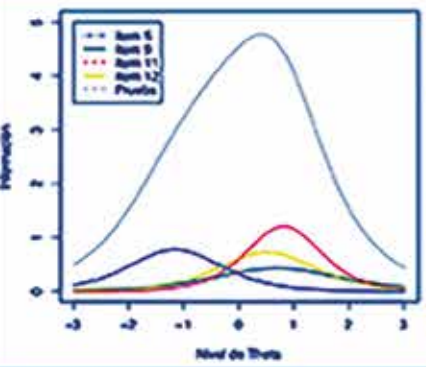
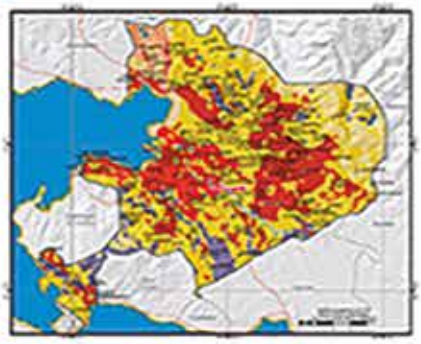
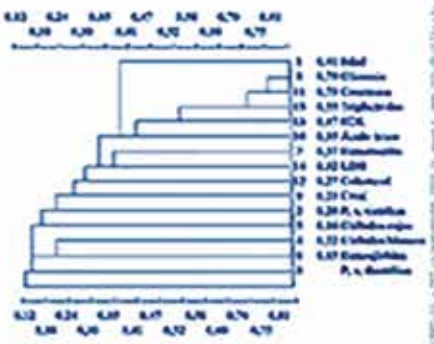
Universidad Mayor
de San Andrés

Varianza

Revista del Instituto de Estadística Teórica y Aplicada







CARRERA DE ESTADÍSTICA
UMSA - FCPN

UMSA
FCPN
CARRERA
ESTADÍSTICA

IETA
Instituto de Estadística
Teórica y Aplicada

Varianza

Revista de la Carrera de Estadística

Publicación del Instituto de Estadística Teórica y Aplicada

UMSA
FCPN
CARRERA
ESTADÍSTICA

IETA
Instituto de Estadística
Teórica y Aplicada

Número 16

Octubre, 2019

ISSN 9876-6789
REVISTA VARIANZA
Nº 16 - Octubre, 2019

PRESENTACIÓN

La presente Revista Varianza es la número 16, edición anual que se ha venido elaborando con la finalidad de difundir los proyectos de investigación e interacción social de los docentes y estudiantes de nuestra universidad, plasmados en el resumen de un artículo. La estadística es y será un medio instrumental importante para cualquier ciencia del conocimiento humano, por lo tanto, se convierte en una herramienta difícil de eludir.

Al transcurrir los años de esta nueva era del siglo veintiuno, la estadística se desarrolla y seguirá desarrollándose en materia de tecnología de aplicación esencial en la investigación, por ello, es fundamental el incremento de recursos humanos dedicados a esta ilustre actividad. La Revista Varianza debe traducir siempre el conocimiento de los grandes pensadores de la estadística científica - aplicada, para incentivar a nuevos postulantes universitarios, como principal objetivo.

Es así, que queda el compromiso de que las próximas ediciones de la revista se mejore y alcance así mayor difusión, basada en una publicación líder del quehacer científico de la estadística a nivel nacional e internacional para el bien de nuestra carrera, la universidad, de investigadores y futuros estudiantes que requieren de un eficaz instrumento de orientación en la rama de la estadística.

M. Sc. Fernando Oday Rivero Suguiura
DIRECTOR CARRERA DE ESTADÍSTICA

Carrera de Estadística
Instituto de Estadística Teórica y Aplicada (I.E.T.A.)
Facultad de Ciencias Puras y Naturales
Universidad Mayor de San Andrés

La Paz - Bolivia
Edificio Antiguo - Planta Baja
Teléfonos: 2442100 -2612844
Correos electrónicos:estadistica@umsa.bo - ieta@umsa.bo

Dedicado a los docentes y estudiantes
que hacen investigación estadística

Contenido

POPULARIDAD PRESIDENCIAL EN AMÉRICA LATINA ANÁLISIS DE SENTIMIENTO EN TWITTER*

Lic. Chirino Gutiérrez, Álvaro Limber

✉ *achirino@aru.org.bo*

RESUMEN

Este artículo presenta una propuesta para medir la popularidad de ocho presidentes de países de América Latina empleando la información de Twitter mediante el uso de scraping web en R. Las medidas están basadas en base a seguidores, favoritos, *retweets* y un análisis de sentimiento de los *tweets* de los usuarios hacia los presidentes. La información corresponde al mes de septiembre de 2019.

PALABRAS CLAVE

Twitter, presidentes América Latina, scraping, estadística, minería de texto, análisis de sentimiento.

ABSTRACT

This article presents a proposal to measure the opinions of eight presidents of Latin American countries using the information of Twitter through the use of web scraping in R. The measures are based on the base of followers, favorites, retweets and an analysis of feelings of the tweets of the users towards the presidents. The information corresponding to the month of September 2019.

KEYWORDS

Twitter, Latin American presidents, scraping, statistics, text mining, sentiment analysis.

1. MOTIVACIÓN

Actualmente las redes sociales se han convertido en una ventana para que las personas interactúen con una fluidez sin precedente, la llegada del internet, los teléfonos inteligentes han acelerado la comunicación. Una de las redes sociales más populares es el *Twitter*, esta plataforma permite a sus usuarios crear miniblogs o mensajes limitados a 140 caracteres y publicarlos de manera pública o privada, no existe una interacción horizontal entre los usuarios, la relación que existe es de carácter vertical dado que los usuarios deciden a que usuario seguir.

La manera que existe para interactuar es mediante los *tags* (etiquetas) por ejemplo el tag #Bolivia está orientado a etiquetar el mensaje con contenido relacionado al *tag*, otro ejemplo son los tags como @evoespueblo que usan los usuarios para crear mensajes que incluyen al usuario del *tag*.

El *Twitter* es una de las herramientas empleadas por los presidentes de los distintos países, en América Latina todos los presidentes tienen cuentas activas, en este documento se exploran los datos de ocho presidentes; Mauricio Macri (Argentina), Evo Morales (Bolivia), Mario Abdo (Paraguay), Nicolás Maduro (Venezuela), Sebastián Piñera (Chile), Martin Vizcarra (Perú), Iván Duque (Colombia) y Lenin Moreno (Ecuador). La cuenta más antigua corresponde a Sebastián Piñera en febrero de 2008 y la más reciente cuenta corresponde a Evo Morales en abril de 2016.

Este documento presenta medidas de popularidad presidencial de ocho presidentes latinoamericanos, la información proviene de las cuentas de *Twitter* y fue obtenida

* Este documento está en el marco del proyecto de investigación: "Aplicación del Web Scraping en la Estadística"

mediante el uso del *scraping web* en R. En 2 se presentan los objetivos y alcances del trabajo, en 3 se detalla la metodología, en 5 presentan los resultados y finalmente en 6 se describen los hallazgos y recomendaciones del estudio.

2. OBJETIVOS Y ALCANCES

El objetivo central del documento es **desarrollar criterios para medir popularidad a nivel de los presidentes de ocho países latinoamericanos empleando información proveniente del Twitter**.

La información del *Twitter* es extraída mediante el uso de *Scraping web* usando el software estadístico R.

En cuanto a los alcances:

- La cobertura temporal de la información extraída del *Twitter* corresponde al mes de septiembre de 2019.
- Se define al español como el criterio de búsqueda para los presidentes.
- Se emplean el lexicón nrc[1] para el análisis de sentimiento.

3. METODOLOGÍA

El *scraping* web con R permite tener acceso a diferentes datos dentro del *Twitter*, a partir de esto se definen las siguientes medidas para medir la popularidad de los ocho presidentes.

- Basado en seguidores, favoritos y *retweets*
  1. Porcentaje de seguidores respecto la población total del país (*followers*).
  2. Número de favoritos por cada 100.000 seguidores basado en el promedio de los últimos 20 *tweets* (*fav20*).
  3. Número de favoritos por cada 100.000 seguidores basado en el promedio de los últimos 200 *tweets* (*fav200*).
  4. Número de *retweets* por cada 100.000 seguidores basado en el promedio de los últimos 20 *tweets* (*retweets20*).
  5. Número de *retweets* por cada 100.000 seguidores basado en el promedio de los últimos 200 *tweets* (*retweets200*).
- Basado en los mensajes de los presidentes y de los usuarios hacia los presidentes
  1. *Wordclouds* de los tweets presidenciales
  2. Análisis de sentimiento en base a los *tweets* de los usuarios hacia los presidentes.

Donde:

$$followers_i = \frac{S_i}{P_{2019,i}} \quad (1)$$

$$fa\bar{v}20_i = \frac{fa\bar{v}20_i}{S_i} * 100.000 \quad (2)$$

$$fa\bar{v}200_i = \frac{fa\bar{v}200_i}{S_i} * 100.000 \quad (3)$$

$$retweets20_i = \frac{re\bar{t}20_i}{S_i} * 100.000 \quad (4)$$

$$retweets200_i = \frac{re\bar{t}200_i}{S_i} * 100.000 \quad (5)$$

Con S_i la cantidad de seguidores registrados a septiembre de 2019 del presidente de *i*, $P_{2019,i}$ es la población proyectada del país del presidente *i* para el 2019. $fa\bar{v}20_i$ y $fa\bar{v}200_i$ el promedio de favoritos de los últimos 20 y 200 tweets presidenciales respectivamente. $re\bar{t}20_i$ y $re\bar{t}200_i$ el promedio de *retweets* de los últimos 20 y 200 *tweets* presidenciales respectivamente.

3.1. *WORDCLOUDS*

Los *wordclouds* presidenciales son nubes de palabras que se construyen a partir de

[1] El NRC Emotion Lexicon es una lista de palabras en inglés y sus asociaciones con ocho emociones básicas https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

las frecuencias de ocurrencias de palabras provenientes de los *tweets* que publican los presidentes.

3.2. ANÁLISIS DE SENTIMIENTO

Para el análisis de sentimientos de los *tweets* de los usuarios hacia los presidentes se siguen los siguientes pasos:

1. Se extraen los *tweets* de los usuarios hacia los presidentes, no se toma en cuenta los retweets
2. Se eliminan caracteres, números y stopword de cada *tweet*. Quedando únicamente las palabras de interés
3. En base al *lexicón nrc* se identifican ocho emociones para cada palabra, los sentimientos son: enojo, expectación, disgusto, miedo, alegría, tristeza, sorpresa y confianza.
4. Se agregan las ocho emociones para todos los *tweets* y se genera la proporción de emociones global.

4. DATOS

Los datos provienen de las siguientes fuentes:

1. Las cuentas de *Twitter* de ocho presidentes, la información corresponde a los tweets y actualizaciones hasta el mes de septiembre de 2019. El total de *tweets* explorados es de 20.699, no incluye los *retweets* de los presidentes.
2. Los tweets provenientes de los usuarios en donde se emplea un *tag* para alguno de los ocho presidentes. El total de *tweets* explorados es de 275.536, no incluye los *retweets* que realizan los usuarios y corresponde al mes de septiembre.
3. Las proyecciones de población provienen de los datos del Banco Mundial mediante la librería wbstats de R.

La información fue recolectada empleando R con las librerías *twitteR*, *wordcloud, rvest, tidyverse, syuzhet y wbstats*.

