ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO

Aplicación de un Modelo Matemático mediante Estadística Multivariante para identificar los estratos sociales de los hogares de la ciudad de Ambato en el cuarto trimestre del año 2019

ALEX ROLANDO MOYOTA PAGUAY

Trabajo de Titulación modalidad Proyectos de Investigación y Desarrollo, presentado ante el Instituto de Posgrado y Educación Continua de la ESPOCH, como requisito parcial para la obtención del grado de:

MAGÍSTER EN MATEMÁTICA, MENCIÓN MODELACIÓN Y DOCENCIA

Riobamba – Ecuador

Julio 2023

DECLARACIÓN DE AUTENTICIDAD

Yo, Alex Rolando Moyota Paguay, declaro que el presente **Trabajo de Titulación Proyectos de Investigación y Desarrollo**, es de mi autoría y que los resultados del mismo son auténticos y originales. Los textos constantes en el documento que provienen de otra fuente están debidamente citados y referenciados.

Como autor, asumo la responsabilidad legal y académica de los contenidos de este Trabajo de Titulación de Maestría, el patrimonio intelectual pertenece a la Escuela Superior Politécnica de Chimborazo.

Riobamba, julio de 2023

ALEX ROLANDO MOYOTA PAGUAY

No. Cédula: 0604063958

ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO

EL TRIBUNAL DEL TRABAJO DE TITULACIÓN CERTIFICA QUE:

El **Trabajo de Titulación modalidad Proyectos de Investigación y Desarrollo, titulado:** Aplicación de un Modelo Matemático mediante Estadística Multivariante para identificar los estratos sociales de los hogares de la ciudad de Ambato en el cuarto trimestre del año 2019, de responsabilidad del señor Alex Rolando Moyota Paguay, ha sido minuciosamente revisado por los Miembros del Tribunal del trabajo de titulación, el mismo que cumple con los requisitos científicos, técnicos, legales, en tal virtud el Tribunal autoriza su presentación.

Firmado electrónicamente por: AMALIA ISABEL ESCUDERO VILLA

Ing. Amalia Isabel Escudero Villa; Ph.D.

PRESIDENTA

Firmado electrónicamente por: LUIS MARCELO CORTEZ BONILLA

Mat. Luis Marcelo Cortez Bonilla; Mgtr.

DIRECTOR

Firmado electrónicamente por: JOSE LUIS PEREZ ROJAS

Ing. José Luis Pérez Rojas; Mgtr.

MIEMBRO

Firmado electrónicamente por: BLADIMIR ENRIQUE URGILES RODRIGUEZ

Ing. Bladimir Enrique Urgiles Rodríguez; Mgtr.

MIEMBRO

Riobamba, julio de 2023

DEDICATORIA

Dedico este trabajo con todo mi corazón y de manera muy especial a mi hermana Irlanda por ser la principal promotora en la edificación de mi vida profesional, gracias por inculcarme el ejemplo del esfuerzo y deseos de superación, en ella tengo el espejo en el cual me quiero reflejar pues sus virtudes infinitas y su gran corazón me llevan a admirarla cada día más.

Gracias DIOS por concederme la mejor de las hermanas.

A mi Padre, a mi Madre y a mi hermano que son las personas que me motivan a ser cada día mejor.

Alex

AGRADECIMIENTO

Agradezco a DIOS, por guiarme en cada paso y decisión que he tomado en el trascurso de mi vida, siendo mi fortaleza en aquellos momentos de dificultad y debilidad. A mis Padres por todo su apoyo incondicional y exhaustivo para lograr escalar con éxito un peldaño más en mi vida profesional. Un agradecimiento muy especial a mi amigo Edy gracias por compartir tus conocimientos y experiencias, por tu valioso aporte y ayuda desinteresada para lograr concluir con éxito este trabajo.

Alex

TABLA DE CONTENIDO

Páginas

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

El objetivo de esta investigación fue aplicar un modelo matemático mediante Estadística Multivariante que permita identificar los estratos sociales de los hogares de la ciudad de Ambato en el cuarto trimestre del año 2019. Para esta investigación se seleccionó la base de datos de las Encuestas de Hogares ENEMDU como fuente de información y se aplicó mediante el *software* libre R un análisis de componentes principales (ACP), este modelo permitió identificar las variables que mejor resumen la información de la base de datos, para segmentar de manera eficiente los estratos sociales de los hogares de la ciudad de Ambato. Estas variables son: las horas trabajadas por semana, satisfacción laboral, antigüedad laboral y la dependencia económica. Mediante estas variables, se definen los umbrales a partir de los cuartiles de corte identificados en las curvas de densidad de cada una de las variables; así, estas variables permiten estratificar los hogares ambateños e identificar a qué segmento pertenecen, utilizando la escala que este estudio ha generado. Se comprobó la hipótesis mediante una prueba estadística chi-cuadrado. Se concluyó que el Análisis de Componentes Principales es un modelo estadístico del Análisis Multivariante que permitió estratificar la ciudad de Ambato. Finalmente, se recomienda utilizar este modelo como referente para que sea replicado en las diferentes urbanizaciones en general, como insumo de análisis socioeconómico para ayudar a la ejecución y análisis en políticas públicas.

Palabras claves: <ESTRATIFICACIÓN SOCIAL>, <ANÁLISIS MULTIVARIANTE>, <ANÁLISIS DE COMPONENTES PRINCIPALES>, <CURVAS DE DENSIDAD>, <PRUEBA CHI-CUADRADO>.

SUMMARY

This research aimed to apply a mathematical model through Multivariate Statistics that allows identifying the social strata of households in the city of Ambato in the fourth quarter of 2019. For this research, the ENEMDU Household Surveys database was selected. as a source of information, and principal component analysis (PCA) was applied using the free software R. This model made it possible to identify the variables that best summarize the information in the database, to efficiently segment the social strata of the households of the city of Ambato. These variables are the hours worked per week, job satisfaction, job seniority, and economic dependency. Through these variables, the thresholds are defined from the cut-off quartiles identified in the density curves of each variable. Thus, these variables make it possible to stratify "Ambateños" households and identify which segment they belong to using the scale that this study has generated. The hypothesis was verified by employing a statistical chi-square test. It was concluded that the Principal Components Analysis is a statistical model of the Multivariate Analysis that allowed to stratify the city of Ambato. Finally, it is recommended to use this model as a reference to be replicated in the different urbanizations in general, as an input for socioeconomic analysis to help the execution and analysis of public policies.

Key words: <SOCIAL STRATIFICATION>, <MULTIVARIATE ANALYSIS>, <PRINCIPAL COMPONENT ANALYSIS>, < DENSITY CURVES >, <CHI-SQUARE TEST>.

CAPÍTULO I

1 INTRODUCCIÓN

1.1 Situación problemática

La estratificación social viene siendo un instrumento de clasificación que existe desde sociedades antiguas en las cuales predominaban las diferencias sociales (Fachelli, 2009). Como lo hace notar Sémbler R. (2006), la diversidad de enfoques clásicos sobre la estratificación social permitió introducir conceptos como clase social, status, estructura ocupacional, etc.

Los principales estudios de estratificación social realizados a nivel internacional están fundamentados en diferentes metodologías, variables y enfoques. (Hardy, 2014). Para la región de América Latina, variables como los años de estudios del jefe de hogar, ingresos, acceso a servicios y bienes de hogar son las más relevantes. (Mendes, 2015)

En el caso del Ecuador, el Instituto Nacional de Estadística y Censos (INEC) realizó la Encuesta de Estratificación de Nivel Socioeconómico (NSE) en 2011, a los hogares urbanos de Quito, Guayaquil, Cuenca, Ambato y Machala, la cual permite identificar a los grupos socioeconómicos relevantes y sus características a través del análisis en cada una de las dimensiones (características de la vivienda, posesión de bienes, acceso a tecnología, hábitos de consumo, nivel de educación, actividad económica del hogar). (INEC, 2011)

Tradicionalmente los hogares se han segmentado por variables ligadas a su condición económica: nivel de ingreso, gastos, consumo. Este tipo de estratificación es útil, pero deja de lado elementos clave del comportamiento de un hogar, tanto a nivel de consumidor, así como sujeto político.

En este contexto, lo que se pretende mediante el presente estudio es aplicar una herramienta estadística capaz de segmentar los hogares de acuerdo al comportamiento de variables e individuos en espacios multidimensionales. Esta propuesta desglosa resultados de una herramienta que permite capturar las múltiples dimensiones de la estratificación de hogares, pero desde dimensiones que los propios datos configuran.

Para identificar los estratos sociales de los hogares de la ciudad de Ambato, se aplicará un Modelo Estadístico Multivariante por medio del software libre R, utilizando la base de datos de la Encuesta Nacional de Empleo, Desempleo y Subempleo (ENEMDU).

1.2 Formulación del problema

¿La Estadística Multivariante, será un instrumento válido y confiable para identificar los estratos sociales de los hogares de la ciudad de Ambato en el cuarto trimestre del año 2019?

1.3 Preguntas directrices o específicas de la investigación

¿De qué manera las variables socioeconómicas influyen en la estratificación de la ciudad de Ambato en el cuarto trimestre del año 2019?

¿Permitirá el Análisis Geométrico de Datos (AGD) de las variables seleccionadas, la segmentación de los estratos sociales de los hogares de la ciudad de Ambato?

¿Cómo ayudará el modelo de regresión más *scoring*, en la identificación de los estratos sociales de los hogares de la ciudad de Ambato durante el cuarto trimestre del año 2019?

1.4 Justificación de la investigación

Según varias investigaciones y experiencias en diferentes países de la región, tradicionalmente los hogares se han segmentado por variables ligadas a su condición económica; este tipo de estatificación es útil, pero deja de lado elementos clave del comportamiento de un hogar. Estas distinciones provocan la necesidad de contar con herramientas de segmentación que capturen agrupaciones sociales más allá de dimensiones tradicionales como la económica.

El aporte que tendrá este trabajo es proporcionar de un instrumento a través de un análisis multivariado, que permita capturar las múltiples dimensiones de la estratificación de hogares, pero desde dimensiones que los propios datos configuren. Para ello se realizará un análisis de las características demográficas de la base de datos de la Encuesta Nacional de Empleo, Desempleo y Subempleo (ENEMDU) del 2019.

La presente investigación es muy importante porque permitirá identificar los estratos sociales de los hogares de la ciudad de Ambato para el año 2019, a partir de variables seleccionadas solo después de un proceso de análisis multivariado, además que será un referente para futuras investigaciones, y de esta manera generar información para que el ejecutor de políticas públicas pueda entender fácilmente al grupo que se está acercándose y como debe hacerlo.

1.5 Objetivos de la investigación

1.5.1 Objetivo general

Aplicar un Modelo Matemático mediante Estadística Multivariante para identificar los estratos sociales de los hogares de la ciudad de Ambato en el cuarto trimestre del año 2019.

1.5.2 Objetivos específicos

a) Seleccionar las variables socioeconómicas de los hogares de la ciudad de Ambato, analizando la base de datos ENEMDU año 2019.
b) Aplicar un Modelo Multivariante para determinar un número reducido de variables que capturen la mayor cantidad de información, para la segmentación de los estratos sociales.

c) Establecer los umbrales para identificar los estratos de los hogares de la ciudad de Ambato en el año 2019.
d) Validar el Modelo Multivariante aplicado, para determinar si es adecuado para la estratificación de los hogares de la ciudad de Ambato.

1.6 Hipótesis

La Estadística Multivariante permitirá identificar los estratos sociales de los hogares de la ciudad de Ambato en el cuarto trimestre del año 2019.

CAPÍTULO II

2 MARCO TEÓRICO

2.1 Antecedentes del estudio

Con el objetivo de construir un indicador del ingreso para estratificar los hogares y segmentos del censo 2000 en Costa Rica, Madrigal (2004) planteo un modelo de regresión multivariable con base en la Encuesta de hogares de propósitos múltiples 2001, donde la variable dependiente es el logaritmo natural del ingreso per cápita del hogar, en función de las variables independientes (educación, ocupación, pertenencias de la familia, hacinamiento, dependientes y ocupados) que están asociadas a la estratificación. Una ecuación para los hogares en los que el jefe se encuentra ocupado, y otra en los que no; las ecuaciones resultantes se aplican a los hogares del censo 2000 y usando el procedimiento Cluster con el algoritmo *K-means* se definen tres niveles de ingreso.

En el caso de Argentina, Fachelli (2009) desarrollo un nuevo modelo de estatificación social y nuevo instrumento para su medición, con el objetivo de conformar un modelo de estratificación social desde una perspectiva multidimensional que, incorporando los aportes tradicionales de estratificación, como la clasificación ocupacional y el ingreso, considera pertinente tomar en cuenta otras dimensiones de análisis. Las técnicas multivariadas utilizadas en esta investigación es el Análisis de Correspondencia Múltiples y el Análisis de Clasificación. Los resultados obtenidos mostraron reducir la complejidad y la diversidad de una sociedad a un número relativamente pequeño y significativo de estratos.

Para el caso del Ecuador el Instituto Nacional de Estadísticas y Censos (INEC), con el objetivo de identificar los grupos socioeconómicos y sus características; en el 2011, realizó la Encuesta Nacional de Estratificación del Nivel Socioeconómico (NSE) para los hogares urbanos de las ciudades de Quito, Guayaquil, Cuenca, Ambato y Machala. Los resultados de la NSE tienen el objetivo de servir como instrumento para una adecuada segmentación del mercado de consumo identificando variables que permitan clasificar y caracterizar los niveles socioeconómicos de los hogares ecuatorianos (INEC, 2011).

Astudillo & Salazar (2020) desarrollaron un nuevo enfoque para la estratificación socioeconómica del Ecuador, en esta investigación se busca construir un criterio de estratificación social basado en los patrones de consumo de los hogares, utilizando los datos de la Encuesta Nacional de Ingresos y Gastos de los Hogares Urbanos y Rurales del Ecuador (ENIGHUR 2011-2012). Aplicando una técnica multivariante K-means, los resultados obtenidos en esta investigación sugieren que el consumo puede acercar a los hogares considerando las preferencias

reveladas; así mismo, este criterio permite identificar cómo influyen las características del jefe de hogar en la estratificación socioeconómica.

De acuerdo al análisis de las investigaciones realizadas, se puede observar que existe varios enfoques para la estratificación socioeconómica de un grupo social, además que existe un amplio conjunto de técnicas multivariantes que segmentan los hogares por variables netamente ligadas a su condición económica (nivel de ingreso, gastos, consumo). Este tipo de estratificaciones es útil, pero deja de lado elementos clave del comportamiento de un hogar; por lo mismo lo que se pretende con el presente trabajo de titulación es segmentar los hogares de acuerdo al comportamiento de variables e individuos en espacios multidimensionales (más allá de lo económico) esto implica contar con herramientas estadísticas capaces de capturar esta complejidad.

2.2 Bases teóricas

2.2.1 Análisis Multivariante

Peña (2002) define el análisis multivariante como un grupo de técnicas que sirven para examinar simultáneamente un conjunto de datos, con el objetivo de estudiar las variables de los elementos u objetos de una población; para describir una realidad compleja.

De igual manera, Meneses (2019) define al análisis multivariante como el conjunto de técnicas estadísticas que tienen como objetivo analizar e interpretar las relaciones entre distintas variables de manera simultánea, mediante la construcción de modelos estadísticos complejos que permiten distinguir la contribución independiente de cada una de ellas en el sistema de relaciones y, de este modo, describir, explicar o predecir los fenómenos que son objeto de interés para la investigación.

Por consiguiente; el análisis multivariante, es una de las herramientas más empleadas por la comunidad científica, para el estudio estadístico de múltiples variables medidas en los elementos de una población. Para este caso de estudio a través de un modelo multivariado lo que se requiere es identificar los grupos socioeconómicos más relevantes de la ciudad de Ambato, para lo cual se requiere disponer de una base de datos bien estructurada.

2.2.1.1 Técnicas de Análisis Multivariante

Habiendo definido el análisis multivariante como el marco analítico general que permite modelar las diversas relaciones existentes entre las diferentes variables involucradas en una investigación, Meneses (2019) ante la diversidad de técnicas disponibles establece en la Tabla 1-2, una clasificación general de algunas de las utilizadas más frecuentemente a partir de los dos grandes contextos de dependencia e interdependencia.

Tabla 1-2: Clasificación de las técnicas de análisis multivariante en función de los objetivos de la investigación y de las características de los datos.

Objetivo general	Escenario de aplicación	Características de los datos	Técnica multivariante
Analizar relaciones de interdependencia para describir la estructura de los datos	Identificación de grupos de características similares	Diversas variables cuantitativas	Análisis de componentes principales
			Análisis factorial
		Diversas variables cualitativas	Análisis de correspondencias
	Identificación de grupos de individuos similares	Diversas variables cuantitativas o cualitativas	Análisis de conglomerados
	Identificación de grupos de objetos similares	Diversas variables cuantitativas o cualitativas	Escalamiento multidimensional
Analizar relaciones de dependencia para hacer explicaciones o predicciones	Explicación de la variabilidad de los individuos	Una variable dependiente cuantitativa	Regresión múltiple
		Dos o más variables dependientes cuantitativas	Correlación canónica
	Explicación de la variabilidad de los grupos de individuos	Una variable dependiente cuantitativa	ANOVA de dos o más factores o ANCOVA
		Dos o más variables dependientes cuantitativas	MANOVA o MANCOVA
	Predicción de la pertenencia de los individuos a grupos	Una variable dependiente cualitativa	Análisis discriminante
			Regresión logística
Analizar relaciones de dependencia e interdependencia simultáneamente	Evaluación del ajuste de modelos concatenados	Diversas variables cuantitativas	Ecuaciones estructurales

Fuente: Meneses, 2019.
Realizado por: Moyota, Alex, 2022.

