# Metodología

## Notación

La tabla 1 contiene elementos, representación y ejemplos de la manera como se presentan los elementos algebraicos abordados en la metodología.

Tabla

Descripción generada automáticamente

A lo largo del artículo se utilizarán letras para hacer referencia a parámetros necesarios, se los enuncia a continuación en la tabla 2:

Tabla

Descripción generada automáticamente

## Bases metodológicas

### Análisis de Correspondencias Simple

El tratamiento multivariante de variables cualitativas requiere un proceso metodológico distinto, uno de los más representativos es el Análisis de Correspondencias (Benzecri. 1973). Según (López 2004), este análisis implica estudios de similaridad o disimilaridad entre categorías, se debe cuantificar la diferencia o distancia entre ellas sumando las diferencias cuadráticas relativas entre las frecuencias de las distribuciones de las variables analizadas, lo que conduce al concepto de la . Así, el análisis de correspondencias puede considerase como un análisis de componentes principales aplicado a variables cualitativas que, al no poder utilizar correlaciones, se basa en la distancia no euclídea de la.

En el enfoque francés del análisis de correspondencias, que se caracteriza por el énfasis en la geometría, el análisis de una tabla cruzada se llama análisis de correspondencias (CA) y el análisis de una colección de matrices indicadoras, se denomina análisis de correspondencias múltiples (MCA) (Michailidis and Leeuw 1998). En contextos anglosajones, el MCA es conocido como Análisis de Homogeneidad o Escalamiento Dual, especialmente en psicometría.

### Análisis de Correspondencias Múltiples (MCA)

El análisis de correspondencias múltiples (MCA) es una generalización del análisis de correspondencias simple o binario, donde se incluyen más variables cualitativas. Se obtiene al realizar el análisis de correspondencias simple a una tabla disyuntiva completa, conocida como la tabla de Burt.

Se tiene una matriz de datos con variables cualitativas, cada una con categorías ( >1). En el ejemplo que se desarrolla para esta investigación, se dispone de una base de datos (*Datak10Contaminated*) constituida por 10 tablas, cada una tiene 10 variables y cada variable, 3 categorías (Alto, Medio y Bajo).

Tabla

Descripción generada automáticamente

Esta matriz es equivalente a la matriz disyuntiva , que desglosa las variables en cada una de sus modalidades y se registra la ocurrencia de eventos de forma binaria.

Tabla

Descripción generada automáticamente

La tabla de Burt viene dada por:

**B = Z′Z**

La construcción de la matriz de Burt se da por la superposición de tablas. En las tablas ubicadas en la diagonal se encuentran matrices diagonales que contienen las frecuencias marginales de cada una de las variables. Fuera de la diagonal de la matriz de Burt se encuentran las tablas cruzadas por pares de variables.  
Para realizar el análisis de correspondencias múltiples se parte de la matriz de Burt, obtenida con la ecuación . Esta matriz está formada por las frecuencias absolutas, éstas se transforman en frecuencias relativas, dividiendo los valores de la matriz por la frecuencia total, dando lugar a una nueva matriz que se denominará **P**.

Se obtienen las marginales de las filas *(mf)* y de las columnas *(mc)* de la matriz **P** (Tabla ). A estos vectores se los conoce también como *Masas de fila y columna*, respectivamente.

Se obtiene la matriz de residuos estandarizados **S**.

**S**=**D** −12 (**P**−**mf mc**′) **D** −12

donde:

* es una matriz diagonal que contiene las masas de las filas.
* es una matriz diagonal que contiene las masas de las columnas

Tabla

Descripción generada automáticamente

Se aplica descomposición singular (SVD) a la matriz **S** (Ecuación ):

**S** = **UDV**′

donde:

* y son matrices ortogonales.
* es una matriz diagonal que contiene los valores singulares.

Para encontrar las coordenadas estandarizadas se aplica lo siguiente:

Texto, Logotipo

Descripción generada automáticamente

Logotipo

Descripción generada automáticamente

Para los fines necesarios, se utilizará las coordenadas de las columnas (Tabla 8).

