}
 d_{HS}^2(W_X, W_Y) &= \|W_X - W_Y\|_{HS}^2 = \|W_X\|_{HS}^2 + \|W_Y\|_{HS}^2 - 2\langle W_X | W_Y \rangle_{HS} \\
 &= \left(\frac{\sum_{j=1}^J \sum_{j=1}^J \chi_{X_j, X_j}^2}{(J \times J)n} + 1 \right) + \left(\frac{\sum_{i=1}^I \sum_{j=1}^J \chi_{Y_i, Y_i}^2}{(I \times I)n} + 1 \right) - 2 \left(\frac{\sum_{i=1}^I \sum_{j=1}^J \chi_{Y_i, X_j}^2}{(I \times J)n} + 1 \right) \\
 &= \chi_{F_{YY}}^2 + \chi_{F_{XX}}^2 - 2\chi_{F_{YX}}^2
 \end{aligned}$$

El problema planteado y resuelto en el trabajo de esta tesis origina nuevos problemas, con lo cual éste se puede convertir en una línea de investigación, debido a que, por una parte, la estructura de datos categóricos formulada para la adaptación del STATIS no es exhaustiva, y por otro, la distancia d_{HS-C} propuesta para medir las diferencias y similitudes entre TCM, puede ser desagregada en las contribuciones debidas a las variables categóricas que definen las TCM, y en las contribuciones de los individuos. En este sentido, dada las representaciones gráficas de los individuos en los planos intraestructura, se puede pensar en valores atípicos multivariantes en el caso de datos categóricos.

REFERENCIAS BIBLIOGRÁFICA

- Agresti A., Analysis of ordinal categorical data. Wiley. New York. (1984)
- Baccalá N., Contribuciones Al Análisis De Matrices De Datos Múltiple: Tipología de las Variables. Universidad de Salamanca Departamento de Estadística. (2004)
- Bécue M., Pagès J., Álvarez L., Hernández M., Análisis factorial múltiple para tablas de contingencia: estudio de la mortalidad en las comunidades autónomas de España. 27 Congreso Nacional de Estadística e Investigación operativa Lleida, 8-11 de abril (2003)
- Bécue M., Pagès J., Multiple factor analysis and clustering of a mixture of quantitative, categorical and frequency data. Computational Statistics & Data Analysis 52, 3255-3268, (2004)
- Bécue M., Pagès J., A principal axes method for comparing contingency tables: MFACT. Computational Statistics & Data Analysis. Volume 45, Issue 3, Pages 481-503. (2004)
- Bécue M., Pagès J., Analyse factorielle multiple intra-tableaux. Application à l'analyse simultanée de plusieurs questions ouvertes. JADT 2000: 5es Journées Internationales d'Analyse Statistique des Données Textuelles PAGE
- Bécue M., Pagès J., Pardo C., Contingency table with a double partition on rows and columns. Visualization and comparison of the partial and global structures. <http://conferences.telecom-bretagne.eu/asmda2005/IMG/pdf/proceedings/355.pdf>
- Bécue M., Pagès J., Multiple factor analysis and clustering of a mixture of quantitative categorical and frequency data. Computational Statistics & Data Analysis 52, 3255 -3268, (2008)
- Benzécri J., Analyse des Données, Vol1: Analyse des correspondances (Data analyse, vol. 1: Correspondence analyse). Dunod, Paris. (1973)
- Benzécri J., Analyse de l'inertie intraclasse par l'analyse d'un tableau de correspondances. Les Cahiers de l'Analyse des Donn'ees, 8(3):351-358, (1983)

Carrasco L., Normas de vectores y matrices. Computación Científica I, Guía de Ejercicios No 2 Valparaíso, 14-04, (2004)

Carroll, J.D., Generalization of canonical correlation analysis to three or more sets of variables. Proceedings of the 76th Convention of the American Psychology Association 3, 227-228, (1968)

Christensen, R. (1990). LOG-LINEAR MODELS. Springer Verlag. New York. (1990)

Coppi, Bolasco. Multiway data analysis. Nort-Hollan. (1989)

Dazy F., Le Barzic J., L'Analyse Des Dones Évolutives. Methode est Applications. Éditions Technip, París. (1996)

Djahuri M., Operators averaging and its application in qualitative variables selection. JMS Vol. 3 No. 1, hal. 41-49, (1998)

Escoufier Y., Opérateur associé à un tableau de données. Annales de l'INSEE, n°22-23,(1976)

Escofier, B. y Pagès, J., Análisis factoriales simples y múltiples. Objetivos, métodos e interpretación. Servicio editorial de la Universidad de País Vasco, (1992)

Fernández A., Modroño H., Calvo L. ACM y Statis dual ponderado. Dos técnicas complementarias para analizar una visión de la cultura de la Universidad. Estadística Española. Vol. 46, Núm. 156, págs. 205-228, (2004)

