Máster en Ingeniería Computacional y Matemática

Trabajo Final de Máster

Análisis multivariante para la detección de patrones de calidad de vida de los hogares: El caso de la provincia de Loja en 2014

Franco Hernán Salcedo López
Máster en Ingeniería Computacional y Matemática
Análisis Multivariante de Datos

Agusti Solanas
Profesor responsable de la asignatura

Junio 2017

C) Copyright

Dedicada a las personas que en mi vida representan la fuerza e impulso para seguir adelante: mi madre, mis hermanos, mis hijas, mi esposa.

AGRADECIMIENTO

Con enorme afecto quiero agradecer la ayuda del profesor Agustí Solanas por el aprendizaje de los fundamentos de las técnicas de análisis multivariante, desde que tuve la oportunidad de tomar la unidad de análisis multivariante, así como la revisión y conducción del Trabajo Final de Máster.

Agradecer a todo el personal docente y administrativo de la Universitat Rovira i Virgili y de la Universidad Oberta de Catalunya del Máster en Ingeniería Computacional y Matemática, por todo el apoyo y los conocimientos adquiridos.

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Análisis multivariante para la detección de patrones de la calidad de vida de los hogares: El caso de la provincia de Loja en 2014*
Nombre del autor:	*Franco Hernán Salcedo López*
Nombre del consultor/a:	*Agusti Solanas*
Nombre del PRA:	
Fecha de entrega (mm/aaaa):	06/2017
Titulación::	*Máster en Ingeniería Computacional y Matemática*
Área del Trabajo Final:	*Análisis Multivariante de Datos*
Idioma del trabajo:	*Español*
Palabras clave	*Estadística multivariante, Detección de patrones, políticas públicas.*

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados i conclusiones del trabajo.*

El desarrollo de un país se relaciona con los indicadores de las condiciones de vida de sus pobladores. Estos indicadores son de mucho interés para los gobiernos, ya que les permite formular mejoras en las políticas públicas para atender y mejorar las condiciones y calidad de vida de sus habitantes.

Ante esta realidad, la presente memoria desarrolla e identifica patrones específicos que permiten evaluar y determinar el nivel de percepción de las condiciones de vida de los hogares, de la provincia de Loja, en base a los resultados de las Encuestas de las Condiciones de Vida ECV6R 2013-2014, realizada por el INEC (Instituto Nacional de

Estadísticas y Censos – Ecuador) y aplicando técnicas de análisis multivariante como el Análisis de Correspondencias Múltiples (ACM) en la primera etapa, que permitió reducir la dimensión de los datos con la mínima pérdida de información, identificando las posibles variables latentes, que facilitan la interpretación de los datos y, el Análisis Factorial Exploratorio en la segunda para poder inferir si las propiedades de la reducción de la dimensión en los datos puede extenderse o generalizarse a la población de donde provienen los datos originales.

Aplicando la técnica de ACM, se redujeron de 17 variables observables a 4, obteniéndose una inercia total de 2.4 y obteniéndose una inercia acumulada del 49.1% entre las 4 variables principales, que es un valor aceptable en estudios de ciencias sociales.

La prueba de significancia de la matriz de correlación policórica, arrojó un p_valor de 1 > 0.05, lo que sirvió para realizar el Análisis Factorial Exploratorio, el mismo que determinó dos factores relevantes que explican un 68% de la varianza total. Para mejorar la interpretación de los factores obtenidos, se hizo necesario realizar la rotación de factores con el método de rotación ortogonal "varimax".

Se sugiere que se extienda el estudio a través de técnicas como el Análisis Factorial Confirmatorio y Modelado de Ecuaciones Estructurales, a fin de poder encontrar relaciones entre los factores latentes encontrados con el Análisis Factorial Exploratorio que permita modelar de forma más significativa la calidad de vida de los hogares de la provincia de Loja.

Abstract (in English, 250 words or less):

The development of a country is related to the indicators of the living conditions of its inhabitants. These indicators become of great interest to governments, since they allow them to formulate improvements in public policies to take care of and improve the conditions and quality of life of its inhabitants.

Given this reality, this master thesis develops and identifies specific patterns that allow the evaluation and determination of the level of perception of the living conditions of the homes of the province of Loja, based on the results of the Surveys of Living Conditions ECV6R 2013-2014, carried out by the INEC (National Institute of Statistics and Censuses - Ecuador) and applying techniques of multivariate analysis such as Multiple Correspondence Analysis (ACM) in the first stage that allowed reducing the dimension of data with minimal loss of information, identifying the possible latent variables that facilitate the interpretation of the data and, the Exploratory Factor Analysis, in the second stage to be able to infer whether of the reduction of the dimension in the data can be extended or generalized to the population where they come from the original data.

Applying the ACM technique, the dataset was reduced from 17 observable variables to 4, obtaining a total inertia of 2.4 and an accumulated inertia of 49.1% among the 4 main variables, which is an acceptable value in social science studies.

The test of significance of the polychoric correlation matrix yielded a p_value of 1> 0.05, which served to perform the Exploratory Factor Analysis, which determined two relevant factors explaining 68% of the total variance. To improve the interpretation of the obtained factors, it was necessary to perform the factor rotation with the orthogonal rotation method "varimax".

It is suggested for the study to be expanded by applying techniques such as Confirmatory Factor Analysis and Modeling of Structural Equations, in order to be able to find relationships between the latent factors found with the Exploratory Factor Analysis, which enables the creation of a more meaningful model of the quality of life of the homes of the province of Loja.

Índice

Lista de Tablas

Lista de Figuras

1. Introducción.

1.1 Contexto y justificación del trabajo

El nivel de vida de una población se relaciona con el grado de satisfacción de las necesidades humanas fundamentales que alcanza a cubrir. Estas necesidades incluyen las de subsistencia, protección, afecto, entendimiento, participación, ocio, creación, identidad y libertad. Para satisfacer estas necesidades propias de todo ser humano existen ciertos factores que varían de acuerdo a las características sociales, económicas y culturales propias de cada realidad regional y local. La desigualdad en la distribución de los recursos económicos y sociales en una región define los niveles de vida para la población, en la que muchos ni siquiera alcanzan a satisfacer las necesidades básicas de subsistencia como la nutrición, protección, ambiente, salud, etc., que definen el nivel de vida que alcanzan por la interacción de factores económicos, sociales y políticos [2].

El crecimiento de un país se mide por las condiciones de vida de la población, que inciden directamente en la calidad de vida de los pobladores. Estas mediciones son de mucho interés para los gobiernos ya que les permite formular mejoras en las políticas públicas que tienda a mejorar las condiciones y calidad de vida, para alcanzar mayor desarrollo en la región [1].

Para una sociedad es importante saber qué nivel de bienestar psicosocial tenemos frente a las condiciones de vida, y ello relacionarlo con la política pública nacional, a fin de que esta información sea utilizada por los involucrados en las políticas públicas del gobierno ecuatoriano, y ejecutar acciones que permitan elevar las condiciones de vida especialmente en las provincias fronterizas, como el caso de la provincia de Loja, en donde la situación económica, social y política difiere significativamente del resto de provincias, por ejemplo, falta de empresas productivas, deficiencias técnicas y tecnológicas del talento humano causado por la falta de pertinencia de la educación superior, bajos niveles de inversión en investigación más desarrollo (I+D), bajos niveles

de acceso a las tecnologías de información y comunicación (TIC) y falta de instrumentos financieros adecuados para actividades productivas incluyendo las diferencias en el sector agroindustrial que, entre otros factores, se debe a que la mayor parte del territorio de la provincia de Loja no es apto para actividades agroproductivas [15].

Ante esta realidad, se requiere desarrollar e identificar patrones específicos que permitan evaluar y determinar el nivel de percepción de las condiciones de vida de los hogares, de la provincia de Loja, en base a los resultados de las Encuestas de las Condiciones de Vida ECV6R 2013-2014, realizada por el INEC (Instituto Nacional de Estadísticas y Censos – Ecuador)[1]; encuesta que estimó el nivel de extrema pobreza y pobreza, de acuerdo al nivel de ingreso o consumo. Con esta información generada por el INEC, y previa solicitud para fines de estudio, se pudieron obtener los datos que permitirán descubrir información relevante para poder determinar la percepción del nivel de vida en los hogares de la provincia de Loja y sus cantones.

Determinar patrones del nivel de vida de los hogares utilizando técnicas multivariantes permite a los involucrados en las políticas públicas de los gobiernos tomar las mejores decisiones, esclarecer la distribución real de los datos, relacionar y predecir comportamientos que permanecen ocultos y alcanzar un mejor entendimiento de las condiciones de vida de una sociedad.

1.2 Objetivos del trabajo

General

Aplicar técnicas de análisis de datos multivariantes para la detección de patrones en las condiciones de vida de los hogares de la Provincia de Loja en el año 2014.

[1] *http://www.ecuadorencifras.gob.ec//documentos/web-inec/ECV/ECV_2015/*

Específicos

Realizar un análisis exploratorio de datos, para describir e interpretar los resultados de las Encuestas de las Condiciones de Vida ECV6R 2013-2014 en la Provincia de Loja en el año 2014.

Representar cada uno de los valores de las Encuestas de las Condiciones de Vida ECV6R 2013-2014 en la Provincia de Loja en el año 2014 en un plano donde la posición relativa de los datos refleje el grado de asociación entre cada uno de los datos de estudio.

Identificar un conjunto de dimensiones o características que se encuentren latentes dentro del conjunto de variables originales, que permitan describir e inferir resultados generalizados, a través del Análisis Multifactorial de las Encuestas de las Condiciones de Vida ECV6R 2013-2014 en la Provincia de Loja en el año 2014.

1.3 Enfoque y método seguido

La metodología que se presenta en el presente trabajo de fin de máster sigue un enfoque de tipo descriptivo y exploratorio, con información cualitativa, orientado a la descripción e interpretación de los resultados de los datos de las Encuestas de las Condiciones de Vida ECV6R 2013-2014 en la Provincia de Loja en el año 2014.

Para alcanzar los objetivos propuestos se utilizará el Análisis de Correspondencias Múltiples (ACM), que permitirá reducir la dimensión de los datos con la mínima pérdida de información, identificando las posibles variables latentes, que facilitan la interpretación de los datos; a continuación se aplicará el Análisis Factorial Exploratorio (AFE), a fin de poder inferir si las propiedades de la reducción de la dimensión en los datos, puede extenderse o generalizarse a la población de donde provienen los datos originales.

En el presente trabajo de fin de máster se analizan datos relacionadas con aspectos socioeconómicos, sociodemográficos, bienestar psicosocial, gastos, ingresos, percepción del nivel de vida, capital social, entre otros. La información obtenida se considera de tipo secundaria, la misma que es el resultado de la Encuestas de las Condiciones de Vida ECV6R 2013-2014, cuyos resultados toman valores de tipo binarios, de tipo categórico en función de una escala de posibles respuestas.

Como herramienta tecnológica para el análisis, descripción e interpretación de los datos, se usará el software RStudio versión 3.3.3, utilizando varias librerías, entre ellas, *ca, dudi.acm, factanal, y paquetes polycor, psych*, los que permitirán describir, analizar e interpretar los datos en estudio.

1.4 Planificación del trabajo

Para alcanzar los objetivos propuestos en el presente TFM, se empleó el análisis sistemático de los datos de la Encuestas de las Condiciones de Vida ECV6R 2013-2014, obtenidos por el Instituto Nacional de Estadísticas y Censos (INEC), con la planificación propuesta:

Tabla 1. ***Planificación de actividades del TFM***

Hasta la primera semana de Enero de 2017	Descripción del Problema
Hasta la tercera semana de Enero de 2017	Objetivos
Hasta la tercera semana de Febrero de 2017	Metodología Revisión del Literatura
Hasta la tercera semana de Marzo de 2017	Descripción de lo que se pretende alcanzar. Identificación, selección y preparación de los Datos.
Hasta la primera semana de Abril de 2017	Selección de herramientas y técnicas adecuadas de Análisis Multivariante y Análisis Factorial, para alcanzar los objetivos propuestos.
Hasta la primera semana de Mayo de 2017	Análisis e Interpretación de los factores encontrados.
Hasta la segunda semana de Junio de 2017	Difusión del conocimiento descubierto. Conclusiones

Nota: Metas para la culminación del TFM.

1.5 Breve resumen de los resultados obtenidos

Aplicar el Análisis de Correspondencias Múltiples permitió reducir de 17 variables observables a 4, con una inercia total de 2.4 obteniéndose una inercia acumulada del 49.1% entre las 4 variables principales, que es un valor aceptable en estudios de ciencias sociales.

La prueba de significancia de la matriz de correlación policórica, arrojó un p_valor de 1 > 0.05, lo que sirvió para realizar el Análisis Factorial Exploratorio, el mismo que determinó dos factores relevantes que explican un 68%, de la varianza total.

Las técnicas y herramientas utilizadas para el desarrollo del presente estudio, contribuyeron de manera significativa a alcanzar los objetivos propuestos en el presente TFM, permitiendo describir, analizar e interpretar los datos de la Sección 11 de la Encuestas de las Condiciones de Vida Sexta Ronda (ECV6R) realizada por el INEC-Ecuador, entre el período noviembre 2013 - octubre 2014.

1.6 Organización del trabajo de fin de máster

El contenido de la presente memoria se divide en tres grandes bloques, el primero relacionado con los ítems básicos de todo trabajo investigativo, el segundo con el desarrollo y obtención de los resultados de la aplicación de las técnicas y herramientas para realizar el análisis multivariante, y finalmente se describen los principales hallazgos y sugerencias en el tercer bloque.

Los tres bloques propuestos se han estructurado de la siguiente manera:

En el bloque uno se presenta la estructuración del contexto y justificación del trabajo, seguido de los objetivos y metodología aplicada, para concluir con la planificación de las actividades programadas.

El segundo bloque de la memoria empieza con la revisión de la literatura, en donde se describe las técnicas multivariantes aplicadas para alcanzar los objetivos propuestos. Se identifican y seleccionan los datos de estudio a partir de los cuales se pretende generar los objetivos propuestos. A continuación se seleccionan las técnicas multivariantes que se ajusten a la naturaleza de los datos de estudio que, en este caso, son de carácter categórico, para lo cual es recomendable aplicar el Análisis de Correspondencias Múltiples. Finalmente dentro de este bloque, se aplica la técnica de Análisis Factorial Exploratorio para determinar los factores latentes que subyacen a las variables observables. Todas las técnicas propuestas están implementadas con RStudio versión 3.3.3.

El tercer bloque termina con una descripción de los aspectos más relevantes del diseño y desarrollo de la investigación, considerando los resultados obtenidos. Se valora el logro de los objetivos planteados al inicio de la investigación y se analiza y critica la planificación y metodología propuesta en relación al alcance y determinación de los resultados propuestos inicialmente.

2. Desarrollo del TFM.

2.1 Revisión de la literatura

Aspectos de la calidad de vida

Uno de los componentes para medir el nivel de vida de una población es a través de las expectativas del bienestar psicosocial de las personas, como la autoestima, aspectos emocionales y autoeficacia para resolver los problemas cotidianos. [13].

Para una sociedad es importante saber qué nivel de bienestar psicosocial tenemos, frente a las condiciones de vida, y ello relacionarlo con la política pública nacional; a fin de que esta información sea utilizada por los involucrados en las políticas públicas del gobierno ecuatoriano, y ejecutar acciones que permitan elevar las condiciones de vida especialmente a las provincias fronterizas, como el caso de la provincia de Loja, en donde la situación económica, social y política difiere significativamente al resto de provincias. [15].

La definición de la calidad de vida como un término multidimensional asociado a las políticas sociales, asociado al bienestar, significa tener buenas condiciones de vida *objetivas* y un alto grado de *bienestar subjetivo*, que incluye satisfacción colectiva de necesidades alcanzadas a través de las políticas sociales junto a la satisfacción individual de las necesidades. [13].

La calidad de vida, contiene dos dimensiones: la primera vista desde una evaluación del nivel de vida basada en *indicadores objetivos;* y, la percepción individual, equiparada con el término *bienestar.* [13].

El término calidad de vida es un concepto multidimensional e incluye aspectos del bienestar y de las políticas sociales: materiales y no materiales, objetivos y subjetivos, individuales y colectivos. [13].

El término *calidad de vida* incluye dimensiones ambientales, psicosociales, además dimensiones tradicionales que miden el nivel de vida, lo demográfico, lo económico, lo propiamente social, lo cultural y lo político. [2].

La calidad de vida implica la conjunción de ideales, propósitos, necesidades básicas y recursos, que constituyen realidades y contextos en que las personas habitan y construyen espacios de relaciones que sirven de referentes comparativos frente a otros sujetos. [1].