Técnicas de dependencia

Estas técnicas están enfocadas en modelos de regresión, es decir una variable o conjunto de variables se identifica como la variable dependiente que debe predecirse o explicarse por otras variables conocidas como variables independientes. A estas técnicas de dependencia se las identifica mejor como un análisis multivariable.

Técnicas de Interdependencia

Este tipo de técnicas involucra el análisis simultáneo de todas las variables en el conjunto, sin distinción entre variables dependientes y variables independientes. En este tipo de análisis lo que interesa es la relación de las variables, las estructuras que tienen entre sí. A estas técnicas de interdependencia se las identifica mejor como un análisis multivariado.

Agregando a lo anterior el Análisis Multivariante es un conjunto de metodologías que nos permiten abordar datos desde una perspectiva más integral y holística con respecto a lo que se puede hacer con los modelos simples.

Un aspecto muy importante de los modelos multivariantes, es que podemos considerar toda la información disponible dentro de un conjunto de datos y maximizarlo para poder obtener conclusiones que sean relevantes, para ello una de las técnicas más utilizadas es el Análisis de Componentes Principales (ACP) que se puede aplicar bajo diferentes contextos tanto de investigación como de desarrollo científico.

2.2.2 Análisis de Componentes Principales (ACP)

Para autores como Everitt & Hothorn (2011) y Jolliffe (2002), el análisis de componentes principales es quizás la técnica de análisis multivariante más antigua y mejor conocida. Fue introducido por primera vez por Karl Pearson (1901) y desarrollado de manera independiente por Hotelling (1933). Al igual que otros métodos multivariantes, su uso era limitado hasta la llegada de las computadoras, con ello actualmente ya se dispone de estos métodos en la mayoría de software estadísticos.

La idea central del análisis de componentes principales (PCA) es reducir la dimensionalidad de un conjunto de datos que consta de una gran cantidad de variables interrelacionadas, conservando la mayor parte de la variación presente en el conjunto de datos. Esto se logra mediante la conversión a un nuevo conjunto de variables, los componentes principales, que son combinaciones lineales de las variables originales, que no están correlacionadas entre sí y ordenados de manera que los primeros retengan la mayor parte de la variación presente en todas las variables originales. (Jolliffe, 2002)

De acuerdo con varios autores, se puede decir que el análisis de componentes principales está enfocado principalmente en reducir un gran número de variables multivariantes en un menor número de nuevas variables llamadas (componentes) en las cuales se concentre la mayor cantidad de información, como se ilustra en la Figura 1-2.

Tabla de Datos					**Componentes**			
$\begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}$				$\longrightarrow$	$\begin{bmatrix} C_{11} & C_{12} & \cdots & C_{1p} \\ C_{21} & C_{22} & \cdots & C_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ C_{n1} & C_{n2} & \cdots & C_{np} \end{bmatrix}$			
100% de la información					**75%**	**19%**	$\cdots$	**0,06%**

Figura 1-2: Trasformación de las variables originales en componentes.
Fuente: Base de datos ENEMDU 2019.
Realizado por: Moyota, Alex, 2022.

Desde el punto de vista de Peña (2002), el análisis de componentes principales tiene una doble utilidad; por un lado, permite representar los datos en R^2, para poder identificar segmentos, clúster, similitudes y disimilitudes, y, por otro lado, convertir variables inicialmente correlacionadas en nuevas variables compuestas no correlacionadas, facilitando la interpretación de los datos.

En síntesis, una de las técnicas de análisis exploratorio de datos más conocidas es el análisis de componentes principales (ACP), una de las más importantes en la minería de datos porque permite agrupar, segmentar, analizar las similitudes entre las variables de una base de datos, así como su importancia radica en que nos permite reducir la dimensión de las variables de una base de datos.

Notaciones y Símbolos.

De acuerdo con la simbología común de diferentes autores, los conceptos básicos del álgebra matricial necesarios para un Análisis de Componentes Principales se presentan a continuación.

Matriz de Datos

La base para el uso de ACP es la estructura de correlación (interdependencia) entre las variables cuantitativas identificadas en una población, en la que cada individuo se define en términos de estas variables (León González et al., 2008). La matriz de datos X, de dimensiones $(n \times p)$, contiene las medidas de p variables cuantitativas tomadas sobre n individuos; como se muestra a continuación:

$$\mathrm{X} = \left(x_{ij}\right) = \begin{bmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & \cdots & x_{ij} & \ldots & x_{ip} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix} \quad (1)$$

Donde x_{ij} es el elemento genérico de está matriz, que representa el valor de la variable j sobre el individuo i. Es decir:

Datos x_{ij} , donde: $i = 1, \ldots, n$ representa el individuo;
$j = 1, \ldots, p$ representa la variable.

La varianza

La varianza se puede definir como la medida que nos indica qué tan disperso o distribuido se encuentra un conjunto de valores. La varianza se calcula como el promedio de los cuadrados de las diferencias de los valores con respecto de la media de este conjunto de valores:

$$s_j^2 = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$$

Observación: Algunos autores definen la varianza muestral usando $n - 1$ en lugar de n en el denominador. Hay razones teóricas para hacerlo, especialmente cuando n es pequeño.

Desviación estándar

La desviación estándar se obtiene como la raíz de la varianza muestral:

$$s_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}$$

Matriz de Varianzas y Covarianzas

La covarianza es la medida de que tanto cambian dos variables juntas. Es una medida de fuerza de la correlación entre dos conjuntos de variables. Si la covarianza entre dos variables es cero, las variables no están correlacionadas. Sin embargo, el que dos variables no estén correlacionadas, no significa que sean independientes, dado que la correlación es sólo una medida de dependencia lineal y para dos variables (x_j, x_k) la covarianza se calcula con la siguiente ecuación:

$$s_{jk} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$$

Para una variable multivariante esta información se puede presentar claramente en la matriz de varianzas y covarianzas. Definimos esta matriz de la siguiente manera:

$$S = \frac{1}{n}\sum_{i=1}^{n}(\mathrm{X}_i - \bar{\mathrm{X}})\,(\mathrm{X}_i - \bar{\mathrm{X}})' \tag{2}$$

Que, al multiplicar los vectores:

$$\begin{bmatrix} x_{i1} - \bar{x}_1 \\ \vdots \\ x_{ip} - \bar{x}_p \end{bmatrix}\left[x_{i1} - \bar{x}_1, \cdots, x_{ip} - \bar{x}_p\right] = \begin{bmatrix} (x_{i1} - \bar{x}_1)^2 & \cdots & (x_{i1} - \bar{x}_1)(x_{ip} - \bar{x}_p) \\ \vdots & \ddots & \vdots \\ (x_{ip} - \bar{x}_p)(x_{i1} - \bar{x}_1) & \cdots & (x_{ip} - \bar{x}_p)^2 \end{bmatrix}$$

Se obtiene la matriz de varianzas y covarianzas S:

$$S = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix}$$

Una matriz cuadrada y simétrica de orden $p \times p$, que contiene en la diagonal principal a las varianzas y fuera de la diagonal las covarianzas entre cada par de valores del conjunto de datos.

Correlación y matriz de correlación

El coeficiente de correlación entre dos variables x_j, x_k nos indica la dependencia lineal que existe entre estas dos variables y se obtiene como:

$$r_{jk} = \frac{s_{jk}}{\sqrt{s_{ii}s_{kk}}}$$

Denotaremos por R a la matriz de correlación, una matriz cuadrada y simétrica que muestra la dependencia lineal por pares entre las variables:

$$R = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}$$

Notemos que la correlación de una variable consigo mismo siempre va ser uno, por lo tanto, la diagonal principal de la matriz de correlaciones va ser siempre uno.

Vectores Propios y Valores Propios

Un vector propio (o autovector) de una matriz A de $n \times n$ se puede definir como un vector $u \in R^n$, diferente de 0, tal que para cierto escalar $\lambda \in R$.

$$Au = \lambda u$$

Donde λ, es un escalar llamado valor propio (o autovalor) de A, es decir, λ es un valor propio de la matriz A si existe una solución no trivial de $Au = \lambda$u, y a u se lo denomina vector propio asociado al valor propio λ.

En inglés un vector propio se conoce como *eigenvector* y la palabra *"eigen"* en alemán significa pertenece a, o que es peculiar a, por lo que el vector propio de una matriz es un vector que pertenece y caracteriza la estructura de los datos. (Hartman, 2011)

2.2.2.1 Cálculo de los Componentes

Cálculo del primer componente

El primer componente principal será la combinación lineal de las variables de partida con la máxima varianza. Los valores de esta primera componente para todos los n individuos estará representado por el vector z_1, que en notación matricial se tiene:

$$z_1 = Xa_1$$

Se busca la primera componente principal z_1 que tenga la máxima varianza. Como las variables originales tienen media cero, el vector z_1 también tendrá media nula. Dado S como la matriz de varianzas y covarianzas de las observaciones, la varianza de z_1 será:

$$Var(z_1) = \frac{1}{n} z_1' z_1 = \frac{1}{n} a_1' X' X a_1 = a_1' S a_1 \quad (3)$$

Podemos maximizar la varianza sin límite incrementando el módulo del vector a_1, es decir, en las ecuaciones de las componentes principales existe un factor de escala arbitraria (existen infinitas soluciones en las mismas direcciones del espacio); para esto es conveniente que los vectores

directores tengan un módulo de 1. Para que la maximización de la varianza de z_1 dada por la ecuación (3) tenga solución debemos imponer una restricción al módulo del vector a_1.

$$a_1' a_1 = 1$$

Es decir, tomamos un vector unitario en la dirección de la primera componente principal. Para maximizar la expresión (3), introduciremos esta restricción por el multiplicador de Lagrange:

$$L = a_1' S a_1 - \lambda(a_1' a_1 - 1)$$

derivando e igualando a cero:

$$\frac{\partial L}{\partial a_1} = 2Sa_1 - 2\lambda a_1 = 0$$

se tiene:

$$Sa_1 = \lambda a_1 \tag{4}$$

Lo que implica que a_1 es un vector propio de la matriz de varianzas y covarianzas S, y λ su correspondiente valor propio asociado.

La matriz de covarianzas S tiene p valores propios $\lambda_1, \dots, \lambda_p$ que supondremos distintos y ordenados de forma decreciente $\lambda_1 \geq \dots \geq \lambda_p \geq 0$. Ahora para determinar qué valor propio de la matriz S es la solución de la ecuación (4), se multiplica por a_1' esta ecuación obteniendo la expresión:

$$Var(z_1) = a_1' S a_1 = \lambda a_1' a_1 = \lambda$$

Por la ecuación (3) se puede interpretar que el valor propio λ es la varianza muestral de la primera componente principal z_1. Como esta es la cantidad que queremos maximizar, λ debe ser λ_1 el valor propio más grande de la matriz de covarianza S y a_1 el vector propio asociado que define los coeficientes de cada variable en el primer componente principal.

Cálculo de segundo componente

La segunda componente principal z_2 se obtiene al maximizar la suma de varianzas de $z_1 = Xa_1$ y $z_2 = Xa_2$, que viene a ser la traza de la matriz de covarianzas del mejor plano de proyección de las variables X, donde los vectores a_1 y a_2 definen este plano.

Utilizando multiplicadores de Lagrange se incorpora las restricciones de que las direcciones deben de tener módulo unitario esto es: $a_i' a_i = 1$, donde $i = 1,2$.

$$\varphi = a_1' S a_1 + a_2' S a_2 - \lambda_1(a_1' a_1 - 1) - \lambda_2(a_2' a_2 - 1) \tag{5}$$

Derivando e igualando a cero:

$$\frac{\partial \varphi}{\partial a_1} = 2Sa_1 - 2\lambda_1 a_1 = 0$$

$$\frac{\partial \varphi}{\partial a_2} = 2Sa_2 - 2\lambda_2 a_2 = 0$$

Se tiene que:

$$Sa_1 = \lambda_1 a_1 \quad (6)$$
$$Sa_2 = \lambda_2 a_2 \quad (7)$$

Lo que implica que a_1 y a_2 son autovectores de S, y λ_1 y λ_2 sus correspondientes autovalores asociados.

Sustituyendo (6) y (7) en (5), se obtiene la varianza máxima expresada como la suma de los autovalores λ_1 y λ_2 :

$$\varphi = \lambda_1 + \lambda_2 \quad (8)$$

De la ecuación (7), es claro que a_2 es un vector propio de S, que define los coeficientes de cada variable en la segunda componente principal. Siguiendo el mismo argumento, podemos obtener las sucesivas componentes principales a partir de los correspondientes valores propios y autovectores.

Generalización

En definitiva, las componentes principales se obtendrán de la descomposición de la matriz de varianzas y covarianzas S, en valores propios (autovalores) mediante:

$$|S - \lambda I| = 0 \quad (9)$$

Y sus vectores propios (autovectores) asociados son:

$$(S - \lambda_i I)a_i = 0 \quad (10)$$

Sea Z una matriz cuyas columnas son los valores de p componentes en los n individuos, estas nuevas variables se relacionan con las variables originales por:

$$Z = XA$$

El cálculo de componentes principales corresponde a aplicar una transformación ortogonal A las variables X (ejes originales) para obtener nuevas variables Z no correlacionadas. Esta operación puede entenderse como la selección de nuevos ejes de coordenadas, que coincidan con los "ejes naturales" de los datos. (Peña, 2002)

Propiedades de los componentes principales

Peña (2002), menciona las siguientes propiedades de los componentes como nuevas variables.

Conservan la varianza original es decir que la suma de las varianzas de las componentes z_i es igual a la suma de las varianzas de las variables iniciales x_i, pero sus distribuciones son muy diferentes en los dos conjuntos, por tanto:

$$\sum_{i=1}^{p} Var(z_i) = \sum \lambda_i = \sum_{i=1}^{p} Var(x_i)$$

A si mismo la varianza generalizada de los componentes es igual a la varianza original.

$$|S_x| = \lambda_1 \dots \lambda_p = \prod\prod_{i=1}^{p} Var\,(z_i) = |S_z|$$

La proporción de variabilidad explicada, es decir la cantidad de información recogida por cada una de las componentes o de otra manera la cantidad de varianza absorbida por cada una de las componentes, es el cociente entre el valor propio correspondiente a esa componente y la suma de todos los valores propios.

$$\frac{\lambda_h}{\sum_{i=1}^{p} \lambda_i}$$

De manera similar se obtiene la proporción de variabilidad total explicada para las q primeras componentes:

$$\frac{\sum_{j=1}^{q} \lambda_j}{\sum_{i=1}^{p} \lambda_i}$$

Las covarianzas entre cada componente principal y las variables X vienen dadas por el producto de las coordenadas del vector propio a_i que define el componente por el valor propio λ_i, donde a_i es el vector de coeficientes de la componente z_i.

$$Cov(z_i; x_1, \dots x_p) = \lambda_i a_i = (\lambda_i a_{i1}, \dots, \lambda_i a_{ip})$$

Entre una componente principal z_i y una variable x_j, la correlación está dada por:

$$Corr(z_i, x_j) = a_{ij} \frac{\sqrt{\lambda_i}}{s_j}$$

Donde a_{ij} es el coeficiente de la variable, $\sqrt{\lambda_i}$ es la desviación típica del componente y s_j la desviación típica de la variable.

2.2.2.2 Selección de componentes

En esta sección se sugieren algunas alternativas para seleccionar el número de componentes principales que recogen la mayor parte de la información, es decir, la mayor variabilidad de las variables originales estandarizadas. Existen varios criterios de selección de Componentes Principales, siendo los dos más comunes el criterio Scree Plot y el criterio de porcentaje de varianza acumulada.

Scree Plot

Cattell (1966) sugiere realizar un gráfico de los valores propios λ_i frente al número de componentes i llamado diagrama scree. Se seleccionan ejes hasta que se vea un decrecimiento brusco en la magnitud de los valores propios, es decir, hasta que todos los demás componentes tengan el mismo valor de λ_i. La idea es buscar una "rodilla" o un punto de inflexión en la curva, a partir de la cual los valores propios son aproximadamente iguales, entonces el número de componentes seleccionados es el valor de i correspondiente a al "rodilla" en la gráfica. El criterio es mantener algunos componentes excluyendo los asociados a valores pequeños. (Peña, 2002)

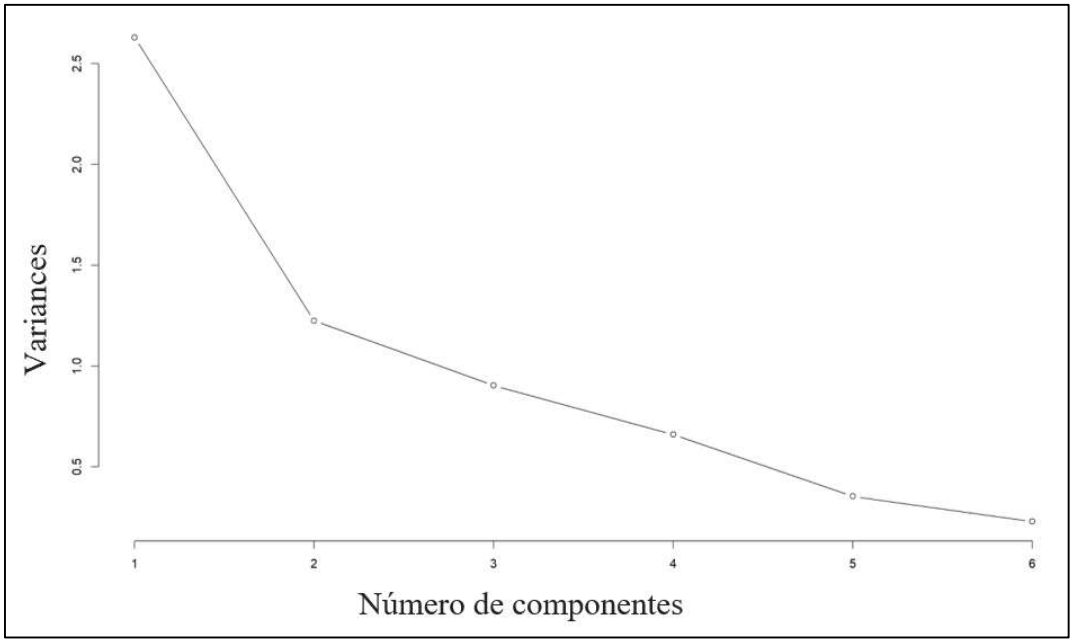

Figura 2-2: Gráfica Scree Plot para determinar el número de CP.
Fuente: INEC - Base de datos ENEMDU 2019.
Realizado por: Moyota, Alex, 2022.