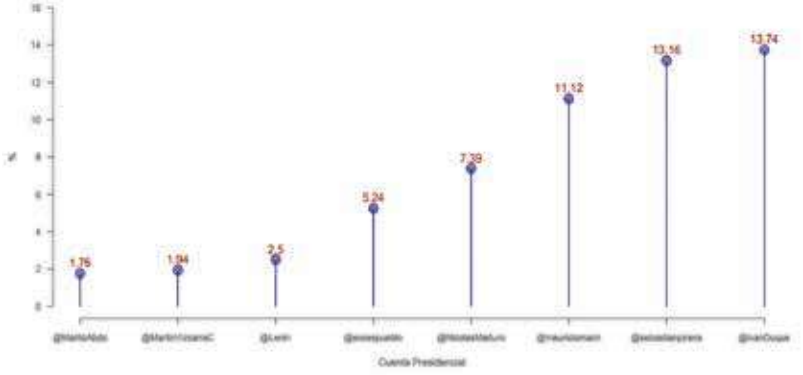
5. RESULTADOS

Siguiendo la metodología descrita y en base a los datos recolectados, en esta sección se presentan los resultados.

5.1. BASADO EN SEGUIDORES, FAVORITOS Y RETWEETS

La Gráfica Nº 1 presenta el indicador *follower* por presidente, los resultados se presentan de forma ascendente, se aprecia que los presidentes con más seguidores respecto la población del país son Iván Duque (Colombia) y Sebastián Piñera (Chile) mientras los con menos seguidores

Gráfica Nº 1
Porcentaje de seguidores respecto la población total del país.



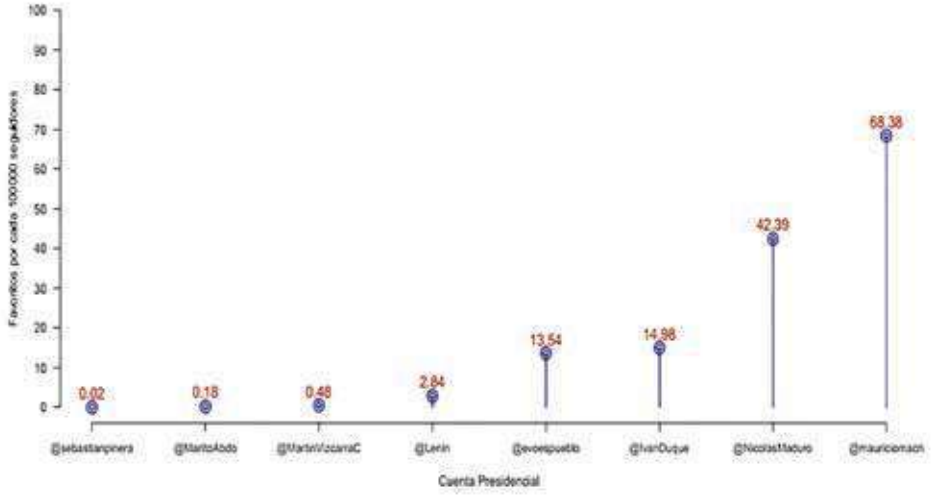
Fuente: Elaboración del Autor

son Mario Abdo (Paraguay) y Martin Vizcarra (Perú).

Las Gráficas Nº 2 y Nº 3 presentan los indicadores basados en la cantidad de favoritos, en ambas figuras se aprecia que los presidentes con los valores más altos son Nicolás Maduro (Venezuela) y Mauricio Macri (Argentina) mientras los más bajos son Sebastián Piñera (Chile) Martin Viscarra (Perú) y Mario Abdo (Paraguay). Notar que la cantidad de favoritos

Gráfica Nº 2

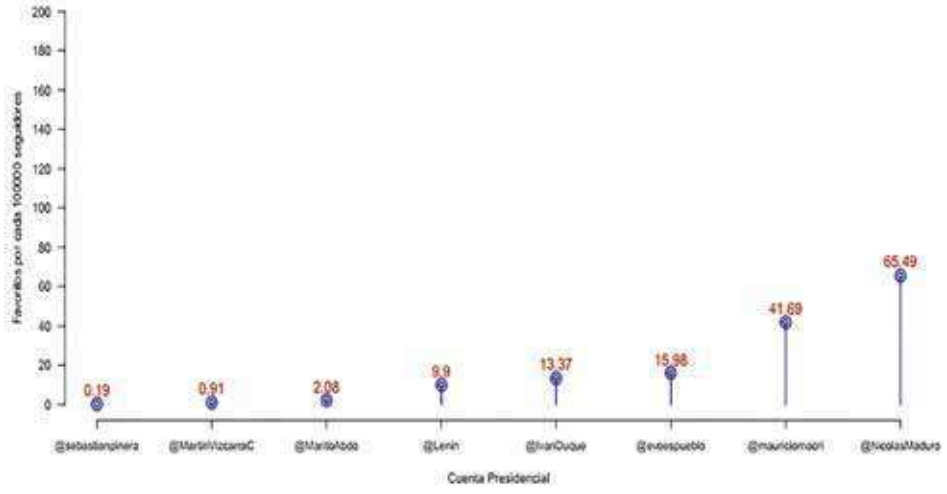
Promedio del Número de Favoritos por cada 100.000 seguidores basado en los últimos 20 tweets presidenciales.



Fuente: Elaboración del Autor

Gráfica Nº 3

Promedio del Número de Favoritos por cada 100.000 seguidores basado en los últimos 200 tweets presidenciales.

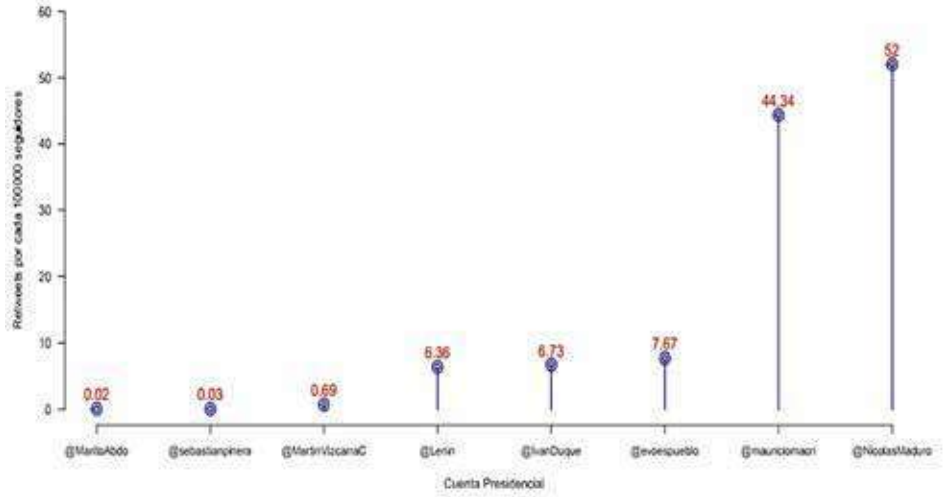


Fuente: Elaboración del Autor

por cada 100.000 seguidores es muy bajo en todos los casos, logrando en el mejor de los casos 69 favoritos..

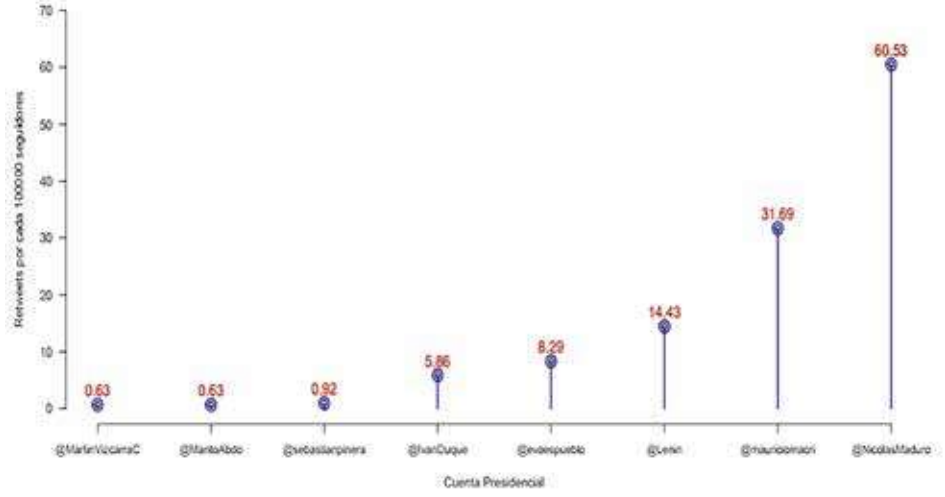
Las Gráficas Nº 4 y Nº 5 presentan los indicadores basados en la cantidad de retweets, el

Gráfica Nº 4
Numero de Retweets por cada 100.000 seguidores basado en el promedio de los últimos 20 tweets presidenciales



Fuente: Elaboración del Autor

Gráfica Nº 5
Numero de Retweets por cada 100.000 seguidores basado en el promedio de los últimos 200 tweets presidenciales



Fuente: Elaboración del Autor

comportamiento es similar a los indicadores de favoritos tanto en los presidentes con valores más altos y bajos.

5.2. BASADO EN LOS MENSAJES DE LOS PRESIDENTES Y DE LOS USUARIOS HACIA LOS PRESIDENTES

En la Gráfica Nº 6 se presenta las nubes de palabras basadas en los *tweets* de los ocho presidentes, mientras que en la Gráfica Nº 7 se presenta la nube de palabras de los usuarios.

Para el análisis de sentimiento se empleó el grafico de telaraña que permite visualizar las ocho emociones, esto se presenta en la Gráfica Nº 8. El patrón recurrente es que la emoción con mayor frecuencia es la confianza seguida de la tristeza y el miedo, la presencia marcada de la confianza puede deberse en parte a los *tweets* de las distintas entidades públicas de los distintos países,