Los niveles de vida *objetivos* pueden ser medidos cuantitativamente; en cambio los aspectos *subjetivos* se relacionan con la percepción de las personas de la satisfacción de sus necesidades y para medirlo se emplean métodos cualitativos. [2].

Análisis multivariante de datos

El análisis multivariante es el conjunto de métodos o técnicas diseñadas para el análisis e interpretación de la información contenida en un conjunto de variables sin perder la interacción o grado en que se afectan unas con otras. [4]. Las técnicas de análisis multivariantes tienen aplicaciones en todos los campos de las ciencias: en la Biología, para resolver problemas de clasificación; en el Marketing y las Ciencias Sociales, para encontrar indicadores; en las Ciencias de la Computación, para resumir la información y diseñar sistemas de clasificación y reconocimiento de patrones. [14].

Dentro de las técnicas multivariadas desde la perspectiva de los objetivos de análisis y el tipo de datos obtenidos, se destacan:

1. Simplificación de la estructura de datos. Reducen el espacio de las variables en estudio, transformando algunas variables a una menor dimensión.

2. Clasificación. Considera los individuos y las variables dispersas en un multiespacio, tratando de ubicarlos espacialmente.

3. Interdependencia. Se estudia la interdependencia entre las variables. Examinándose desde la independencia total de las variables, hasta la dependencia de alguna con respecto al resto. Estas técnicas buscan el *cómo* y el *por qué* se relacionan o asocian un conjunto de variables. Entre éstas técnicas están: El Análisis de Componentes Principales, Análisis de Correspondencias, Análisis de Conglomerados, Escalamiento Multidimensional, Modelos de Ecuaciones Estructurales.

4. Dependencia. Técnica que permite encontrar la asociación entre dos conjuntos de variables, en donde uno se considera como la realización de mediciones dependientes de otro conjunto de variables. Entre estas técnicas están: Regresión Múltiple, Análisis Discriminante, Análisis de Correlación Canónica, Análisis de Varianza Multivariado, entre otros.

De acuerdo a los datos de estudio tomados de la Encuestas de las Condiciones de Vida ECV6R 2013-2014, en los que los datos son de carácter cualitativo, tanto binarios dicotómicos como categóricos multicotómicos, de tipo nominal como ordinal, de los cuáles se desea detectar patrones de nivel de vida de los hogares de la provincia de Loja, se requiere utilizar la técnica de Análisis de Correspondencias múltiples, para la reducción de información contenida en una tabla de contingencias, luego de lo cual para poder generalizar a la población de estudio, se utiliza la técnica de Análisis Factorial.

Análisis de correspondencias múltiples

El Análisis de Correspondencias (AC), es una técnica exploratoria multivariante que permite mostrar simultáneamente las puntuaciones de las categorías de las filas y columnas en una tabla de contingencias bidireccional como coordenadas de puntos en el espacio vectorial. El AC tiene como objetivo aclarar la relación entre las variables fila y columna de la tabla de contingencias y descubrir una explicación de baja dimensión para las posibles desviaciones de la independencia de esas variables. [11].

La tabla de contingencia, es una tabla en doble entrada en donde se recoge las frecuencias de aparición de dos o más variables cualitativas de dimensiones *I* filas x *J* columnas, que han sido obtenidas cruzando las I categorías de A con las J categorías de B, en donde f_{ij} es el número de veces en que aparece la intersección $A_i \cap B_j$, lo que genera la tabla de contingencia I x J. [3].

Tabla 2. ***Tabla de contingencia I x J***

	B_1	B_2	....	B_J	
A_1	f_{11}	f_{12}	....	f_{1J}	$f_{1.}$
A_2	f_{21}	f_{22}	....	f_{2J}	$f_{2.}$
⋮	⋮	⋮	.....	⋮	⋮
A_I	f_{I1}	f_{I2}	....	f_{IJ}	$f_{I.}$
	$f_{.1}$	$f_{.2}$	....	$f_{.J}$	n

Nota: Tabla que representa las frecuencias de aparición de dos variables categóricas.

En donde:

$n = \sum_{ij} f_{ij}$

$f_{i\bullet} = \sum_j f_{ij}$ frecuencia marginal de A*i*.

$f_{\bullet j} = \sum_j f_{ij}$ frecuencia marginal de B*j*.

Desde luego, la Tabla 2 es el resultado de manipular o computar los datos iniciales que se explican de forma binaria, en la que cada fila, solo una categoría tomara el valor de 1 que representa la modalidad seleccionada y tomara el valor de 0 para el resto de modalidades, como se muestra en la Tabla 3.

Tabla 3. ***Tabla de datos inicial***

	A_1	A_2	.....	A_I	B_1	B_2	.....	B_J
1	0	0	.....	1	1	0	.....	0
2	1	0	.....	0	0	1	.....	0
⋮	⋮	⋮	.....	⋮	⋮	⋮	.....	⋮
n	0	0	.....	1	1	0	.....	0

Nota: Tabla de datos inicial, con datos binarios.

Para el análisis de las Tablas 2 y 3 que representan el cruce de dos variables categóricas, el análisis de correspondencia que debería aplicarse es el Análisis de Correspondencia Simple (ACS); para tablas de mayor dimensión, el análisis de correspondencias que se aplica es el Análisis de Correspondencias Múltiples (ACM). Las variantes de estas técnicas son el escalamiento dual, el promedio recíproco, el mapeo perceptivo y el análisis del espacio social. En general el AC es aplicable cuando las variables son discretas con varias categorías y, por lo tanto, es muy adecuado para analizar grandes tablas de contingencias; aunque también se puede utilizar para variables continuas, que pueden ser segmentados en un número de rangos, pero esta discretización de una variable continua puede implicar cierta pérdida de información [11].

El AC se desarrolla sobre tablas de datos; si la tabla contiene frecuencias respecto a las modalidades de dos variables, entonces se debe aplicar el análisis de correspondencias binarias o simples; si la tabla de datos contiene información sobre varias variables, se aplica el análisis que se conoce como correspondencias múltiples [5].

Para aplicar la técnica del ACM, es necesario cuantificar los datos de las variables categóricas, asignándoles valores numéricos, de tal forma que los

individuos están próximos si tienen las mismas modalidades y, los individuos están alejados si no han respondido de la misma manera; estas proximidades permiten determinar las distancias entre individuos que matemáticamente se obtiene sumando las diferencias entre modalidades. [5].

Así entonces, si se desea obtener la distancia entre el individuo *i* con el individuo *i´* en la modalidad *k*, la distancia sería:

$$d^2{}_{i,i'} = \sum_{k=1}^{K} \frac{(x_{ik} - x_{i'k})^2}{I_k}$$

En donde:

x_{ik} = observación del individuo *i* en la modalidad *k*.
$x_{i'k}$ = observación del individuo *i´* en la modalidad *k*.
I_k = número de individuos que toman la modalidad *k*.

El ACM puede aplicarse sobre una tabla de contingencia, pero también se puede aplicar sobre una *matriz de Burt* B. Esta matriz B, es una matriz con características simétricas, que genera coordenadas principales en una escala reducida en comparación con la AC a la matriz binaria. [8].

Para comprender de manera parsimoniosa la técnica del AC, empecemos considerando una tabla de contingencia o matriz de datos $\mathbb{X}$, desde dos espacios vectoriales, el espacio fila ($\mathbb{R}^p$) y el espacio columna ($\mathbb{R}^n$), en donde un elemento de la matriz de datos $\mathbb{X}$, n_{ij} representa el número de individuos de la fila *i* y la columna *j*. [5].

Para encontrar el número total de individuos por fila se obtiene matemáticamente así:

$$n_{i.} = \sum_{j=1}^{p} n_{ij}, \ para\ i=1,\ldots\ldots,n.$$

De la misma manera, podemos obtener el número total de individuos por columna:

$$n_{.j} = \sum_{i=1}^{n} n_{ij},\ para\ j{=}1{,}\ldots\ldots{,}p.$$

El número total de individuos de la tabla de contingencia, se denota:

$$N = \sum_{i=1}^{n}\sum_{j=1}^{p} n_{ij} = \sum_{i=1}^{n} n_{i.} = \sum_{j=1}^{p} n_{.j}$$

Las frecuencias relativas absolutas marginales tanto de filas como de columnas, se denotan como:

$$\boldsymbol{F} = f_{ij} = \frac{n_{ij}}{N};\ f_{i.} = \sum_{j=1}^{p} f_{ij} = \frac{n_{i.}}{N};\ f_{.j} = \sum_{i=1}^{n} f_{ij} = \frac{n_{.j}}{N}$$

Para describir y comprender la técnica de AC consideremos los datos sobre el consumo de cuatro marcas de un producto relacionadas con tres segmentos de consumidores, en donde se relacionan el número de personas que compran la marca del producto *i* y que pertenecen al segmento de mercado *j*.

La Tabla 4 y Tabla 5 presenta la tabla de contingencia y tabla de frecuencias relativas respectivamente.

Tabla 4. ***Tabla de contingencia relacionando marca de un producto y segmento de mercado***

	SEGMENTO			
MARCA	**1**	**2**	**3**	**TOTAL**
A	30	30	155	**215**
B	30	130	30	**190**
C	80	30	30	**140**
D	80	30	5	**115**
TOTAL	**220**	**220**	**220**	**660**

Nota: Tabla de contingencia de los datos de ejemplo.

Tabla 5. ***Tabla de frecuencias relativas***

	SEGMENTO				
MARCA	**1**	**2**	**3**	**TOTAL**	
A	0,045	0,045	0,235	**0,326**	f1+
B	0,045	0,197	0,045	**0,288**	f2+
C	0,121	0,045	0,045	**0,212**	f3+
D	0,121	0,045	0,008	**0,174**	f4+
TOTAL	**0,333**	**0,333**	**0,333**	**1,000**	
	f+1	f+2	f+3		

Nota: Tabla de frecuencias relativas de los datos de ejemplo.

Las frecuencias relativas condicionales de columnas con respecto a perfiles fila y de fila con respecto a perfiles columnas, se denotan como:

$$f_{i|j} = \frac{n_{ij}}{n_{.j}} = \frac{f_{ij}}{f_{.j}} \; ; \; f_{j|i} = \frac{n_{ij}}{n_{i.}} = \frac{f_{ij}}{f_{i.}} \, , para \; i = 1, \ldots . , n; j = 1, \ldots . , p$$

Siguiendo con el ejemplo propuesto, las frecuencias relativas condicionales se muestran en la Tabla 6 y Tabla 7.

Tabla 6. ***Tabla de perfil fila***

	SEGMENTO			
MARCA	**1**	**2**	**3**	**TOTAL**
A	0,140	0,140	0,721	**1,000**
B	0,158	0,684	0,158	**1,000**
C	0,571	0,214	0,214	**1,000**
D	0,696	0,261	0,043	**1,000**
Centro de Gravedad	**0,333**	**0,333**	**0,333**	**1,000**

Nota: Tabla de perfil fila, la marca del producto condicionada al segmento de mercado.

Tabla 7. ***Tabla de perfil columna***

	SEGMENTO			
MARCA	**1**	**2**	**3**	**Centro de Gravedad**
A	0,136	0,136	0,705	**0,326**
B	0,136	0,591	0,136	**0,288**
C	0,364	0,136	0,136	**0,212**
D	0,364	0,136	0,023	**0,174**
TOTAL	**1,000**	**1,000**	**1,000**	**1,000**

Nota: Tabla de perfil columna, el segmento de mercado condicionada a marca del producto.

En el espacio vectorial fila ($\mathbb{R}^p$) o nube de puntos fila, el i-ésimo vector del perfil fila tiene como coordenadas:

$$\left(\frac{n_{i1}}{n_{i.}}, \ldots\ldots, \frac{n_{ip}}{n_{i.}}\right) = \left(\frac{f_{i1}}{f_{i.}}, \ldots\ldots, \frac{f_{ip}}{f_{i.}}\right) = \left(f_{1|i}, \ldots\ldots, f_{p|i}\right); i = 1, \ldots\ldots, n$$

La nube de puntos del perfil fila, queda determinada por la matriz diagonal $\mathbf{D_n^{-1}F}$, en donde $\mathbf{D_n}$= **Diagonal** ($\boldsymbol{f_{i.}}$) y contiene las frecuencias marginales por fila o pesos $f_{i.}$.

El punto medio ponderado denominado *centroide*, *baricentro* o *centro de gravedad* de la nube de puntos fila, será el vector fila $\mathbf{G_f}$ cuyas coordenadas son las frecuencias marginales:

$$\mathbf{G_f} = \left(f_{.1}, \ldots\ldots, f_{.p}\right)$$

En el espacio fila ($\mathbb{R}^n$) o nube de puntos columna, el j-ésimo vector del perfil columna tiene como coordenadas:

$$\left(\frac{n_{1j}}{n_{.j}}, \ldots\ldots, \frac{n_{nj}}{n_{.j}}\right) = \left(\frac{f_{1j}}{f_{.j}}, \ldots\ldots, \frac{f_{nj}}{f_{.j}}\right) = \left(f_{1|j}, \ldots\ldots, f_{n|j}\right); j = 1, \ldots\ldots, p$$

La nube de puntos del perfil columna, queda determinada por la matriz diagonal $\boldsymbol{F}\mathbf{D_p^{-1}}$, donde $\mathbf{D_p}$= **Diagonal**($\boldsymbol{f_{.j}}$) y contiene las frecuencias marginales por columna o pesos $f_{.j}$.

El punto medio ponderado denominado *centroide*, *baricentro* o *centro de gravedad* para el caso de la nube de puntos columna, será el vector columna $\mathbf{G_c}$ cuyas coordenadas son las frecuencias marginales:

$$\mathbf{G_c} = \left(f_{1.}, \ldots\ldots, f_{n.}\right)$$

En el AC uno de los objetivos es obtener un pequeño número de dimensiones denominados factores, en donde la primera dimensión explica la mayor parte de la asociación total entre filas y columnas, y que la métrica para cuantificar dicha asociación es el coeficiente conocido como ***ji-cuadrado*** χ^2; la segunda dimensión explica la mayor parte del resto de la asociación no explicada por la primera; la tercera dimensión explica la mayor parte del resto de asociación que no fue explicada por la segunda y así sucesivamente con el resto de dimensiones [5].

Para determinar la asociación entre dos individuos, la distancia χ^2 entre dos perfiles fila *i* e *i´*, es:

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

Y para determinar la distancia χ^2 entre dos perfiles columna *j* e *j´*, es:

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

La inercia de la nube de puntos fila se encuentra utilizando el criterio de la distancia ***ji-cuadrado*** χ^2. Para el ejemplo propuesto, la inercia total de la nube de puntos fila, es la suma de las inercias de los puntos fila, tal como se muestra en la Tabla 8.

Tabla 8. ***Tabla de inercia de la nube de puntos fila***

	SEGMENTO 1	SEGMENTO 2	SEGMENTO 2	SUMA	INERCIAS
Inercia A	0,112673517	0,112673517	0,450694069	0,6760411	**0,220225511**
Inercia B	0,092336103	0,369344414	0,092336103	0,55401662	**0,159489633**
Inercia C	0,170068027	0,042517007	0,042517007	0,25510204	**0,054112554**
Inercia D	0,393824827	0,015752993	0,252047889	0,66162571	**0,115283267**
INERCIA TOTAL					**0,549110966**

Nota: Tabla de inercias total fila, como resultado de la suma de las inercias de las marcas.

Así mismo, la inercia de la nube de puntos columna se encuentra utilizando el criterio de la distancia ***ji-cuadrado*** χ^2. La inercia total de la nube de puntos columna, es la suma de las inercias de los puntos columna, tal como se muestra en la Tabla 9.

Tabla 9. ***Tabla de inercia de la nube de puntos columna***

	INERCIA 1	INERCIA 2	INERCIA 3	INERCIA TOTAL
MARCA A	0,11011276	0,11011276	0,44045102	
MARCA B	0,07974482	0,31897927	0,07974482	
MARCA A	0,10822511	0,02705628	0,02705628	
MARCA B	0,20586298	0,00823452	0,13175231	
SUMA	0,50394566	0,46438282	0,67900442	
INERCIAS	**0,16798189**	**0,15479427**	**0,22633481**	**0,549110966**

Nota: Tabla de inercias total columna, como resultado de la suma de las inercias de los segmentos de mercado.

Se evidencia que la inercia total es la misma, tanto si se la obtiene con el perfil fila como en el perfil columna, y representa la variabilidad total de la tabla.