Porcentaje de variabilidad deseado.

Este criterio plantea retener las componentes necesarias para explicar un determinado porcentaje de la varianza total de las variables originales; por lo general, se sugieren valores entre el 70% y el 90% (Everitt & Hothorn, 2011). De acuerdo a la literatura en la práctica, es deseable retener al menos un 80% de la variabilidad, aunque esto no es una ley; esta regla es arbitraria y debe aplicarse con cierta cautela. En realidad, es el investigador quien puede elegir un numero a priori que considere un buen porcentaje de explicación a partir del conocimiento que tiene de sus datos.

Los gráficos de porcentaje de variabilidad indican qué cantidad de la variabilidad total es estimada por cada componente principal.

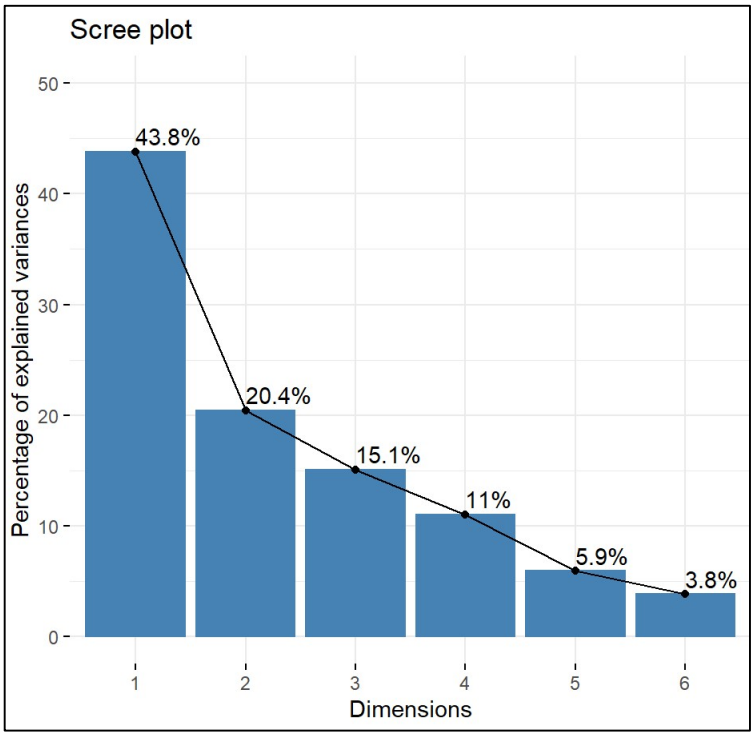

Figura 3-2: Scree plot.
Fuente: INEC - Base de datos ENEMDU 2019.
Realizado por: Moyota, Alex, 2022.

2.2.3 Estratificación social

De acuerdo con Marín García (2021) la estratificación social es la forma en la que se clasifican los individuos de una determinada sociedad, de acuerdo a características económicas, políticas, sociales y culturales. Existe una variedad de variables socioeconómicas, como ingresos, educación, gastos o consumo, capacidad de pago entre otras, que se utilizan para estratificar los hogares dependiendo cual sea el objetivo de la estratificación, por lo que cada estrato variará de acuerdo con los criterios considerados. (López & Sepúlveda, 2014)

Este tipo de procesos genera información que ayuda a tomar decisiones de carácter económico, social y cultural a los formuladores y/o ejecutores de políticas públicas tanto a nivel global como local. (Argüello, 1991)

2.2.4 La estratificación socioeconómica en el mundo

A nivel internacional, existen diferentes criterios para definir los estratos socioeconómicos de una población en base a diversas metodologías, variables y enfoques (Astudillo Macas & Salazar Ortiz, 2020). En la Tabla 2-2 se presentan algunos de los criterios existentes en todo el mundo y en general, en la mayoría de los países, se puede apreciar que las variables mayormente utilizadas son la ocupación y el nivel educativo del jefe de hogar. Para los países europeos, la ocupación del

jefe hogar parece ser la principal variable relevante para segmentar la sociedad. Por otra parte, para países de América Latina, el nivel de ingreso, posesión de bienes, los ingresos actuales del hogar y el lugar de residencia parecen ser mucho más relevantes (Mendes, 2015).

Tabla 2-2: Comparación de variables utilizadas para estratificación social en el mundo.

Países/Variables	Ocupación del jefe del hogar	Educación del jefe de hogar	Nivel de escolaridad del cónyuge	Posesión de bienes	Existencia de un empleado interno	Ingreso actual de la familia	Número de personas que contribuyen a los ingresos	Número de miembros de hogar	Nivel de vida	Lugar de residencia
Argentina	x	x		x						
Chile	x	x		x		x				x
Venezuela	x	x	x	x		x	x			x
Perú		x		x	x					x
Uruguay	x	x			x					
Costa Rica	x	x		x	x	x				
El Salvador	x	x		x	x	x				
Honduras	x	x		x	x	x				
Guatemala	x	x		x	x	x				
Nicaragua	x	x		x	x	x				
Puerto Rico	x	x								
México		x		x						
Portugal	x	x								
Italia	x	x				x	x	x	x	
Reino Unido	x									
Alemania	x									
Francia	x									
Rusia	x									
Japón						x				

Fuente: Recuperado de Astudillo Macas & Salazar Ortiz (2020).
Realizado por: Moyota, Alex, 2022.

2.2.5 Prueba estadística ji-cuadrado (o chi cuadrado)

El estadístico ji-cuadrado (X^2), que tiene una distribución de probabilidad del mismo nombre, es una técnica estadística (no paramétrica) que sirve para poner a prueba hipótesis referidas a distribuciones de frecuencias (McHugh, 2013). De manera general (Mendivelso & Rodríguez, 2018) resaltan que, esta prueba contrasta las frecuencias observadas con las esperadas de acuerdo con la hipótesis nula.

Fue Karl Pearson quien propuso el siguiente estadístico:

$$X^2 = \sum_{i=1}^{k} \left[\frac{(O_i - E_i)^2}{E_i} \right] \tag{11}$$

Donde:

O_i: Frecuencias observadas.

E_i: Frecuencias esperadas o teóricas.

Como en cualquier prueba de contraste estadístico, se intenta rechazar la hipótesis nula y aceptar en consecuencia, la hipótesis alternativa.

CAPÍTULO III

3 METODOLOGÍA DE INVESTIGACIÓN

En este capítulo se realizará la identificación de las variables que sinteticen la mayor parte de la información contenida en la base de datos generada por la ENEMDU para el 2019; esto se conseguirá a partir de un análisis de componentes principales. Una vez identificado las variables se podrá segmentar los hogares de la ciudad de Ambato. Para cerrar este capítulo se detallará brevemente los paquetes y librerías del software estadístico RStudio que se han empleado para este fin.

3.1 Selección de fuente de información

Para esta investigación se seleccionó la base de datos de las Encuestas de Hogares ENEMDU como fuente de información. La Encuesta Nacional de Empleo, Desempleo y Subempleo (ENEMDU) es una encuesta que es llevada a cabo por el INEC con el objetivo brindar una visión sobre la situación del empleo en el país, las características del mercado laboral, la actividad económica y las fuentes de ingresos de la población ecuatoriana. La ENEMDU es una encuesta de aplicación continua, y la información que de ella se obtiene también ayuda a determinar la magnitud de determinados fenómenos sociodemográficos, aportando datos e indicadores que muestran la situación en la que viven mujeres y hombres dentro de sus hogares y viviendas. (INEC, 2019).

La base de datos generada por la ENEMDU tiene una representatividad nacional, urbana, rural, para cinco ciudades principales (Quito, Guayaquil, Cuenca, Machala, Ambato) y el resto del país; para la población de 15 años en adelante.

Cabe mencionar que la ENEMDU emplea un muestreo probabilístico estratificado bietápico de elementos, con un factor de expansión que permite inferir los datos muestrales al total de hogares existentes en la ciudad de Ambato.

Por lo tanto, para este trabajo los datos estadísticos se obtuvieron del banco de datos abiertos del INEC. Repositorio de donde se descargó la base de datos acumulada del cuarto trimestre del 2019 a nivel nacional, para luego filtrar los datos correspondientes de la ciudad de Ambato; a partir de lo cual se obtuvo 3 333 encuestas de hogares que incluyen variables de población, características ocupacionales, nivel educativo, nivel de ingreso, gastos y consumo.

3.2 Análisis descriptivo de variables a utilizar

Para esta investigación de acuerdo a la literatura revisada en la sección 2.1 y 2.2.4, las variables de la encuesta de hogares ENEMDU del cuarto trimestre 2019 que pueden representar factores

de estratificación son seis variables continuas sin orden de importancia que se detallan en la siguiente tabla.

Tabla 1-3: Descripción de variables.

Código	Variable	Descripción
p10a	Nivel educativo	Nivel educativo aprobado por el jefe de hogar (ninguno, centro de alfabetización, primaria, etc.).
p24	Horas trabajadas por semana	Horas trabajadas la semana anterior en todos los trabajos que tenga el jefe del hogar.
p43	Dependencia económica	Tipo de relación laboral del jefe de hogar (con nombramiento, contrato permanente, contrato temporal, etc.).
p45	Antigüedad laboral	Número de años en el trabajo actual.
p59	Satisfacción laboral	Satisfacción del jefe de hogar con el trabajo actual (contento, poco contento, descontento pero conforme, totalmente descontento, no sabe no responde).
ingpc	Ingreso per cápita	Agregado de ingreso económico mensual para el jefe de hogar.

Fuente: INEC - Base de datos ENEMDU 2019.
Realizado por: Moyota, Alex, 2022.

3.3 Selección de la técnica de Análisis Multivariante

En este estudio se aplicará un análisis de componentes principales como una herramienta estadística que permita reducir la dimensión de la base de datos y quedarnos solamente con unas cuantas variables que sintetizan la mayor parte de la información contenida en los datos. Esto se consigue aplicando el método matemático explicado en la sección 2.2.2 y 2.2.2.1, de manera que se puede segmentar los hogares de la ciudad de Ambato de acuerdo al comportamiento de las variables e individuos en espacios multidimensionales, logrando una estratificación desde dimensiones que los propios datos configuren.

Las aplicaciones del ACP son numerosas y entre ellas se pueden citar la estratificación multivariada a partir de una combinación lineal de las variables que proporcionan la máxima variabilidad, como se menciona en los diferentes apartados del capítulo 2, bajo este contexto se seleccionan las dimensiones solo después de un proceso de análisis multivariado siendo el ACP el método que se ajusta a las demandas de esta investigación. Utilizando las variables continuas que se filtraron conceptualmente en el apartado 3.2, el análisis de componentes principales, arroja que, las variables que mejor resumen la estratificación de la ciudad de Ambato están representadas con las contribuciones capturadas en la principal dimensión.

3.4 Metodología aplicada en el software R

Para efectuar los diferentes cálculos numéricos y estadísticos detallados en las secciones 2.2.2 y 2.2.2.1 del análisis de componentes principales se utilizará el lenguaje estadístico R, mediante su entorno gráfico RStudio. En este apartado se analizarán los módulos o paquetes (packages) que se deben instalar para poder efectuar un análisis exploratorio de los datos (matriz de correlación, gráficas de correlación), análisis de componentes principales (ACP) y la representación gráfica de las componentes. Los paquetes que se requieren para poder analizar los datos e identificar aquellas variables que estratifican a la ciudad de Ambato son dplyr, stats, corrplot, ggplot2 y factoextra las mismas que se detallan a continuación.

3.4.1 Paquete dplyr

Este paquete al ser un miembro central del tidyverse se cargará con este paquete, pero también se lo puede cargar usando dplyr. El paquete dplyr presenta cinco funciones clave y es una gramática que está diseñado para la manipulación de datos, de manera que sea más fácil operar en un grupo particular de la base de datos completa. Para obtener información más detallada sobre el uso y funcionamiento de este paquete y sus funciones, consulte la página web de (Wickham et al., 2022).

3.4.2 Paquete stats

El sistema R base o core o núcleo de esta plataforma, además de las funciones más fundamentales, incluye un conjunto de paquetes, entre ellos el paquete de estadísticas R (stats) y que representan un gran porcentaje de la funcionalidad del sistema (Verdugo Chaura, 2022). El paquete stats contiene funciones para cálculos estadísticos y la generación de números aleatorios.

Para el desarrollo de los cálculos estadísticos del apartado 2.2.2 se usarán específicamente las funciones `cor, prcomp, summary, biplot y heatmap`. Para obtener una lista completa de funciones de este paquete, escriba `library (help = "stats")` directamente en la consola de R.

3.4.2.1 Función cor

Está función calcula la covarianza o correlación de la matriz de datos.

Sintaxis

```
cor(x, y = NULL, use = "everything", method = c("pearson",
"kendall", "spearman"))
```

Argumentos

`x:`	un vector numérico, matriz o data frame.
`y:`	`NULL` (default) o un vector, matriz o data frame con dimensiones compatibles con `x`. El valor predeterminado es equivalente a `y = x` (pero más eficiente).
`use:`	una cadena de caracteres opcional que proporciona un método para calcular las covarianzas en presencia de valores faltantes. Debe ser (una abreviatura de) una de las cadenas `"everything"`, `"all.obs"`, `"complete.obs"`, `"na.or.complete"`, o `"pairwise.complete.obs"`.
`method:`	una cadena de caracteres que indica qué coeficiente de correlación (o covarianza) se va a calcular. Uno de `"pearson"` (predeterminado), `"kendall"` o `"spearman"`: se puede abreviar.

Valor

Nos devuleve una matriz de correlación.

3.4.2.2 Función prcomp

La función `prcomp`, realiza un análisis de componentes principales en la matriz de datos dada.

Esta función realiza el cálculo de los componentes explicados en la sección 2.2.2.1.

Sintaxis

```
prcomp(x, center = TRUE, scale. = TRUE, ...)
```

Argumentos

`x:`	una matriz numérica (o data frame) que proporciona los datos para el análisis de componentes principales.
`center:`	un valor lógico que indica si las variables deben cambiarse para que estén centradas en cero. Alternativamente, se puede proporcionar un vector de longitud igual al número de columnas de `x`. El valor se pasa a escala.
`scale.:`	un valor lógico que indica si las variables deben escalarse para tener una varianza unitaria antes de que se lleve a cabo el análisis. El valor predeterminado es falso (`FALSE`) para mantener la coherencia con `S`, pero en general se recomienda escalar. Alternativamente, se puede proporcionar un vector de longitud igual al número de columnas de `x`. El valor se pasa a escala.

Valor

Devuelve los resultados como un objeto de clase `prcomp`.

3.4.2.3 Función summary

`summary` es una función genérica utilizada para producir resúmenes de resultados de los resultados de varias funciones de ajuste de modelos. La función invoca métodos particulares que dependen de la clase del primer argumento.

Sintaxis

```
summary(object, ...)
```

Argumentos

`object:` un objeto para el que se desea un resumen.

Valor

Nos devuelve resúmenes de objetos

3.4.2.4 Función biplot

Esta función traza un biplot de datos multivariados en el dispositivo de gráficos actual.

Sintaxis

```
biplot(x, ...)
```

Argumentos

`x:` un objeto para el que se desea trazar un biplot de datos multivariados.

Valor

Devuelve la gráfica de PCA

3.4.2.5 Función heatmap

La función `heatmap` permite crear mapas de calor en R a partir de una matriz. Un mapa de calor es una imagen en falso color (básicamente una imagen `(t(x))`) con un dendrograma agregado en el lado izquierdo y en la parte superior. Normalmente, se lleva a cabo el reordenamiento de las filas y columnas de acuerdo con algún conjunto de valores (medias de fila o columna) dentro de las restricciones impuestas por el dendrograma. Para mayor información de los argumentos que incluye esta función escriba `help("heatmap")` directamente en la consola.

Sintaxis

```
heatmap (x, symm=TRUE, …)
```

Argumentos

`x:` matriz numérica de los valores a graficar

`symm:` indicación lógica si `x` debe tratarse simétricamente; solo puede ser cierto cuando `x` es una matriz cuadrada.