Foucart T., Une nouvelle aproche de la méthode STATIS. Revue de Statistique Appliquée, Tome 31, n° 2 P. 61-75, (1983)

Greenacre M., Correspondence Analysis in Practice. Academic Press. (1993)

Goitisoló B., El Análisis Simultáneo. Propuesta y Aplicación de un Nuevo Método de Análisis Factorial de Tablas de Contingencia. Tesis Doctoral. (2002)

Gower J., Generalized Procrustes Analysis, Psychometrika 40(1), 33-51, (1975)

- Israëls A., Eigenvalues Techniques for qualitative data. DSWO Press, Leiden (1987)
- Kettenring, J.R., Canonical analysis of several sets of variables. *Biometrika*, vol. 58, n° 3. (1976)
- Kiers, Comparison of “ANGLO-SAXON” and French Three mode methods. *Statistique et Analyse de dones*. Vol. 13 n° 3 p. 14-32, (1988)
- Lavit C., *Analyse Conjointe de Tableaux Quantitatifs*. Editotial Masson. (1988)
- L’Hermier des Plantes, H. Structuration des tableaux à trois indices de la statistique: théorie et application d’une méthode d’analyse conjointe. Tesis de doctorado, Université des Sciences et Techniques du Languedoc, Montpellier. (1976)
- Ramírez G., Vásquez M., Apuntes de clases y materiales presentados en las cátedras de Postgrado Estadística U.C.V., (2003)
- Rodríguez, Galindo-Villardón, Vicente-Villardón. Comparison and integration of subspaces from a biplot perspective JSPI., (2001)
- Robert P., Escoufier Y., A Unifying Tool for Linear Multivariate Statistical Methods: The RV-Coefficient. *Appl. Statist.*, 25, N°. 3, p.257, (1976)
- Torres M., Impacto de la crisis venezolana en los patrones alimentarios en la población infantil. Tesis Doctoral, Doctorado en Ciencias Sociales, UCV, (2005)
- Tucker L., An Inter-Battery Method of Factor analysis. *Psychometrika* 23(2), (1958)
- Vivien M., Sabatier R., A generalization of STATIS-ACT strategy: DO-ACT for two multiblocks tables. *Computational Statistics & Data Analysis*. 46, 155-171, (2004)

ANEXOS

Anexo 5.1 Programa para simular los datos

Sintaxis para generar datos simulados correspondientes a una muestra con distribución normal multivariante. Los datos corresponden a la caracterización de $n=300$ individuos, según 15 variables cuantitativas, con una estructura de correlación como se especifica a continuación:

- a) La estructura de correlación entre las cinco primeras variables X_1, \dots, X_5 es nula.
- b) La estructura de correlaciones entre las variables X_6, X_7, \dots, X_{10} es moderada.
- c) La estructura de correlación entre las variables $X_{11}, X_{12}, \dots, X_{15}$ es muy fuerte. Matriz de correlación R: Anexo 2.

```
clear;
x = randn(300,15);
x(:,11) = sum(x,2);
x(:,12) = sum(x,2);
x(:,13) = sum(x,2);
x(:,14) = sum(x,2);
x(:,15) = sum(x,2);
x(:,6) = sum(x,2);
[r,p] = corrcoef(x); % Compute sample correlation and p-values.
[i,j] = find(p<0.05); % Find significant correlations.
end;
```

Con esta sintaxis se genera una matriz de orden 300×15 cuyas filas denotan 300 individuos caracterizados por 15 variables, con la siguiente matriz de correlación R

Anexo 5.3 Programa en Matlab para realizar el ACB. Datos de entrada: la tabla de contingencia múltiple

```
clear;
F=[---];
Dx=inv(sqrt(diag(sum(F/(5*5*157)))));
Dy=inv(sqrt(diag(sum(F'/(5*5*157)))));
YY=Dy*(F/(5*5*157))*Dx*Dx*(F'/(5*5*157))*Dy;
[U,D,V]=svd(YY);
ChiTotal=trace(YY)-1;
CHI=sum(diag(D))-1;
End.
```