En el AC se define la *inercia total* o simplemente *inercia*, al valor $\mathbf{X^2}/n$, donde n es el total de la tabla. Este valor es una medida de la *varianza total* de la tabla independientemente de su tamaño. En estadística, entre las varias definiciones, se denomina como "*coeficiente medio cuadrático de contingencias*". Su raíz cuadrada se denomina "*coeficiente phi*" (ϕ), y por tanto se puede expresar como inercia a ϕ^2, relacionado con el supuesto de homogeneidad o de independencia ($\mathbf{X^2}$). Geométricamente, la inercia mide lo lejos que se hallan los perfiles fila o perfiles columnas de su perfil medio. Considerando que el perfil medio simboliza el supuesto o hipótesis de homogeneidad de los perfiles. [8].

Para poder realizar la prueba de independencia en el ejemplo que se está tratando, consideremos los datos de frecuencias relativas de la Tabla 5, tomando las frecuencias observadas f_{ij} y las que deberíamos observar a través del cálculo:

$$\chi^2 = \sum_{i=1}^{n}\sum_{j=1}^{p}\frac{(f_{ij} - f_{i.}f_{.j})^2}{f_{i.}f_{.j}}$$

Tabla 10. ***Tabla de prueba de independencia***

	SEGMENTO		
MARCA	**1**	**2**	**3**
A	0,036704252	0,036704252	0,146817007
B	0,026581606	0,106326422	0,026581606
C	0,036075036	0,009018759	0,009018759
D	0,068620993	0,00274484	0,043917435
		χ2	**362,4132373**
	INERCIA TOTAL		**0,549110966**

Nota: El valor de inercia total, como el resultado de dividir χ^2/ 660.

A partir de χ^2 considerando una probabilidad del 5% y 6 grados de libertad, se obtiene un p-valor de 12,59 el cual es menor que 362,41; por lo que se rechaza la hipótesis nula de independencia, lo que significa que existen al menos dos segmentos en los que los consumidores es diferente y así mismo, existen al menos dos marcas en los que los segmentos de mercado son diferentes.

Otro de los objetivos del AC es encontrar una tipología de individuos en un subespacio $\mathbb{R}^q$ de menor dimensión que $\mathbb{R}^p$, que conserve la máxima información de la nube original; es decir se necesita encontrar un subespacio $\boldsymbol{H}$, tal que la inercia de los puntos proyectados se maximice la expresión:

$$\sum_i f_{i.}\, d^2_{\boldsymbol{H}}(i, \boldsymbol{G}_f)$$

donde $d^2_{\boldsymbol{H}}(i, \boldsymbol{G}_f)$ es la distancia al cuadrado entre el perfil fila i y su respectivo centroide $\boldsymbol{G}_f$, el cual está contenido en $\boldsymbol{H}$. [5].

El AC busca la recta que esté en la dirección de un vector unitario $\boldsymbol{u_1}$, el cuál recoja la *máxima inercia proyectada*. Luego se busca otra que sea ortogonal a la primera y en la dirección de un segundo vector unitario $\boldsymbol{u_2}$ que recoja la *máxima inercia restante proyectada*. A continuación se busca una tercera recta que sea ortogonal a las dos primeras y en la dirección al vector unitario $\boldsymbol{u_3}$ que reúna la *máxima inercia restante proyectada* y así sucesivamente. El subespacio $\boldsymbol{H}$ que se desea encontrar se genera por los vectores unitarios $\boldsymbol{u_i}$. [5].

Los vectores $\boldsymbol{u_1}$, $\boldsymbol{u_2}$, …., $\boldsymbol{u_p}$, que determinan la posición y dirección de los *ejes principales* son generados por los respectivos valores propios de la matriz:

$$\boldsymbol{S} = F' D_n^{-1} F \, D_p^{-1}$$

en el que $\boldsymbol{\lambda_1 \geq \lambda_2 \geq \ldots\ldots \geq \lambda_p}$, son soluciones del sistema:

$$\boldsymbol{S}\mathbf{u} = \boldsymbol{\lambda}\mathbf{u}$$

La inercia recogida en cada eje corresponde al valor propio asociado al eje y así lograríamos encontrar la inercia total:

$$\boldsymbol{I_T} = \boldsymbol{\lambda_1} + \boldsymbol{\lambda_2} + \cdots \ldots . + \boldsymbol{\lambda_p}$$

La línea que une el origen con el centro de gravedad **G** fila o columna, es un vector propio de la matriz **S** con relación al valor propio **λ =1**, el cual tiene la forma $g = (f_{.1}, \ldots \ldots, f_{.p})$ en el espacio fila, que para el AC no se lo considera.

El análisis de correspondencias múltiple (ACM) es un análisis de correspondencias simple aplicado no solo a una tabla de contingencia sino a una *tabla disyuntiva completa*, en el que una variable categórica asigna a cada individuo de una población una modalidad, y, en consecuencia, particiona de manera disyuntiva y exhaustiva a los individuos de la población. El ACM se desarrolla sobre la *tabla disyuntiva completa*, para luego equipararla con el análisis de la tabla de Burt (**B**) [5].

Una *tabla disyuntiva completa*, es una matriz $\mathbb{X}$, con *n*-filas y *p*-columnas, que describe las *k*–respuestas de los *n*-individuos a través de un código binario (0 o 1) [5].

Cada una de las tablas $\mathbb{X}_j$, j= 1,….,k, describe la partición de los n individuos de acuerdo con sus respuestas a la pregunta j, de manera que $\mathbb{X}_j = (x_{im})$, donde:

$$x_{im} = \begin{cases} 1, & si\ el\ i-ésimo\ individuo\ tiene\ la\ modalidad\ m\ de\ la\ pregunta\ j, \\ 0, & si\ el\ i-ésimo\ individuo\ no\ tiene\ la\ modalidad\ m\ de\ la\ pregunta\ j \end{cases}$$

A partir de la *tabla disyuntiva completa* $\mathbb{X}$ se construye una tabla simétrica **B (Tabla de contingencia BURT)** de tamaño (*p x p*) que contiene las frecuencias para los cruces entre todas las *k* variables [5].

$$\mathbf{B} = \mathbb{X}'\,\mathbb{X}$$

El término general de **B** se define:

$$b_{jj'} = \sum_{i=1}^{n} x_{ij} x_{ij'}$$

Los marginales son:

$$b_j = \sum_{j'=1}^{p} b_{ij'} = kx_{.j}\ , para\ todo\ j\ \leq p$$

La frecuencia total es:

$$b = x^2 x_{.j}$$

La tabla **B**, está conformada por k^2 bloques, donde:

El bloque $X'_j X_{j'}$ de tamaño (p_j x $p_{j'}$) es la tabla de contingencia que cruza las respuestas a las preguntas j y j´.

El j-ésimo bloque cuadrado $X'_j X_j$, se obtiene mediante el cruce de cada variable consigo misma.

Sobre la diagonal de **B**, se encuentran matrices diagonales:

$$\boldsymbol{D}_j = P\mathbb{X}_j' P\mathbb{X}_j\ ; j = 1, \ldots., k$$

Los elementos de la diagonal son las frecuencias de las modalidades de la pregunta *j*.

La matriz **D**, de tamaño *p x p,* está conformada por k^2 bloques, en la que en la diagonal están las frecuencias correspondientes:

$$d_{jj} = b_{jj} = x_{.j}$$
$$d_{jj'} = 0 \text{ para todo } j \neq j'.$$

Para ilustrar comprender la construcción de la tabla disyuntiva completa y la tabla de contingencia BURT, consideremos la siguiente tabla de datos de muestra, que se presenta en la Tabla 11.

Tabla 11. ***Tabla de datos para representar la tabla disyuntiva***

Individuos	Genero	Años	Ingreso
1	Mujer	5	Medio
2	Mujer	3	Alto
3	Hombre	4	Bajo
4	Mujer	1	Bajo
5	Mujer	2	Medio
6	Hombre	5	Alto
7	Mujer	2	Medio
8	Hombre	3	Bajo
9	Hombre	1	Alto
10	Mujer	4	Medio

Nota: Ejemplo de datos de diez observaciones considerando tres variables.

Tabla 12. ***Tabla disyuntiva de ejemplo***

		Género		Años					Ingreso		
		M	H	1	2	3	4	5	B	M	A
Género	M	1	0	0	0	0	0	1	0	1	0
	H	1	0	0	0	1	0	0	0	0	1
Años	1	0	1	0	0	0	1	0	1	0	0
	2	1	0	1	0	0	0	0	1	0	0
	3	1	0	0	1	0	0	0	0	1	0
	4	0	1	0	0	0	0	1	0	0	1
	5	1	0	0	1	0	0	0	0	1	0
Ingresos	B	0	1	0	0	1	0	0	1	0	0
	M	0	1	1	0	0	0	0	0	0	1
	A	1	0	0	0	0	1	0	0	1	0

Nota: Tabla disyuntiva transformada en formato binario.

La Tabla 12 muestra las variables género, años e ingreso son sus respectivas categorías de respuesta y transformado a formato binario.

Para encontrar la tabla de BURT, es necesario realizar la multiplicación de la matriz disyuntiva con la matriz disyuntiva transpuesta. La Tabla 13 y Tabla 14, se muestra la matriz disyuntiva transpuesta y la tabla de BURT.

Tabla 13. ***Tabla disyuntiva transpuesta de ejemplo***

		Género		Años					Ingreso		
		M	**H**	**1**	**2**	**3**	**4**	**5**	**B**	**M**	**A**
Género	**M**	1	1	0	1	1	0	1	0	0	1
	H	0	0	1	0	0	1	0	1	1	0
Años	**1**	0	0	0	1	0	0	0	0	1	0
	2	0	0	0	0	1	0	1	0	0	0
	3	0	1	0	0	0	0	0	1	0	0
	4	0	0	1	0	0	0	0	0	0	1
	5	1	0	0	0	0	1	0	0	0	0
	B	0	0	1	1	0	0	0	1	0	0
Ingresos	**M**	1	0	0	0	1	0	1	0	0	1
	A	0	1	0	0	0	1	0	0	1	0

Tabla 14. ***Tabla de BURT de ejemplo***

		Género		Años					Ingreso		
		M	**H**	**1**	**2**	**3**	**4**	**5**	**B**	**M**	**A**
Género	M	6	0	1	2	1	1	1	1	4	1
	H	0	4	1	0	1	1	1	2	0	2
Años	1	1	1	2	0	0	0	0	1	0	1
	2	2	0	0	2	0	0	0	0	2	0
	3	1	1	0	0	2	0	0	1	0	1
	4	1	1	0	0	0	2	0	1	1	0
	5	1	1	0	0	0	0	2	0	1	1
Ingresos	B	1	2	1	0	1	1	0	3	0	0
	M	4	0	0	2	0	1	1	0	4	0
	A	1	2	1	0	1	0	1	0	0	3

Nota: Matriz de tamaño 3 x 3, está conformada por $3^2 = 9$ bloques.

Los ejes factoriales se encuentran a través de los valores y vectores propios de la matriz:

$$\boldsymbol{S} = F' D_n^{-1} F \, D_p^{-1} = \frac{1}{k} \mathbb{X}' \mathbb{X} \, \boldsymbol{D}$$

Cuyo término general es:

$$\boldsymbol{s_{jj'}} = \frac{1}{kx_{.j'}} \sum_{i=1}^{n} x_{ij} x_{ij'}$$

En el espacio fila o de los individuos en $\mathbb{R}^p$, la ecuación del α–ésimo eje factorial $\boldsymbol{u_\alpha}$, es:

$$\frac{1}{k}\, \mathbb{X}'\mathbb{X} D^{-1} \boldsymbol{u_\alpha} = \lambda_\alpha \boldsymbol{u_\alpha}$$

En el espacio columna o de modalidades en $\mathbb{R}^n$, la ecuación del α–ésimo eje factorial $\boldsymbol{\psi_\alpha}$, es:

$$\frac{1}{k}\, \mathbb{X} D^{-1} \mathbb{X}' \boldsymbol{\psi_\alpha} = \lambda_\alpha \boldsymbol{\psi_\alpha}$$

donde los factores $\boldsymbol{\varphi_\alpha}$ y $\boldsymbol{\psi_\alpha}$ de norma λ_α son las coordenadas de los puntos fila y columna sobre el eje factorial α [5].

Análisis factorial

Peña (2002) manifiesta que:

> *El Análisis Factorial está relacionado con las componentes principales, pero existen ciertas diferencias. En primer lugar, los componentes principales se construyen para explicar las varianzas, mientras que los factores se construyen para explicar las covarianzas o correlaciones entre las variables. En segundo lugar, componentes principales es una herramienta descriptiva, mientras que el análisis factorial presupone un modelo estadístico formal de generación de datos.*

El Análisis Factorial (AF) tiene por objetivo encontrar asociaciones entre variables observables, que informan de dimensiones que no podemos observar

directamente. A partir de unas variables que si se pueden observar y cuantificar, el AF pretende inferir y cuantificar nuevas variables inobservables o latentes, pero que resumen el comportamiento del conjunto de variables. [14].

El análisis factorial es un método multivariante que expresa *p* variables observables como una combinación lineal de *m* variables latentes, hipotéticas o inobservables, denominadas *factores* [3].

Para aplicar la técnica de AF es necesario realizar la prueba de independencia a la matriz de correlaciones entre variables, ya que si las variables no tienen ninguna relación o su relación es cercana a cero, no tiene caso empezar la técnica de AF. Si la prueba de independencia de la matriz de correlación entre variables es diferente de cero, entonces es un indicador de la aplicabilidad de dicha técnica, lo que se obtendría los factores comunes a partir de la matriz de correlaciones entre variables:

$$\mathbf{R} = \begin{pmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & & 1 \end{pmatrix}$$

Peña (2002) interpreta la hipótesis básica del modelo factorial como:

> *Si observamos un vector de variables* ***x****, de dimensiones (p x 1), en elementos de una población. El modelo de Análisis Factorial establece que este vector de datos observados se genera mediante la relación:*
>
> $$\mathbf{x} = \boldsymbol{\mu} + \Lambda\mathbf{f} + \mathbf{u}$$
>
> dónde:
>
> 1. ***f****, es un vector (m x 1) de variables latentes, o factores no observados. Supongamos que sigue una distribución* $\mathbf{N_m(0,I)}$*, es decir, los factores son de media cero e independientes entre sí y con distribución normal.*

2. ***Λ**, una matriz (p x m), de constantes desconocidas (m < p), que contiene los coeficientes que describen como los factores **f**, afectan a las variables observadas, **x**, y se denomina matriz de carga.*

3. ***u**, un vector (p x 1) de perturbaciones no observadas, que recoge el efecto de todas las variables distintas de los factores que influyen sobre **x**. Se supone que **u** tiene una distribución $\mathbf{N_p(0, \psi)}$, donde **ψ** es diagonal, y que las perturbaciones están incorreladas con los factores **f**.*

*Con estas tres hipótesis, se deduce que: (a). **μ** es la media de las variables **x**, ya que los factores como las perturbaciones tienen media cero; y (b). **x**, tiene distribución normal, al ser suma de variables normales, y llamando **V** a su matriz de covarianzas.*

$$\mathbf{x \sim N_p(\mu, V)}$$

El modelo de análisis multifactorial de *m* factores comunes, considera que las *p* variables observables u originales $X_1, \ldots, X_p$ depende de *m* variables latentes o inobservables $F_1,\ldots,F_m$ llamados factores comunes y *p* factores únicos $U_1,\ldots,U_p$.[3]

De acuerdo al sistema lineal, se puede apreciar la relación de las variables observables con las variables latentes o inobservables comunes que se denominan *comunalidades*, junto a un factor único relacionado directamente con la variable observable, denominado *unicidades.*

$$X_1 = a_{11}F_1 + \ldots\ldots + a_{1m}F_m + d_1U_1$$
$$X_2 = a_{21}F_1 + \ldots\ldots + a_{2m}F_m + d_2U_2$$
$$\ldots\ldots$$
$$X_p = a_{p1}F_1 + \ldots\ldots + a_{pm}F_m + d_pU_p$$

Este sistema tiene las siguientes hipótesis del modelo multifactorial:

1. Los factores comunes y los factores únicos están incorrelados dos a dos

$$cor(F_i, F_j) = 0,\ i \neq j = 1, ..., m,$$
$$cor(U_i, U_j) = 0,\ i \neq j = 1, ..., p,$$

2. Los factores comunes están incorrelados con los factores únicos

$$cor(F_i, U_j) = 0,\ i = 1, ..., m,\ j = 1, ..., p,$$

3. Los factores comunes como los factores únicos, son variables reducidas, con media 0 y varianza 1.

La matriz factorial **A** (*p x m)* contiene coeficientes a_{ij} que representan las *saturaciones* entre cada variable X_i y el factor F_j. [3].

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & & a_{pm} \end{pmatrix}$$

Si **X** = $(X_1, ..., X_p)'$ es el vector columna de las variables, **F** = $(F_1, ..., F_m)'$ y ***U***=$((U_1, ..., U_p)'$ el modelo factorial expresado en forma matricial es:

$$\mathbf{X = AF + DU}$$

donde, **D** = diag(d1,….,dp) es la matriz diagonal con las saturaciones entre variables y factores únicos. [3].