`...:` argumentos adicionales pasados a la imagen, por ejemplo, `col` especificando los colores.

Valor

Dibuja un mapa de calor.

3.4.3 Paquete PerformanceAnalytics

Este paquete proporciona una colección de funciones econométricas para análisis de riesgo y desempeño de instrumentos financieros o carteras. De este paquete se usará específicamente la

función `chart.Correlation` para obtener una gráfica de la matriz de correlación (observar capítulo 2). Para visualizar una lista completa de todas las funciones de este paquete, escriba `library (help = "PerformanceAnalytics")` directamente en la consola de R.

3.4.3.1 Función chart.Correlation

Esta función muestra una visualización de una Matriz de Correlación. En la parte superior el valor (absoluto) de la correlación más el resultado de la prueba `cor.test` como estrellas. En la parte inferior muestra los diagramas de dispersión bivariados, con una línea ajustada.

Sintaxis

```
chart.Correlation(R,histogram=TRUE,method=c("pearson","kendall",
"spearman"), ...)
```

Argumentos

`R:`	datos para el eje `x`, pueden tomar matrices, vectores o series temporales
`histogram:`	`TRUE/FALSE` mostrar o no un histograma.
`method:`	una cadena de caracteres que indica qué coeficiente de correlación (o covarianza) se va a calcular. Se puede abreviar uno de `"pearson"` (predeterminado), `"kendall"` o `"spearman"`.
`… :`	cualquier otro parámetro.

Valor

Devuelve un gráfico de matriz de correlación.

3.4.4 Paquete ggplot2

`ggplot2` es un lenguaje de gráficos que funciona por capas que al final se compilan sobre una única imagen para generar como resultados gráficos de alta calidad. Exactamente tiene siete capas (`Theme, Coordinates, Statistics, Facets, Geometries, Aesthetics y Data`) que nos permite crear visualizaciones elegantes de los datos y obtener gráficas y figuras de gran calidad con un mejor aspecto (Wickham, 2016).

Para obtener mayor información de este paquete y sus funciones escriba `library(help="ggplot2")` directamente en la consola de R. Si se desea una comprensión más profunda de las aplicaciones de `ggplot2` se puede revisar el libro de recetas de R Graphics (Chang, 2022) mismo que me servirá de guía para realizar este trabajo.

3.4.5 Paquete factoextra

`factoextra` es un paquete R que proporciona algunas funciones que facilitan la extracción y visualización de la salida de análisis exploratorios de datos multivariantes, incluidos 'PCA' (Análisis de componentes principales), 'CA' (Análisis de correspondencia), 'MCA' (Análisis de correspondencia múltiple), 'FAMD' (Análisis factorial de datos mixtos), 'MFA' (Análisis factorial

múltiple) y 'HMFA' (Análisis jerárquico factorial múltiple) de diferentes paquetes de R. También contiene funciones para simplificar algunos pasos del análisis de agrupamiento y proporciona una elegante visualización de datos basada en 'ggplot2' con menos escritura.

Para obtener una lista completa de las funciones de este paquete, escriba `library (help = "factoextra")` directamente en la consola de R.

3.4.5.1 Función fviz_pca

A partir de la literatura revisada en la sección 2.2.2 el análisis de componentes principales (PCA) reduce la dimensionalidad de los datos multivariados a dos o tres que se pueden visualizar gráficamente con una pérdida mínima de información. La función `fviz_pca()` proporciona una visualización elegante basada en `ggplot2` de los individuos/variables de la salida del análisis de componentes principales (PCA) de la función `prcomp` integrada en el paquete `stats` como se revisó en la sección 3.4.2.2.

Se utilizan las siguientes funciones, del paquete `factoextra`:

`fviz_pca_ind():`	Gráfico de individuos
`fviz_pca_var():`	Gráfico de variables
`fviz_pca_biplot():`	Biplot de individuos y variables

Para obtener una lista completa de los argumentos de estas funciones, escriba `help("fviz_pca")` directamente en la consola de R.

Sintaxis

```
fviz_pca_ind(X, col.ind = "black", gradient.cols = NULL, repel = FALSE, …)

fviz_pca_var(X, col.var = "black", gradient.cols = NULL, repel = FALSE, …)
```

Argumentos

`X:`	un objeto de clase prcomp del paquete [`stats`]
`col.ind, col.var:`	color para individuos y variables, respectivamente. Puede ser una variable continua o una variable factorial. Los valores posibles incluyen también: `"cos2", "contrib", "coord", "x"` o `"y"`. En este caso, los colores de los individuos/variables se controlan automáticamente por sus cualidades de representación `("cos2")`, contribuciones `("contrib")`, coordenadas `(x^2+y^2, "coord")`, valores x `(" x")` o valores de y `("y")`. Para usar la coloración automática `(por cos2, contrib, ....)`, asegúrese de que `habillage ="none"`.
`gradient.cols:`	vector de colores que se usará para el degradado de n colores. Los valores permitidos incluyen las paletas de colores brewer y ggsci.
`repel:`	un valor booleano `(TRUE/FALSE)`, si usar `ggrepel` para evitar sobretrazar las etiquetas de texto o no.

... : argumentos adicionales.

Valor

Esta función nos permite la visualización del análisis de componentes principales en una gráfica de alta calidad generada por el paquete `ggplot2`.

3.4.5.2 Función get_pca

Esta función extrae todos los resultados (coordenadas, coseno cuadrado, correlaciones, contribuciones) para las variables/individuos activos de las salidas de un Análisis de Componentes Principales (PCA). Esta función nos permite extraer los resultados para las variables e individuos o por separado es decir solo individuos o solo variables a partir de la siguiente sintaxis respectivamente.

`get_pca()`: Extraer los resultados para variables e individuos
`get_pca_ind()`: Extraiga los resultados solo para individuos
`get_pca_var()`: Extraer los resultados solo para variables

Como el objetivo de esta investigación es determinar las variables que mejor resumen la estratificación de la ciudad de Ambato, en este paso nos interesa es extraer solamente los resultados para variables. Para mayor información sobre los argumentos que incluye esta función escriba `help("get_pca")` directamente en la consola.

Sintaxis

```
get_pca_var(res.pca)
```

Argumento

`res.pca`: un objeto devuelto por la función `PCA` del paquete [FactoMineR] o de la función `prcomp` y `princomp` del paquete [stats].

Valor

Esta función nos arroja una lista de matrices que contienen todos los resultados para las variables activas, incluyendo las coordenadas, coseno cuadrado y contribuciones de las variables.

Todos los paquetes (packages) y sus funciones descritas en este capítulo, que servirán para realizar los cálculos y procesamiento de datos multivariantes están disponibles desde la denominada Red Exhaustiva de Archivos R (Comprehensive R Archive Network, CRAN). Es muy importante hacer notar la variedad de argumentos que nos ofrecen cada una de las funciones utilizadas en esta investigación, pero solo se emplearan los argumentos descritos en cada uno de los apartados.

CAPÍTULO IV

4 RESULTADOS Y DISCUCIÓN

Este capítulo consta de cuatro secciones que presentan los resultados y sus respectivos análisis. La primera sección describe la correlación existente entre las variables que se detallan en el apartado 3.2, a partir de un análisis exploratorio de datos empleando métodos gráficos y no gráficos. En la segunda sección se identifican las variables que explican en mayor medida la varianza, obtenidas de un análisis de componentes principales (ACP). En una tercera sección se definen los umbrales de cada una de las variables seleccionadas. Finalmente, en la cuarta sección mediante una prueba de chi cuadrado se espera determinar si el modelo multivariante aplicado es adecuado para la segmentación de los estratos sociales de los hogares de la ciudad de Ambato.

4.1 Análisis exploratorio de las variables socioeconómicas

Una condición para llevar a cabo un análisis multivariado es que, las variables deben estar correlacionadas. Para analizar la correlación entre las variables socioeconómicas seleccionadas en el capítulo 3, se realizará un análisis exploratorio gráfico y no gráfico de la base de datos, utilizando los paquetes y funciones detalladas en la sección 3.4.

Al aplicar un análisis exploratorio no gráfico a la base de datos ENEMDU del año 2019 utilizando la sintaxis del apartado 3.4.2.1, se obtuvo como resultado la matriz de correlación Figura 1-4, en la que se puede observar que los valores de la diagonal principal son iguales a uno, esto debido a que es la comparación de una variable consigo misma, el resto de valores como se puede apreciar tienen una correlación positiva. De la literatura revisada el máximo valor que puede tomar una correlación positiva es igual a uno, si ponemos atención a los mayores valores de está matriz podemos identificar las variables que podrían estar más relacionadas positivamente, por ejemplo, 0.75139 es la correlación que existe entre las variables p24 (horas trabajadas por semana) y p59 (satisfacción laboral), es decir existe una buena correlación entre estas dos variables. Al no tener valores negativos en la matriz se hace notar que no existe una correlación negativa entre las variables.

Los valores de correlación iguales o cercanos a cero como por ejemplo 0.09425 nos indica que la variable ingpc (ingreso per cápita) no tiene nada que ver o no se relaciona con la variable p43 (dependencia económica). En forma general en esta matriz se puede observar que existe una cierta correlación entre las variables estudiadas.

RStudio Source Editor

n

Filter

	p10a	p24	p43	p45	p59	ingpc
p10a	1.00000000	0.2732250	0.21560953	0.05083629	0.19248241	0.30593357
p24	0.27322502	1.0000000	0.49579441	0.57541212	0.75139132	0.12644708
p43	0.21560953	0.4957944	1.00000000	0.11533004	0.44501697	0.09425153
p45	0.05083629	0.5754121	0.11533004	1.00000000	0.52727243	0.10517243
p59	0.19248241	0.7513913	0.44501697	0.52727243	1.00000000	0.02550958
ingpc	0.30593357	0.1264471	0.09425153	0.10517243	0.02550958	1.00000000

Figura 1-4: Matriz de correlación.
Fuente: INEC - Base de datos ENEMDU 2019.
Realizado por: Moyota, Alex, 2022.

Debido a la dificultad de ver tantos números en la matriz de correlación y para corroborar si existe correlación entre las variables de estudio se ha empleado un método gráfico. Utilizando la sintaxis del apartado 3.4.2.5 el software R nos imprime un mapa de calor Figura 2-4 que visualmente nos puede ayudar a definir mejor si los datos están correlacionados.

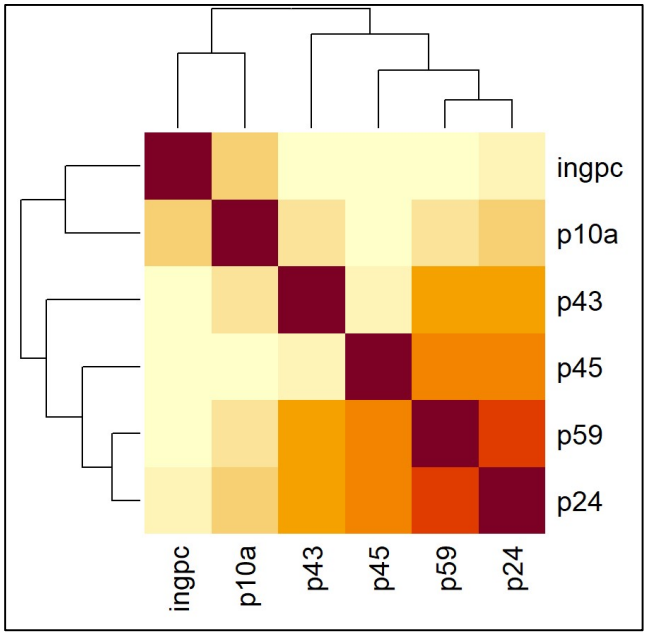

Figura 2-4: Mapa de calor (heatmap).
Fuente: INEC - Base de datos ENEMDU 2019.
Realizado por: Moyota, Alex, 2022.

En el mapa de calor, visualmente se puede observar que las variables p24 (horas trabajadas por semana), p59 (satisfacción laboral), p45 (antigüedad laboral) y p43 (dependencia económica) son las que mayormente están correlacionadas. Probablemente las variables que están más

correlacionadas van hacer las más importantes para este estudio, porque son las que están variando de manera conjunta entre sí.

El requerimiento más importante para realizar un análisis multivariado es revisar si en efecto hay correlaciones entre las variables, la opción no gráfica que es la matriz de correlación nos muestra cierta incertidumbre en algunas variables, pero con la ayuda de este mapa de calor (heatmap) se puede afirmar que vale la pena realizar un análisis multivariado.

4.2 Resultado del Análisis de Componentes Principales (ACP)

Al correr las líneas de código que se detallan en la sección 3.4.2.2 y 3.4.2.3 se ejecuta la técnica multivariante ACP a la base de datos ENEMDU 2019 seleccionada en el apartado 3.1, mediante el cual se obtuvo un resumen de los resultados del análisis de componentes principales (ACP) Tabla 1-4 que muestra la desviación estándar, proporción de la varianza explicada y la proporción de varianza acumulada.

Tabla 1-4: Importancia de los componentes.

```
> cprin_enemdu_persona_201912_1 <-prcomp(enemdu_persona_201912_1,center=TRUE,
scale.=TRUE) #prcomp genera PCA

> summary(cprin_enemdu_persona_201912_1)

Importance of components:
                          PC1    PC2    PC3    PC4     PC5     PC6
Standard deviation     1.6216 1.1066 0.9505 0.8126 0.59476 0.47791
Proportion of Variance 0.4383 0.2041 0.1506 0.1100 0.05896 0.03807
Cumulative Proportion  0.4383 0.6423 0.7929 0.9030 0.96193 1.00000
```

Fuente: INEC - Base de datos ENEMDU 2019.
Realizado por: Moyota, Alex, 2022.

Como se muestra en la Tabla 2-4 este análisis genera tantas componentes como variables originales existen. Entonces siendo así: PC1, PC2, PC3, PC4, PC5 y PC6 son nuevas variables que resultan de la combinación lineal de todas las variables originales, pero como podemos observar estas variables tienen pesos diferentes entre sí. El peso o *"loadings"* es el efecto que tiene cada una de las variables en cada una de las componentes como se revisó en el capítulo 2.

Por lo tanto, lo que estamos observando son operaciones matriciales que van a producir nuevas variables llamadas CP1, CP2, CP3, …, etc. que se van alimentar de la combinación lineal de todas las variables originales con diferentes pesos, los mismos que nos van a revelar cuales son las variables más significativas.

Tabla 2-4: Loadings, pesos de cada variable.

```
> cprin_enemdu_persona_201912_1
Standard deviations (1, .., p=6):
[1] 1.6215936 1.1065630 0.9505496 0.8125715 0.5947584 0.4779099
Rotation (n x k) = (6 x 6):
             PC1          PC2          PC3          PC4         PC5          PC6
p10a  -0.2484125  0.62940311  0.133612507 -0.70388028  0.1496767 -0.08024998
p24   -0.5612165 -0.10488121 -0.002749102 -0.02583429 -0.1802098  0.80055321
p43   -0.3861859  0.08390478  0.664395111  0.44508077  0.4274728 -0.14686622
p45   -0.4142401 -0.27678391 -0.591852542 -0.06390039  0.5992158 -0.19586596
p59   -0.5299257 -0.21795058  0.039745729 -0.06882203 -0.6115387 -0.53979669
ingpc -0.1475518  0.67948499 -0.434568892  0.54494818 -0.1713158 -0.03689000
```

Fuente: INEC - Base de datos ENEMDU 2019.
Realizado por: Moyota, Alex, 2022.

También debemos notar en referencia a lo explicado en el marco teórico que las componentes generadas se constituyen matemáticamente como los eigenvectores que son nuevas proyecciones de las variables originales que están dados por la combinación de todas las variables originales con diferentes pesos (eigenvalores). Entonces a partir de estos resultados podemos decir que ese eigenvalor va representar el peso de la variable determinada para toda la variación de ese vector o eigenvector.

4.2.1 Análisis de individuos y variables de la salida del ACP

La función `fviz_pca()` detallada en el apartado 3.4.5.1 permite visualizar gráficamente los componentes identificados y nivel de asociación entre ellos y de los individuos/variables de la salida del análisis de componentes principales Figura 3-4.

En esta etapa se realiza dos interpretaciones: (1) la asociación entre variables para identificar dimensiones conceptuales de segmentación, y (2) la *"clusterización"* de individuos respecto a las dimensiones conceptuales para identificar concentraciones de la segmentación.