Tabla

Descripción generada automáticamente

### Generalización a tablas

Si se tienen *k* tablas, con la misma estructura de la tabla , como se visualiza en la figura 1, se aborda el enfoque del análisis factorial múltiple (MFA). Escofier and Pagès (1994) indica que el MFA utiliza análisis de correspondencias múltiples cuando se trata de variables cualitativas. El procedimiento implica la realización de un MCA por cada tabla y dividirlo por su primer valor propio con la finalidad de obtener grupos normalizados. Posteriormente se consideran todas las tablas y se realiza un MCA global.

Diagrama, Tabla

Descripción generada automáticamente

La generalización a *k* tablas del procedimiento del MCA, se presenta en la Figura 2

Diagrama

Descripción generada automáticamente

Se llama a cada tabla de coordenadas. Con la finalidad de detectar la magnitud de las variables latentes, su aporte neto a las variables, se aplica valor absoluto a los elementos de la matriz .

### Aporte del Análisis Factorial Múltiple (MFA)

Una vez que se tienen las coordenadas de las columnas, se procede a realizar la normalización mediante un procedimiento similar al que se utiliza en el Análisis Factorial Múltiple. Sea el primer valor propio obtenido de la descomposición singular de la k-ésima tabla C. Se normaliza la tabla multiplicándola por . Con esto se obtiene la tabla , que corresponde a la tabla de coordenadas normalizadas.  
Individualmente, para el caso de la matriz k, se tendría la siguiente expresión.

Diagrama, Esquemático

Descripción generada automáticamente

Aglomerando las matrices normalizadas en una sola, se tiene la matriz . Esta contiene todos los elementos de las *k* tablas.

Imagen que contiene Forma

Descripción generada automáticamente

La normalización que realiza el MFA se encarga de ponderar las *k* tablas, con el objetivo de evitar alguna descompensación al momento de realizar el análisis conjunto de las tablas.

## Gráfico de control T2Qv

### Obtención del gráfico de control

Para definir el gráfico de control Hotelling se deben tomar las siguientes consideraciones:

* La tabla (Ecuación 7) se denomina Consenso, sirve como referente para el escenario *bajo control.*
* El estadístico Hotelling se calcula con los vectores de media, sin embargo, se espera incluir la variabilidad de todas las tablas, incluyendo las que estén potencialmente contaminadas, es decir, las que se representarían como puntos fuera de control en el gráfico T2Qv, esto desemboca en un problema al momento de definir la matriz consenso, dado que se espera que esta capture el comportamiento “en control”. Otros autores en su búsqueda de la matriz consenso, resuelven esto omitiendo las tablas que difieren del comportamiento normal, sin embargo, otra alternativa es tomar conceptos de robustez, y considerar el vector de medianas, mismo que no se deja influenciar por los valores atípicos.
* De la matriz consenso se obtiene y .
* Cada matriz tiene el mismo número de filas (n) y columnas (p) (individuos y variables).
* El vector de medianas está atado a la tabla , es decir, el gráfico de control estará en función de las diferencias entre las matrices y la matriz consenso .
* Las matrices siguen una distribución normal multivariante con vector de centralidad y matriz de covarianzas .

El estadístico viene dado por:

Tomando en cuenta las consideraciones previas, se obtiene el estadístico

Se sabe que, bajo control, el se distribuye como una Chi-cuadrado con grados de libertad . En este caso se puede aplicar este principio, ya que se utiliza la matriz consenso (), que representa al escenario bajo control.

Dado que este gráfico de control está basado en distancias de Mahalanobis ponderadas, sólo tiene límite de control superior. Este viene dado por la ecuación 9.

Texto

Descripción generada automáticamente

donde es el número de dimensiones y es la significancia predeterminada, inicialmente se considera 0.0027.

### Interpretación de puntos fuera de control

El gráfico multivariante de Hotelling para variables cualitativas es capaz de señalar que el proceso salió de control, pero no permite reconocer el momento ni las causas por las que ocurrió esto. Es obvio que, más allá de reconocer el estado del proceso, interesa saber cuándo y por qué salió de control. Es importante tener en cuenta que cada punto representado en el gráfico de Hotelling representa a una tabla (muestra), constituida por un grupo de individuos (observaciones) y p variables que pueden tener muchas categorías, algunas de éstas pueden mostrar un comportamiento anómalo. Por consiguiente, es necesario analizar con detenimiento que está pasando con los datos de las tablas reportadas.