Matemáticamente, la *comunalidad* es el cuadrado de la carga factorial y que coincide con el coeficiente de correlación R^2 de la regresión entre variables. La *comunalidad* nos dice qué porcentaje de variabilidad de la correspondiente variable observable viene explicada por el factor latente; y, el valor de 1- R^2 se le denomina varianza única o específica. [9].

En el modelo de AF se verifica que:

$$var(X_i) = a_{i1}^2 + \cdots . + a_{im}^2 + d_i^2$$

donde a_{ij}^2 es la parte de la variabilidad de la variable X_i debida al factor común F_j; y d_i^2 es la parte de la variabilidad explicada exclusivamente por el factor único U_i.

Separando la parte común y la parte única de la variable *Xi*, se tiene:

$\boldsymbol{h_i^2} = a_{i1}^2 + \cdots . + a_{im}^2$, como la *comunalidad*.

$\boldsymbol{d_i^2}$ = *unicidad.*

Para cada variable observable, se tiene:

$$\boldsymbol{Variabilidad = comunalidad + unicidad}$$

La *comunalidad* es la parte de la variabilidad de las variables que sólo es explicada por los factores comunes; si suponemos que las variables observables son reducidas, entonces:

$$1 = h_i^2 + d_i^2$$

La matriz de correlaciones reducida, se obtiene a partir de la matriz de correlaciones **R**, substituyendo los unos de la diagonal por las *comunalidades*.

$$\mathbf{R}^* = \begin{pmatrix} h_1^2 & r_{12} & \dots & r_{1p} \\ r_{21} & h_2^2 & \dots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & & h_p^2 \end{pmatrix}$$

Verificándose que:

$$\mathbf{R = R^* + D^2}$$

El número m de factores comunes está limitado por un valor máximo m_a. Consideramos que hay $p(p-1)/2$ correlaciones diferentes y $p \times m$ saturaciones.[3].

Si **A,** es una matriz factorial con factores **F**, también lo será **AT**, con factores:

$$\bar{\mathbf{F}} = \mathbf{T}'\mathbf{F}$$

donde **T** es matriz ortogonal. Como $\mathbf{TT'=I}$, se tiene $m(m-1)/2$ restricciones y el número de parámetros libres de **A** será $p.m - m(m-1)/2$.

El número de correlaciones menos el número de parámetros libres será:

$$d = p(p-1)/2 - [\, p.m - m(m-1)/2] = \tfrac{1}{2}\,[(p-m)^2 - p - m]$$

haciendo d = 0, se obtiene una ecuación de segundo grado que resolviéndola se prueba que:

$$m \leq m_a = \tfrac{1}{2}\,(2p + 1 - \sqrt{8p + 1})$$

Si:

$$\begin{cases} m > m_a, & \textit{existen más saturaciones libres que correlaciones,} \\ m = m_a, & \textit{el modelo es determinado y se puede encontrar } \boldsymbol{A} \textit{ a partir de } \boldsymbol{R} \\ m < m_a, & \textit{se plantea la estimación estadística de } \boldsymbol{A}\textit{, donde } d > 0 \textit{ es el} \\ & \qquad \textit{número de grados de libertad del modelo} \end{cases}$$

El número máximo de m* factores comunes en función de p es:

p	2	3	4	5	6	7	8	9	10	20	30	40
m^*	0	1	1	2	3	3	4	5	6	14	22	31

Matriz policórica

Para realizar el análisis factorial, se requiere aplicar el principio de correlación entre variables; sin embargo, las correlaciones de Pearson son correlaciones producto-momento, es decir que no consideran la naturaleza ordinal de los datos y aplicando la matriz de correlaciones puede arrojar datos distorsionados. Si en un estudio las variables indicadores fuesen de naturaleza dicotómica, las correlaciones policóricas se conocen como correlaciones tetracóricas. [6]. Los métodos estadísticos que asumen distribuciones continuas se aplican a menudo a las medidas observadas que son de escala ordinal, en estos casos existe la posibilidad de un desajuste crítico entre los supuestos que subyacen al modelo estadístico y las características empíricas de los datos a analizar. [7].

Si los datos de interés en una investigación son de naturaleza ordinal, la matriz de correlaciones para poder estimar los factores, debería ser la matriz de correlaciones policóricas, ya que una matriz de correlaciones policórica estima la relación lineal entre dos variables latentes continuas que subyacen a dos variables observadas ordinales que son indicadores reflejo de las latentes. [7].

En el presente TFM considerando la naturaleza de los datos categóricos, parece razonable utilizar la correlación policórica, a fin de evitar riesgos de obtener una distribución asimétrica de los datos, elevada kurtosis y no obtener una distribución normal al calcular la matriz de correlación, que como se dijo, es el primer paso para realizar el análisis factorial.

Al tratar las variables ordinales, éstas indican que para cada variable $\boldsymbol{x}$ con $\boldsymbol{s}$ categorías se asume que hay una variable latente $\boldsymbol{\xi \sim N(0,1)}$ con media igual a cero y desviación estándar igual a uno, cuya relación con la variable ordinal univariada es explicada como: [12].

Si $\boldsymbol{x = i}$ entonces $\boldsymbol{x \in}$ categoría $\boldsymbol{i}$, valores bajos de $\boldsymbol{x}$ indican que $\boldsymbol{x}$ clasificó en categoría baja y que los valores asociados de $\boldsymbol{\xi}$ son también valores pequeños.

$$\boldsymbol{x = i} \quad si \quad a_{i-1} < \boldsymbol{\xi} \leq a_i$$

Donde $\boldsymbol{a_i}$, son los umbrales para las variables ordinales, los cuales son determinados por la expresión:

$$\boldsymbol{a_i} = \phi_1^{-1}\left(\sum_j^i {n_j}/{N}\right) = \phi_1^{-1} P_i$$

n_j Número de observaciones en la categoría *j*

N Número total de observaciones de la variable *x*

P_i Proporción acumulada hasta la categoría *i*

ϕ_1^{-1} Inversa de la función de distribución normal univariada.

El $\boldsymbol{z}$ correspondiente a $\boldsymbol{x = i}$ es la media de $\boldsymbol{\xi}$ en el intervalo

$$\boldsymbol{a_{i-1} < \xi \leq a_i}$$

Por ejemplo, si tomamos una variable ordinal **S** con 5 categorías con un número de observaciones de muestra, con la que queremos obtener una matriz de datos policóricos, partimos de la Tabla 15.

Tabla 15. ***Tabla de ejemplo para obtener medidas policóricas***

S	Observ	Frec. Relativa	Frec. Acumul
1	4	0,10	0,10
2	15	0,36	0,45
3	6	0,14	0,60
4	6	0,14	0,74
5	11	0,26	1,00
SUMA	42		

Nota: Matriz de datos de muestra.

Si se desea por ejemplo encontrar el nuevo valor para X = 3

El nuevo valor debería estar comprendida entre:

$$a_2 < \xi \leq a_3$$

Así:

$$\boldsymbol{a_2} = \phi_1^{-1}\left(\sum_j^2 {}^{n_j}/_N\right) = \phi_1^{-1}\left(\frac{4+15}{42}\right) = \phi_1^{-1}(0.45) = -0.12$$

$$\boldsymbol{a_3} = \phi_1^{-1}\left(\sum_j^3 {}^{n_j}/_N\right) = \phi_1^{-1}\left(\frac{4+15+6}{42}\right) = \phi_1^{-1}(0.60) = 0.24$$

Entonces el nuevo valor policórico para X = 3, estará entre:

$$-0.12 < \xi \leq 0.24$$

2.2 Descripción de lo que se pretende alcanzar

El presente estudio propuesto en el TFM, consiste en comprender las semejanzas entre individuos desde el punto de vista del conjunto de variables, en otras palabras se establece una tipología de individuos homogéneos. Se compara los individuos según la presencia o ausencia de las modalidades que se consideraron. Se busca obtener variables sintéticas que resuma la información contenida en las diversas variables consideradas. Para alcanzar lo propuesto se utilizará el Análisis de Correspondencias Múltiples (ACM), que permita reducir la dimensión de los datos con la mínima pérdida de información, identificando las variables latentes, que facilitan la interpretación de los datos; posteriormente identificando los factores latentes, se aplicará el Análisis Factorial Exploratorio (AFE), a fin de poder obtener un índice o indicador de calidad de vida de la provincia de Loja.

2.3 Identificación, selección y preparación de los datos

Para realizar el análisis exploratorio de datos, se ha tomado los resultados de las Encuestas de las Condiciones de Vida Sexta Ronda (ECV6R) realizada por el INEC, entre el período noviembre 2013 - octubre 2014. El proceso estadístico, utilizó una encuesta dirigida a hogares por muestreo probabilístico que seleccionó una parte de las viviendas ubicadas en las áreas urbanas y rurales del Ecuador. [10].

La ECV es una encuesta multipropósito de cobertura nacional que buscó recoger información específica sobre las principales variables asociadas al bienestar de los hogares ecuatorianos, para lo cual se recogió información de las 24 provincias del país Continental e Insular, y las ciudades de Quito, Guayaquil Cuenca y Machala. [10].

El diseño muestral de la ECV es de tipo probabilístico, con un nivel de confianza del 95% con un error relativo que no supere el 10%. De acuerdo al diseño muestral, los resultados pueden generalizarse para toda la población. El diseño es estratificado y proporcional al tamaño de la población; también es bietápico, donde la unidad de selección es la vivienda y la unidad de observación es el hogar. [10].

La ECV se divide en secciones entre las que tenemos:

- Sección 1. Datos de la vivienda y del hogar
- Sección 2. Registro de los miembros del hogar
- Sección 3. Salud
- Sección 4. Hábitos, prácticas y uso del tiempo
- Sección 5. Educación
- Sección 6. Migración
- Sección 7. Actividades económicas

- Sección 8. Fecundidad y salud materna
- Sección 9. Bienestar psicosocial
- Sección 10. Gastos, otros ingresos y equipamiento del hogar
- Sección 11. Percepción del nivel de vida, capital social, inseguridad ciudadana y retorno migratorio
- Sección 12. Negocios del hogar y trabajadores independientes
- Sección 13. Actividades agropecuarias

La sección de interés para la elaboración del presente TFM, es la Sección 11, en la que se tienen datos sobre la percepción del nivel de vida; para ello se ha seleccionado de toda la base de datos a nivel nacional, datos de la provincia de Loja, que corresponden a un tamaño muestral de 98 Unidades Primarias de Muestreo (UPM), dando un total de 1.176 viviendas encuestadas.

De estas 98 UPM se han tomado las variables de interés para el TFM, que se detallan en la Tabla 16.

Tabla 16. ***Variables seleccionadas para el estudio multivariante***

ITEM	VARIABLE	DESCRIPCION	RANGO DE RESPUESTAS
1	COD_CIUDAD	CIUDAD	provincia+ciudad+canton+parroquia
2	INGRESO_VIVEN	Cómo viven con los ingresos que tienen	1="Bien"; 2=" Mas o menos"; 3="Mal"
3	CON_SIST_ECON	En la actual situac. económica,..principalmente	1="Ahorran dinero"; 2="Equilibran ingresos y gastos"; 3="Gastan ahorros"; 4="Obligados a endeudarse"
4	SU_HOGAR_ES	Usted considera que su hogar es	1="Muy pobre"; 2="Pobre"; 3="Más o menos pobre"; 4="No pobre"
5	NIVEL_VIDA	El nivel de vida de su hogar en los ult. 12 meses	1="Mejoró"; 2="Está igual"; 3="Empeoró"
6	SU_HOGAR_CON_OTROS	En relación al resto de hogares, considera que su hogar es	1="Más pobre"; 2="Igual"; 3="Más rico"

Nota: Recodificación de los datos, para el Análisis de Correspondencias Múltiples.

Utilizando el paquete estadístico R se puede describir los datos utilizados para el presente TFM. La estructura de los datos denominado "DATOS_LOJA_2014" contiene las siguientes variables de tipo factor:

```
> str(DATOS_LOJA_2014)

Classes 'tbl_df', 'tbl' and 'data.frame':     1168 obs. of  6 variables:
 $ COD_CIUDAD        : num  110150 110150 110150 110150 110150 ...
 $ INGRESO_VIVEN     : Factor w/ 3 levels "1","2","3": 2 2 2 2 2 2 2 2 2 2 ...
 $ CON_SIST_ECON     : Factor w/ 4 levels "1","2","3","4": 2 2 2 2 2 2 2 2 2 2 ...
 $ SU_HOGAR_ES       : Factor w/ 4 levels "1","2","3","4": 3 1 3 2 3 3 3 2 2 3 ...
 $ NIVEL_VIDA        : Factor w/ 3 levels "1","2","3": 2 2 2 2 1 2 2 3 2 2 ...
 $ SU_HOGAR_CON_OTROS: Factor w/ 3 levels "1","2","3": 2 1 2 2 1 2 2 1 2 2 ...
```

```
> head(DATOS_LOJA_2014)

  COD_CIUDAD INGRESO_VIVEN CON_SIST_ECON SU_HOGAR_ES NIVEL_VIDA SU_HOGAR_CON_OTROS
1     110150             2             2           3          2                  2
2     110150             2             2           1          2                  1
3     110150             2             2           3          2                  2
4     110150             2             2           2          2                  2
5     110150             2             2           3          1                  1
6     110150             2             2           3          2                  2
```

```
> tail(DATOS_LOJA_2014)

     COD_CIUDAD INGRESO_VIVEN CON_SIST_ECON SU_HOGAR_ES NIVEL_VIDA SU_HOGAR_CON_OTROS
1163     111650             2             2           2          2                  2
1164     111650             2             2           3          2                  2
1165     111650             2             2           2          2                  2
1166     111650             2             2           2          2                  1
1167     111650             2             2           2          2                  2
1168     111650             2             2           3          2                  2
```

Los datos transformados en una Tabla Binaria para realizar el ACM se genera con el paquete *dudi.acm {ade4}*. El resultado de aplicar *acm.disjonctif* se puede observar una parte de los datos en el Anexo 1; y la Tabla de Burt aplicando *acm.burt* una parte de la tabla se muestra en el Anexo 2.

Para el ACM y para el Análisis Factorial Exploratorio (AFE) se reduce la estructura de los datos "DATOS_LOJA_2014" en "datosestudio" y "datos" cuyas estructuras son:

```
> str(datosestudio)

'data.frame': 1168 obs. of  5 variables:
 $ INGRESO_VIVEN     : Factor w/ 3 levels "Bien","Mas o menos",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ CON_SIST_ECON     : Factor w/ 4 levels "Ahorran dinero",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ SU_HOGAR_ES       : Factor w/ 4 levels "Muy pobre","Pobre",..: 3 1 3 2 3 3 3 2 2 3 ...
 $ NIVEL_VIDA        : Factor w/ 3 levels "Mejoró","Está igual",..: 2 2 2 2 1 2 2 3 2 2 ...
 $ SU_HOGAR_CON_OTROS: Factor w/ 3 levels "Más pobre","Igual",..: 2 1 2 2 1 2 2 1 2 2 ...
```

```
> head(datosestudio)
```

```
  INGRESO_VIVEN                CON_SIST_ECON       SU_HOGAR_ES NIVEL_VIDA SU_HOGAR_CON_OTROS
1   Mas o menos Equilibran ingresos y gastos Más o menos pobre Está igual              Igual
2   Mas o menos Equilibran ingresos y gastos         Muy pobre Está igual          Más pobre
3   Mas o menos Equilibran ingresos y gastos Más o menos pobre Está igual              Igual
4   Mas o menos Equilibran ingresos y gastos             Pobre Está igual              Igual
5   Mas o menos Equilibran ingresos y gastos Más o menos pobre     Mejoró          Más pobre
6   Mas o menos Equilibran ingresos y gastos Más o menos pobre Está igual              Igual
```

```
> tail(datosestudio)
```

```
     INGRESO_VIVEN                CON_SIST_ECON       SU_HOGAR_ES NIVEL_VIDA SU_HOGAR_CON_OTROS
1163   Mas o menos Equilibran ingresos y gastos             Pobre Está igual              Igual
1164   Mas o menos Equilibran ingresos y gastos Más o menos pobre Está igual              Igual
1165   Mas o menos Equilibran ingresos y gastos             Pobre Está igual              Igual
1166   Mas o menos Equilibran ingresos y gastos             Pobre Está igual          Más pobre
1167   Mas o menos Equilibran ingresos y gastos             Pobre Está igual              Igual
1168   Mas o menos Equilibran ingresos y gastos Más o menos pobre Está igual              Igual
```

```
> str(datos)
```

```
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':     1168 obs. of  5 variables:
 $ INGRESO_VIVEN     : Factor w/ 3 levels "1","2","3": 2 2 2 2 2 2 2 2 2 2 ...
 $ CON_SIST_ECON     : Factor w/ 4 levels "1","2","3","4": 2 2 2 2 2 2 2 2 2 2 ...
 $ SU_HOGAR_ES       : Factor w/ 4 levels "1","2","3","4": 3 1 3 2 3 3 3 2 2 3 ...
 $ NIVEL_VIDA        : Factor w/ 3 levels "1","2","3": 2 2 2 2 1 2 2 3 2 2 ...
 $ SU_HOGAR_CON_OTROS: Factor w/ 3 levels "1","2","3": 2 1 2 2 1 2 2 1 2 2 ...
```

```
> head(datos)
```

```
  INGRESO_VIVEN CON_SIST_ECON SU_HOGAR_ES NIVEL_VIDA SU_HOGAR_CON_OTROS
1             2             2           3          2                   2
2             2             2           1          2                   1
3             2             2           3          2                   2
4             2             2           2          2                   2
5             2             2           3          1                   1
6             2             2           3          2                   2
```

```
> tail(datos)
```

```
     INGRESO_VIVEN CON_SIST_ECON SU_HOGAR_ES NIVEL_VIDA SU_HOGAR_CON_OTROS
1163             2             2           2          2                   2
1164             2             2           3          2                   2
1165             2             2           2          2                   2
1166             2             2           2          2                   1
1167             2             2           2          2                   2
1168             2             2           3          2
```

2.4 Selección de herramientas y técnicas adecuadas de análisis multivariante y análisis factorial para alcanzar los objetivos propuestos.