Gráfica Nº 6
Wordcloud de los Tweets presedenciales

(a) Mauricio Macri

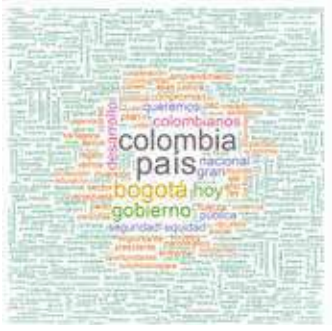
(b) Evo Morales

(c) Mario Abdo

(d) Nicolás Maduro

(e) Sebastian Piñera

(f) Martín Viscarra

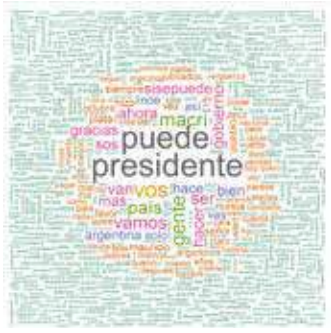


(g) Iván Duque

(h) Lenin Moreno

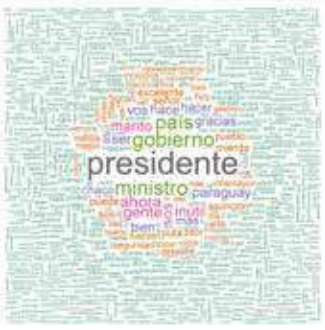
Fuente: Elaboración del Autor

Gráfica Nº 7
Wordcloud de los Tweets de los usuarios hacia los presidentes

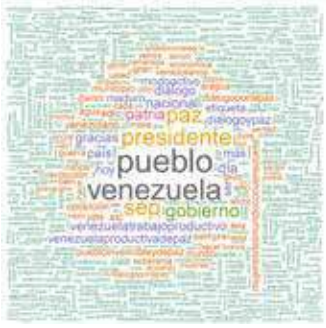


(a) Mauricio Macri

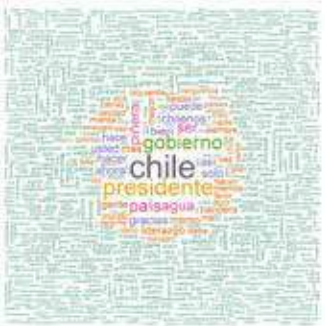
(b) Evo Morales



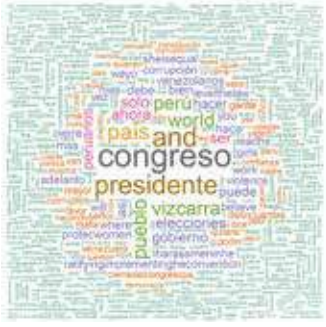
(c) Mario Abdo



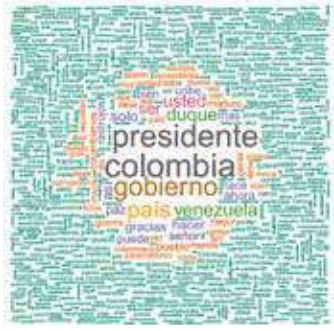
(d) Nicolás Maduro



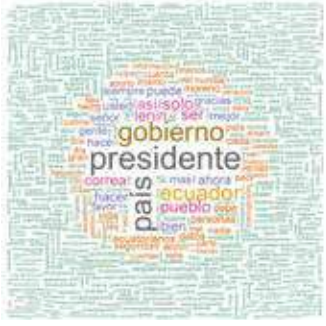
(e) Sebastian Piñera



(f) Martín Viscarra



(g) Iván Duque



(h) Lenin Moreno

Fuente: Elaboración del Autor

Gráfica Nº 8
Emociones hacia los presidentes de parte de los usuarios

(a) Mauricio Macri

(b) Evo Morales

(c) Mario Abdo

(d) Nicolás Maduro

(e) Sebastian Piñera

(f) Martín Viscarra

(g) Iván Duque

(h) Lenin Moreno

Fuente: Elaboración del Autor

para siguientes aproximaciones se podría explorar la relación entre la emoción y los horarios en los que se publican.

6. HALLAZGOS Y RECOMENDACIONES

Los hallazgos del documentos pueden resumirse en (1) El *Twitter* es un espacio de generación de información masiva que brinda la oportunidad de explorar diversos aspectos de la población vinculada a ella, (2) el R es una de las herramientas estadísticas actuales con mayor versatilidad para explorar las nuevas tendencias de las ciencia de datos y (3) el análisis de sentimiento es una técnica dentro de la minería de texto que aún necesita adaptaciones para el contexto boliviano. En cuanto a las recomendaciones para futuros estudios (1) incorporar el análisis de sentimiento diferenciando las horas y días para evaluar si las emociones están condicionadas a ciertos momentos del tiempo y (2) a partir de los indicadores es posible realizar un monitoreo diario.

BIBLIOGRAFÍA

Chen, L.-P. (2019). Text mining in practice with R.

Feldman, R., y Sanger, J. (2006). The Text Mining Handbook.

Iacus, S. M. (2015). Automated Data Collection with R - A Practical Guide to Web Scraping and Text Mining (Vol. 68) (n.o Book Review 3).

Naldi, M. (2019). A review of sentiment computation methods with R packages., 1–11. Descargado de http://arxiv.org/abs/1901 .08319

Singh, S., y Choudhary, S. S. (2017). Social Media Data Analysis:Twitter Sentimental Analysis Using R Language. International Journal of Advances in Electronics and Computer Science, 4 (11), 13–17. Descargado de http:// iraj.in

Zhao, Y., y Cen, Y. (2013). Data Mining Applications with R.

MODELO DE OJIVA NORMAL DE DOS PARÁMETROS: UNA ALTERNATIVA PARA EL ANÁLISIS DE INSTRUMENTOS DE MEDICIÓN

Dr(c). Chocotea Poca, Omar[1,2] & Lic. Villa Cabero, Dafne Maritza[3]

✉ *omar.chocotea@postgrado.uv.cl* ✉ *mvilla@doc.emi.edu.bo*

RESUMEN

Los modelos de la Teoría de Respuesta al ítem, TRI, (ítem Response Theory, en inglés) son una alternativa más eficiente frente a la Teoría Clásica de los Test, TC o TCT, para el análisis de la calidad técnica de instrumentos de medición. Comparada con la TC, la TRI permite obtener más información sobre los ítems de la prueba y sobre el constructo o variable latente que interesa medir en los examinados; sin embargo, hay todavía problemas de estimación no resueltos, que provocan algunas veces, la imposibilidad de ajustar ciertos modelos en conjuntos específicos de datos. El presente artículo centra su atención en presentar la teoría del modelo de ojiva normal de dos parámetros, donde los parámetros y también el rasgo latente juegan un papel importante. Dadas las ventajas del modelo, el problema se encuentra en elegir el mejor método de estimación de los parámetros, al ser un modelo no lineal, se deben utilizar procesos iterativos de simulación. Las estimaciones de los parámetros fueron obtenidas mediante el método de simulación de Monte Carlo vía Cadenas de Markov (Markov Chain Monte Carlo: MCMC) en WinBUGS, donde las correspondientes corridas se hicieron en BRugs, que es una librería de R. Una vez obtenidos los valores de las estimaciones de los parámetros se pueden obtener las probabilidades de respuesta del ítem dado el rasgo latente del estudiante, las respectivas Curvas Características del Ítem, Función de Información de la prueba y Error Típico de Información. Para esta aplicación se utilizó el examen del curso de verano 2005.

PALABRAS CLAVE

Medición, Psicometría, Teoría de respuesta al ítem, Análisis de ítems, Estimación de modelos.

ABSTRACT

Models of item response theory, IRT (Item Response Theory, in English) are more efficient compared to the Classical Theory Test, TC or TCT, for the analysis of the technical quality of measuring instruments alternative. Compared with CT, TRI allows more information on test items and on the construct or latent variable of interest measured in examinees; however, there are still unresolved problems estimate, causing sometimes impossible to adjust certain models in specific datasets. This article focuses on presenting the theory model warhead normal two parameters, where the parameters and the latent trait also play an important role. Given the advantages of the model, the problem lies in choosing the best method of parameter estimation, being a nonlinear model, you must use iterative simulation processes. The parameter estimates were obtained by the method of Monte Carlo simulation via Markov Chain (Markov Chain Monte Carlo: MCMC) in WinBUGS, where the corresponding runs were made in Brugs, which is a library of R. Once obtained values of the parameter estimates can be obtained probabilities item response given student latent trait, the respective Item Characteristic Curves, Test Information Function and Standard Error Information. For this application the examination of the summer course 2005 was used.

KEYWORDS

Measurement, Psychometrics, Item response theory, Item analysis, Model estimation.

[1] Instituto de Estadística, Universidad de Valparaíso, Chile.
[2] Carreras de Estadística/Sociología, Universidad Mayor de San Andrés, Bolivia
[3] Carreras de Ingeniería Comercial/Financiera, Escuela Militar de Ingeniería, Bolivia

1. INTRODUCCIÓN

El origen de la TRI es antiguo, data de los años cuarenta del siglo pasado, en pleno apogeo de la TC (ver Lawley, 1940, 1944). No obstante, dada la complejidad e imposibilidad de llevar a cabo los cálculos requeridos para las estimaciones, no comenzó a difundirse y utilizarse hasta la generalización de los ordenadores con amplias capacidades de cálculo. La razón de su éxito y rápida expansión radica en que permite analizar aspectos de los test que no son posibles o son difíciles de justificar bajo la TC.

1.1. TEORÍA DE RESPUESTA AL ÍTEM.

Una forma de establecer las relaciones entre las medidas observadas y el constructo es relacionando con éste las puntuaciones de cada uno de los ítems y por medio de los patrones de las respuestas obtener una estimación del valor del sujeto en el constructo. Esta aproximación se refleja en un O. Chocotea & M. D. Villa conjunto de modelos etiquetados de forma general como la TRI.

1.2. CARACTERÍSTICAS

Bajo el nombre genérico de la TRI se agrupan muchos modelos distintos. Aunque se diferencian en algunos aspectos, todos tienen en común una serie de rasgos básicos, especialmente el de ser modelos estructurales que establecen una relación matemática formalizada entre la respuesta a un ítem concreto y el nivel de rasgo o aptitud de un sujeto. El hecho de presentar un conjunto de aspectos comunes es lo que hace que aparezcan como un cuerpo teórico unificado. A continuación, se mencionan los postulados básicos que caracterizan a los modelos de la TRI:

a) Parten de la existencia de rasgos o aptitudes latentes que permiten predecir o explicar la conducta de un examinado ante un ítem de un test. La TRI parte de la premisa de que el comportamiento de un sujeto ante un ítem puede explicarse en términos de una o varias características del sujeto denominadas rasgos o aptitudes latentes, que no pueden observarse directamente. Por ejemplo, la puntuación de un sujeto en un test de inteligencia (comportamiento observable) es resultado o función de una aptitud o rasgo (inteligencia) que posee el sujeto y que no podemos observar, pero que se manifiesta a través de ciertas conductas (respuestas a los ítems del test).

b) La relación entre el rendimiento o la conducta de un examinado en un ítem y el conjunto de rasgos responsables de dicho rendimiento pueden describirse mediante una función monótona creciente, denominada función característica del ítem o curva característica del ítem (CCI).

1.3. CURVA CARACTERÍSTICA DEL ÍTEM

Mediante ella se representa una relación funcional entre la proporción de respuestas correctas a un ítem y el nivel del atributo. En general, y por razones de sus orígenes en los test de aptitudes, el rasgo suele denominarse aptitud, aunque conviene indicar que el rasgo puede ser cualquier atributo o constructo en el que se manifiesten diferencias individuales, tales como rendimientos académicos, variables de personalidad, actitudes, intereses, etc., no limitándose los modelos a la inteligencia y rendimiento académico. En adelante nos referiremos a estos indistintamente como rasgo, aptitud o atributo.

1.4. VENTAJAS DE LOS MODELOS DE LA TRI.

Al cumplir un conjunto de supuestos los diferentes modelos de la TRI, los cuales se describen líneas más abajo, estos poseen una serie de ventajas sobre la TC y que se derivan de los procedimientos de estimación

que utilizan:

a) Los modelos de la TRI, a diferencia de los de la TC, son falsables. En cualquier aplicación de la TRI es esencial evaluar el ajuste del modelo a los datos.

b) Los ítems pueden ser descritos por unas propiedades o parámetros que se pueden estimar.

c) Asumiendo la existencia de un amplio universo o población de ítems para la medida del mismo rasgo, la cantidad de rasgo que posee un sujeto particular es independiente del conjunto concreto de ítems utilizado en su estimación.

d) A diferencia de la TC que caracteriza la precisión por medio de un único valor para todo el conjunto de puntuaciones (el coeficiente de fiabilidad), la TRI la caracterizará por medio de una función denominada función de información, que indicará cómo son de precisas las puntuaciones en los diferentes niveles de aptitud.

1.5. SUPUESTOS

Unidimensionalidad

De forma general se asume que hay un conjunto de rasgos responsables de la actuación del examinado en la prueba, el cual define un espacio dimensional latente y pudiendo representarse en el la posición de cada uno de los examinados y de los diferentes ítems. Sin embargo, en las aplicaciones de la TRI se supone que una única aptitud o rasgo es suficiente para explicar los resultados de los examinados y las relaciones entre los ítems.

Independencia local

La independencia local significa que si se mantienen constantes las aptitudes que explican el rendimiento de la prueba, las respuestas de los examinados a un par de ítems cualesquiera, son estadísticamente independientes. La independencia local se deriva de la unidimensionalidad ya que simplemente significa que la respuesta a un ítem solo depende de sus parámetros ya que no está influida por el orden de presentación de los ítems. Cuando se cumple el supuesto de unidimensionalidad, se obtiene la independencia local. En este sentido, los dos supuestos son equivalentes (ver Lord, 1980; Lord & Novick, 1968).