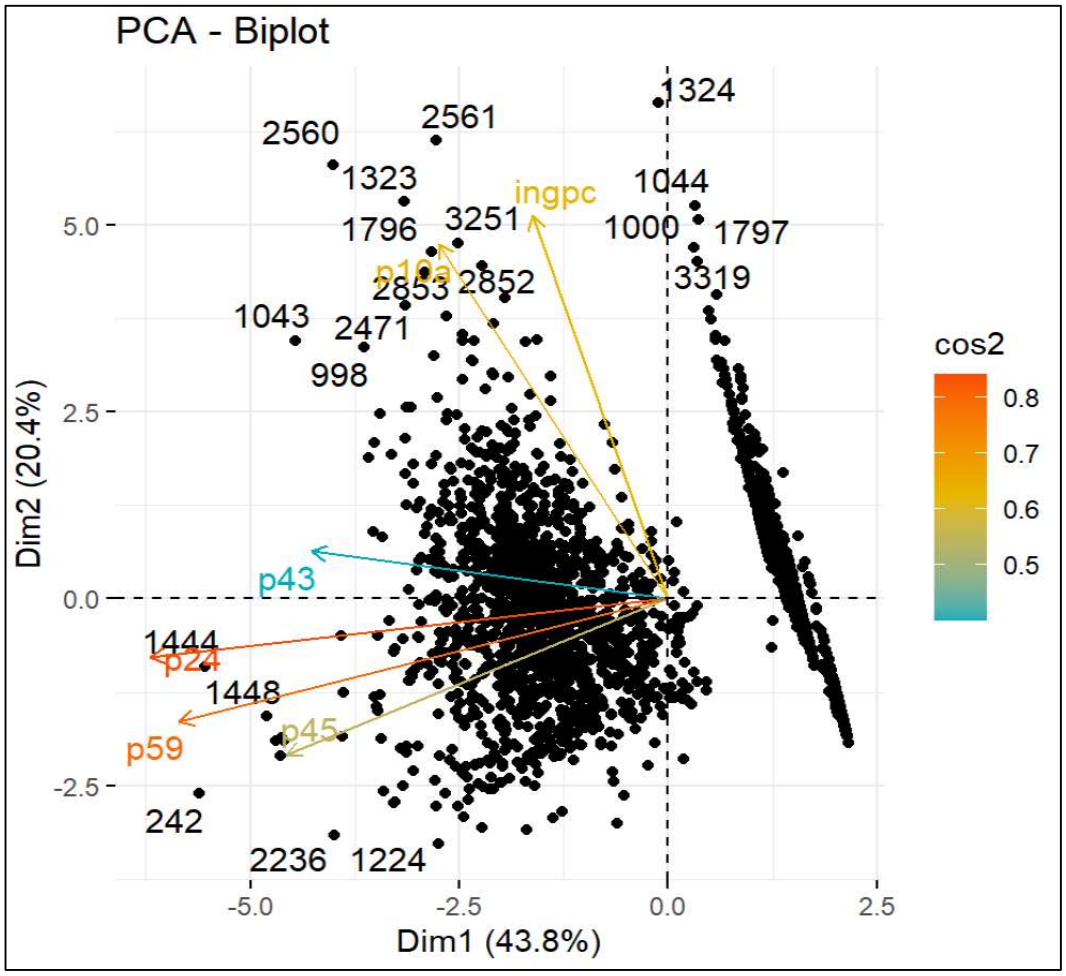

Figura 3-4: PCA - Biplot.
Fuente: INEC - Base de datos ENEMDU 2019.
Realizado por: Moyota, Alex, 2022.

4.2.1.1 Asociación entre variables para identificar dimensiones conceptuales de segmentación

Para este paso primero limpiamos a los "individuos" o casos de la segmentación para concentrarnos en identificar e interpretar la segmentación de variables.

En la Figura 4-4 se observa que la contribución con la primera dimensión esta graficada de acuerdo a la intensidad del color (más rojo mayor contribución, más azul menor contribución). En este punto de análisis nos interesa saber qué variables están contribuyendo más y que significado conceptual tienen las variables que se agrupan entre sí.

Observando la gráfica de la Figura 4-4 y la Tabla 2-4 las variables de mayor contribución para la dimensión 1 (Dim1) son: Horas trabajadas por semana (p24), Satisfacción laboral (p59), Antigüedad laboral (p45) y Dependencia económica (p43). Preliminarmente llamaremos a esta dimensión "Capital humano del jefe de hogar".

Las variables de mayor contribución para la dimensión 2 (Dim2) son: Ingreso per cápita (ingpc), Nivel educativo (p10a), Antigüedad laboral (p45) y Satisfacción laboral (p59). Tentativamente llamaremos a esta dimensión: "Capital económico del jefe de hogar"

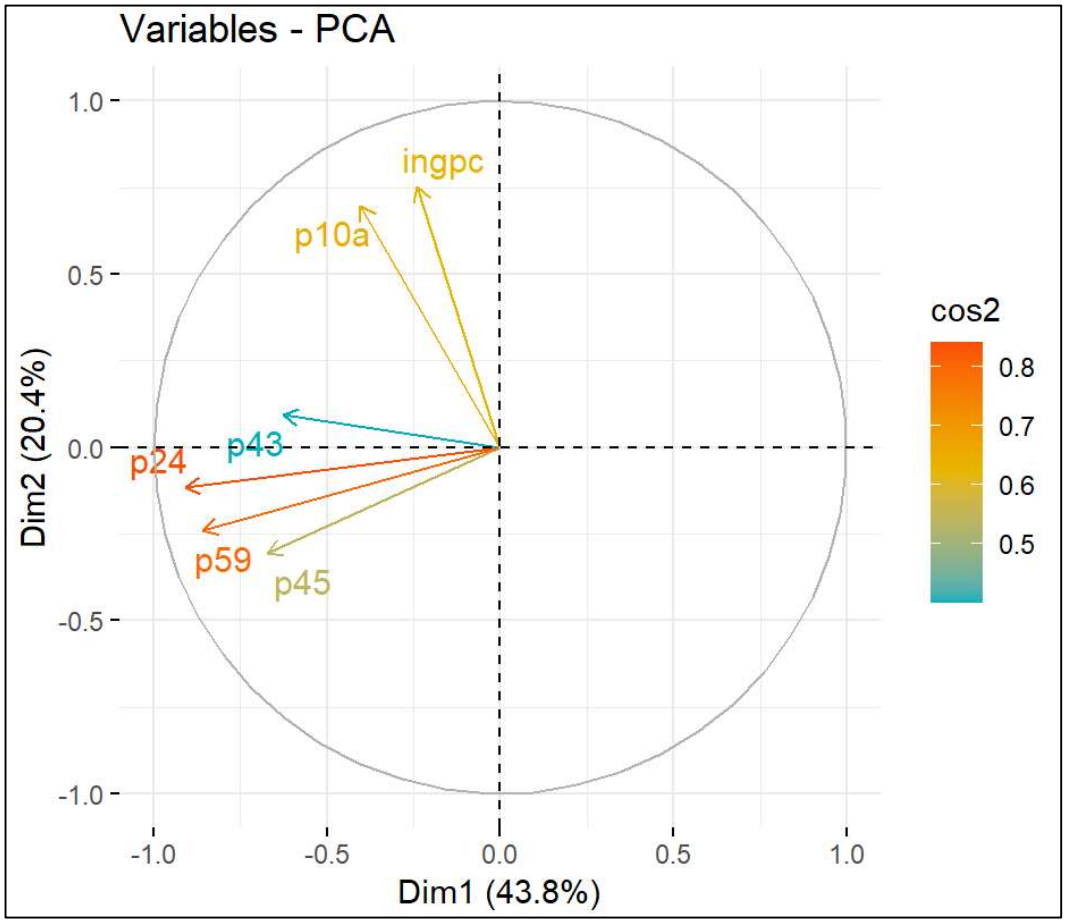

Figura 4-4: Círculo de correlación.
Fuente: INEC - Base de datos ENEMDU 2019.
Realizado por: Moyota, Alex, 2022.

En suma, tenemos conceptualmente dos dimensiones que están estratificando a la población de Ambato: (1) El capital humano de los jefes de hogar y (2) el capital económico de los jefes de hogar. La segmentación que se realice puede utilizar estas conceptualizaciones para interpretar cómo clasificar los hogares de Ambato a partir de estas dimensiones.

4.2.1.2 Clusterización de individuos respecto a las dimensiones conceptuales para identificar concentraciones de la segmentación.

Ahora bien, aún es necesario identificar dónde se distribuyen los individuos alrededor de estas dimensiones Figura 5-4 y sobre todo el paso más importante: identificar un reducido número de variables que capturen la mayor cantidad de información de la estratificación que los propios datos están revelando.

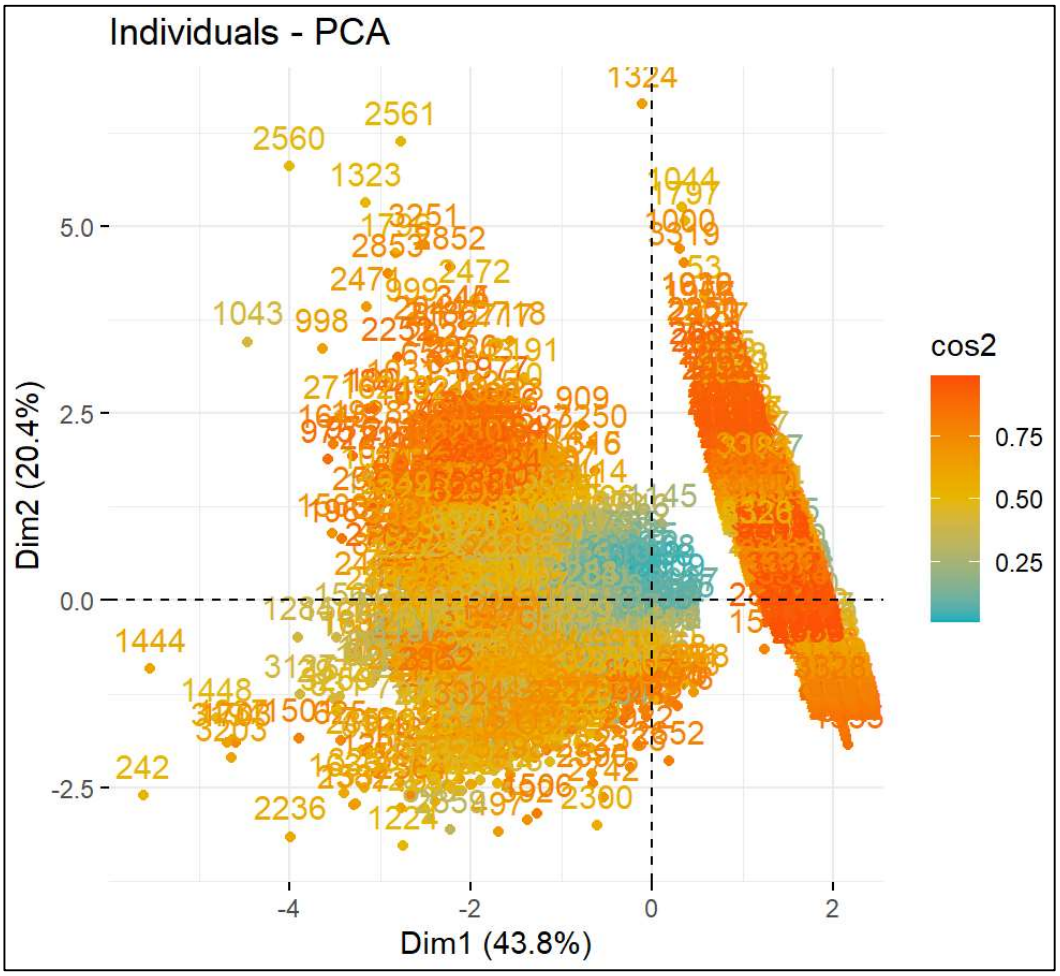

Figura 5-4: Representación de los individuos respecto a las dimensiones conceptuales.
Fuente: INEC - Base de datos ENEMDU 2019.
Realizado por: Moyota, Alex, 2022.

En la Figura 6-4 el racimo más visible que se puede visualizar se agrupa hacia la sección negativa de la primera dimensión y levemente hacia la sección negativa de la segunda. Es decir, preliminarmente se observa que gran parte de Ambato tiene capitales humanos y económicos menores que el promedio; se puede interpretar que este racimo es el de las clases más populares.

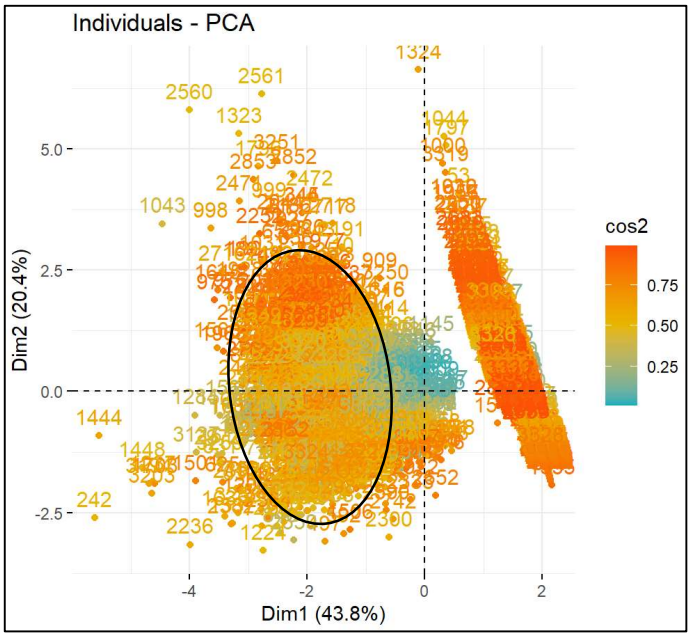

Figura 6-4: Identificación de la clase baja.
Fuente: INEC - Base de datos ENEMDU 2019.
Realizado por: Moyota, Alex, 2022.

El segundo racimo Figura 7-4 es menos poblado y se ubica ligeramente a la derecha del primer racimo, mostrando mayor capital humano, pero relativamente igual capital económico que el primero. Es decir, la preliminar clase media se diferencia de la clase popular por diferenciales en capital humano, no en capital económico.

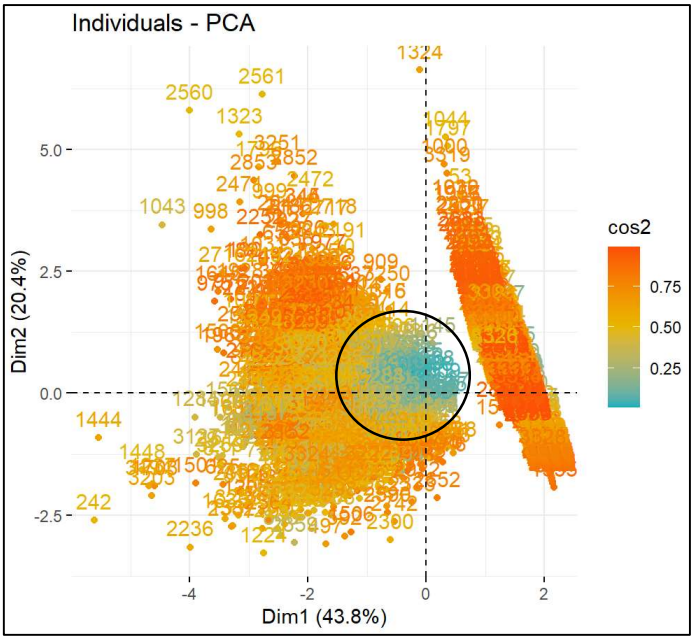

Figura 7-4: Identificación de la clase media.
Fuente: INEC - Base de datos ENEMDU 2019.
Realizado por: Moyota, Alex, 2022.

El tercer racimo aparece claramente identificado al separarse de los dos estratos anteriores Figura 8-4. En este caso presenta mayores niveles de capital humano y mayores niveles económicos. Adicionalmente la brecha con los dos racimos significa que existen considerables niveles de desigualdad puesto que en este tercer estrato hay pocos y con mayores niveles de capital humano y económico al compararse con el resto.

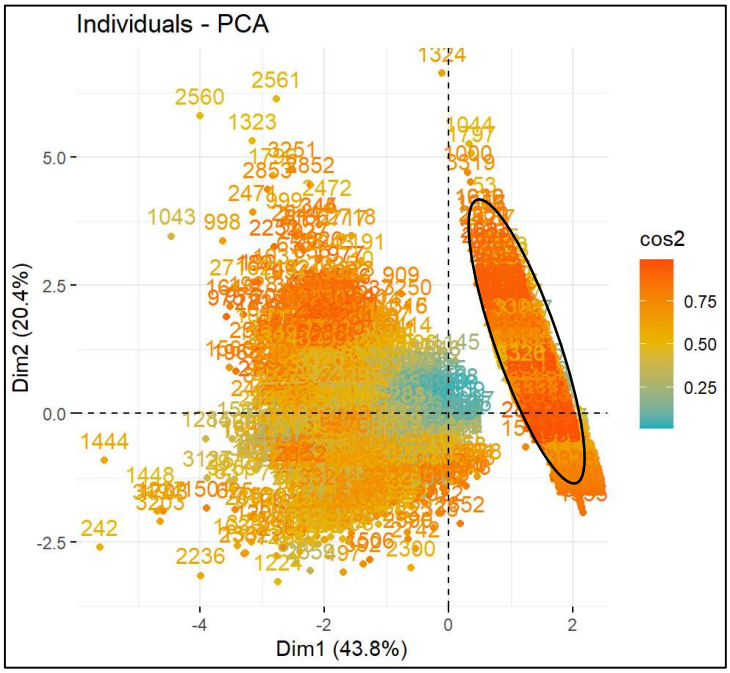

Figura 8-4: Identificación de la clase alta.
Fuente: INEC - Base de datos ENEMDU 2019.
Realizado por: Moyota, Alex, 2022.

De forma agregada Figura 9-4 los tres racimos presentan información que facilita la interpretación de los estratos en la ciudad de Ambato. Podríamos crear un *scoring* con las 6 variables para identificar a qué estrato pertenece cada individuo de la base de datos; sin embargo, la intención de esta investigación es identificar variables clave que permitan segmentar con menos variables, pero respetando la estructura que acabamos de encontrar.