Dentro de las técnicas de Análisis Multivariante, y de acuerdo a la naturaleza de los datos a tratar, que son de tipo categórico, se debe utilizar el Análisis de Correspondencias Múltiples (ACM) para reducir la dimensión de los datos y determinar los factores latentes relacionados con los patrones de calidad de vida. Luego de describir y determinar los factores latentes, se procede a aplicar la técnica de Análisis Factorial Exploratorio (AFE), que permite inferir si con los factores latentes encontrados se pude generalizarse a la población de estudio.

Para el análisis, descripción e interpretación de los datos se procesará estadísticamente con el software RStudio versión 3.3.3, utilizando varias librerías, entre ellas *ca, mjca, dudi.acm, factanal, y paquetes polycor, psych* que permiten describir, analizar e interpretar los datos en estudio. A continuación se detalla la descripción de cada uno de los paquetes a utilizar:

ca

El comando *ca* permite realizar el análisis de correspondencia simple, mediante la siguiente sintaxis:

ca(obj, nd = NA, suprow = NA, supcol = NA, subsetrow = NA, subsetcol = NA, ...)

En donde:

obj: Matriz de datos que contiene factores.
nd: Número de factores de salida.
suprow: Índice de filas suplementarias.
supcol: Índice de columnas suplementarias.
subsetrow: Índice de categorías de subconjuntos fila.
subsetcol: Índice de categorías de subconjuntos columna

mjca

El comando *mjca* permite realizar el análisis de correspondencia múltiple y conjunta, mediante la siguiente sintaxis:

mjca(obj, nd = 2, lambda = "adjusted", supcol = NA, subsetcol = NA,
ps = "", maxit = 50, epsilon = 0.0001)

En donde:

obj: Matriz de datos que contiene factores.
nd: Número de factores de salida; si nd=NA se incluyen las dimensiones posibles.
lambda: Método de escalamiento. Pudiendo ser “indicator”, “Burt”, “ajusted” y “JCA”. “JCA” para realizar el análisis de correspondencia conjunta.
supcol: Índice de columnas suplementarias.
subsetcol: Índice de categorías de subconjuntos.
ps: separador, utilizado para combinar nombres de variables y categorías.
maxit: número máximo de iteraciones.
epsilon: criterio de convergencia.

dudi.acm

Parte de la librería *FactoClass,* que utiliza el paquete estadístico *ade4* para realizar el análisis factorial de datos.

dudi.acm permite realizar el análisis de correspondencia múltiple a una tabla de factores.
acm.burt permite obtener la tabla cruzada de Burt.

acm.disjonctif permite obtener la tabla disyuntiva completa de la tabla de factores.
boxplot.acm permite generar la gráfica para su interpretación.

Sintaxis:

dudi.acm (df, row.w = rep(1, nrow(df)), scannf = TRUE, nf = 2)
acm.burt (df1, df2, counts = rep(1, nrow(df1)))
acm.disjonctif (df)

df, df1, df2: Marcos de datos que contienen sólo factores.

row.w: Vector de pesos de fila, por defecto, ponderación uniforme.

scannf: Un valor lógico que indica si se debe mostrar el gráfico de la barra de valores propios.

nf: Si scannf =FALSE, un número entero que indica el número de ejes guardados.

x: Un objeto de clase acm.

xax: El número de factores a mostrar.

factanal

El comando *factanal* permite realizar el análisis factorial a una matriz de covarianza, mediante la siguiente sintaxis:

factanal(x, factors, data = NULL, covmat = NULL, n.obs = NA,
subset, na.action, start = NULL,
scores = c("none", "regression", "Bartlett"),
rotation = METHOD control = NULL, ...)

x: matriz de covarianza.

factors: número de factores a recuperar.

data: marco de datos (opcional)

covmat: matriz de covarianza devuelta.

n.obs:	número de observaciones.
subset:	subconjunto de casos a utilizar.
na.action:	acción a ejecutar, si x es una fórmula.
star:	matriz de valores iniciales.
scores:	resultados a producir. Por defecto "regression" para el resultado de Thompson´s.
rotation:	rotación de factores.
control:	lista de valores de control. (nstart, trace, lower, opt, rotate)

Paquete <polycor>

Calcula la matriz de correlaciones policóricas.

Author : John Fox [aut, cre]

URL: https://r-forge.r-project.org/projects/polycor/, http://CRAN.R-project.org/package=polycor

Paquete psych:

Procedimientos para la investigación psicológica, psicométrica y de personalidad

Herramienta para análisis multivariante en psicología, en temas de personalidad, utilizando análisis factorial, análisis de componentes principales y análisis de conglomerados.

Author : William Revelle

URL: http://personality-project.org/r/psych http://personality-project.org/r/psych-manual.pdf

2.5 Análisis e interpretación de los factores encontrados.

Para realizar el análisis de correspondencias utilizamos la función *dudi.acm* de la librería *FactoClass,* del paquete estadístico *ade4*.

Tabla disyuntiva completa:

La tabla disyuntiva completa en estudio del presente TFM es una tabla de 1.168 observaciones por 17 variables, en donde se utiliza un código indicador para cada categoría (1 si el individuo la asume, 0 caso contrario).

```
Z1 <-acm.disjonctif(datosestudio)
```

Centro de Gravedad en porcentaje:

Las frecuencias marginales de la tabla disyuntiva completa se obtienen así:

```
G <- (colSums(Z1)/ nrow(Z1)/5)*100
```

Tabla 17. ***Centros de gravedad.***

Variable	Porcentaje
INGRESO_VIVEN.Bien	2,071917808
INGRESO_VIVEN.Mas o menos	14,67465753
INGRESO_VIVEN.Mal	3,253424658
CON_SIST_ECON.Ahorran dinero	1,130136986
CON_SIST_ECON.Equilibran ingresos y gastos	16,30136986
CON_SIST_ECON.Gastan ahorros	0,702054795
CON_SIST_ECON.Obligados a endeudarse	1,866438356
SU_HOGAR_ES.Muy pobre	1,797945205
SU_HOGAR_ES.Pobre	9,640410959
SU_HOGAR_ES.Más o menos pobre	6,660958904
SU_HOGAR_ES.No pobre	1,900684932
NIVEL_VIDA.Mejoró	1,335616438
NIVEL_VIDA.Está igual	15,65068493
NIVEL_VIDA.Empeoró	3,01369863
SU_HOGAR_CON_OTROS.Más pobre	5,993150685
SU_HOGAR_CON_OTROS.Igual	13,80136986
SU_HOGAR_CON_OTROS.Más rico	0,205479452
	100

Nota: Porcentaje de los centros de gravedad (CG), del Análisis de Correspondencias Múltiples. Fuente: Datos alcanzados en el estudio

Distancia entre individuos:

Para encontrar la distancia entre filas aplicamos el criterio de χ^2.

```
> n1 <-nrow(Z1) # número de filas
> n1
[1] 1168
```

```
> Dp1<-diag(colSums(Z1)) # suma de las Diagonales principales de Z1
```

```
> Dp1
```

Tabla 18. ***Suma de las diagonales principales***

V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17
121	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	857	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	190	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	66	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	952	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	41	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	109	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	105	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	563	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	389	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	111	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	78	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	914	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	176	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	350	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	806	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12

Nota: Suma de las diagonales de los datos “datosestudio

```
> s1<-ncol(datosestudio) # número de columnas
> s1
[1] 5
```

Así la distancia entre individuos será encontrada, aplicando el criterio de χ^2.

```
> X1 <-sqrt(n1/s1)* as.matrix(Z1)%*%solve( sqrt(Dp1))
```

```
> selin<-seq(25,445,25) # intervalo de visualización de los datos
```

```
> Dis1<-dist(X1[selin ,])
```

```
> round(Dis1,1)
```

La Tabla 19 muestra una parte de las distancias entre individuos entre las 5 variables. Se aprecia, que las observaciones 75, 200, 225, 425, entre otras, tienen distancia "cero", es decir asumen las mismas categorías para las 5 variables de estudio.

Tabla 19. ***Distancia entre observaciones***

25	50	75	100	125	150	175	200	225	250	275	300	325	350	375	400
50	1.0														
75	0.0	1.0													
100	2.2	1.9	2.2												
125	1.9	1.9	1.9	2.7											
150	1.6	1.6	1.6	2.5	1.0										
175	3.6	3.6	3.6	3.0	3.0	3.2									
200	0.0	1.0	0.0	2.2	1.9	1.6	3.6								
225	0.0	1.0	0.0	2.2	1.9	1.6	3.6	0.0							
250	2.6	2.6	2.6	3.2	3.0	2.9	4.2	2.6	2.6						
275	2.9	2.9	2.9	2.2	2.6	2.4	2.1	2.9	2.9	3.7					
300	1.6	1.6	1.6	2.5	1.0	0.0	3.2	1.6	1.6	2.9	2.4				
325	1.4	1.0	1.4	2.2	1.6	1.9	3.5	1.4	1.4	2.8	3.1	1.9			
350	1.0	1.4	1.0	2.4	1.6	1.9	3.4	1.0	1.0	2.8	3.1	1.9	1.0		
375	0.0	1.0	0.0	2.2	1.9	1.6	3.6	0.0	0.0	2.6	2.9	1.6	1.4	1.0	
400	1.0	1.4	1.0	2.4	1.6	1.9	3.4	1.0	1.0	2.8	3.1	1.9	1.0	0.0	1.0
425	0.0	1.0	0.0	2.2	1.9	1.6	3.6	0.0	0.0	2.6	2.9	1.6	1.4	1.0	0.0	1.0

Determinación de los factores principales:

Para determinar los factores principales utilizamos la función *mjca* de la librería *ca*, tomando el argumento *Lambda = "indicador"* que proporciona un análisis de correspondencia múltiple basado en la matriz de indicadores con inercias correspondientes (autovalores).

```
> acm<-mjca(datosestudio,lambda="indicator",nd=4)

> summary(acm)
```

Tabla 20. ***Inercias principales***

```
Principal inertias (eigenvalues):

 dim    value      %    cum%   scree plot
 1      0.420494  17.5  17.5   ****
 2      0.314005  13.1  30.6   ***
 3      0.238656   9.9  40.5   **
 4      0.204761   8.5  49.1   **
 5      0.197658   8.2  57.3   **
 6      0.181067   7.5  64.9   **
 7      0.178522   7.4  72.3   **
 8      0.152382   6.3  78.6   **
 9      0.146053   6.1  84.7   **
 10     0.139161   5.8  90.5   *
 11     0.129099   5.4  95.9   *
 12     0.098144   4.1 100.0   *
        -------- -----
 Total: 2.400000 100.0
```

Nota: Valores propios obtenidos del ACM con la función mjca.

Utilizando la función *mjca* (multiple and joint correspondence analysis), utilizando el argumento número de dimensiones nd=4, método de escalamiento lambda = "indicador", el cual permite realizar el análisis de correspondencia múltiple basado en la matriz de indicadores con sus inercias correspondientes. El resultado se observa en la Tabla 20 donde se muestra la inercia explicada de las 12 nuevas dimensiones, la inercia total que es de 2.4, el porcentaje de inercia de cada factor y la inercia acumulada; en la Figura 1, se representa un diagrama de barras, con los valores de las inercias encontradas.

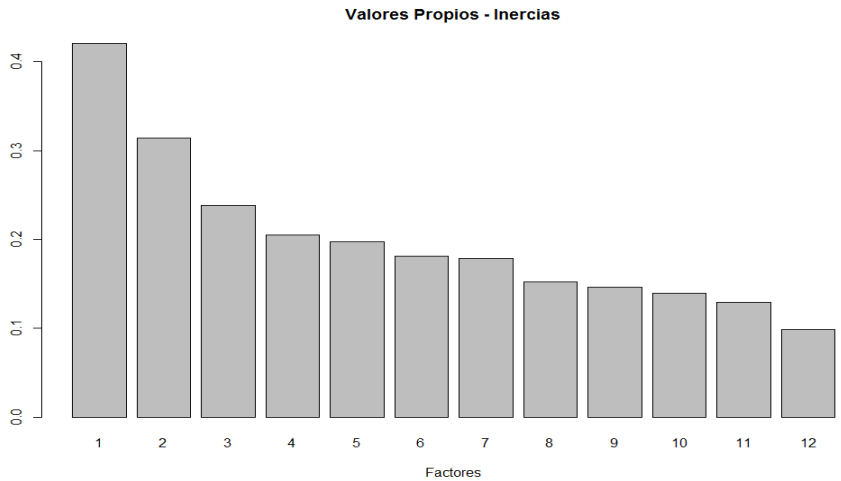

Figura 1. Diagrama de barras de los valores propios / inercias obtenidas del ACM.

El detalle de los 4 factores encontrados se muestran en la Tabla 21 en la que se observa para cada dimensión, las masas (mass) o pesos de cada variable, la calidad (qlt) del análisis del ACM que representa la suma de los cuadrados de las correlaciones de las 4 primeras dimensiones, que a su vez representa la suma de las correlaciones; las inercias (inr) de cada factor y las contribuciones de cada factor; Para cada dimensión se observa las coordenadas de las 4 primeras dimensiones (k = 1..4); junto con las coordenadas, se encuentran las correlaciones al cuadrado (cor) y las contribuciones (ctr) de cada variable. Los valores de estas tablas están expresados en tanto por mil.

Tabla 21. ***Factores encontrados***

variables.categoria	mass	qlt	inr	k=1	cor	ctr	k=2	cor	ctr	k=3	cor	ctr	k=4	cor	ctr
INGRESO_VIVEN:Bien	21	446	75	1034	123	53	-1573	286	163	546	34	26	-153	3	2
INGRESO_VIVEN:Mas o menos	147	518	23	196	106	13	374	386	65	-98	26	6	6	0	0
INGRESO_VIVEN:Mal	33	555	82	-1541	461	184	-686	91	49	92	2	1	70	1	1
CON_SIST_ECON:Ahorran dinero	11	480	81	1672	168	75	-2245	302	181	225	3	2	-343	7	7
CON_SIST_ECON:Equilibran ingresos y gastos	163	464	14	48	10	1	272	326	38	166	122	19	36	6	1
CON_SIST_ECON:Gastan ahorros	7	691	69	-397	6	3	-712	18	11	-1446	76	62	4029	591	557
CON_SIST_ECON:Obligados a endeudarse	19	608	74	-1279	168	73	-747	57	33	-1046	113	86	-1620	270	239
SU_HOGAR_ES:Muy pobre	18	502	84	-1837	333	144	-1051	109	63	-157	2	2	763	58	51
SU_HOGAR_ES:Pobre	96	488	40	-268	67	16	304	86	28	549	280	122	-244	55	28
SU_HOGAR_ES:Más o menos pobre	67	625	54	489	119	38	333	55	24	-936	437	244	165	14	9
SU_HOGAR_ES:No pobre	19	554	81	1382	201	86	-1716	309	178	645	44	33	-63	0	0
NIVEL_VIDA:Mejoró	13	304	71	1084	84	37	-978	68	41	-1457	152	119	8	0	0
NIVEL_VIDA:Está igual	157	554	17	115	48	5	203	148	21	299	321	58	101	37	8
NIVEL_VIDA:Empeoró	30	468	71	-1078	206	83	-620	68	37	-905	145	103	-528	49	41
SU_HOGAR_CON_OTROS:Más pobre	60	503	65	-954	390	130	-145	9	4	419	75	44	260	29	20
SU_HOGAR_CON_OTROS:Igual	138	521	28	396	350	52	107	26	5	-216	104	27	-136	41	12
SU_HOGAR_CON_OTROS:Más rico	2	186	72	1213	15	7	-2963	91	57	2312	55	46	1538	25	24

Nota: Primeros 4 factores que determinan una inercia acumulada del 49.1%.