2. MODELO

Sea y_{ij} una variable aleatoria que representa la respuesta binaria de un examinado *i (i <= n)* en el ítem *j (j <= k)*. Para una respuesta correcta $y_{ij}=1$, y para una respuesta incorrecta $y_{ij}=0$. Suponiendo un espacio latente unidimensional (ver por ejemplo Villa, 2006, p. 25), la probabilidad de responder correctamente está dada

$$p_j(\theta_i) = \Pr\left[y_{ij} = 1 \backslash \theta_i, \alpha_i, \beta_i\right] = \Phi\left(\alpha_i(\theta_i - \beta_i)\right) \qquad (1)$$

donde $\Phi(\cdot)$ denota la función de distribución acumulada de la distribución normal estándar, y $q_i(\theta_i)=1- p_j(\theta_i)$.

Sea

$$y' = (y_{11}, y_{12}, \dots, y_{nk})$$
$$\theta' = (\theta_1, \theta_2, \dots \dots, \theta_n)$$
$$\alpha' = (\alpha_1, \alpha_2, \dots \dots, \alpha_k)$$
$$\beta' = (\beta_1, \beta_2, \dots \dots \dots, \beta_k)$$

entonces, la función de verosimilitud es

$$L(y \backslash \theta, \alpha, \beta) = \prod_{i \le n} \prod_{j \le k} [p_j(\theta_i)]^{y_{ij}} [q_j(\theta_i)]^{1-y_{ij}} \quad (2)$$

La información concerniente a los parámetros de interés θ,α y β es contenida en la distribución a posteriori de estos parámetros, y esta es

$$\pi(\theta, \alpha, \beta \backslash y) \propto L(y \backslash \theta, \alpha, \beta)\pi(\theta, \alpha, \beta) \qquad (3)$$

donde la densidad conjunta π(θ, α, β) es la distribución a priori del vector de parámetros θ,α y β. También asumiremos independencia

entre los parámetros, es decir

$$\pi(\theta,\alpha,\beta) = \pi(\theta)\pi(\alpha)\pi(\beta) \quad (4)$$

De acuerdo con Vega (2006), asumiremos que

$$\theta_i \sim N(0,1) \quad (5)$$
$$\alpha_j \sim LN(1,4) \quad (6)$$
$$\beta_j \sim N(0,1) \quad (7)$$

3. APLICACIÓN

3.1. DESCRIPCIÓN DE LOS DATOS

Ilustremos el análisis con los datos de la prueba tomada en el curso de verano de la Carrera de Informática a la asignatura Estadística II el 2005. La evaluación se efectuó a 117 estudiantes, con un total de 12 ítems, donde se llega a tener 12 dificultades, 12 discriminaciones y 117 rasgos latentes.

Entonces, nuestra base de datos es especificada por el marco de datos, de algunas variables que pueden llegar a ser las más importantes para poder medir la discriminación que llega a tener la habilidad. Con respecto a la edad, la edad mínima es de 19 años y la edad máxima es de 35 años, son 80 hombres y 37 mujeres, la mayoría egreso del colegio que se encuentra en la ciudad (108), y 51 trabajan. El más antiguo que ingreso a la carrera lo hizo en 1990 y los nuevos ingresaron el año 2003, 49 estudiantes aprobaron del colegio particular y 66 del fiscal, los que si aprobaron la materia de Matemática son 72 y los que no aprobaron son 45 estudiantes.

3.2. FORMULACIÓN DEL PROBLEMA

Dado el modelo (1), donde se tiene que hallar la probabilidad de respuesta del ítem dadas las habilidades de los estudiantes, para luego poder hallar las CCI de los ítems, la curva característica de la prueba, y las funciones de respuesta del ítem y la función de información de la prueba. Para luego poder hallar las funciones de respuesta de la prueba.

3.3. ESTIMACIÓN DE LOS PARÁMETROS DE LOS ÍTEMS

Utilizando WinBugs y BRugs se obtienen las siguientes estimaciones, primeramente para los parámetros para luego poder tener el de la habilidad, donde no se toman todos los valores de la estimación (ver Vega, 2006).

Cuadro Nº 1
Información de la densidad a posteriori de los parámetros de los ítems, parámetro de dificultad

Ítem	Media	DE	2,5%	Mediana	97,5%
1	-1,054e-01	0,1814	-4,600e-01	-1,048e-01	0,24900
2	-3,172e-01	0,1815	-6,755e-01	-3,162e-01	0,03539
3	3,504e-01	0,1836	-7,309e-03	3,496e-01	0,71160
4	-4,707e-01	0,1878	-8,395e-01	-4,684e-01	-0,11020
5	-1,155e+00	0,2231	-1,617e+00	-1,148e+00	-0,73550
6	-4,614e-01	0,1845	-8,289e-01	-4,610e-01	-0,10280
7	3,522e-02	0,1796	-3,158e-01	3,390e-02	0,39710
8	-1,358e-03	0,1836	-3,612e-01	-1,800e-03	0,35800
9	6,958e-01	0,1911	3,242e-01	6,926e-01	1,07300
10	1,008e-01	0,1784	-2,461e-01	9,889e-02	0,45180
11	8,162e-01	0,2139	4,266e-01	8,090e-01	1,26200
12	5,128e-01	0,1898	1,456e-01	5,110e-01	0,89360

Fuente: Elaboración Propia

La dificultad más alta corresponde a $\hat{\beta}_{11} = 0{,}8162$ y el ítem más sencillo 11 corresponde $\hat{\beta}_5 = -1{,}155$, según el Cuadro Nº 1 de toda la prueba.

Cuadro Nº 2
Información de la densidad a posteriori de los parámetros de los ítems, parámetro de discriminación

Ítem	Media	DE	2,5%	Mediana	97,5%
1	6,885e-01	0,5593	2,378e-02	5,540e-01	2,08200
2	6,275e-01	0,5125	2,318e-02	5,068e-01	1,90700
3	6,346e-01	0,5188	2,259e-02	5,082e-01	1,93500
4	7,146e-01	0,5421	2,995e-02	6,050e-01	2,01100
5	1,032e+00	0,6558	6,244e-02	9,458e-01	2,52200
6	5,863e-01	0,5276	1,706e-02	4,303e-01	1,93500
7	6,360e-01	0,5124	2,445e-02	5,175e-01	1,91600
8	9,032e-01	0,6094	4,971e-02	8,027e-01	2,33200
9	7,674e-01	0,5771	3,507e-02	6,457e-01	2,15100
10	6,750e-01	0,5208	2,838e-02	5,661e-01	1,97100
11	1,292e+00	0,7590	8,323e-02	1,243e+00	2,91300
12	1,001e+00	0,6302	6,620e-02	9,190e-01	2,42500

Fuente: Elaboración Propia

En el Cuadro Nº 2 se pueden observar las estimaciones de los parámetros de discriminación, donde los ítems más discriminatorios son los siguientes, $\hat{\alpha}_{11} = 1,292$ con una mediana de 1,243 que llega a ser la más alta, $\hat{\alpha}_{12} = 1,001$ con una mediana de 0,9190, $\hat{\alpha}_{5} = 1,032$, los menos discriminatorios son $\hat{\alpha}_{2} = 0,6275$, $\hat{\alpha}_{1} = 0,9032$ y $\hat{\alpha}_{4} = 0,7146$ que tiene una mediana de 0,6050.

3.4. ESTIMACIÓN PUNTUAL DE LA HABILIDAD

Cuadro Nº 3
Información de la densidad a posteriori de los parámetros de los ítems, parámetro de discriminación

Ítem	Media	DE	2,5%	Mediana	97,5%
5	-3,877e-03	0,3638	-7,351e-01	-8,502e-03	0,78330
6	-1,001e-01	0,3607	-9,275e-01	-4,955e-02	0,57810
7	1,516e-01	0,3746	-5,362e-01	8,954e-02	1,01600
8	6,123e-03	0,3483	-7,476e-01	7,980e-03	0,72030
9	1,970e-01	0,4050	-4,497e-01	1,090e-01	1,16600
10	2,511e-01	0,3983	-3,453e-01	1,578e-01	1,21600
11	1,520e-01	0,3747	-5,130e-01	8,434e-02	1,01900
12	7,190e-03	0,3519	-7,288e-01	4,307e-03	0,75100
13	1,964e-01	0,3962	-4,340e-01	1,080e-01	1,14600
14	-3,446e-01	0,4438	-1,415e+00	-2,402e-01	0,25650
15	-5,756e-02	0,3514	-8,299e-01	-2,970e-02	0,63960
16	-9,604e-02	0,3614	-9,308e-01	-4,513e-02	0,57100
17	-1,274e-01	0,3818	-1,046e+00	-5,721e-02	0,53880
18	-1,769e-01	0,3728	-1,076e+00	-1,006e-01	0,44150
19	-3,139e-01	0,4276	-1,350e+00	-2,116e-01	0,28940

20	-1,064e-01	0,3752	-9,839e-01	-5,241e-02	0,56850
21	4,047e-02	0,3417	-6,687e-01	2,277e-02	0,78780
22	2,296e-02	0,3677	-7,098e-01	4,728e-03	0,84470
23	-5,706e-02	0,3505	-8,234e-01	-2,992e-02	0,64060
24	-3,650e-01	0,4661	-1,523e+00	-2,497e-01	0,24250
25	2,475e-01	0,4149	-3,739e-01	1,455e-01	1,27200
26	5,043e-02	0,3621	-6,669e-01	2,692e-02	0,83090
27	-1,767e-01	0,3694	-1,058e+00	-1,017e-01	0,44240
28	4,375e-02	0,3632	-7,061e-01	2,422e-02	0,81350
29	-1,064e-01	0,3621	-9,198e-01	-5,141e-02	0,56180
30	-3,673e-02	0,3594	-8,358e-01	-1,403e-02	0,67760

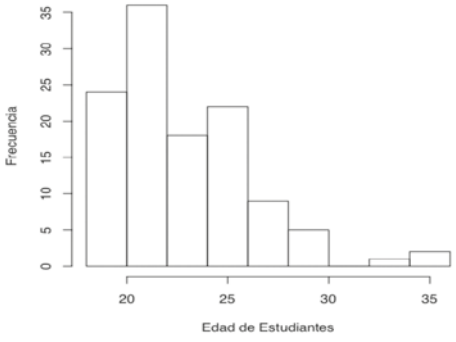
Fuente: Elaboración Propia

Los valores encontrados en el Cuadro Nº 3, reflejan que $\hat{\theta}_{30}$=−0,03673 es un poco más hábil que $\hat{\theta}_{29}$=−0,1064 pero ambas habilidades no son muy buenas como $\hat{\theta}_{22}$ = 0,1520 que llega a ser la más alta habilidad de este Cuadro pero se tiene la habilidad más alta de la prueba que corresponde a $\hat{\theta}_{57}$ = 0,4328 (ver Vega, 2006).

3.5. ANÁLISIS DEL PARÁMETRO DE DISCRIMINACIÓN

La Figura Nº 2 del nuevo modelo que se llama habilidad y que llega a ser nuestro parámetro de discriminación, con respecto a la edad, se ha dividido la edad en cuatro grupos donde se puede ver que el grupo de los jóvenes tiene más habilidad que el grupo de los mayores y también se ve que se tiene más alumnos con habilidades altas entre las edades de 19 y 23 años.