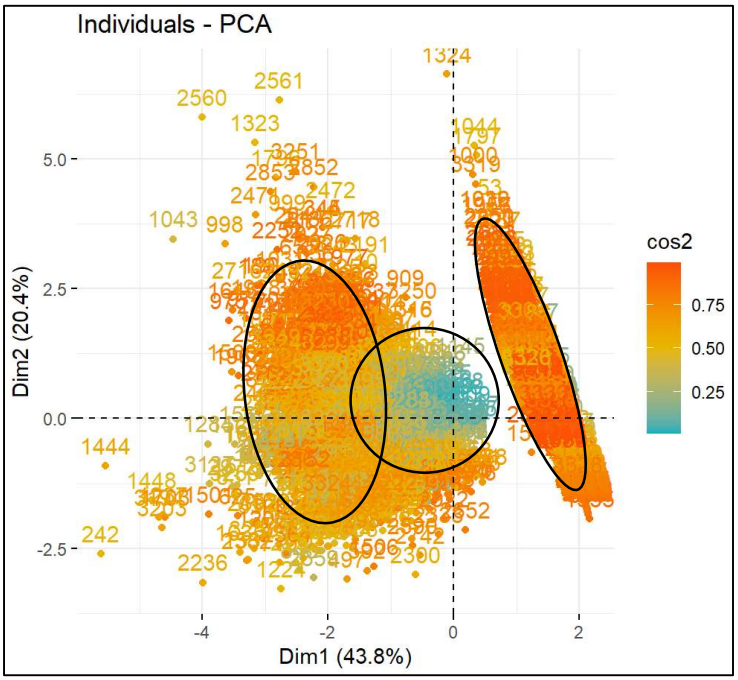

Figura 9-4: Representación de los tres racimos.
Fuente: INEC - Base de datos ENEMDU 2019.
Realizado por: Moyota, Alex, 2022.

4.2.2 Selección de componentes

Para ello analizamos los componentes Tabla 2-4 que el análisis ha encontrado, que capturan cuantitativa y cualitativamente la varianza suficiente para poder identificar qué variables resumirían mejor la estratificación.

4.2.2.1 Análisis del criterio Scree Plot

Para seleccionar el número de componentes a partir de este criterio observamos el diagrama Scree Figura 10-4 y de manera visual observamos que la "rodilla" o punto de inflexión en la curva se genera entre los ejes 2 y 3. A partir de este criterio se seleccionan 3 componentes.

4.2.2.2 Análisis del criterio de porcentaje de varianza acumulada

En base a la literatura revisada, este criterio plantea que la proporción de varianza acumulada sean valores entre el 70% y el 90% (Everitt & Hothorn, 2011), siendo así CP1, CP2 y CP3 nos explican el 79% de la variación. Con este resultado ratificamos que son tres las componentes con las mayores variaciones, siendo estas las más importantes (es decir, las más principales).

El gráfico de porcentaje de varianza explicada Figura 10-4 nos indica qué cantidad de la variabilidad total es estimada por cada componente principal. Observamos que la varianza explicada por cada componente muestra que son tres los componentes que explican más del 15%. Es decir, si analizamos las variables que contribuyen a estos tres componentes podemos identificar las variables clave que nos permitan estratificar la ciudad de Ambato.

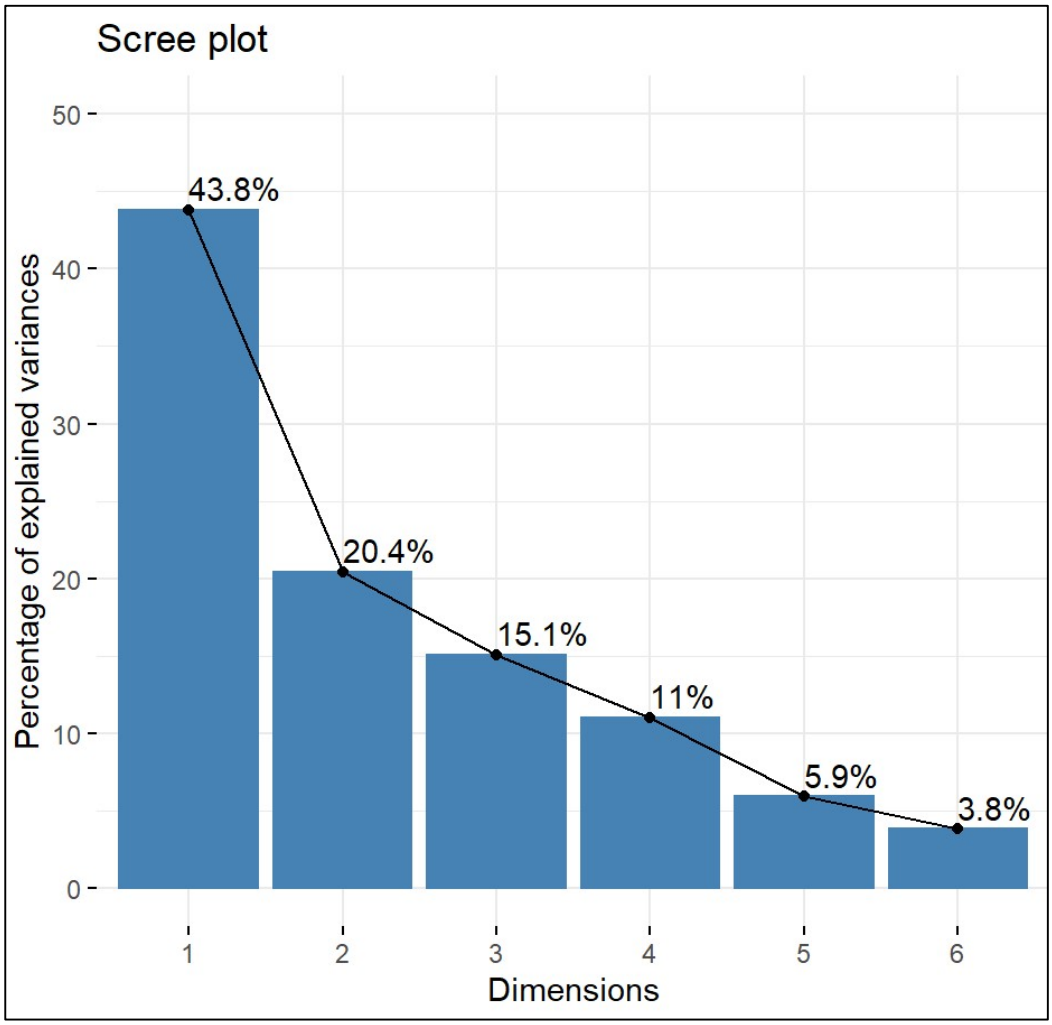

Figura 10-4: Proporción de variabilidad explicada por cada componente.
Fuente: INEC - Base de datos ENEMDU 2019.
Realizado por: Moyota, Alex, 2022.

Finalmente, con los resultados de estos dos criterios podemos estar seguros que con los tres principales componentes (CP1, CP2 y CP3) estamos capturando una cantidad relevante de estratificación.

4.2.3 Análisis de la composición lineal de componentes

Ahora, cualitativamente necesitamos explorar cada uno de los tres componentes para identificar qué variables presentan mayor coherencia y repetición. Por referencias de literatura previa se suele capturar los mayores aportes a la primera dimensión, pero estos deben ser coherentes con el resto de "aportes mayores" del resto. Vemos en la Tabla 3-4 que en este caso sí es así.

Tabla 3-4: Composición lineal de componentes para identificar las variables que presentan mayor coherencia y repetición.

PC1 = -0.5612165**p24**-0.5299257**p59**-0.4142401**p45**-0.3861859**p43**-0.2484125**p10a**-0.1475518**ingpc**

PC1 = -0.5612165 **Horas trabajadas** -0.5299257 **Satisfacción laboral** -0.4142401 **Antigüedad laboral** -0.3861859 **Dependencia económica** -0.2484125 **Nivel educativo** -0.1475518 **Ingreso per cápita**

PC2 = 0.67948499**ingpc**+0.6294031**p10a**-0.2767839**p45**-0.21795058**p59**-0.1048812**p24**+0.08390478**p43**

PC2 = 0.67948499 **Ingreso per cápita** +0.6294031 **Nivel educativo** -0.2767839 **Antigüedad laboral** -0.21795058 **Satisfacción laboral** -0.1048812 **Horas trabajadas** +0.08390478 **Dependencia económica**

PC3 = 0.6643951**p43**-0.59185254**p45**-0.43456889**ingpc**+0.1336125**p10a**+0.039745729**p59**-0.002749**p24**

PC3 = 0.6643951 **Dependencia económica** -0.59185254 **Antigüedad laboral** -0.43456889 **Ingreso per cápita** +0.1336125 **Nivel educativo** +0.039745729 **Satisfacción laboral** -0.002749 **Horas trabajadas**

Fuente: INEC - Base de datos ENEMDU 2019.
Realizado por: Moyota, Alex, 2022.

4.2.4 Selección de las variables de estratificación

Finalmente podemos, cualitativa y cuantitativamente, asegurar que las variables que mejor resumen la estratificación están representadas con las contribuciones capturadas en la principal dimensión. Si la intención de la investigación es volver parsimonioso y eficiente la estratificación entonces una buena regla es capturar aquellas variables que contribuyen más de 90% de la varianza. Como podemos ver en la Tabla 4-4 el ejercicio arroja entonces 4 variables clave de segmentación: horas trabajadas por semana, satisfacción laboral, antigüedad laboral y dependencia económica.

Tabla 4-4: Contribuciones de las variables.

Código	Descripción de variable	%	
p24	Horas trabajadas por semana	31.496394	91,652%
p59	Satisfacción laboral	28.082129	
p45	Antigüedad laboral	17.159489	
p43	Dependencia económica	14.913955	
p10a	Nivel educativo	6.170879	
ingpc	Ingreso per cápita	2.177153	

Fuente: INEC - Base de datos ENEMDU 2019.
Realizado por: Moyota, Alex, 2022.

4.2.5 Análisis de las curvas de densidad para definir umbrales

Una vez que contamos con las variables de estratificación resta crear una escala que asigne puntajes de acuerdo a las condiciones de cada individuo.

Cabe recordar que las variables capturadas son las que el propio ejercicio ha depurado, por lo que una vez obtenida el emparejamiento de cada individuo con su score o puntaje de estratificación se podría reconstruir una estructura de segmentación que intenta replicar la realidad de la sociedad ambateña.

Para proceder con este paso es necesario definir umbrales en cada una de las 4 variables seleccionadas, pero este proceso, nuevamente, no debe ser discrecional del investigador, sino que debe responder a la distribución de cada una de estas variables para el caso ambateño.

Por ello se analizan las distribuciones de cada una, se identifican los cuartiles de corte y se definen los umbrales. Finalmente se asignan valores de puntaje de acuerdo a esos cortes.

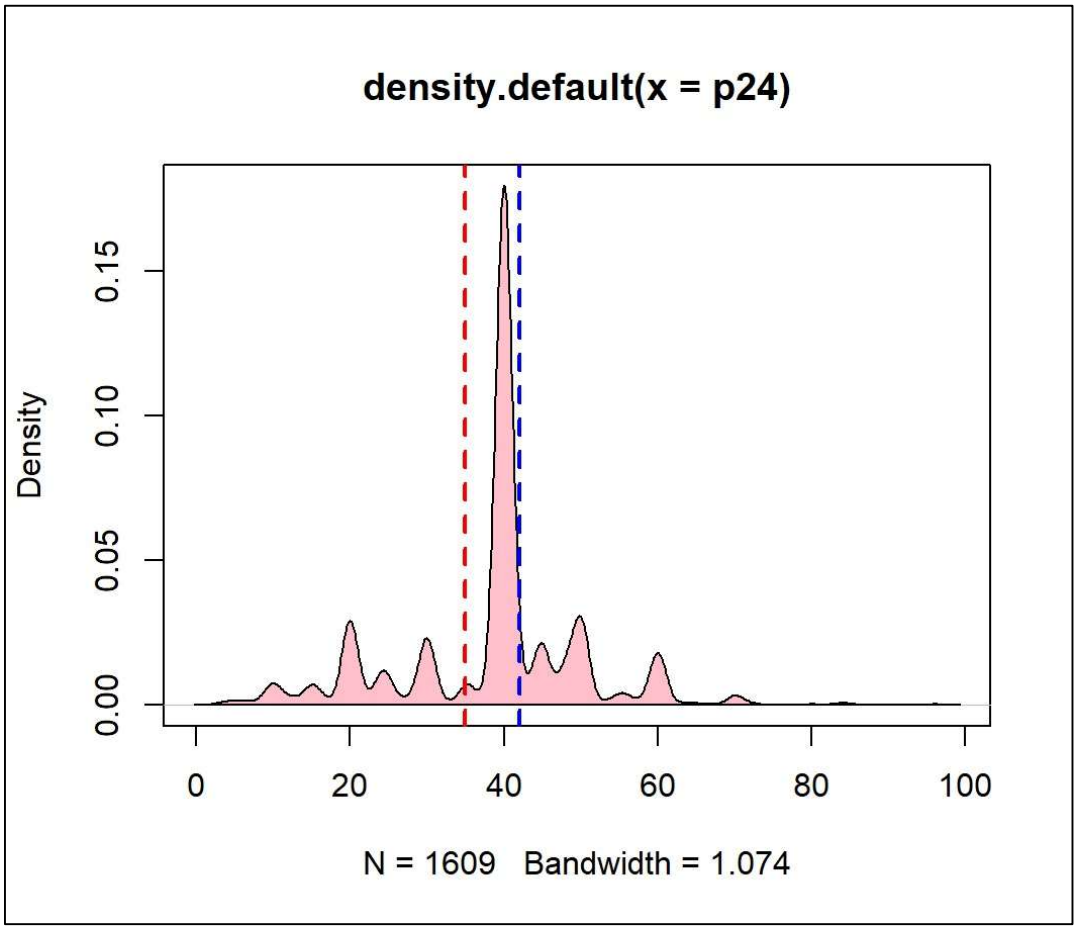

Figura 11-4: Densidad de horas trabajadas por semana (p24).
Fuente: INEC - Base de datos ENEMDU 2019.
Realizado por: Moyota, Alex, 2022.

Tabla 5-4: Cortes de estratificación: Horas trabajadas por semana (p24).

Estrato 1	Estrato 2	Estrato 3
≤ 35	$35 < x < 42$	≥ 42
Menos de 35 horas trabajadas por semana	Entre 35 y 42 horas trabajadas por semana	Más de 42 horas trabajadas por semana

Fuente: INEC - Base de datos ENEMDU 2019.
Realizado por: Moyota, Alex, 2022.

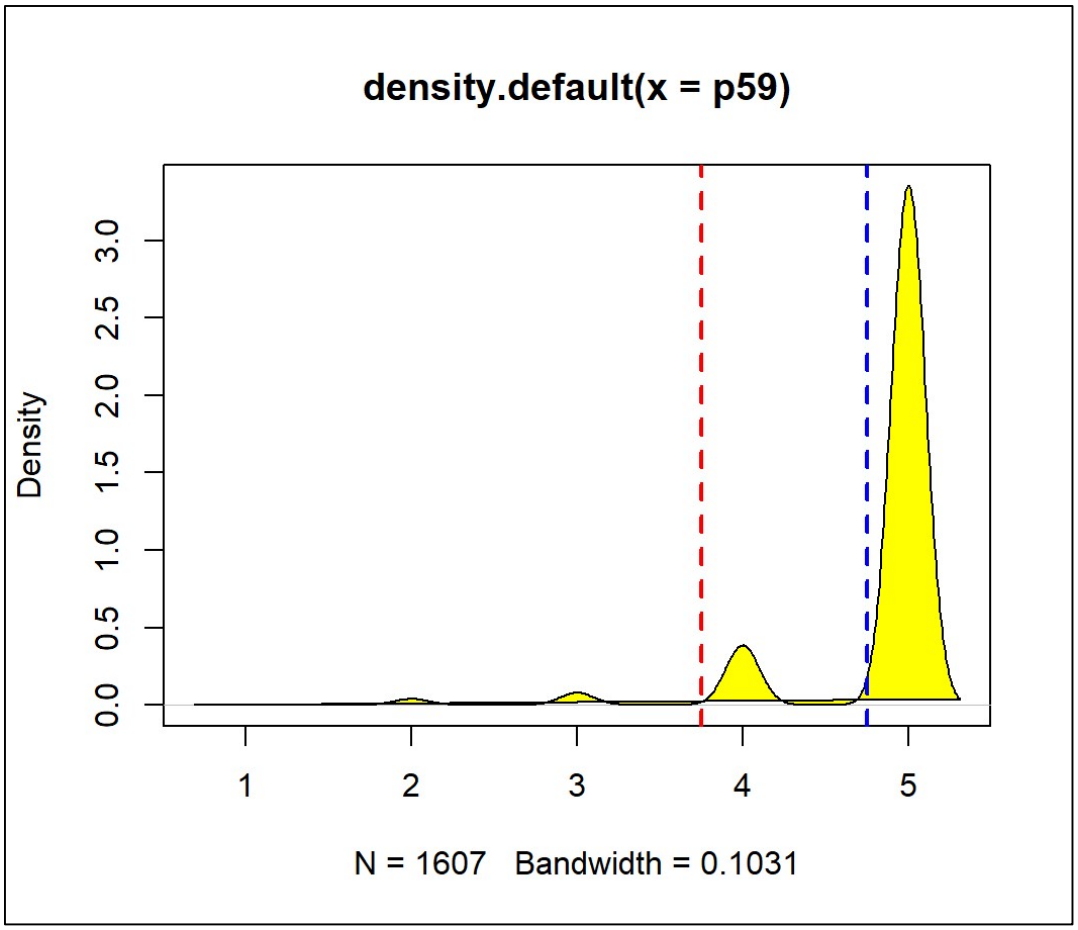

Figura 12-4: Densidad de satisfacción laboral (p59).
Fuente: INEC - Base de datos ENEMDU 2019.
Realizado por: Moyota, Alex, 2022.