Este análisis se realiza comparando la ubicación de los puntos que representan las categorías de las variables en el MCA de la tabla consenso y la ubicación de los puntos en los gráficos MCA de cada tabla reportada como fuera de control. Las categorías que están incidiendo en el estado fuera de control son aquellas cuya ubicación en ambas tablas comparadas muestra diferencias importantes. Para cuantificar la magnitud del comportamiento anómalo de estas categorías se calcula las distancias Chi-cuadrado entre las masas de las columnas de la tabla reportada como fuera de control y las de la tabla consenso, tomada como referente. Mientras mayor es el valor del estadístico, mayor es su incidencia en el desplazamiento de la media del proceso que, finalmente, pueden llevarlo a un estado fuera de control.

De esta manera, la metodología propuesta en esta investigación permite explicar cuándo y por qué el proceso salió de control.

**Tabla de diferencias Chi Cuadrado**

Consiste en un gráfico interactivo de barras que representa a las distancias χ^2 entre las masas de columna de las variables de la tabla consenso y la tabla i, que podría ser la que está fuera de control u otra que se quisiera analizar. Las barras que denotan mayor altura son las que más están incidiendo en la variación de la tendencia central del proceso y, por consiguiente, su salida de control. Para proporcionar mayor detalle, este gráfico interactivo también ofrece, mediante un gráfico circular anidado, una representación de la distribución de las categorías de la variable observada, correspondiente a la tabla i.

**5. Análisis de sensibilidad**

Como se ha mencionado, en el gráfico T2Qv un punto fuera de control se interpreta como una tabla (ki) que incluye una cantidad o una proporción de variables contaminadas. En estos casos, se espera que los puntos en el gráfico T2Qv generalicen el comportamiento de estas diferencias en su distribución y así se supere el límite de control superior (UCL). La ubicación de este límite de control varía en función del número de dimensiones que se representen, así, cuando es alto se logra un desempeño óptimo, mientras que, se introduce inestabilidad y se pierde confiabilidad en los resultados al disminuir el número de dimensiones de entre las que se puede representar.

El gráfico de control propuesto es capaz de detectar un punto fuera de control, aún con un bajo número de variables contaminadas, cuando se trabaja con un alto número de dimensiones. Se recomienda p - 1, tal que p es el número total de dimensiones de la matriz inicial (Tabla 3). Cuando se disminuye el número de dimensiones también disminuye la altura del límite de control superior (UCL), en consecuencia, se incrementa el número de puntos fuera de control, aunque no necesariamente las variables expresen diferencias significativas en su valores, crece la probabilidad de obtener error tipo I.

Gráfico, Gráfico de superficie

Descripción generada automáticamentePor consiguiente, la pregunta que surge es hasta cuántas dimensiones se puede disminuir en el análisis sin perder confiabilidad en el resultado. La importancia de esta pregunta radica en la necesidad de disponer un gráfico confiable, que identifique puntos fuera de control aún si se ha aplicado a los datos una técnica de una reducción de dimensiones, sin caer en casos de falso positivo.

Gráfico, Gráfico de barras

Descripción generada automáticamenteGráfico, Histograma

Descripción generada automáticamente

Figura 17. Curvas de nivel y superficie de respuesta obtenidas con el gráfico T2Qv

El análisis de sensibilidad utiliza curvas de nivel y superficies de respuesta (figura 17) para representar el número de puntos fuera de control, considerando el porcentaje de variables contaminadas de la ki tabla y el número de dimensiones representadas. Los datos de prueba utilizados en el modelo se registran en 10 tablas, cada una de ellas incluye 10 variables y cada variable tiene tres categorías: alto, medio y bajo. La tabla 10 tiene una distribución diferente de las demás, esta es la tabla contaminada.

Se observa que el modelo es capaz de identificar un punto fuera de control trabajando con 9 dimensiones (p-1), aún con un porcentaje bajo de variables contaminadas. Cuando el número de dimensiones disminuye a 8 y el porcentaje de variables contaminadas es cercano a 100%, detecta correctamente 1 punto fuera de control. Se observa además que cuando el número de dimensiones es menor se pierde estabilidad y se mitiga la potencia de la prueba. En consecuencia, el análisis de sensibilidad ratifica que el gráfico de control T2Qv tiene un buen rendimiento cuando trabaja con altas dimensiones.