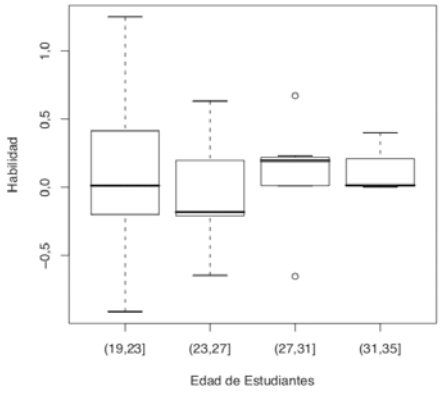
Figura Nº 1
Histograma de las edades de Estudiantes de Informática



Fuente: Elaboración propia

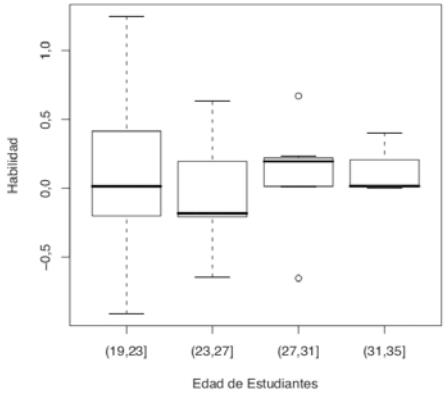
Se tiene en la Figura Nº 1 el total de edades de los estudiantes del curso de verano de la gestión 2005, la materia de Estadística II, de la Facultad de Ciencias Puras y Naturales.

Figura Nº 2
Habilidad vs. Edad



Fuente: Elaboración propia

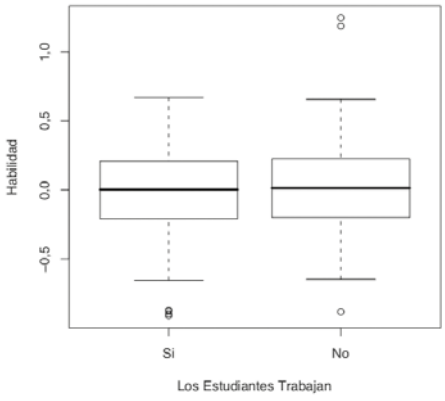
Figura Nº 3
Habilidad vs. Sexo



Fuente: Elaboración propia

Se tiene a la habilidad con respecto al sexo y se puede ver que los hombres son más hábiles que las mujeres según la figura del boxplot de la habilidad ya que la mediana de los hombres es mayor que de las mujeres con el sexo según la Figura Nº 3.

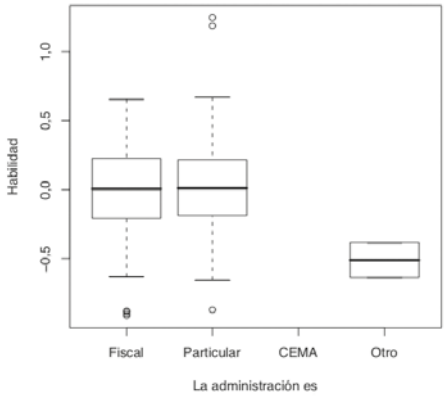
Figura Nº 4
Habilidades vs. Personas que trabajan



Fuente: Elaboración propia

Ahora en la Figura Nº 4, a aquellos estudiantes que trabajan con respecto a la habilidad y se nota que no existe diferencia significativa entre ellos.

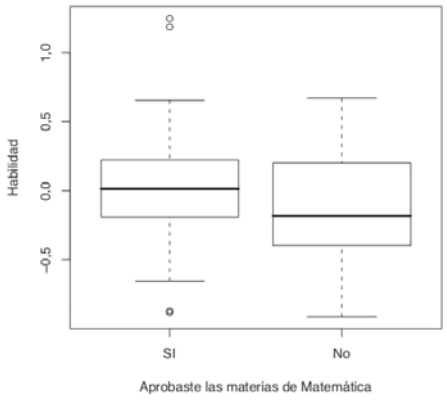
Figura Nº 5
Habilidad con respecto a la administración del colegio



Fuente: Elaboración propia

Ahora en la Figura Nº 5, si la administración del colegio afecta de alguna manera en la habilidad del estudiante y se nota que los colegios de administración pública son los que tienen un poco más de habilidad con respecto a los de colegios particulares pero no tiene mucha diferencia significativa.

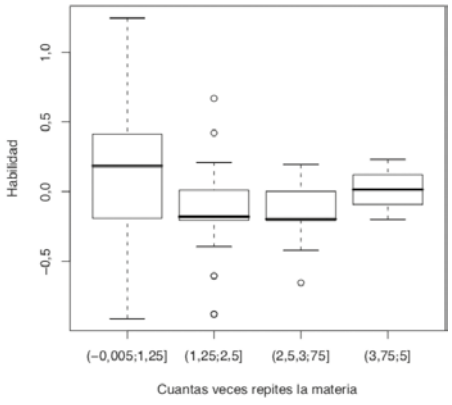
Figura Nº 6
Materias aprobadas de Matemática



Fuente: Elaboración propia

Cómo se puede apreciar en la Figura Nº 6, la mayoría de los estudiantes aprobaron todas las materias de matemáticas, lo cual puede llegar a influir en la habilidad de cada uno de los estudiantes.

Figura Nº 7
Habilidad con respecto a cuantas veces repites la materia

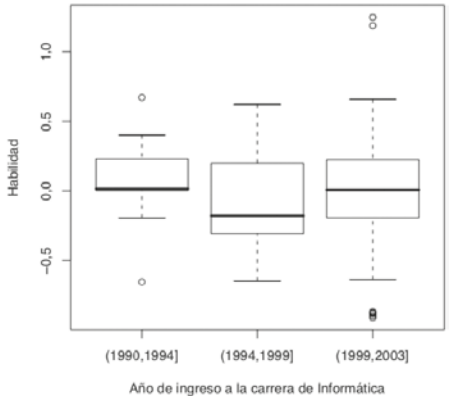


Fuente: Elaboración propia

El haber reprobado con más de una ves dice mucho que sea una persona hábil, y la Figura Nº 7 refleja que cuanto más veces repites una materia tienes un poco mas de habilidad para

poder aprobar el curso.

Figura Nº 8
Habilidad con respecto a cuantas veces repites la materia



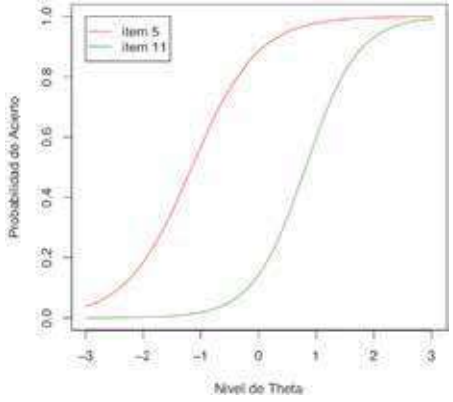
Fuente: Elaboración propia

Se puede ver que los estudiantes que ingresaron entre los años 1990-1997, llevan un poco más de ventaja en relación a los que ingresaron entre los años 1997-2003, pero los que tienen mas habilidad llegan a ser los de los años 1997-2003, eso muestra la Figura Nº 8.

3.6. CURVAS CARACTERÍSTICAS

Curvas características del ítem (CCI)

Figura Nº 9
Curvas características de los ítems 5 y 11

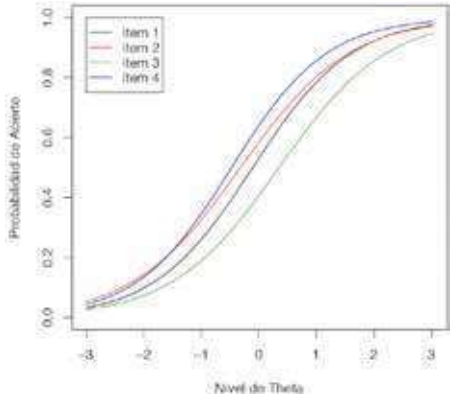


Fuente: Elaboración propia

Se tiene las curvas características de los ítems 5 y 11, donde se tiene dos ítems con menor y mayor dificultad. Y se ve que el ítem de color rojo corresponde al ítem 5 y llega ser el mas fácil de toda la prueba, el de color verde corresponde al ítem 11 que es el más difícil. Donde ambos ítems llegan a ser los más discriminantes de toda la prueba según la Figura anterior y de sus respectivas estimaciones.

Se elige al azar dos estudiantes con habilidad θ_{25}=0,2475, donde tiene una probabilidad de responder el ítem 5 de aproximadamente 0,92 que es el ítem más sencillo, pero para el ítem 11 que es el más difícil será de 0,25 aproximadamente. Ahora para la habilidad negativa de $\theta_{58} = -0{,}3722$ la probabilidad de respuesta para el ítem difícil será de 0,35, para el ítem fácil es de 0,75 aproximadamente.

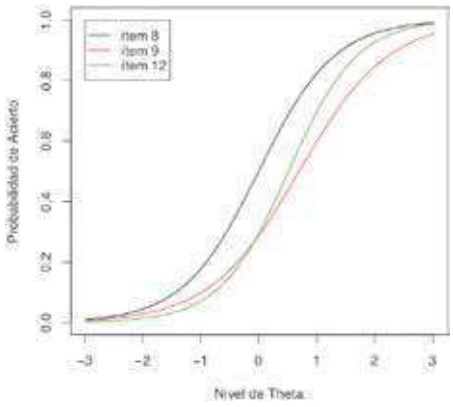
Figura Nº 10
Curvas características de los ítems 1, 2, 3 y 4



Fuente: Elaboración propia

Las probabilidades de acierto para θ_1=0,1640 del ítem 1 es de 0,51, para el ítem 2 es de 0,62, del ítem 3 es 0,41 y para el ítem 4 será de 0,69, por lo que el ítem más difícil corresponde al ítem 3.

Figura Nº 11
Curvas características de los ítems 8, 9 y 12



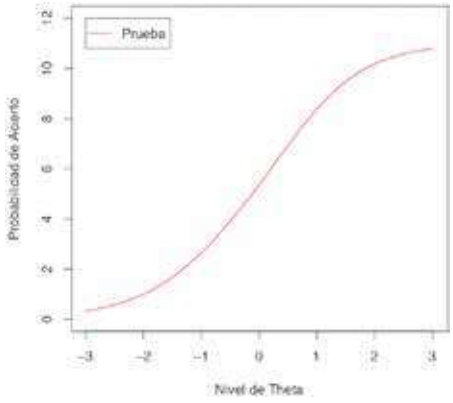
Fuente: Elaboración propia

Para poder determinar probabilidades de respuesta para θ_{15}=−0,05756, donde el ítem 8 será respondido con una probabilidad de 0,44, el ítem 9 con una probabilidad de 0,21 y la misma probabilidad para el ítem 12.

Curva Característica de la Prueba (CCP)

El papel CCP es proporcionar un procedimiento para poder transformar las puntuaciones de habilidad en puntuaciones verdaderas.

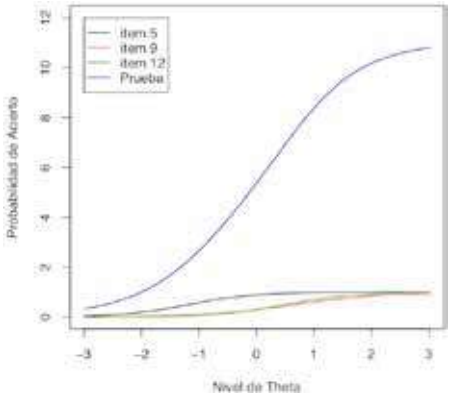
Figura Nº 12
Curva característica de la prueba



Fuente: Elaboración propia

La CCP predice las habilidades de los estudiantes, cuantas respuestas serán las correctas, también hace posible realizar estimaciones a priori.