Tabla 6-4: Cortes de estratificación: Satisfacción laboral (p59).

Estrato 1	Estrato 2	Estrato 3
≤ 3	$3 < x \leq 4$	> 4
Descontento pero conforme y Totalmente descontento	Poco contento	Contento

Fuente: INEC - Base de datos ENEMDU 2019.
Realizado por: Moyota, Alex, 2022.

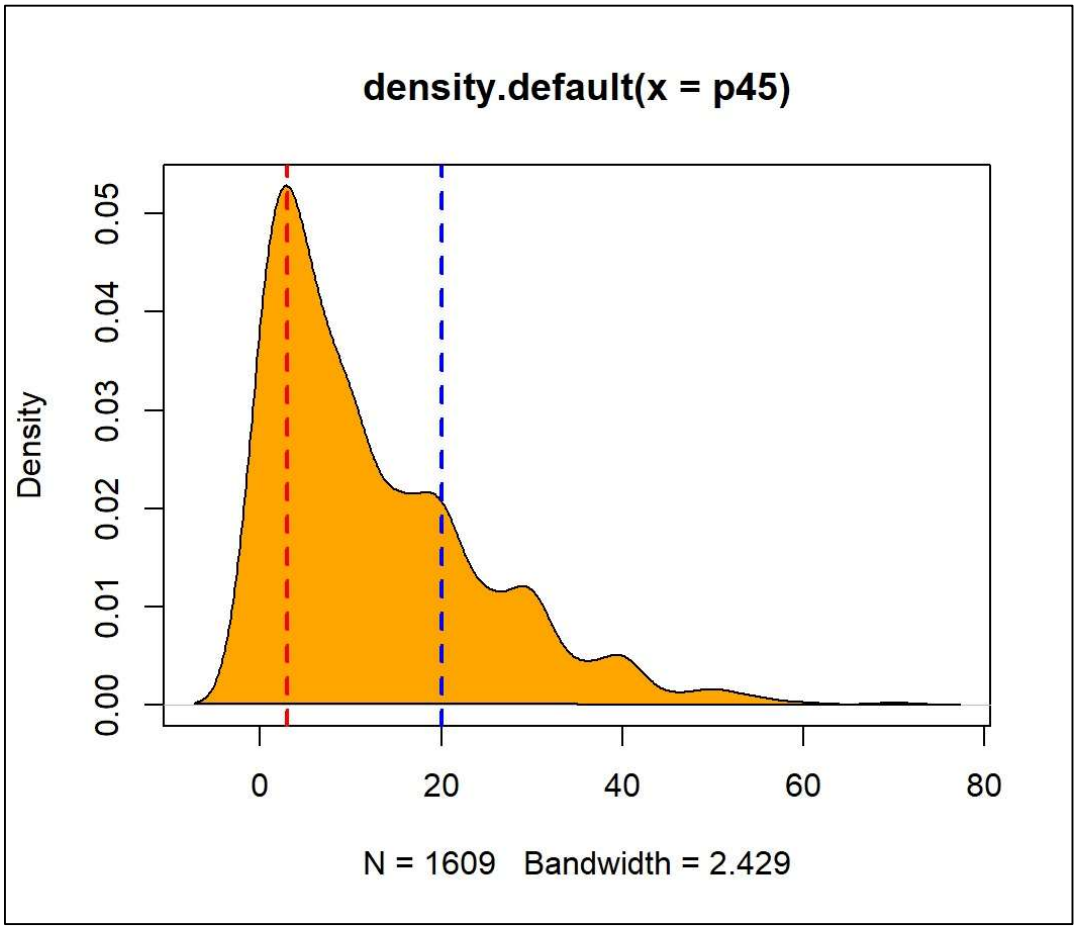

Figura 13-4: Densidad de antigüedad laboral (p45).
Fuente: INEC - Base de datos ENEMDU 2019.
Realizado por: Moyota, Alex, 2022.

Tabla 7-4: Cortes de estratificación: Antigüedad laboral (p45).

Estrato 1	Estrato 2	Estrato 3
≤ 3	$3 < x < 20$	≥ 20
Menos de 3 años de antigüedad laboral	Entre 3 y 20 años de antigüedad laboral	Más de 20 años de antigüedad laboral

Fuente: INEC - Base de datos ENEMDU 2019.
Realizado por: Moyota, Alex, 2022.

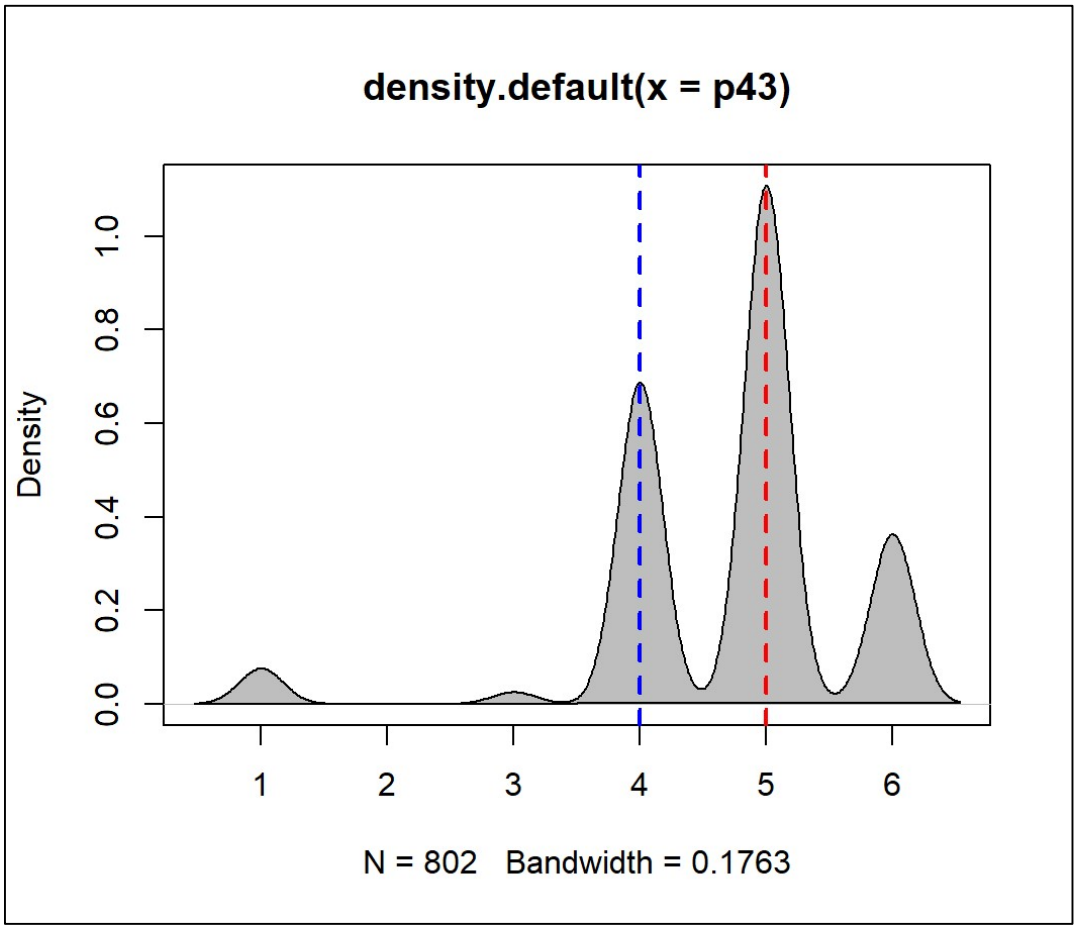

Figura 14-4: Densidad de dependencia económica (p43).
Fuente: INEC - Base de datos ENEMDU 2019.
Realizado por: Moyota, Alex, 2022.

Tabla 8-4: Densidad de dependencia económica (p43).

Estrato 1	Estrato 2	Estrato 3
≤ 4	$4 < x \leq 5$	> 5
Con contrato temporal, ocasional o eventual. Por obra, a destajo. Por horas. Por jornal.	Con contrato permanente / indefinido / estable o de planta	Con nombramiento

Fuente: INEC - Base de datos ENEMDU 2019.
Realizado por: Moyota, Alex, 2022.

Una vez identificados los umbrales resta recalibrar la proporción que cada variable obtiene en base a su contribución (total del 92%), ver Tabla 9-4.

Tabla 9-4: Reajuste de la proporción.

Código	Descripción de variable	%
p24	Horas trabajadas por semana	34.365201
p59	Satisfacción laboral	30.639952
p45	Antigüedad laboral	18.722438
p43	Dependencia económica	16.272372

100%

Fuente: INEC - Base de datos ENEMDU 2019.
Realizado por: Moyota, Alex, 2022.

Finalmente, en la Tabla 10-4 asignamos el puntaje que permite obtener la contribución de cada estrato en cada caso (para facilitar segmentación rápida vía encuestas redondeamos).

Tabla 10-4: Resultado de estratificación.

	Rangos	Puntaje
Horas trabajadas por semana	≤ 35	28
	$35 < x < 42$	30
	≥ 42	34
Satisfacción laboral	≤ 3	18
	4	24
	5	31
Antigüedad laboral	≤ 3	2
	$3 < x < 20$	10
	≥ 20	19
Dependencia económica	≤ 4	10
	5	13
	6	16

Fuente: INEC - Base de datos ENEMDU 2019.
Realizado por: Moyota, Alex, 2022.

4.2.6 Análisis de resultados de la prueba ji-cuadrado (o chi cuadrado)

En esta última sección se presenta los resultados de la prueba X^2 la cual nos va permitir cumplir con el ultimo objetivo de este trabajo que es el de validar si la estadística multivariante permitirá determinar la estratificación de los hogares de la ciudad de Ambato.

Para esto, primero definimos la hipótesis:

H_o**:** La estadística multivariante ***no*** permite identificar los estratos sociales de los hogares de la ciudad de Ambato en el cuarto trimestre del año 2019.

H_1: La estadística multivariante ***si*** permite identificar los estratos sociales de los hogares de la ciudad de Ambato en el cuarto trimestre del año 2019.

Tabla 11-4: Valores observados.

Código	Descripción de variables	N
p24	Horas trabajadas por semana	1609
p59	Satisfacción laboral	1607
p45	Antigüedad laboral	1609
p43	Dependencia económica	802
	Total:	**5627**

Fuente: INEC - Base de datos ENEMDU 2019.
Realizado por: Moyota, Alex, 2022.

En la Tabla 11-4 se muestra el número de hogares (N) para cada variable de estratificación que el análisis de componentes ha capturado. La intención es evaluar si las variables que nos arroja el análisis de componentes están estratificando a los hogares de la ciudad de Ambato para lo cual se considera un nivel de significancia del 5%.

$$\propto = 0{,}05$$

Para llevar a cabo esta prueba estadística también calculamos los grados de libertad (gl).

$\mathrm{gl} = (\mathrm{n} - 1)$; Donde n representa el número de variables.

$\mathrm{gl} = (4 - 1)$

$\mathrm{gl} = 3$

El siguiente paso es establecer el valor de crítico (rechazo) de la hipótesis nula (H_o) para la distribución X^2. Para obtener este valor crítico consultamos una tabla de distribución de probabilidad X^2, con un nivel de significancia de 0,05 y grados de libertad igual a 3, se obtiene un valor crítico igual a 7,8147.

$$X^2_{(3;0,05)} = 7{,}8147$$

Para calcular el estadístico de contraste, primero se obtiene la frecuencia relativa observada porcentual como la relación del número de observaciones de cada variable entre el total de observaciones y la frecuencia relativa esperada porcentual dada por la Tabla 9-4, como se puede observar en la Tabla 12-4.

Ahora, el valor para el estadístico de contraste (frecuencia esperada) se calcula como el producto de la frecuencia relativa esperada de cada variable por el total de frecuencias observadas, estos resultados se muestran en la Tabla 13-4.

Tabla 12-4: Frecuencia relativa observada y esperada porcentual.

Frecuencia relativa observada porcentual	Frecuencia relativa esperada porcentual
29%	34%
29%	31%
29%	19%
14%	16%
100%	100%

Fuente: INEC - Base de datos ENEMDU 2019.
Realizado por: Moyota, Alex, 2022.

Tabla 13-4: Frecuencias esperadas.

Código	Variables	f_o	f_e
p24	Horas trabajadas por semana	1609	1913
p59	Satisfacción laboral	1607	1744
p45	Antigüedad laboral	1609	1069
p43	Dependencia económica	802	900
	Total:	**5627**	**5627**

Fuente: INEC - Base de datos ENEMDU 2019.
Realizado por: Moyota, Alex, 2022.

Con estos valores finalmente calculamos el valor de X^2 usando la fórmula (11) descrita inicialmente en la sección 2.2.5 y los resultados se muestran en la Tabla 14-4.

Tabla 14-4: Chi cuadrado (X^2) calculado.

Código	Variables	f_o	f_e	$f_o - f_e$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
p24	Horas trabajadas por semana	1609	1913	-304	92525	48,36214
p59	Satisfacción laboral	1607	1744	-137	18871	10,81796
p45	Antigüedad laboral	1609	1069	540	291460	272,61382
p43	Dependencia económica	802	900	-98	9667	10,73710
	Total:	**5627**	**5627**	**0**	**Chi-cuadrado:**	**342,531**

Fuente: INEC - Base de datos ENEMDU 2019.
Realizado por: Moyota, Alex, 2022.

Como el valor calculado de X^2 es 342,531 y previamente se estableció que el punto crítico para la distribución X^2 con un alfa de 0,05 y 3 gl es igual a 7,8147 podemos afirmar que nuestro valor del estadístico de prueba está dentro de la zona de rechazo de H_o formulada a priori, como se puede observar en la Figura 15-4.

Por lo tanto, como $X^2_{Calculado} > X^2_{Tabla}$ se rechaza H_o, con lo que podemos afirmar que la estadística multivariante si permite identificar los estratos sociales de los hogares de la ciudad de Ambato en el cuarto trimestre del año 2019.

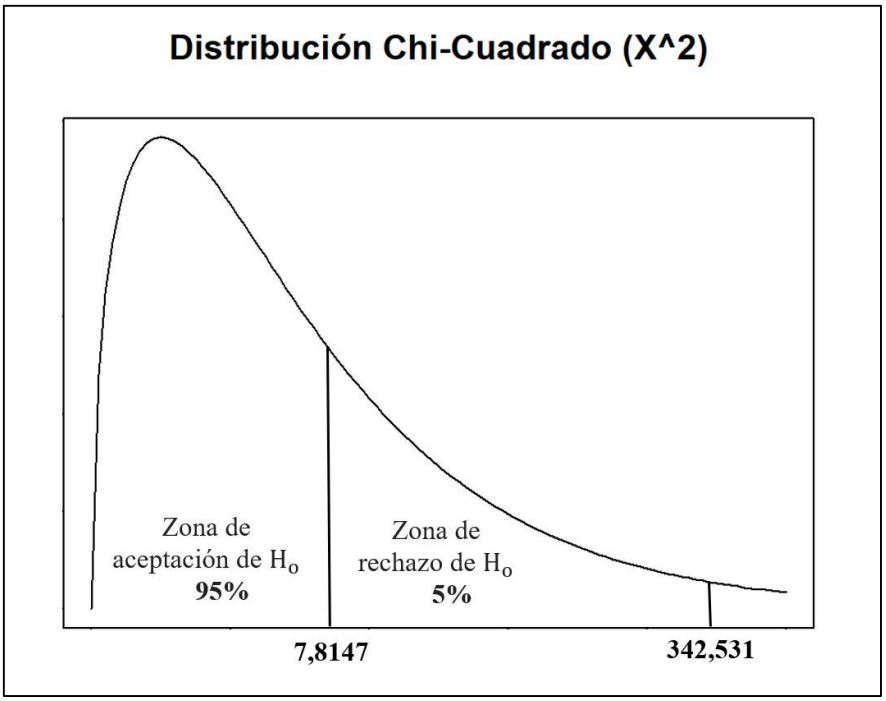

Figura 15-4: Curva de distribución de Chi-cuadrado (X^2).
Fuente: INEC - Base de datos ENEMDU 2019.
Realizado por: Moyota, Alex, 2022.