Para determinar la cantidad de ejes a conservar, aplicamos el criterio que viene dado por el cociente entre la inercia total y el total de los valores propios. Este resultado se debe comparar con los valores de los valores propios o inercias y deberán conservarse todos los ejes que se encuentren por encima de este valor.

Así:

$$\frac{inercia\ total}{total\ de\ factores}$$

$$\frac{2.4}{12} = 0.2$$

De las 17 modalidades de estudio, el ACM permitió reducir a 12 dimensiones para alcanzar el 100% de la información. De estos nuevos 12 ejes, el número de ejes a conservar son los que alcancen un valor mayor o igual a 0.2. Observando la Tabla 20, se deben tomar los 4 primeros factores que alcanzan una inercia acumulada de 49.1% de los datos, el resto de ejes aparecen, pero no contienen información que aporte al estudio. Se puede ver en la Figura 1 que el decrecimiento de los valores propios o inercias tiene un comportamiento regular.

Se puede observar que para el caso de la primera dimensión (k=1), las variables que más aportan a esta dimensión son: INGRESO_VIVEN: Mal con 184 puntos de contribución, SU_HOGAR_ES: Muy pobre, con 144 puntos de contribución, SU_HOGAR_CON_OTROS: Más pobre, con 130 de contribución.

Para el caso de la segunda dimensión (k=2), las modalidades que más contribuyen a este factor son: CON_SIST_ECON: Ahorran dinero, SU_HOGAR_ES: No pobre y INGRESO_VIVEN: Bien.

Para la tercera dimensión, la modalidad que más contribuye es SU_HOGAR_ES: Más o menos pobre; para la cuarta dimensión se destaca CON_SIST_ECON: Gastan ahorros.

Representación gráfica de los factores principales

```
> plot(acmTFM,dim = c (1,2))
```

En la Figura 2 se representa los datos de estudio en dos primeras dimensiones de acuerdo a los 2 factores principales que contribuyen con el 17,5% y 13,1% respectivamente y entre los 2 cubren un 30,6% de la inercia total.

Se observa que el primer eje factorial está altamente determinado por el nivel de contribución de la variable la contribución INGRESOS_VIVEN mal, seguida de la variable SU_HOGAR_ES muy pobre, y, SU_HOGAR_CON_OTROS Más pobre, con contribuciones de 184, 144 y 130 respectivamente.

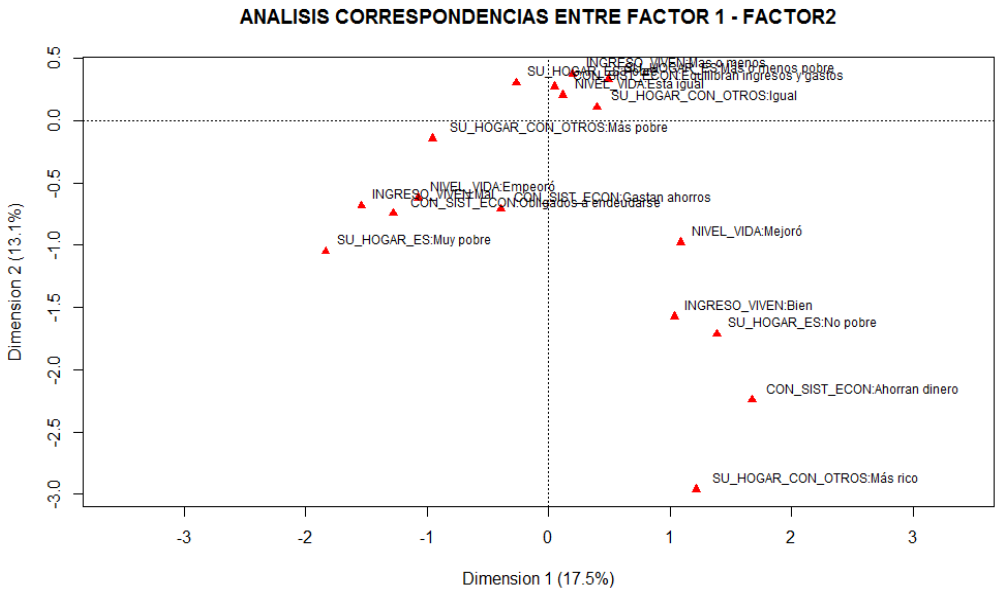

Figura 2. Representación del factor 1 y factor 2 en el ACM.

Para el segundo eje factorial, se observa que está determinado por el nivel de contribución de las variables CON_SIST_ECON Ahorran dinero, seguida de SU_HOGAR_ES no pobre, y, INGRESO_VIVEN bien con contribuciones de 181, 178 y 163 respectivamente.

Los resultados evidencian que en el lado positivo del primer eje factorial, se agrupan las modalidades indicativas de hogares que se consideran con un nivel de vida alta, comparada con la parte izquierda de este eje, en la que se agrupan las modalidades indicativas de hogares que se consideran con un nivel de vida crítica. Observamos que las modalidades agrupadas en este caso, son los hogares en donde el NIVEL_DE_VIDA mejoró, con los INGRESOS_VIVEN bien, SU_HOGAR_ES es considerado no pobre, CON_SIST_ECON logran ahorrar, y, SU_HOGAR_CON_OTROS se consideran más ricos frente a otros hogares.

En el lado negativo del primer eje se agrupan los hogares considerados críticos en donde los hogares que consideran que el: NIVEL_DE_VIDA empeoró, con los INGRESOS_VIVEN mal, SU_HOGAR_ES es considerado muy pobre, CON_SIST_ECON gastan ahorro y están obligados a endeudarse y, SU_HOGAR_CON_OTROS se consideran más pobres frente a otros hogares.

En la parte superior de la de la Figura 2 se agrupan modalidades que evidencian que los hogares que se consideran de un nivel de vida medio. Así notamos que la agrupación entre los hogares, donde: NIVEL_DE_VIDA está igual, con los INGRESOS_VIVEN más o menos, SU_HOGAR_ES es considerado más o menos pobre, CON_SIST_ECON equilibran ingresos y gastos, y, SU_HOGAR_CON_OTROS se consideran igual frente a otros hogares.

Y de forma única y aislada se encuentra SU_HOGAR_ES pobre.

```
> plot(acmTFM,dim = c (1,3))
```

Al relacionar el factor 1 con el factor 3 como se muestra en la Figura 3, se observa que en la parte izquierda del plano, se destaca la calidad de vida baja, frente a la parte derecha que puede inferirse que es una calidad de vida elevada. En el primer eje factorial se distribuye desde izquierda a derecha las variables INGRESO_VIVEN desde mal, hasta INGRESO_VIVEN bien; en la segunda dimensión la variable relevante es CON_SIST_ECON desde la parte inferior que gastan ahorro, hasta la parte superior en donde ahorran dinero.

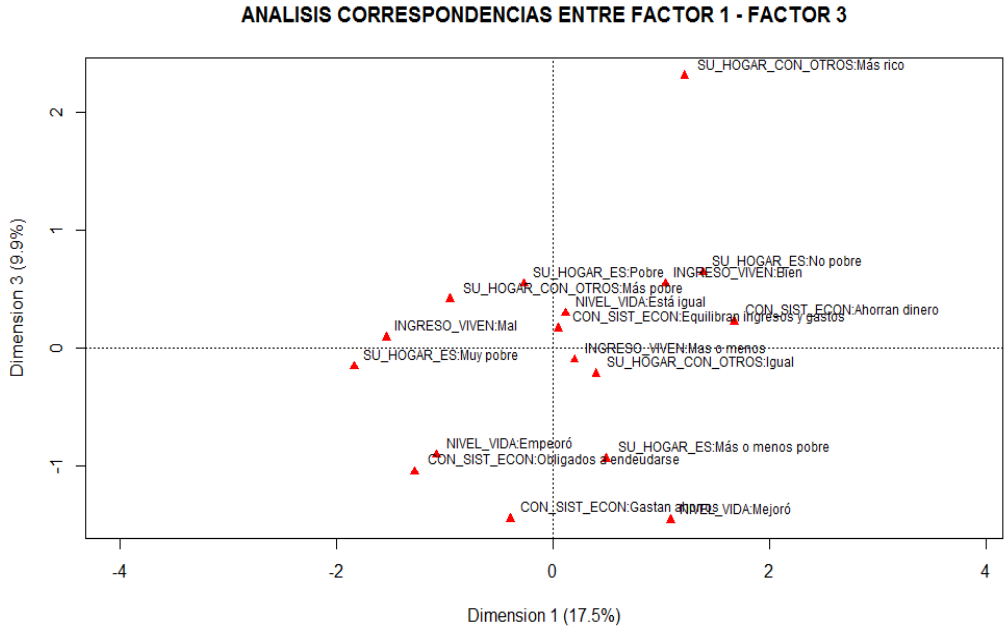

Figura 3. Representación del factor 1 y factor 3 en el ACM.

```
> plot(acmTFM,dim = c (1,4))
```

En el primer eje factorial se distribuye desde izquierda a derecha en función de las variables SU_HOGAR_ES: Muy pobre hasta SU_HOGAR_ES: No pobre; y en el segundo eje factorial la modalidad que se destaca CON_SIST_ECON: Gastan ahorros hasta CON_SIST_ECON: Obligados a endeudarse. Así se observa en la Figura 4.

Al relacionar el factor 1 con el factor 4 se evidencia que en el lado positivo del primer eje factorial se agrupan las modalidades indicativas de hogares que se consideran con un nivel de vida alta. Se observa que en la dimensión 4 se proyecta de manera perfecta la variable CON_SIST_ECON Gastan ahorros, Obligados a endeudarse; con el resto de variables se puede observar las diferentes agrupaciones que se relacionan de manera perfecta de acuerdo al nivel de calidad de vida.

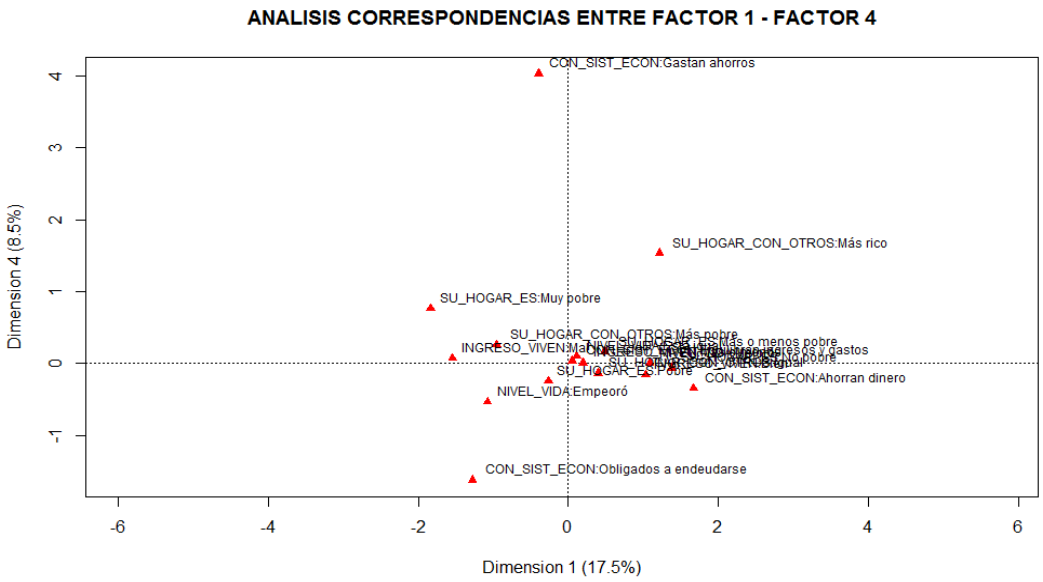

Figura 4. Representación del factor 1 y factor 4 en el ACM.

```
> plot(acmTFM,dim = c (2,3))
```

En la Figura 5 se observa que en la parte suprior izquierda y al centro del primer eje factorial se destacan las modalidades de hogares que se consideran de nivel de vida alta, frente a la parte baja izquierda se agrupan las modalidades de los hogares considerados de nivel de vida crítica. En el cuadrante superior derecho, se ubican las variables que representan la calidad de vida normal; y, en la parte inferior derecha se ubican las variables que representan la calidad de vida consoladora.

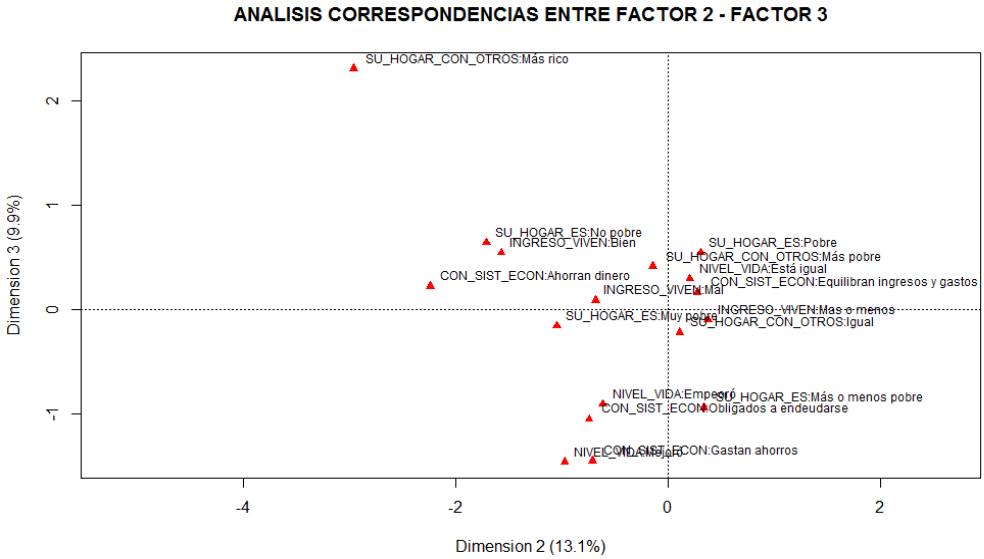

Figura 5. Representación del factor 2 y factor 3 en el ACM.

```
> plot(acmTFM, dim = c (2,4))
```

En la Figura 6 se observa que en el primer eje factorial se distribuye desde izquierda a derecha las modalidades SU_HOGAR_CON_OTROS desde más rico hasta más pobre; y, en la segunda dimensión la variable relevante es CON_SIST_ECON desde la parte inferior que está obligado a endeudarse hasta la parte superior en donde gastan ahorro.

La relación entre el factor 2 y 4, no permite identificar claramente las agrupaciones de modalidades, ya que prácticamente todas las modalidades están agrupadas en el centro del plano.

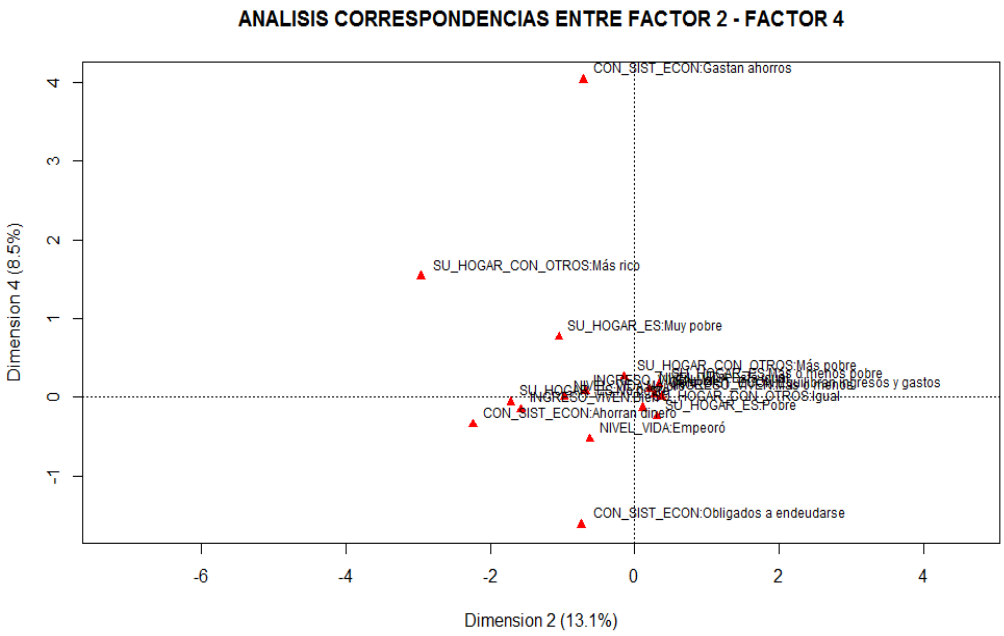

Figura 6. Representación del factor 2 y factor 4 en el ACM.

```
> plot(acmTFM,dim = c (3,4))
```

En la Figura7 la distribución de las modalidades tiene un comportamiento similar al de la Figura6. Aquí la distribución del primer eje desde izquierda a derecha se presenta la modalidad SU_HOGAR_CON_OTROS desde más pobre hasta más rico; y, en la segunda dimensión la variable relevante es CON_SIST_ECON desde la parte inferior que está obligado a endeudarse hasta la parte superior en donde gastan ahorro.