Figura Nº 13
Curvas características de los ítems 5, 9, 12 y de la prueba

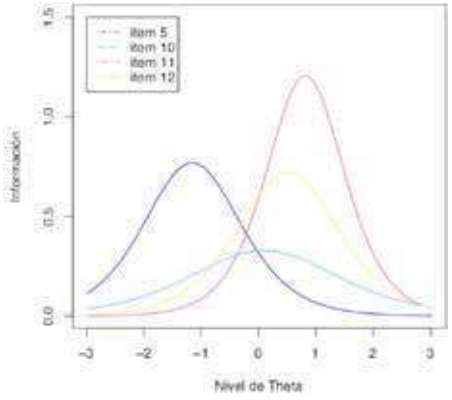


Fuente: Elaboración propia

3.7. FUNCIONES DE INFORMACIÓN

Función de Información del Ítem (FII)

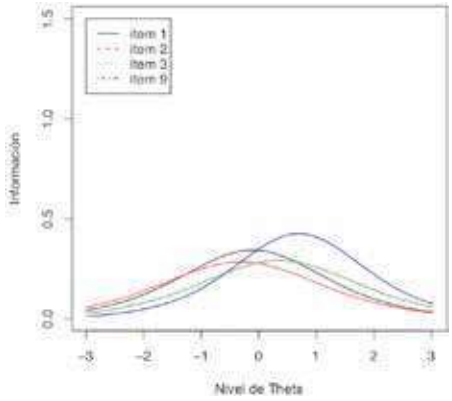
Figura Nº 14
Funciones de información de los ítems 5, 10, 11 y 12



Fuente: Elaboración propia

El ítem 11 aporta con mayor información con una $\theta \approx 1$, el de menor información corresponde al ítem 5.

Figura Nº 15
Funciones de información de los ítems 1, 2, 3 y 9

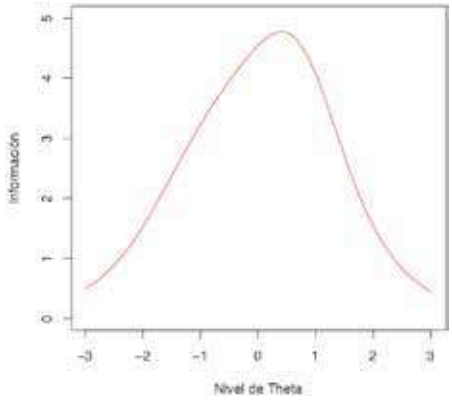


Fuente: Elaboración propia

La mayor información corresponde al ítem 9 y al ítem 3 para una θ≈1.

Función de Información de la Prueba (FIP)

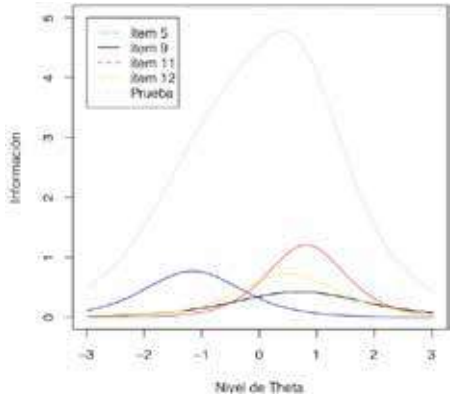
Figura Nº 16
Función de información de la prueba



Fuente: Elaboración propia

La FIP es la curva más alta que de los ítems por que mide en forma general.

Figura Nº 17
FIP y los ítems 5, 9, 11 y 12



Fuente: Elaboración propia

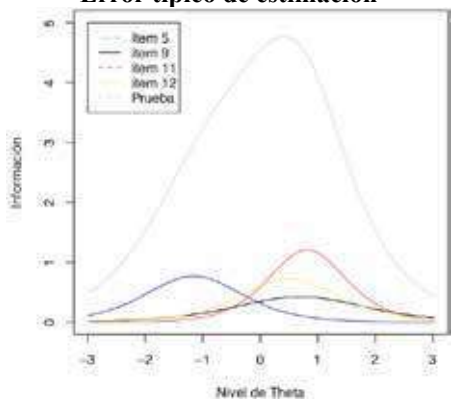
Se ve claramente que es mayor la FIP a los ítems 5, 9, 11 y 12. Donde siempre la FIP será mayor que la FII.

Error típico de estimación

Conceptualmente este error típico ET(θ), no es un estadístico, sino una función de α. Para una prueba cualquiera se tiene muchos errores típicos de estimación. Con este concepto, dejan tener utilidad los conceptos de fiabilidad y de generalizabilidad de una prueba, ya que una prueba puede ser fiable, tener un poco error en ciertos niveles de α y poco fiables en otros. Para los datos de aplicación, los errores típicos de estimación en los diferentes niveles de θ son los siguientes,

$$ET(\theta) = \frac{1}{\sqrt{I(\theta)}} \quad (8)$$

Figura Nº 18
Error típico de estimación



Fuente: Elaboración propia

3.8. AJUSTE DEL MODELO

Para los modelos de la TRI, se tiene muchos métodos de ajuste o de bondad de ajuste a los datos, del tipo chi-cuadrado. El que se utiliza es el método de ajuste Q_y de Yen (1981).

El estadístico de se distribuye según una chi-cuadrado con $m-k$ grados de libertad donde m es el número de intervalos en que se dividió la aptitud y $k*$ es el número de parámetros del modelo. Si el valor del estadístico de Yen (1981), supera el valor crítico de $\chi^2_{m-k*,1-\alpha}$, se rechaza la hipótesis nula de que la CCI se ajusta a los datos. Cuando se tiene muchos ítems que no se ajustan, podemos sospechar que se ha elegido un modelo inadecuado e intentar reanalizar los datos bajo otro modelo alternativo.

Como se ve en el Cuadro Nº 4 la mayoría de los ítems se ajustan a Q_y, pero el ítem 5, como el ítem 11 no se ajustan. Para la prueba se tiene que Q_y = 3,52 tiene que ser menor al calculado $\chi^2_{m-k*,1-\alpha}$= 3,940. Por lo tanto, la prueba se ajusta.

Cuadro Nº 4
Resultados del ajuste del modelo

Ítem	χ^2	Q_y
1	0,711	0,0254722
2	0,352	0,02466778
3	0,103	0,005913995
4	0,103	0,09653338
5	0,103	1,822082
6	0,103	0,01296998
7	0,103	0,012344662
8	0,103	0,02590525
9	0,103	0,04286298
10	0,103	0,01179184
11	0,103	1,273712
12	0,103	0,1709525

Fuente: Elaboración propia

4. CONCLUSIONES Y RECOMENDACIONES

4.1. CONCLUSIONES

El modelo de ojiva normal de dos parámetros es un modelo que ayuda a resolver una serie de problemas en la medición educacional. La principal ventaja que ofrece esta teoría es la invarianza de los parámetros que describen los ítems (dificultad, discriminación), y de los parámetros que describen a las personas. La utilización del modelo de ojiva normal de dos parámetros en el campo de la evaluación educacional es sin duda un aporte significativo que facilitara y perfeccionara la tarea de diseño e implementación de pruebas, especialmente aquellas de gran escala.

El modelo de ojiva normal de dos parámetros nos permite un ajuste adecuado a los datos, quedando atrás los métodos tradicionales, pues nos permite llegar a conclusiones claras de acuerdo con los objetivos del estudio. El empleo del método MCMC, resulta de mayor importancia a la hora de obtener resultados

confiables, se aplico WinBUGS y BRugs. Cuando se utiliza TRI, es muy difícil para un profesor predecir el puntaje que tendrá un alumno en la prueba real, ya que lo más seguro es que no tendrá los parámetros de las preguntas que él mismo ha diseñado para el ensayo, ni las herramientas para estimar el puntaje a partir de éstos.

4.2. RECOMENDACIONES

TRI es una teoría que se funda en una serie de supuestos que se cumplan. Es importante evaluar que los supuestos se cumplan de manera adecuada, es decir, éstas no sean tan importantes como para invalidar las aplicaciones de TRI y afectar significativamente la propiedad de invarianza de los parámetros. Para ello, es fundamental que expertos en el tema realicen las pruebas adecuadas para testear si efectivamente se está cumpliendo la unidimensionalidad y si el modelo está ajustándose a los datos experimentales. Por ejemplo, si se detecta que un determinado tipo de pregunta es fuente de multidimensionalidad, se debe evaluar el efecto desde el punto de vista del contenido y propósitos del test, y cuáles serían los efectos de su eliminación. Es posible que en muchos casos sea preferible cambiar el modelo estadístico a utilizar antes que desechar definitivamente el tipo de pregunta. Es muy importante que los expertos encargados de implementar el modelo estén al tanto de las nuevas investigaciones y vayan avanzando juntamente con los progresos que se vayan dando en la teoría.

Agradecimientos

El primer autor fue parcialmente apoyado por la beca FIB-UV de la Universidad de Valparaíso, de Chile. Los autores agradecen al editor por sus útiles comentarios.

Apéndice

Código R2OpenBugs.

```
> model <- function() {
+     for (i in 1:n) {
+         for (j in 1:k) {
+             p[i, j] <- phi(alfa[j] * theta[i] - beta[j])
+             y[i, j] ~ dbern(p[i, j])
+         }
+         theta[i] ~ dnorm(0, 1)
+     }
+     for (j in 1:k) {
+         alfa[j] ~ dlnorm(0, 4)
+         beta[j] ~ dnorm(0, 1)
+     }
+ }
```

BIBLIOGRAFÍA

Lawley, D. N. (1940). VI.—the estimation of factor loadings by the method of maximum likelihood., *"Proceedings of the Royal Society of Edinburgh"*, 60(1), 64–82. 1

Lawley, D. N. (1944). X.—the factorial analysis of multiple item tests. *"Proceedings of the Royal Society of Edinburgh. Section A. Mathematical and Physical Sciences"*, 62(1), 74–82. 1

Lord, A. B. F. M. (1980). *"Applications of Item Response Theory to Practical Testing Problems"*.Lawrence Erlbaum Associates. 3

Lord, A. B. F. M. & Novick, M. R. (1968). *"Statistical Theories of Mental Test Scores"*. Addison-Wesley Publishing Company. 3

Vega, G. F. (2006). *"El modelo logístico de dos parámetros"*. Tesis de Licenciatura, Universidad Mayor de San Andrés, La Paz. 3, 4, 5

Villa, M. D. (2006). *"El modelo rasch y aplicaciones"*. Tesis de Licenciatura, Universidad Mayor de San Andrés, La Paz. 3

Yen, W. M. (1981)., *"Using simulation results to choose a latent trait model"*. Applied Psychological Measurement, 5(2), 245–262. 10

CONSECUENCIAS DE ALTA MULTICOLINEALIDAD EN UN MODELO DE REGRESIÓN LINEAL

Dr. (C) Coa Clemente, Ramiro[1]

✉ *clementecoa@gmail.com*

RESUMEN

En este artículo se revisan e ilustran algunas consecuencias de la alta multicolinealidad entre covariables presentes en la parte sistemática de un modelo de regresión lineal. Con este propósito, se comparan dos modelos. En el primero no existe el problema de multicolinealidad, es decir las covariables son linealmente independientes. En el segundo modelo se tiene el problema de alta multicolinealidad, es decir las covariables están muy asociadas linealmente. Se analizan cuatro tipos de consecuencias: (i) sobre la magnitud de los coeficientes de regresión, (ii) sobre las sumas de cuadrados adicionales, (iii) sobre la magnitud de los errores estándar para los estimadores de coeficientes y (iv) sobre pruebas estadísticas de los coeficientes. En presencia de alta multicolinealidad entre las covariables del modelo, estas consecuencias podrían conducir a inferencias estadísticas erróneas y consecuentemente a conclusiones incorrectas.

PALABRAS CLAVE

Multicolinealidad, multicolinealidad perfecta, alta multicolinealidad o multicolinealidad imperfecta.