CONCLUSIONES

a) Para este estudio se seleccionó seis variables socioeconómicas (horas trabajadas, satisfacción laboral, antigüedad laboral, nivel educativo, dependencia económica e ingreso per cápita) de la base de datos ENEMDU 2019 que de acuerdo a la bibliografía revisada pueden representar factores de estratificación de los hogares de la ciudad de Ambato.

b) En esta investigación se aplicó un Análisis de Componentes Principales (ACP) que permitió identificar cuatro variables de estratificación que recogen el 92% de la variación de los factores estratificantes: horas trabajadas por semana (34%), satisfacción laboral (31%), antigüedad laboral (19%) y dependencia económica (16%).

c) Se estableció los umbrales a partir de los cuartiles de corte identificados en las curvas de densidad en cada una de las variables seleccionadas, logrando identificar los estratos sociales de los hogares de la ciudad de Ambato: alto, medio y bajo.

d) Se validó el Modelo Multivariante aplicando una prueba estadística chi-cuadrado, de manera que $X^2_{Calculado}$ resulto mayor que X^2_{Tabla} argumento con el que se concluye que el Análisis de Componentes Principales si permite identificar los estratos sociales de los hogares de la ciudad de Ambato en el cuarto trimestre del año 2019.

e) En esta investigación también se puede concluir que son dos factores conceptuales que están segmentando la ciudad de Ambato: (i) el capital humano del jefe del hogar, que puede ser operacionalizado a través de las horas trabajadas por semana, satisfacción laboral, Antigüedad laboral y la dependencia económica de hogar, y (ii) el capital económico del jefe de hogar, operacionalizado a través del ingreso per cápita, el nivel educativo, la antigüedad laboral y en menor medida la satisfacción laboral. Cualquier interpretación cuantitativa puede ser contextualizada utilizando la narrativa que entienda la estratificación sobre estas dos dimensiones.

f) Finalmente, en base a los resultados se calculó los niveles de estratos sociales para la ciudad de Ambato, donde la clase baja se ubica por debajo de valores iguales o menores que 58, la clase media se ubica entre valores menores o iguales que 77 y mayores que 58 y por último entre valores menores o iguales a 100 y mayores que 77 se encuentra la clase alta.

g) Estos insumos permiten estratificar los hogares ambateños e identificar a qué segmento pertenecen, utilizando la escala que este análisis ha generado y expuesto en la tabla final de estratificación.

RECOMENDACIONES

a) Se recomienda para futuras investigaciones incrementar las variables de segmentación, permitiendo que la cobertura de la misma se incremente sobre el 92%.
b) Se recomienda utilizar el banco de datos abiertos del INEC para ejecutar este tipo de análisis con el objetivo de estudiar las variables de los individuos de una determinada población para de esta manera replicar la realidad de una sociedad de interés.
c) Una vez construida la estratificación se puede agregar variables complementarias que permitan identificar y sobre todo interpretar las condiciones de cada estrato en relación con otras dimensiones de interés de futuras investigaciones. Por ejemplo, categoría de ocupación (p42) o recibe cursos de capacitación (p44i), entre muchas otras.
d) Es posible parametrizar el ejercicio para que sea proyectable a herramientas como el censo de población o base de datos geoestadísticas. Por ejemplo, asumiendo que el censo 2020 mantiene el cuestionario del 2010, se podría proyectar la estratificación, pero usando las variables que la tabla de estratificación que contenga el cuestionario. Luego habría que recalibrar contribuciones y seguir el mismo procedimiento.
e) La estratificación realizada intenta no ser invasiva y busca que la misma distribución de individuos respecto a las variables arroje la estratificación de la ciudad. Por este motivo si se requiere identificar las variables que contribuyan a la estratificación a nivel nacional se recomienda volver a correr el mismo ejercicio, pero con la base de datos agregada.
f) Se recomienda utilizar este modelo como referente para que sea replicado en las diferentes urbanizaciones en general, como insumo de análisis socioeconómico para ayudar a la ejecución y análisis en políticas públicas.

GLOSARIO

ACP: Análisis de Componentes Principales (Principal Component Analysis, PCA)

CRAN: Red Exhaustiva de Archivos R (Comprehensive R Archive Network)

Dim1: Dimensión 1

ENEMDU: Encuesta Nacional de Empleo, Desempleo y Subempleo

ENIGHUR: Encuesta Nacional de Ingresos y Gastos de los Hogares Urbanos y Rurales del Ecuador

$\mathbf{f_e}$**:** Frecuencia esperada

$\mathbf{f_o}$**:** Frecuencia observada

gl: Grados de libertad

$\mathbf{H_o}$**:** Hipótesis nula

$\mathbf{H_1}$**:** Hipótesis alternativa

INEC: Instituto Nacional de Estadística y Censos

ingpc: Ingreso per cápita

p10a: Nivel educativo

p24: Horas trabajadas por semana

p43: Dependencia económica

p45: Antigüedad laboral

p59: Satisfacción laboral

PC1: Componente principal 1

$\mathbf{X^2}$**:** Prueba Chi-cuadrado

$\propto$**:** Nivel de significancia

BIBLIOGRAFÍA

Argüello, O. (1991). *DESARROLLO ECONOMICO, POLITICAS SOCIALES Y POBLACION (El marco para una política sociodemográfica). Notas de población*. 11–12. https://repositorio.cepal.org/bitstream/handle/11362/12935/NP53-01_es.pdf?sequence=1

Astudillo Macas, S. F., & Salazar Ortiz, E. A. (2020). *UN NUEVO ENFOQUE PARA LA ESTRATIFICACIÓN SOCIOECONÓMICA DEL ECUADOR* [Escuela Politécnica Nacional]. https://bibdigital.epn.edu.ec/bitstream/15000/20973/1/CD%2010496.pdf

Cattell, R. B. (1966). The Scree Test For The Number Of Factors. *Multivariate Behavioral Research*, *1*(2), 245–276. https://doi.org/10.1207/s15327906mbr0102_10

Chang, W. (2022, July 13). *Libro de recetas R Graphics*. https://r-graphics.org/

Everitt, B., & Hothorn, T. (2011). *An Introduction to Applied Multivariate Analysis with R*. Springer New York. https://doi.org/10.1007/978-1-4419-9650-3

Fachelli, Sandra. (2009). Nuevo modelo de estratificación social y nuevo instrumento para su medición. El caso argentino. In *TDX (Tesis Doctorals en Xarxa)*. http://www.tdx.cat/handle/10803/5149

Hardy, C. (2014). *Estratificación social en América Latina: Retos de cohesión social* (LOM Ediciones).

Hartman, G. (2011). *Fundamentals of Matrix Algebra* (Third Edition). https://drive.google.com/file/d/18m2ZaPfomgYTQ3LQxSwu9YlFy3-JvwVw/view

INEC. (2011). *Encuesta de Estratificación del Nivel Socioeconómico NSE 2011*. https://www.ecuadorencifras.gob.ec//documentos/web-inec/Estadisticas_Sociales/Encuesta_Estratificacion_Nivel_Socioeconomico/111220_NSE_Presentacion.pdf

INEC. (2019). *Documento Metodológico - Encuesta Nacional de Empleo, Desempleo y Subempleo (ENEMDU). Instituto Nacional de Estadística y Censos, Quito - Ecuador.* https://www.ecuadorencifras.gob.ec/documentos/web-inec/EMPLEO/2018/Septiembre-2018/ENEMDU_Metodologia%20Encuesta%20Nacional%20de%20Empleo%20Desempleo%20y%20Subempleo.pdf

Jolliffe, I. T. (2002). *Principal Component Analysis* (Second Edition). Springer-Verlag. https://doi.org/10.1007/b98835

León González, Á., Llinás Solano, H., & Tilano, J. (2008). Análisis multivariado aplicando componentes principales al caso de los desplazados. *Ingeniería & Desarrollo.*, *23*, 119–142. https://www.redalyc.org/pdf/852/85202310.pdf

López, D., & Sepúlveda, C. E. (2014). Modelos de estratificación socioeconómica a partir de la información catastral para la ciudad de Bogotá, D.C. In *Los límites de la estratificación: en busca de alternativas* (pp. 109–112). Editorial Universidad del Rosario. https://doi.org/10.7476/9789587385373.0007

Madrigal, J. (2004). *Estratificación de hogares y segmentos por niveles de ingreso en el censo 2000*. https://ccp.ucr.ac.cr/bvp/pdf/censo2000/libro-censo/1.2-Madrigal-2.doc.pdf

Marín García, A. (2021, March 9). *Estratificación social. Economipedia*. https://economipedia.com/definiciones/estratificacion-social.html

McHugh, M. L. (2013). The Chi-square test of independence. *Biochemia Medica*, 143–149. https://doi.org/10.11613/BM.2013.018

Mendes, T. (2015). ESTRATIFICAÇÃO SOCIOECONÔMICA: UMA PROPOSTA A PARTIR DO CONSUMO. *Curitiba: Universidade Federal Do Paraná*. https://acervodigital.ufpr.br/bitstream/handle/1884/38048/R%20-%20D%20-%20THIAGO%20MENDES%20ROSA.pdf?sequence=3

Mendivelso, F., & Rodríguez, M. (2018). Prueba Chi-Cuadrado de independencia aplicada a tablas 2xN. *Revista Médica Sanitas*, *21*(2), 92–95. https://doi.org/10.26852/01234250.6

Meneses, J. (2019). *Introducción al análisis multivariante*. https://femrecerca.cat/meneses/publication/introduccion-analisis-multivariante/introduccion-analisis-multivariante.pdf

Peña, D. (2002). *Análisis Multivariante de Datos*. McGraw-Hill. https://www.researchgate.net/publication/40944325

Sémbler R., C. (2006). *Estratificación social y clases sociales: una revisión analítica de los sectores medios*. ECLAC. https://repositorio.cepal.org/handle/11362/6130

Verdugo Chaura, D. A. (2022). *Ráster con R*. https://doi.org/10.5281/zenodo.5959446

Wickham, H. (2016). *Elegant graphics for data analysis. En: ggplot2* (Second Edition). Springer International Publishing. https://doi.org/10.1007/978-3-319-24277-4

Wickham, H., François, R., Enrique, L., & Muller, K. (2022). *dplyr: una gramática de la manipulación de datos*. https://dplyr.tidyverse.org/

ANEXOS

Anexo A.

Código R: Análisis exploratorio de variables

```
library(haven)
library(dplyr)
library("corrplot")
library("PerformanceAnalytics")
library("factoextra")
library(ggplot2)
library(factoextra)

## Leer la base de datos formato SAV ##
enemdu_persona_201912 <-read_sav("C:/Users/CLIENTE/Downloads/BASE_DATOS/enemdu_persona_
201912.sav")
View(enemdu_persona_201912)
## Filtrar datos de la ciudad de Ambato ##
enemdu_persona_201912_1 <- enemdu_persona_201912[48368:51700,c(17, 35, 55, 67, 86, 144)]
View(enemdu_persona_201912_1)
## Reemplazar NA con 0 ##
enemdu_persona_201912_1[is.na(enemdu_persona_201912_1)]<-0
View(enemdu_persona_201912_1)
## Graficar BoxPlot para detectar si hay outliers ##
g_caja<-boxplot(enemdu_viv_hog_201912_1$p10a, col="skyblue", frame.plot=F)
g_caja$out
## Análisis exploratorio de datos ##
## Matriz de correlación  ##
n=cor(enemdu_persona_201912_1)
View(n)
n
## Gráficas de correlación ##
corrplot(cor(enemdu_persona_201912_1))
chart.Correlation(enemdu_persona_201912_1, histogram=TRUE, pch=10)
## Mapa de calor (heatmap)##
heatmap(x = cor(enemdu_persona_201912_1), symm = TRUE, main = "heatmap")
## Estandarización de datos ##
View(scale(enemdu_persona_201912_1))
```

Código R: Análisis de Componentes Principales (ACP)

```
## Obtención de los componentes principales – PCA ##
# prcomp genera PCA
cprin_enemdu_persona_201912_1 <- prcomp(enemdu_persona_201912_1,center=TRUE,scale.=TRUE)
# Importancia de componentes ¿dónde me detengo?
```

```
summary(cprin_enemdu_persona_201912_1)
#Loadings, genera los pesos de cada variable
cprin_enemdu_persona_201912_1
```

Código R: Selección de Componentes

```
plot(cprin_enemdu_persona_201912_1, type = "b") #Barras
plot(cprin_enemdu_persona_201912_1, type = "l") #Líneas - Scree Plot
fviz_eig(cprin_enemdu_persona_201912_1, addlabels = FALSE, ylim = c(0, 50)) #Porcentaje
de varianza acumulada
```

Código R: Análisis de individuos y variables de la salida del ACP

```
## biplot devuelve la gráfica de PCA ##
biplot(cprin_enemdu_persona_201912_1, scale=0)#biplot devuelve la gráfica de PCA
eig.val <- get_eigenvalue(cprin_enemdu_persona_201912_1)
eig.val
## Individuals - PCA ##
fviz_pca_ind(cprin_enemdu_persona_201912_1)
fviz_pca_ind(cprin_enemdu_persona_201912_1,
             col.ind = "cos2",
             gradient.cols=c("#00AFBB","#E7B800","#FC4E07"),
             repel = FALSE
## Variables - PCA ##
fviz_pca_var(cprin_enemdu_persona_201912_1, col.var = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE) # evita sobrelapamiento de texto
## PCA – Biplot ##
fviz_pca_biplot(cprin_enemdu_persona_201912_1, col.var = "cos2",
                gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
                repel = TRUE) # evita sobrelapamiento de texto
```

Código R: Curvas de densidad

```
## Curva de densidad p24 ##
library(dplyr)
enemdu_persona_201912_2 <- enemdu_persona_201912[48368:51700,c(35)]
View(enemdu_persona_201912_2)
datos2 <- enemdu_persona_201912_2[!is.na(enemdu_persona_201912_2$p24),]
datos2

attach(datos2)
quantile(p24) #Calcula los cuartiles
plot(density(p24)) #Gráfica de densidad
polygon(density(p24),col="pink")
abline(v=c(35,42),col=c("red","blue"),lty=c(2,2), lwd=c(2,2))
```

```
## Curva de densidad p43 ##

library(dplyr)
enemdu_persona_201912_3 <- enemdu_persona_201912[48368:51700,c(55)]
View(enemdu_persona_201912_3)
datos3 <- enemdu_persona_201912_3[!is.na(enemdu_persona_201912_3$p43),]
datos3

attach(datos3)
quantile(p43) #Calcula los cuartiles
plot(density(p43)) #Gráfica de densidad
polygon(density(p43),col="gray")
abline(v=c(2,3),col=c("blue","red"),lty=c(2,2), lwd=c(2,2))

## Recodificación de la variable p43 ##

library(car)
library(dplyr)
enemdu_persona_201912_3 <- enemdu_persona_201912[48368:51700,c(55)]
View(enemdu_persona_201912_3)
datos3 <- enemdu_persona_201912_3[!is.na(enemdu_persona_201912_3$p43),]
datos3

datos3$p43 <- recode(datos3$p43, "1=6;2=5;3=4;4=3;5=2;6=1")
datos3

attach(datos3)
quantile(p43) #Calcula los cuartiles
plot(density(p43)) #Gráfica de densidad
polygon(density(p43),col="gray")
abline(v=c(4,5),col=c("blue","red"),lty=c(2,2), lwd=c(2,2))

## Curva de densidad p45 ##

library(dplyr)
enemdu_persona_201912_4 <- enemdu_persona_201912[48368:51700,c(67)]
View(enemdu_persona_201912_4)
datos4 <- enemdu_persona_201912_4[!is.na(enemdu_persona_201912_4$p45),]
datos4

attach(datos4)
quantile(p45) #Calcula los cuartiles
plot(density(p45)) #Gráfica de densidad
polygon(density(p45),col="orange")
abline(v=c(3,20),col=c("red","blue"),lty=c(2,2), lwd=c(2,2))
```

```
## Curva de densidad p59 ##

library(dplyr)
enemdu_persona_201912_5 <- enemdu_persona_201912[48368:51700,c(86)]
View(enemdu_persona_201912_5)
datos5 <- enemdu_persona_201912_5[!is.na(enemdu_persona_201912_5$p59),]
datos5

attach(datos5)
quantile(p59) #Calcula los cuartiles
plot(density(p59)) #Gráfica de densidad
polygon(density(p59),col="yellow")
abline(v=c(1,1),col=c("red","red"),lty=c(2,2), lwd=c(2,2))

## Recodificación de la variable p59 ##

library(car)
library(dplyr)
enemdu_persona_201912_5 <- enemdu_persona_201912[48368:51700,c(86)]
View(enemdu_persona_201912_5)
datos5 <- enemdu_persona_201912_5[!is.na(enemdu_persona_201912_5$p59),]
datos5

datos5$p59 <- recode(datos5$p59, "1=5;2=4;3=3;4=2;5=1")
datos5

attach(datos5)
quantile(p59) #Calcula los cuartiles
plot(density(p59)) #Gráfica de densidad
polygon(density(p59),col="yellow")
abline(v=c(3.75,4.75),col=c("red","blue"),lty=c(2,2), lwd=c(2,2))
```

Código R: Chi – Cuadrado (X^2)

```
curve(dchisq(x,df=3),from=0, to=10, main="Distribución Chi-Cuadrado (X^2)")
```

espoch | Dirección de Bibliotecas y Recursos del Aprendizaje

UNIDAD DE PROCESOS TÉCNICOS Y ANÁLISIS BIBLIOGRÁFICO Y DOCUMENTAL

REVISIÓN DE NORMAS TÉCNICAS, RESUMEN Y BIBLIOGRAFÍA

Fecha de entrega: *19* / 07 / 2023

INFORMACIÓN DEL AUTOR/A (S)
Nombres – Apellidos*:* *Alex Rolando Moyota Paguay*
INFORMACIÓN INSTITUCIONAL
Instituto de Posgrado y Educación Continua
Título a optar: *Magíster en Matemática mención Modelación y Docencia*
f. Analista de Biblioteca responsable: Lic. Luis Caminos Vargas Mgs.

Firmado electrónicamente por:
LUIS ALBERTO
CAMINOS VARGAS

UNIDAD DE PROCESOS TÉCNICOS DBRA ESPOCH

0073-DBRA-UTP-IPEC-2023