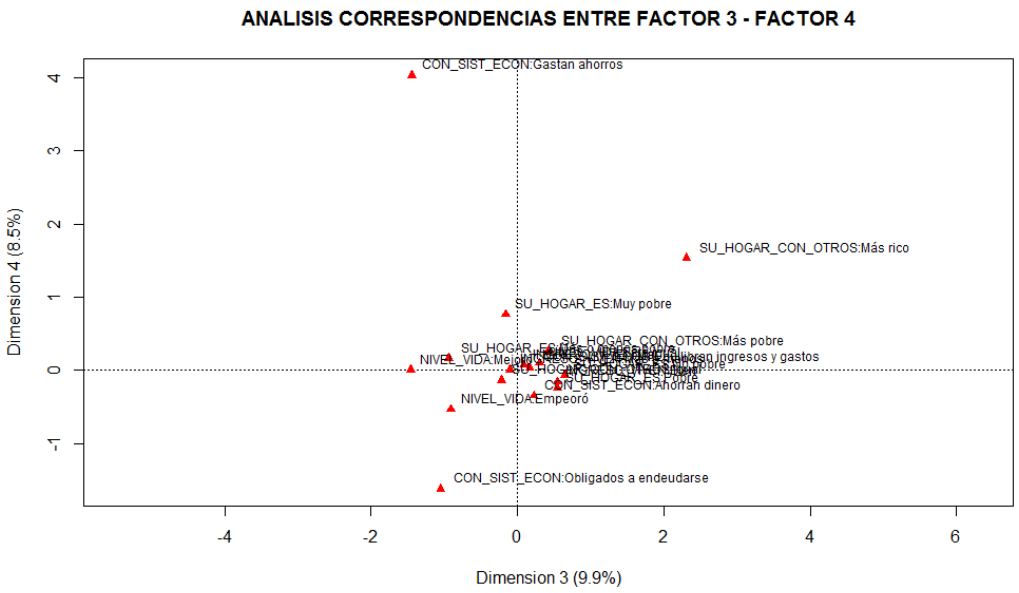

Figura 7. Representación del factor 3 y factor 4 en el AFE.

2.6 Difusión del conocimiento descubierto.

Análisis factorial exploratorio

Para poder describir o explicar la estructura de las relaciones de las variables de estudio, se utiliza la técnica estadística conocida como Análisis Factorial Exploratorio (AFE) a fin de reducir o simplificar los datos, con los cuáles se puede utilizar como descriptores resumidas de los datos denominadas *latentes o inobservables* que se generan a partir de las variables originales u observables que deben estar correlacionadas entre sí, si no fuera el caso, no habría motivo de realizar un AFE.

Para el AFE en el presente TFM, se inicia el estudio con la matriz de correlaciones; sin embargo considerando la naturaleza de los datos de estudio que son de tipo categórico, y evitar resultados defectuosos, se utiliza la matriz de correlaciones policórica.

Matriz de correlaciones policórica.

```
> dfOrdTFM <- data.frame(datosestudio)
> ordNumTFM <- data.matrix(dfOrdTFM)
```

```
> library(polycor)
> pcTFM <- hetcor(dfOrdTFM, ML=TRUE) # matriz de correl policórica
> pcTFM
```

Tabla 22. ***Matriz de correlaciones policórica***

	INGRESO_VIVEN	CON_SIST_ECON	SU_HOGAR_ES	NIVEL_VIDA	SU_HOGAR_ CON_OTROS
INGRESO_VIVEN	1	Polychoric	Polychoric	Polychoric	Polychoric
CON_SIST_ECON	0,3953027	1	Polychoric	Polychoric	Polychoric
SU_HOGAR_ES	-0,5436655	-0,3521021	1	Polychoric	Polychoric
NIVEL_VIDA	0,32615392	0,35564554	-0,280671	1	Polychoric
SU_HOGAR_ CON_OTROS	-0,3659554	-0,2425026	0,57304718	-0,2247976	1

Nota: Matriz de correlación policórica, utilizando el paquete "policor".

Prueba de Significación de la matriz de correlaciones policórica.

Para realizar la prueba de significación de la matriz de correlaciones se considera las siguientes hipótesis:

H_0 = Las variables son independientes (Correlación cero)
H_a = Las variables son dependientes (Correlación significativa)

Si consideramos una probabilidad del 5% y una confiabilidad del 95% de la que se deduce que si el valor obtenido de la prueba de significancia es menor del 5%, entonces debería aceptarse la H_0.

Aplicando la prueba chi-cuadrado de Pearson en el software R, se obtiene el valor de chi-cuadrado, los grados de libertad y el p_valor. Estos valores permiten identificar la existencia o no de algún tipo de relación entre las variables analizadas

```
> chisq.test(sapply(datos,as.numeric))

        Pearson's Chi-squared test

data:  sapply(datos, as.numeric)
X-squared = 774.13, df = 4668, p-value = 1
```

De acuerdo a los resultados: p-valor > 0.05, entonces el valor está en la zona de rechazo, por lo que se rechaza H_0 y se acepta H_a. Con este resultado se puede continuar con la aplicación de la técnica de AFE.

Valores y vectores propios (Eigen valores).

```
> VTFM<-eigen(pcTFM$correlations)
> VTFM
```

Tabla 23. ***Matriz de valores y vectores propios***

```
$values
[1] 2.4874234 0.9143885 0.6582451 0.5694160 0.3705270

$vectors
           [,1]        [,2]        [,3]        [,4]        [,5]
[1,]  0.4864300  0.01072613  0.30958213  0.71243086 -0.39983999
[2,]  0.4109818 -0.47653847  0.56540732 -0.53267004 -0.02412997
[3,] -0.5143535 -0.35439140 -0.02355052 -0.06958196 -0.77746442
[4,]  0.3685652 -0.60760244 -0.69667927  0.08633824  0.04650486
[5,] -0.4405452 -0.52727796  0.31393765  0.44318051  0.48263040
```

Nota: Matriz de valores y vectores propios a partir de la matriz de correlación policórica.

Extracción de los pesos factoriales.

Para realizar la extracción de los pesos factoriales se ha implementado una rutina en R, a fin de encontrar los pesos factoriales, varianza explicada por cada factor y las comunalidades. (Ver Anexo N.3).

```
> factoresTFM<-pesos_factores(VTFM,5,3)
> factoresTFM

$peso_factores
           [,1]        [,2]        [,3]        [,4]        [,5]
[1,] -0.7671763 -0.01025671 -0.25117112 -0.53759788  0.24338632
[2,] -0.6481827  0.45568354 -0.45872800  0.40195099  0.01468814
[3,]  0.8112161  0.33888204  0.01910708  0.05250631  0.47324983
[4,] -0.5812850  0.58101171  0.56523196 -0.06515054 -0.02830794
[5,]  0.6948088  0.50420250 -0.25470486 -0.33442249 -0.29378162
```

```
$varianza_expl

         [,1]       [,2]       [,3]      [,4]       [,5]
v_var  2.487423  0.9143885  0.6582451  0.569416   0.370527
%     49.748468 18.2877699 13.1649023 11.388320   7.410540
%acum 49.748468 68.0362375 81.2011397 92.589460 100.000000


$comunalidad

[1] 0.6517516 0.8382197 0.7732777 0.9949541 0.8018540
```

```
> factoresTFM$peso_factores
```

Tabla 24. ***Matriz de factores encontrados***

	[Factor1]	[Factor 2]	[Factor 3]	[Factor 4]	[Factor 5]
[1,]	-0.7671763	-0.01025671	-0.25117112	-0.53759788	0.24338632
[2,]	-0.6481827	0.45568354	-0.45872800	0.40195099	0.01468814
[3,]	0.8112161	0.33888204	0.01910708	0.05250631	0.47324983
[4,]	-0.5812850	0.58101171	0.56523196	-0.06515054	-0.02830794
[5,]	0.6948088	0.50420250	-0.25470486	-0.33442249	-0.29378162

Nota: Matriz de factores encontrados, como producto de vectorpropio[i]*sqrt(valorpropio).

Para obtener los pesos factoriales del primer factor (a1), multiplicamos cada elemento del vector (V_1) por la raíz cuadrada de λ_1, e invirtiendo el signo. Los cinco factores se muestran en la Tabla 24.

```
> -VTFM$vectors[,1]*sqrt(VTFM$values[1])

[1] -0.7671763 -0.6481827  0.8112161 -0.5812850  0.6948088
```

Varianza explicada por cada factor

Para calcular la varianza explicada por cada factor, se suman los pesos factoriales al cuadrado de cada variable. Para el primer factor, la varianza explicada del factor 1 sería:

```
> apply(factoresTFM$peso_factores^2,2,sum)

[λ1] 2.4874234 0.9143885 0.6582451 0.5694160 0.3705270
```

Para el resto de factores, se muestran en la Tabla 25.

```
> factoresTFM$varianza_expl
```

Tabla 25. ***Varianza explicada por cada factor***

	[Factor1]	[Factor 2]	[Factor 3]	[Factor 4]	[Factor 5]
λ_1	2.487423	0.9143885	0.6582451	0.569416	0.370527
%	49.748468	18.2877699	13.1649023	11.388320	7.410540
%acum	49.748468	68.0362375	81.2011397	92.589460	100.000000

Nota: Matriz de varianzas explicadas.

Considerando que las 5 variables están estandarizadas, la varianza total será 5, por lo que el porcentaje de varianza que explica cada factor, se calcula al respecto de 5; por ejemplo: 2.4874/5=0.4974, lo que implica un 49.75%.

Los factores explican la varianza de la matriz de correlaciones de forma decreciente. El factor 1 es el de mayor varianza (λ_1=2.49), a continuación el factor 2 (λ_2=0.914), y así sucesivamente.

Si observamos la matriz de varianzas explicadas, el primer factor explica casi la mitad de la varianza total de la matriz de correlación, y el segundo factor explica casi la quinta parte, y entre los dos se explica el 68% de la varianza total de la matriz de correlación R, el resto de factores explican el 32% de la varianza total. Es decir podemos considerar los factores 1 y factor 2, como los factores significativos.

Comunalidades

Considerando los 3 primeros factores, la comunalidad para cada variable se muestra en la Tabla 26.

```
> factoresTFM$comunalidad
  [1] 0.6517516 0.8382197 0.7732777 0.9949541 0.8018540
```

El cálculo de la primera comunalidad es:

```
> comunaTFM<-apply((factoresTFM$peso_factores[,1:3])^2, 1, sum)
```

```
> comunaTFM

[1] 0.6517516 0.8382197 0.7732777 0.9949541 0.8018540
```

Tabla 26. ***Comunalidad explicada por los 3 primeros factores***

Variable	Comunalidad
1	0.652
2	0.838
3	0.773
4	0.995
5	0.801

Rotación de factores y su representación gráfica

Para el obtener la rotación de factores y el diagrama, usamos el paquete y librería psych.

```
> library(psych)

> faPCTFM <- fa(r=pcTFM$correlations, nfactors=2, n.obs=N, rotate="vari
max")

Factor Analysis using method =  minres
Call: fa(r = pcTFM$correlations, nfactors = 2, n.obs = N, rotate = "var
imax")
Standardized loadings (pattern matrix) based upon correlation matrix
                     MR1   MR2   h2    u2 com
INGRESO_VIVEN      -0.46  0.48 0.44 0.556 2.0
CON_SIST_ECON      -0.21  0.62 0.42 0.576 1.2
SU_HOGAR_ES         0.92 -0.25 0.91 0.092 1.2
NIVEL_VIDA         -0.17  0.52 0.30 0.702 1.2
SU_HOGAR_CON_OTROS  0.56 -0.22 0.37 0.634 1.3


                       MR1  MR2
SS loadings           1.44 1.00
Proportion Var        0.29 0.20
Cumulative Var        0.29 0.49
Proportion Explained  0.59 0.41
Cumulative Proportion 0.59 1.00

Mean item complexity =  1.4
Test of the hypothesis that 2 factors are sufficient.
```

```
> faPCTFM$loadings

Loadings:
                   MR1    MR2
INGRESO_VIVEN      -0.458  0.484
CON_SIST_ECON      -0.206  0.618
SU_HOGAR_ES         0.918 -0.254
NIVEL_VIDA         -0.169  0.519
SU_HOGAR_CON_OTROS  0.562 -0.224

                 MR1   MR2
SS loadings    1.440 1.000
Proportion Var 0.288 0.200
Cumulative Var 0.288 0.488
```

Se observa que para la primera variable latente MR1, subyacen con las variables observables u originales: SU_HOGAR_ES con 0.9, SU_HOGAR_CON_OTROS con 0.6. Para la segunda variable latente MR2, está contribuyendo a las variables originales: CON_SIST_ECON con un peso de 0.6, NIVEL_DE_VIDA con 0.5, INGRESO_VIVEN con 0.5. La representación se muestra en la Figura 8.

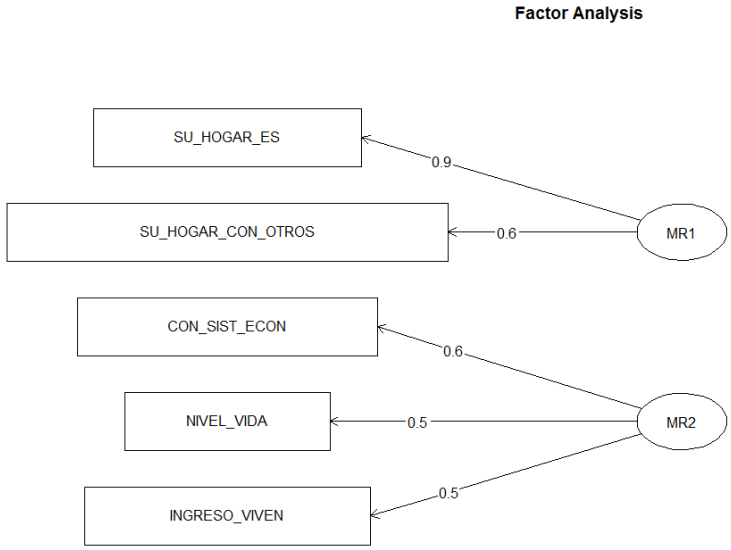

Figura 8. Representación de los dos factores principales del AFE.

Tabla 27. ***Dos factores principales rotados con sus comunalidades y unicidades***

	MR1	MR2	h2	u2	com
INGRESO_VIVEN	-0.46	0.48	0.44	0.556	2.0
CON_SIST_ECON	-0.21	0.62	0.42	0.576	1.2
SU_HOGAR_ES	0.92	-0.25	0.91	0.092	1.2
NIVEL_VIDA	-0.17	0.52	0.30	0.702	1.2
SU_HOGAR_CON_OTROS	0.56	-0.22	0.37	0.634	1.3

Nota: Principales factores, comunalidades y sus respectivas unicidades, luego de haber hecho la rotación "varimax".

Se observa los valores principales mayores que 0.5 que permiten encontrar los factores. Así para la variable SU_HOGAR_ES habría un 91% de su varianza que viene explicada por los factores MR1 y MR2, quedando un 9% de la varianza sin explicar; para SU_HOGAR_CON_OTROS, un 37% de su varianza viene explicada por los factores comunes, quedando el 63% de la misma sin explicar, y así para el resto de variables observables.

De acuerdo a la Figura 8 podríamos explicar las variables observables en forma de ecuaciones estructurales, tomando las raíces cuadradas de los valores de los pesos factoriales de cada factor y la raíz cuadrada de la diferencia de 1 menos el peso factorial para encontrar la puntuación residual.