ABSTRACT

In this article, some consequences of the high multicollinearity among covariates present in the systematic part of a linear regression model are reviewed and illustrated. For this purpose, two models are compared. In the first there is no problem of multicollinearity, that is, the covariates are linearly independent. In the second model there is the problem of high multicollinearity, that is, the covariates are very linearly associated. Analyze four types of consequences: (i) on the magnitude of the regression coefficients, (ii) on the sums of additional squares, (iii) on the magnitude of the standard errors for the coefficient estimators and (iv) on statistical tests of the coefficients. In the presence of high multicollinearity among the covariates of the model, these consequences can lead to erroneous statistical inferences and consequently to incorrect conclusions.

KEYWORDS

Multicollinearity, perfect multicollinearity, high multicollinearity or imperfect multicollinearity.

1. EL PROBLEMA DE MULTICOLINEALIDAD

Consideremos el modelo de regresión lineal múltiple $Y=X\beta+\varepsilon$, donde Y es el vector $n*1$ de variables respuesta, X es la matriz $n*p$ de observaciones, β es el vector $p*1$ de coeficientes y ε es un vector de errores aleatorios con los supuestos clásicos, es decir $\varepsilon \sim N(0,\sigma^2 I_n)$.

Formalmente, la multicolinealidad se define en términos de la dependencia lineal entre las columnas de la matriz X. Recordemos que los vectores $X_1,X_2,...,X_K$ son linealmente dependientes si hay un conjunto de constantes $a_1,a_2,...,a_K$, no todos ceros, tal que $\sum_{k=1}^{K} a_k x_k = 0$. Si esta ecuación se cumple para algún subconjunto de las columnas de X, entonces el rango de la matriz

[1] Ex-Director de Investigación de la Unidad de Analisis y Politica Social (UDAPSO)

$X'X$ es inferior a p, y por tanto no existe una solución única para el vector de coeficientes. En esta situación se tiene un problema denominado multicolinealidad perfecta. El problema más frecuente, sin embargo, es el denominado multicolinealidad imperfecta o alta multicolinealidad. Este se presenta cuando todas o algunas covariables del modelo están altamente correlacionadas, es decir cuando se tiene una dependencia "casi lineal" entre las columnas de X. Por este hecho, la multicolinealidad es un problema principalmente de grado o de nivel. Se podría decir, entonces, que cada conjunto de datos a ser analizado con un modelo de regresión sufre del problema en alguna medida, a no ser que las columnas de X sean ortogonales, en cuyo caso no existe problema de multicolinealidad, ni perfecta ni imperfecta, lo que se da generalmente en un experimento diseñado apropiadamente.

2. CONSECUENCIAS DE LA MULTICOLINEALIDAD IMPERFECTA

En este artículo se examinan cuatro efectos de la presencia de multicolinealidad imperfecta: (i) efectos sobre la magnitud de los coeficientes de regresión, (ii) sobre la suma de cuadrados adicional, (iii) sobre el error estándar de los coeficientes estimados y (iv) sobre las decisiones en pruebas de hipótesis.

Para visualizar apropiadamente estos efectos, la información a ser analizada previamente es transformada mediante la denominada *transformación correlación*. Dos de las interesantes características de esta transformación son: (i) la nueva matriz simétrica $X'X$ representa la matriz de correlaciones entre las columnas de la nueva matriz X y que (ii) el nuevo vector $X'Y$ consiste de las correlaciones entre cada nueva covariable y la nueva variable respuesta. Consecuentemente, el nuevo vector de coeficientes puede ser estimado a partir de la matriz y el vector de correlaciones.

2.1 EFECTOS SOBRE LA MAGNITUD DE LOS COEFICIENTES DE REGRESIÓN

Para ilustrar el efecto de la multicolinealidad imperfecta sobre la magnitud de los coeficientes consideremos dos modelos de regresión, cada uno con dos covariables. En el primer modelo no existe asociación lineal entre las dos covariables, es decir la correlación lineal es 0; en el segundo modelo ambas covariables están altamente asociadas, una correlación lineal de 0,92.

Cuando ambas covariables no están correlacionadas, los estimadores de los coeficientes permanecen invariables. El coeficiente para X_1 es el mismo (0,742) cuando X_1 es la única covariable en el modelo o cuando ambas covariables están en el modelo. Lo mismo ocurre para el coeficiente de X_2. Sin embargo, cuando ambos regresores están fuertemente asociados, los valores de los coeficientes tienen grandes cambios. Cuando X_1 es la única covariable en el modelo, su coeficiente es 0,915; mientras cuando ambas covariables están presentes en el modelo, su coeficiente es 0,294, una reducción de 67,0%. De manera similar, aunque menos acentuada, la reducción para el coeficiente de X_2 es 28,8% (Cuadro Nº 1).

Cuando las covariables están fuertemente asociadas, las magnitudes de sus coeficientes cambian significativamente. Consecuentemente, esos coeficientes no reflejan los efectos reales de sus correspondientes covariables sobre la respuesta, sólo reflejan un efecto parcial o marginal, que podría conducir a conclusiones erradas.

Cuadro Nº 1
Cambios en las magnitudes de los coeficientes

Variables en el Modelo	Modelo 1 (Correlación 0)		Modelo 2 (Correlación 0,92)	
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
X_1	0,742		0,915	
X_2		0,638		0,945
X_1 X_2	0,742	0,638	0,294	0,673
Cambio Relativo (%)	0	0	67,9%	28,8%

Fuente: Elaboración Propia

2.2 EFECTOS SOBRE LAS SUMAS DE CUADRADOS ADICIONALES

Consideremos los dos modelos de regresión anteriores. Recordemos que el primero es un modelo en el que no existe asociación lineal entre los dos regresores (correlación lineal 0), mientras en el segundo ambos regresores se encuentran altamente asociados (correlación lineal de 0,92).

Cuando ambos regresores no están asociados linealmente, la contribución adicional de X_1 a la suma de cuadrados de la regresión (SCReg) es la misma que cuando X_1 es la única covariable presente en el modelo. Esto es, la $SCReg(X_1)$=0,550 $=SCReg(X_1 \mid X_2)$. De manera similar, la $SCReg(X_2)$=0,408=$SCReg(X_2 \mid X_1)$. En cambio, cuando ambas covariables están altamente relacionadas, la contribución marginal de cualquiera de ellas a la suma de cuadrados de la regresión es diferente de la contribución a la suma de cuadrados cuando la misma variable es la única en el modelo; es decir, la $SCReg(X_1)$=0,838≠0,013=$SCReg(X_1 \mid X_2)$ y la $SCReg(X_2)$=0,892≠0,067=$SCReg(X_2 \mid X_1)$ (Cuadro Nº 2).

Cuadro Nº 2
Cambios en las sumas de cuadrados

	Modelo 1 (Correlación 0)	Modelo 2 (Correlación 0,92)
$SCReg(X_1)$	0,550	0,838
$SCReg(X_1 \mid X_2)$	0,550	0,013
$SCReg(X_2)$	0,408	0,892
$SCReg(X_2 \mid X_1)$	0,408	0,067

Fuente: Elaboración Propia

La razón por la que la $SCReg(X_1 \mid X_2)$=0,013 es muy pequeña comparada con la SCReg(X_1)=0,838 es que X_1 y X_2 están altamente correlacionadas. Cuando X_2 se encuentra en el modelo, la contribución marginal de X_1 a la suma de cuadrados de la regresión es comparativamente pequeña debido a que X_2 contiene mucha de la misma información contenida en X_1. Dicho en otros términos, cuando las covariables del modelo están muy asociadas linealmente, no hay una única suma de cuadrados que pueda ser adscrita a cualquiera de estas covariables, por lo que no es posible determinar el efecto

neto de esas covariables en la explicación de la variabilidad de la respuesta.

2.3 EFECTOS SOBRE LA MAGNITUD DE LOS ERRORES ESTÁNDAR DE LOS ESTIMADORES DE COEFICIENTES

Nuevamente consideremos los dos modelos de regresión anteriores. Ahora se analiza el efecto de una fuerte asociación lineal entre las covariables del modelo sobre los errores estándar de sus correspondientes coeficientes estimados.

En ausencia de asociación lineal entre las covariables del modelo, los errores estándar de los coeficientes estimados se reducen. Por ejemplo, cuando el modelo contiene sólo X_1, el error estándar de su coeficiente asociado β_1 es 0,274, valor que se reduce a 0,092 cuando ambas covariables se encuentran en el modelo. Similar comportamiento se observa para X_2 (Cuadro Nº 3). Este hecho es deseable puesto que los estimadores de los coeficientes tienen mayor precisión.

Cuadro Nº 3
Cambios en los errores estándar de los coeficientes estimados

Variables en el Modelo	Modelo 1 (Correlación 0)		Modelo 2 (Correlación 0,92)	
	$ee(\hat{\beta}_1)$	$ee(\hat{\beta}_2)$	$ee(\hat{\beta}_1)$	$ee(\hat{\beta}_2)$
X_1	0,274		0,142	
X_2		0,314		0,945
X_1, X_2	0,092	0,092	0,303	0,303
Cambio	Reducción	Reducción	Incremento	Incremento

Fuente: Elaboración Propia

En cambio, cuando las covariables están fuertemente asociadas, los errores estándar de sus correspondientes coeficientes estimados se incrementan considerablemente. Por ejemplo, el error estándar de $\hat{\beta}_1$se incrementa de 0,142 a 0,303, un incremento de un poco más del doble. El incremento para el error estándar de $\hat{\beta}_2$ es de casi el triple, pasando de 0,116 a 0,303. En consecuencia, el alto grado de multicolinealidad entre las covariables produce un incremento considerable en los errores estándar de los estimadores, lo que conduce a inferencias menos imprecisas e incluso a falsas conclusiones.

2.4 EFECTOS SOBRE PRUEBAS ESTADÍSTICAS DE LOS COEFICIENTES

Un abuso frecuente en el análisis de modelos de regresión lineal es examinar para cada coeficiente de regresión la estadística $t = \hat{\beta}_k / e.e(\hat{\beta}_k)$ que resulta de dividir el estimador del coeficiente por su error estándar, esto con el propósito de decidir si la hipótesis nula H_0: $\beta_k = 0$ es o no rechazada, para un determinado nivel de significancia α. En presencia de alta multicolinealidad, sin embargo, las conclusiones derivadas de dicho examen podrían ser incorrectas. Esta situación es analizada a continuación.

En ausencia del problema de multicolinealidad, el test t para probar individualmente cada una de la dos hipótesis nulas H_0:β_1=0 y H_0:β_2=0 conduce a la misma decisión a la que se llega con el test F, un test adecuado para probar la hipótesis de que simultáneamente ambos coeficientes son nulos, es decir H_0:β_1=β_2=0. En efecto, los valores de las estadísticas t para $\hat{\beta}_1$(8,10) y

para $\hat{\beta}_2$ (6,97) son superiores al valor crítico 2,97, por lo que individualmente se rechazan ambas hipótesis, H_0:β_1=0 y H_0:β_2=0 (Cuadro Nº 4). Estas decisiones son coherentes con la decisión tomada con base en el test *F*. Este test también conduce a rechazar la hipótesis de que simultáneamente ambos coeficiente son nulos H_0:β_1=β_2=0, pues el valor de la estadística *F*(57,06) supera el percentil 95 de la distribución *F*(5,79). En consecuencia, cuando las covariables no están asociados linealmente, ambos tests, *t* y *F*, son coherentes, conducen a las mismas decisiones y, por ende, a las mismas conclusiones.