Anexo 5.4 Programa en Matlab para realizar la Adaptación del STATIS

Los programas son particulares para cada caso presentado en la Tesis. El siguiente es diseñado para cuatro conjuntos de variables diferenciados por el número de variables y de categorías observados sobre 256 unidades de análisis.

```
clear;
M=[...]
F=mat2cell(M,256,[15 21 24 12]);
Oz1=F{1,1};
Oz2=F{1,2};
Oz3=F{1,3};
Oy=F{1,4};
I=eye(256);
Dt=(1/(256))*I;
Mz1=inv(diag(sum(Oz1'*Dt*Oz1)));
Mz2=inv(diag(sum(Oz2'*Dt*Oz2)));
Mz3=inv(diag(sum(Oz3'*Dt*Oz3)));
My=inv(diag(sum(Oy'*Dt*Oy)));
Wz1=Oz1*Mz1*Oz1';
Wz2=Oz2*Mz2*Oz2';
Wz3=Oz3*Mz3*Oz3';
Wy=Oy*My*Oy';
Wz1Wz1=trace(Dt*Wz1*Dt*Wz1);
Wz1Wz2=trace(Dt*Wz1*Dt*Wz2);
Wz1Wz3=trace(Dt*Wz1*Dt*Wz3);
Wz1Wy=trace(Dt*Wz1*Dt*Wy);
Wz2Wz2=trace(Dt*Wz2*Dt*Wz2);
```

```

Wz2Wz3=trace(Dt*Wz2*Dt*Wz3);
Wz2Wy=trace(Dt*Wz2*Dt*Wy);
Wz3Wz3=trace(Dt*Wz3*Dt*Wz3);
Wz3Wy=trace(Dt*Wz3*Dt*Wy);
WyWy=trace(Dt*Wy*Dt*Wy);
S=[Wz1Wz1 Wz1Wz2 Wz1Wz3 Wz1Wy;
Wz1Wz2' Wz2Wz2 Wz2Wz3 Wz2Wy;
Wz1Wz3' Wz2Wz3' Wz3Wz3 Wz3Wy;
Wz1Wy' Wz2Wy' Wz3Wy' WyWy];
[US,DS,VS]=svd(S);
chi1=trace(S);
chi2=sum(diag(DS));
Ud=mat2cell(US,4,[1 1 1 1]);
U1=Ud{1,1};
U2=Ud{1,2};
U11=-1*U1;
a=sqrt(DS(1,1));
b=sqrt(DS(2,2));
lambda=inv(a);
alpha=lambda*(sqrt(S(1,1))+sqrt(S(2,2))+sqrt(S(3,3))+sqrt(S(4,4)));
W=((alpha*U11(1,1)*Wz1)+(alpha*U11(2,1)*Wz2)+(alpha*U11(3,1)*Wz3)+(alpha*U11(4,1)*Wy))/4;
WW=trace(Dt*W*Dt*W);
COORDW=sqrt(WW);
XS=[a*U11
COORDW];
YS=[b*U2
0];
% gscatter(XS,YS);
dz1z2=sqrt(Wz1Wz1+Wz2Wz2-(2*Wz1Wz2));
dz1z3=sqrt(Wz1Wz1+Wz3Wz3-(2*Wz1Wz3));
dz1y=sqrt(Wz1Wz1+WyWy-(2*Wz1Wy));
dz2z3=sqrt(Wz2Wz2+Wz3Wz3-(2*Wz2Wz3));
dz2y=sqrt(Wz2Wz2+WyWy-(2*Wz2Wy));
dz3y=sqrt(Wz3Wz3+WyWy-(2*Wz3Wy));
dez1z2=sqrt(((XS(1,1)-XS(2,1))^2)+((YS(1,1)-YS(2,1))^2));
dez1z3=sqrt(((XS(1,1)-XS(3,1))^2)+((YS(1,1)-YS(3,1))^2));
dez1y=sqrt(((XS(1,1)-XS(4,1))^2)+((YS(1,1)-YS(4,1))^2));
dez2z3=sqrt(((XS(2,1)-XS(3,1))^2)+((YS(2,1)-YS(3,1))^2));
dez2y=sqrt(((XS(2,1)-XS(4,1))^2)+((YS(2,1)-YS(4,1))^2));
dez3y=sqrt(((XS(3,1)-XS(4,1))^2)+((YS(3,1)-YS(4,1))^2));
p11=(dez1z2/dz1z2)*100;
p12=(dez1z3/dz1z3)*100;
p13=(dez1y/dz1y)*100;

```

```

p14=(dez2z3/dz2z3)*100;
p15=(dez2y/dz2y)*100;
p16=(dez3y/dz3y)*100;
[UW,DW,VW]=svd(W);
E=mat2cell(UW,256,[1 1 1 253]);
E1=E{1,1};
E2=E{1,2};
E3=E{1,3};
aw=-sqrt(DW(2,2));
bw=sqrt(DW(3,3));
XW=aw*E2;
YW=bw*E3;
gscatter(XW,YW);
X1=inv(aw)*Wz1*E2;
X2=inv(aw)*Wz2*E2;
X3=inv(aw)*Wz3*E2;
X4=inv(aw)*Wy*E2;
Y1=inv(bw)*Wz1*E3;
Y2=inv(bw)*Wz2*E3;
Y3=inv(bw)*Wz3*E3;
Y4=inv(bw)*Wy*E3;
Xt=[XW
    X1
    X2
    X3
    X4];
Yt=[YW
    Y1
    Y2
    Y3
    Y4];
G1=ones(256,1);
G=[G1;
   2*G1;
   3*G1;
   4*G1;
   5*G1];
% gscatter(Xt,Yt,G);

```