Así, tenemos para la primera variable SU_HOGAR_ES:

1 - 0.92 = 0.08
sqrt(0.92) = 0.96 ; sqrt(0.08) = 0.28

La Tabla 28, representa las ecuaciones y relaciones que se muestran en la Figura 8.

Tabla 28. ***Ecuaciones estructurales del AFE***

SU_HOGAR_ES	= 0.96 MR1 + 0.28
SU_HOGAR_CON_OTROS	= 0.75 MR1 + 0.66
CON_SIST_ECON	= 0.78 MR2 + 0.62
NIVEL_VIDA	= 0.72 MR2 + 0.69
INGRESO_VIVEN	= 0.71 MR2 + 0.71

3. Conclusiones

El estudio presenta como resultado de aplicar el Análisis de Correspondencias Múltiples, reducir de 17 variables observables a 4 variables, con una inercia total de 2.4 obteniéndose una inercia acumulada del 49.1% entre las 4 variables principales, que es un valor aceptable en estudios de las ciencias sociales. Así la primera dimensión tiene una inercia del 17,5%, la segunda dimensión el 13,1%, y así sucesivamente hasta alcanzar la cuarta dimensión el 8,5%.

La aplicación del Análisis de Correspondencias Múltiples permitió comprender las semejanzas, tipología de individuos homogéneos y la desigualdad de la calidad de vida de los hogares de la provincia de Loja, entre las 1168 observaciones de la sección 11 de la Encuestas de las Condiciones de Vida Sexta Ronda (ECV6R) realizada por el INEC-Ecuador, entre el período noviembre 2013 - octubre 2014.

Es importante e interesante realizar la prueba de significancia de la matriz de correlación, ya que permite anticipar la aplicación o no del Análisis Factorial Exploratorio. Si las correlaciones encontradas son bajas, entonces las variables en estudio son casi ya independientes entre sí, por lo que ya no sería necesario hallar ningún factor.

En el Análisis Factorial Exploratorio se consideran los dos primeros factores encontrados como significativos, ya que entre los dos factores explican un 68%, de la varianza total de la matriz de correlaciones policórica.

Es necesario dar un significado teórico adecuado a los factores encontrados en el Análisis Factorial Exploratorio, sin embargo en ocasiones esto no es fácil, por ello es necesario hacer una transformación o rotación de factores para facilitar su interpretación. Así en el presente estudio al aplicar la rotación de factores con el método de rotación ortogonal “varimax”, se obtuvieron dos factores significativos cuyos pesos factoriales para MR1 es de 0.918 para la variable observable SU_HOGAR_ES y SU_HOGAR_CON_OTROS con 0.562. Para

para MR2, con CON_SIST_ECON con un peso de 0.6, NIVEL_DE_VIDA con 0.5, e INGRESO_VIVEN con 0.5.

En el Análisis Factorial, el primer paso para aplicar la técnica, es encontrar es la matriz de correlaciones; sin embargo considerando que los datos del presente TFM son de naturaleza categórica, al aplicar la correlación de Pearson se obtendrían factores distorsionados que generarían un desajuste en los supuestos que subyacen al modelo factorial. Para solventar este problema es necesario consideran la matriz de correlaciones policóricas que permite estimar la relación lineal entre variables latentes continuas que subyacen a las variables observables.

Las técnicas y herramientas utilizadas para el desarrollo del presente estudio, contribuyeron de manera significativa a alcanzar los objetivos propuestos en el presente TFM, permitiendo describir, analizar e interpretar los datos de la sección 11 de la Encuestas de las Condiciones de Vida Sexta Ronda (ECV6R) realizada por el INEC-Ecuador, entre el período noviembre 2013 - octubre 2014, definiendo semejanzas y tipologías entre individuos y la desigualdad de la calidad de vida de los hogares de la provincia de Loja.

Para modelar de forma más significativa la calidad de vida de los hogares de la provincia de Loja, se sugiere que se amplié el estudio a través de técnicas como el Análisis Factorial Confirmatorio y Modelado de Ecuaciones Estructurales, a fin de poder encontrar relaciones entre los factores latentes encontrados con el Análisis Factorial Exploratorio.

4. Glosario

Análisis de Correspondencias Múltiples (ACM)	Técnica exploratoria multivariante que permite mostrar simultáneamente las puntuaciones de información categóricas en filas y columnas en una tabla de contingencias bidireccional como coordenadas de puntos en el espacio vectorial de pocas dimensiones
Análisis Factorial Exploratorio (AFE),	Método multivariante que expresa p variables observables como una combinación lineal de m variables latentes, hipotéticas o inobservables, denominadas factores
Análisis Multivariante	conjunto de métodos o técnicas diseñadas para el análisis e interpretación de la información contenida en un conjunto de variables sin perder la interacción o grado en que se afectan unas con otras
Calidad de Vida	término multidimensional asociado a las políticas sociales asociado al bienestar, que significa tener buenas condiciones de vida objetivas y, un alto grado de bienestar subjetivo
Centroide	Punto medio ponderado.
Comunalidad	Cuadrado del peso factorial y que coincide con el coeficiente de correlación R2; porcentaje de variabilidad de la correspondiente variable que explica el factor latente
Contribución a la inercia	Componente de la inercia explicada por un determinado punto en un eje principal.
ECV6R	Encuestas de las Condiciones de Vida 6 Ronda
Estadístico ji-cuadrado ($\chi 2$)	Medida global de las diferencias entre las frecuencias observadas y las frecuencias esperadas de una tabla de contingencia.
Factor Latente	En el análisis factorial, son variables inobservables que carecen de una escala de medida determinada.
I+D	Investigación y Desarrollo
INEC	Instituto Nacional de Estadísticas y Censos – Ecuador
Inercia	Media ponderada de los cuadrados de las distancias $\chi 2$entre los perfiles fila y su perfil media.
Masa	Suma marginal total de una fila o una columna de una tabla dividida por la suma total de la tabla. La utilizamos como pesos en AC
Matriz Binaria	Codificación de datos multivariantes categóricos en forma de variables binarias.

Matriz de Burt	Tipo de matriz compuesta, que consiste en todas los cruzamientos de Q variables categóricas, incluyendo los cruzamientos de las variables con ellas mismas.
Matriz Policórica	Matriz cuyos ítems son medidas categóricas y que mejor se aproximan a las variables continuas para el estadístico correlacional
Perfil	Valores de una fila o columna de una tabla de contingencia dividida por su total. Los puntos que visualizamos en AC son perfiles
Tabla de contingencia	Clasificación de un conjunto de individuos de acuerdo con el cruce de dos variables categóricas. Por tanto, el total de la tabla es el número total de individuos.
TFM	Trabajo de Fin de Master
TIC	Tecnologías de información y comunicación
Unicidad	Varianza única o específica de la variable observable
Valor Propio	Valor inherente de una matriz cuadrada. Forma parte de la descomposición de una matriz como el producto de matrices más simples. En general, las matrices cuadradas tienen tantos valores propios y vectores propios asociados como su rango. En AC, valor propio es sinónimo de inercia principal.

5. Bibliografía

[1] Arias, Bárbara (2013). *El concepto de la calidad de vida en las teorías de desarrollo*. Scientific journals, Vol.5, Núm 8 (2013). Recuperado de: http://revistas.fuac.edu.co/index.php/criteriojuridicogarantista/article/view/413/397.

[2] Bravo, R. Comisión Económica para América Latina y el Caribe (CEPAL) (2000). *Condiciones de vida y Desigualdad Social. Una propuesta para la selección de indicadores*. (pág. 70). Buenos Aires, Argentina.

[3] Cuadras, C. (2014). *Nuevos Métodos de Análisis Multivariante*. Barcelona. CMC Editions.

[4] De la Garza, J., Morales, B. & González, B. (2013). *Análisis Estadístico Multivariante. Un enfoque teórico y práctico*. México DF. MacGrawHill.

[5] Díaz, L. & Morales, M. (2012). *Análisis Estadístico de Datos Multivariados*. Primera Edición. Bogotá. Ed: Editorial Universidad Nacional de Colombia.

[6] Elosua, P. & Zumbo, B. (2008). *Coeficientes de fiabilidad para escalas de respuesta categórica ordenada.* Psicothema, Vol.20, Núm 4 (2008), pp 896-901. Recuperado de: http://www.psicothema.com/psicothema.asp?id=3572

[7] Flora, D. & Curran, P. (2004). *An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data.* Psychological Methods, Vol. 9, pp 466-491.

[8] Greenacre, M. (2008). *La Práctica del Análisis de Correspondencias.* Primera Edición. Bilbao. Rubes Editorial.

[9] Guijarro, F. (2013). *Estadística Aplicada a la Valoración. Modelos Multivarinates.* Valencia. Editorial Universitat Politécnica de Valencia.

[10] Instituto Nacional de Estadísticas y Censos. INEC. (2015). *Metodología de la Encuesta de Condiciones de Vida ECV*. Recuperado de: http://www.ecuadorencifras.gob.ec/documentos/webinec/ECV/ECV_2015/documentos/Metodologia/Documento%20Metodologico%20ECV%206R.pdf

[11] Izenman, A. (2008). *Modern Multivariate Statistical Techniques. Regression, Classification, and Manifold Learning.* Philadelphia. PA 19122. Springer.

[12] López de Murillo, A., Marín, M. & Arroyo, A. (2003). *Modelo de Ecuación Estructural con Intervención de Variables Ordinales. Cálculo de sus Correlaciones.* Ingeniería y Competitividad, Vol. 4,2 pp 60-67. Recuperado de: http://bibliotecadigital.univalle.edu.co/xmlui/handle/10893/1545

[13] Palomba, R. (2009). *Calidad de Vida. Conceptos y medidas. Taller sobre calidad de vida y redes de apoyo de las personas adultas mayores.* CELADE/División de Población, CEPAL. Santiago, Chile. Recuperado de: http://www.ecuadorencifras.gob.ec/documentos/webinec/ECV/ECV_2015/documentos/Metodologia/Documento%20Metodologico%20ECV%206R.pdf.

[14] Peña, D. (2002). *Análisis de Datos Multivariantes*. Madrid. MacGrawHill.

[15] Secretaria Nacional de Planificación. SENPLADES. (2015). *Agenda Zonal, Zona 7 Sur.* Recuperado de: http://www.planificacion.gob.ec/biblioteca/

6. Anexos

6.1 Anexo 1. Tabla binaria

INGRESO_VIVEN.Bien	INGRESO_VIVEN.Mas o menos	INGRESO_VIVEN.Mal	CON_SIST_ECON.Ahorran dinero	CON_SIST_ECON.Equilibran ingresos y gastos	CON_SIST_ECON.Gastan ahorros	CON_SIST_ECON.Obligados a endeudarse	SU_HOGAR_ES.Muy pobre	SU_HOGAR_ES.Pobre	SU_HOGAR_ES.Más o menos pobre	SU_HOGAR_ES.No pobre	NIVEL_VIDA.Mejoró	NIVEL_VIDA.Está igual	NIVEL_VIDA.Empeoró	SU_HOGAR_CON_OTROS.Más pobre	SU_HOGAR_CON_OTROS.Igual	SU_HOGAR_CON_OTROS.Más rico
0	1	0	0	1	0	0	0	0	1	0	0	1	0	0	1	0
0	1	0	0	1	0	0	1	0	0	0	0	1	0	1	0	0
0	1	0	0	1	0	0	0	0	1	0	0	1	0	0	1	0
0	1	0	0	1	0	0	0	1	0	0	0	1	0	0	1	0
0	1	0	0	1	0	0	0	0	1	0	1	0	0	1	0	0
0	1	0	0	1	0	0	0	0	1	0	0	1	0	0	1	0
0	1	0	0	1	0	0	0	0	1	0	0	1	0	0	1	0
0	1	0	0	1	0	0	0	1	0	0	0	0	1	1	0	0
0	1	0	0	1	0	0	0	1	0	0	0	1	0	0	1	0
0	1	0	0	1	0	0	0	0	1	0	0	1	0	0	1	0
0	1	0	0	1	0	0	0	0	1	0	0	1	0	0	1	0
0	1	0	0	1	0	0	0	1	0	0	0	1	0	0	1	0
0	1	0	0	1	0	0	0	0	1	0	0	1	0	1	0	0
0	1	0	0	1	0	0	0	0	1	0	0	1	0	0	1	0
0	0	1	0	1	0	0	0	0	1	0	0	1	0	0	1	0
0	1	0	0	0	0	1	0	1	0	0	0	1	0	0	1	0
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

6.2 Anexo 2. Tabla de burt

	INGRESO_VIVEN.Bien	INGRESO_VIVEN.Mas o menos	INGRESO_VIVEN.Mal	CON_SIST_ECON.Ahorran dinero	CON_SIST_ECON.Equilibran ingresos y gastos	CON_SIST_ECON.Gastan ahorros	CON_SIST_ECON.Obligados a endeudarse	SU_HOGAR_ES.Muy pobre	SU_HOGAR_ES.Pobre	SU_HOGAR_ES.Más o menos pobre	SU_HOGAR_ES.No pobre	NIVEL_VIDA.Mejoró	NIVEL_VIDA.Está igual	NIVEL_VIDA.Empeoró	SU_HOGAR_CON_OTROS.Más pobre	SU_HOGAR_CON_OTROS.Igual	SU_HOGAR_CON_OTROS.Más rico
INGRESO_VIVEN.Bien	121	0	0	27	84	4	6	5	37	35	44	17	94	10	22	96	3
INGRESO_VIVEN.Mas o menos	0	857	0	37	737	26	57	35	416	339	67	57	698	102	214	636	7
INGRESO_VIVEN.Mal	0	0	190	2	131	11	46	65	110	15	0	4	122	64	114	74	2
CON_SIST_ECON.Ahorran dinero	27	37	2	66	0	0	0	1	11	24	30	15	48	3	5	57	4
CON_SIST_ECON.Equilibran ingresos y gastos	84	737	131	0	952	0	0	74	481	321	76	53	785	114	284	662	6
CON_SIST_ECON.Gastan ahorros	4	26	11	0	0	41	0	7	13	17	4	4	26	11	15	25	1
CON_SIST_ECON.Obligados a endeudarse	6	57	46	0	0	0	109	23	58	27	1	6	55	48	46	62	1
SU_HOGAR_ES.Muy pobre	5	35	65	1	74	7	23	105	0	0	0	1	67	37	80	25	0
SU_HOGAR_ES.Pobre	37	416	110	11	481	13	58	0	563	0	0	17	466	80	215	344	4
SU_HOGAR_ES.Más o menos pobre	35	339	15	24	321	17	27	0	0	389	0	43	298	48	51	336	2
SU_HOGAR_ES.No pobre	44	67	0	30	76	4	1	0	0	0	111	17	83	11	4	101	6
NIVEL_VIDA.Mejoró	17	57	4	15	53	4	6	1	17	43	17	78	0	0	13	64	1
NIVEL_VIDA.Está igual	94	698	122	48	785	26	55	67	466	298	83	0	914	0	259	645	10
NIVEL_VIDA.Empeoró	10	102	64	3	114	11	48	37	80	48	11	0	0	176	78	97	1
SU_HOGAR_CON_OTROS.Más pobre	22	214	114	5	284	15	46	80	215	51	4	13	259	78	350	0	0
SU_HOGAR_CON_OTROS.Igual	96	636	74	57	662	25	62	25	344	336	101	64	645	97	0	806	0
SU_HOGAR_CON_OTROS.Más rico	3	7	2	4	6	1	1	0	4	2	6	1	10	1	0	0	12

6.3 Anexo 3. Rutina en R, para encontrar pesos factoriales, varianza explicada y comunalidades.

```
pesos_factores = function(x,var,num_fac){
#Elaborado por: Franco Salcedo
# mayo 2017

# Variables de entrada
  #x: matriz de valores y vectores propios
  #var: número de variables
  #num_fac : número de factores para alcanzar las comunalidades

#Variables de Salida
  #A:  pesos factorales o factores
  #varianza_expl: varianza explicada por cada factor
  # comunalidad: comunalidad

# pesos factgoriales
A <-matrix(0, nrow = var, ncol = var )
for(i in 1:var){

  A[,i]<- -x$vectors[,i]*sqrt(x$values[i])
}

#varianza explicada por cada factor
#A2<-A^2

v_var<-apply(A^2,2,sum)

# porcentaje
v_por<-v_var/var
g=0

acum_var<- matrix(0, nrow = 1, ncol = var)
for(i in 1:var){
  acum_var[,i]<-g + v_por[i]
  g<-acum_var[,i]
}
# COMUNALIDADES
comuna<-apply((A[,1:num_fac])^2, 1, sum)

VAR_EXPL<-as.matrix(rbind(v_var,v_por*100,acum_var*100))
retorna<-list(peso_factores=A, varianza_expl=VAR_EXPL, comunalidad=comuna)


return(retorna)
}
```