Cuadro Nº 4
Efectos sobre pruebas estadísticas de los coeficientes

Test	Modelo 1 (Correlación 0)		Modelo 2 (Correlación 0,92)	
	Hipotesis	**Hipotesis**	**Hipotesis**	**Hipotesis**
	H_0: β_1=0 H_0: β_2=0	H_0: β_1= β_2=0	H_0: β_1=0 H_0: β_2=0	H_0: β_1= β_2=0
t	$t_{tabla}(0{,}9875; 6) = 2{,}97$ $t_{cal,\hat{\beta}_1} = 8{,}10$ $t_{cal,\hat{\beta}_2} = 6{,}97$ Como: $t_{cal,\hat{\beta}_1} > t_{tabla}$ $t_{cal,\hat{\beta}_2} > t_{tabla}$ Decisión: **Rechazar** ambas H_0		$t_{tabla}(0{,}9875; 8) = 2{,}75$ $t_{cal,\hat{\beta}_1} = 0{,}97$ $t_{cal,\hat{\beta}_2} = 2{,}22$ Como: $t_{cal,\hat{\beta}_1} < t_{tabla}$ $t_{cal,\hat{\beta}_2} < t_{tabla}$ Decisión: **No Rechazar** ambas H_0	
F		$F_{tabla}(0{,}95; 2{,}5) = 5{,}79$ $F_{cal} = 57{,}06$ Como: $F_{cal} > F_{tabla}$ Decisión: **Rechazar** H_0		$F_{tabla}(0{,}95; 2{,}7) = 4{,}74$ $F_{cal} = 33{,}36$ Como: $F_{cal} > F_{tabla}$ Decisión: **Rechazar** H_0

Fuente: Elaboración Propia

En cambio, en presencia de alta multicolinealidad entre ambas covariables, los dos tests, *t* y *F*, conducen a decisiones contradictorias. La prueba *t* conduce a no rechazar individualmente cada una de las dos hipótesis H_0:β_1=0 y H_0:β_2=0; mientras la prueba F, una prueba más apropiada para este problema, conduce a rechazar la hipótesis de que simultáneamente ambos coeficientes son nulos H_0:β_1=β_2=0.

La razón para este resultado contradictorio es que cada una de las dos pruebas t es una prueba marginal. Esto es, un valor pequeño de la $SCReg(X_1 \mid X_2)$ (Cuadro Nº 2) indica que X_1 no proporciona mucha información adicional sobre la que proporciona la covariable X_2, por lo que se arriba a la

conclusión de que $\beta_1=0$. De manera similar se llega a la conclusión de que $\beta_2=0$ porque la *SCReg* $(X_2 \mid X_1)$ es pequeña, reflejando que X_2 no proporciona información adicional substancial cuando X_1 se encuentra en el modelo. En consecuencia, los dos tests, *t* y *F*, conducen a decisiones y consecuentemente a conclusiones contradictorias.

3. CONCLUSIÓN

La presencia de alta multicolinealidad en la parte sistemática del modelo de regresión tiene varias consecuencias. Una primera es la disminución en la magnitud de los coeficientes correspondientes a covariables fuertemente asociadas, razón por lo que esos coeficientes sólo reflejan un efecto parcial o marginal sobre la respuesta, no reflejan los efectos reales de sus correspondientes covariables. Una segunda consecuencia tiene que ver con el hecho de que no hay una única suma de cuadrados que pueda ser adscrita a cualquiera de las covariables altamente asociadas, por lo que no es posible determinar el efecto neto de esas covariables. La tercera consecuencia es el incremento en los errores estándar de los coeficientes estimados, lo que conduce a inferencias menos precisas. Por último, con relación a las pruebas de hipótesis, la presencia de alta multicolinealidad lleva a decisiones contradictorias entre las pruebas *t* y *F*.

BIBLIOGRAFÍA

Greene, W. (1997), "*Econometric Analysis*", 3ra Ed. MacMillan.

Seber, G.A.F. and Lee, A.J, (2003), "*Linear Regression Analysis*", 2da Ed. Wiley.

Sen, A. and Srivastava, M. (1990) "*Regression Analysis: Theory, Methods, and Applications*". Springer-Verlag.

LA INTEGRAL DE HENSTOCK – KURZWEIL EN LA ENSEÑANZA DE LA TEORÍA DE LA PROBABILIDAD

Lic. Esp. Delgado Álvarez, Raúl[1]

✉ *dea_5@hotmail.com*

RESUMEN

En el presente trabajo, se expone la relación entre la función de distribución acumulada y la función de probabilidades o la función de densidad de probabilidades de una variable aleatoria, a través de la integral de Henstock- Kurzweil, que se presenta como una generalización de la integral de Riemann, Riemann –Stieltjes y la de Lebesgue, esta relación se puede comprender mejor con el segundo teorema fundamental del Cálculo que tiene una interpretación más didáctica a través de la H-K integral, que también cumple con la rigurosidad requerida, para la comprensión y pruebas de los denominados teoremas de límites a través de los teoremas de la convergencia, la integral de H-K es de gran utilidad y comprensión como el Teorema de Convergencia Uniforme que se muestra en el presente trabajo.

PALABRAS CLAVE

Funciones de distribución, Integral de Henstock-Kurzweil, Segundo Teorema Fundamental del Cálculo, Convergencia Uniforme.

ABSTRACT

In this paper, the relationship between the cumulative distribution function and the probability function or the probability density function of a random variable is exposed, through the Henstock-Kurzweil integral, which is presented as a generalization of the integral of Riemann, Riemann-Stieltjes and that of Lebesgue, this relationship can be better understood with the second fundamental theorem of Calculus that has a more didactic interpretation through the integral HK, which also meets the required rigor, for understanding and evidence of the so-called limit theorems through convergence theorems, the integral of HK is very useful and understanding as the uniform convergence theorem shown in the present work.

KEYWORDS

Distribution functions, Henstock-Kurzweil Integral, Second Fundamental Calculation Theorem, Uniform Convergence

1. INTRODUCCIÓN

La relación existente entre la función de probabilidad o la densidad de probabilidades y la función de distribución acumulada se muestra frecuentemente en cursos regulares de teoría de la probabilidad, a través de un teorema de la Matemática denominado Teorema fundamental del Cálculo.

$$\frac{d}{dx}\int_{-\infty}^{x} F(t)dt = f(x)$$

La relación enunciada se desarrolla en un curso inicial de Probabilidad, donde se hace mención de la Integral de Riemann, las limitaciones de la Integral de Riemann en la interpretación de funciones de distribuciones mixtas, es decir, discretas en un tramo y

[1] Docente de la materia de Introducción a la teoría de la probabilidad de la Carrera de Estadistica de la Facultad de Ciencias Puras y Naturales de la UMSA.

continuas en otro, en un curso de teoría de la Probabilidad intermedio o de introducción a la teoría de la probabilidad se salva esta interpretación a partir de la Integral de Riemann – Stieltjes:

$$\int_{[a,b]} g(x)dF(x)$$

Que por supuesto, muestra una definición más consistente y general de la integración para tratar variables aleatorias mixtas, sin embargo, una condición para la existencia de la Integral de Riemann Stieltjes es el problema de la discontinuidad, es decir, cuando ambas funciones poseen puntos de discontinuidad en valores iguales, la función integrando en la Teoría de la Probabilidad generalmente es una función de la variable aleatoria como la función identidad en el caso de la esperanza matemática, funciones potenciales para explicar los momentos ordinarios o centrales y en general funciones que difícilmente poseen discontinuidades, en cambio la función integradora expresada por *dF(x)* es la que en caso discreto mostrará discontinuidades o saltos, por lo cual: la función *F(x)* debe ser de variación acotada esto es: $\sum_{i=1}^{n}|F(x_i) - F(x_{i-1})| < M$, M>0

Esta dificultad nos hace pensar en la Integral de Lebesgue, pero es conocido que su tratamiento no es muy amigable por decirlo suavemente, entonces se debe reflexionar desde el inicio, es decir, cuando se habla por primera vez de la relación entre una función y su derivada.

La historia nos remonta al Siglo XVII con Isaac Newton que encontró que la integración es el proceso inverso de la derivación, es decir una función es newton integrable si tiene una anti derivada, posteriormente Agustin Cauchy define la integral de manera constructiva restringiendo a las funciones continuas, esta integral coincide con la integral de Newton, sin embargo debido a la existencia de derivadas no acotadas permanece más general la definición de Newton, luego Berhnard Riemann haciendo uso de puntos dentro de subintervalos, redefine la integral de Cauchy permitiendo la integración de algunas funciones discontinuas, es conocida esta integral por la sencillez para probar teoremas básicos y su manejo es el que se enseña en los cursos del Cálculo Diferencial e Integral hasta la actualidad pero existen algunas dificultades, para citar alguna es que la función de Riemann debe ser acotada y más aún la interpretación incompleta del Teorema Fundamental del Cálculo como proceso inverso al proceso de derivación hace desventajosa esta integral, posteriormente por el año 1902 aparece Henri Lebesgue que utiliza longitud de intervalos en su deducción, lo que supera las limitaciones de la integral de Riemann en cuanto a los teoremas de convergencia, la misma que es la más aceptada por la eficiencia de los mismos y la gran generalidad de funciones Lebesgue integrables, aún sin embargo existen funciones derivables en todo punto no necesariamente acotadas, cuya derivada *F'* no es Lebesgue integrable.

Mostrando una respuesta más satisfactoria al teorema fundamental del Cálculo, Denjoy y Perron definen una integral que permite recuperar a una función a partir de su derivada pero la definición de la misma es muy complicada, a continuación de este proceso en el año 1960 Ralph Henstock y Jaroslav Kurzweil formulan la H-K integral, que resulta un equivalente a la integral de Denjoy y Perron, pero dan una visión más general del Teorema Fundamental del Cálculo, es decir, una integración que retome la idea inicial de Newton.

Es así que relacionar a Newton y a Riemann

parece importante por lo que la nueva integral debe ser más potente, y por supuesto la simplicidad de Riemann debe estar presente para recuperar una función f a partir de su derivada y sin imponer condición algún sobre *F*.

Se observa en la definición de la integral de Riemann, al dividir cada subintervalo no necesariamente tiene la misma longitud y más aún se escoge una etiqueta I_j de manera arbitraria (lo mismo puede estar al inicio como al final o en general, en un punto interior a cada subintervalo), por lo tanto la medida de fineza de una partición está dada por la máxima longitud de los subintervalos sin que dependan de las etiquetas, si se escogen subintervalos de menor longitud que cierta δ>0 significa que acota la norma de la partición en constante y por tanto los intervalos de la partición no siempre tienen una longitud adecuada, entonces es necesario considerar la relación entre las etiquetas y los intervalos para determinar la medida de fineza de la partición, así se observa en la integral de Henstock Kurzweil ilustrando como es que podemos recuperar una función *F*, a partir de su derivada *F´*.

A continuación, se expondrán los argumentos esenciales que en este trabajo son de mínima comprensión para definir la integral H-K, y se pondrá en evidencia la demostración del Segundo Teorema Fundamental del Cálculo y el Teorema de Convergencia Uniforme, pudiendo notar la sencillez de las indicadas pruebas a través de la integral de Henstock-Kurzweil.

2.- DESARROLLO

La descripción de la relación de cobertura, cubiertas de Cousin, el Teorema de Cousin son argumentos previos para la integral de Henstock–Kurzweil, y los teroremas: II Teorema del Cálculo y el Teorema de la Convergencia Uniforme muestran la potencia de la integral H-K.

Relación de cobertura.- Es una familia de parejas ([*c,d*], *x*), donde $x \in [c,d]$.

Sea $F'(x)= f(x)$ para toda x ϵ [a.b], sea ε >0, consideremos la relación de cobertura β que consiste en todos los pares de intervalos y puntos ([*c,d*], *t*) para los cuales ε[*c,d*] con [*c,d*]⊂ [*a,b*] con la propiedad de que $\left|\frac{F(d)-F(c)}{d-c} - f(t)\right| \leq \varepsilon$, obsérvese que la relación de cobertura es muy grande, ya que por definición de diferenciabilidad, para cada punto $t \in$[a,b] existe un $\delta_t > 0$ tal que β contiene a todas las parejas ([*c,d*], *t*) con $d\text{-}c < \delta_t$

Supóngase que β contiene una partición: $\pi=\{([x_{i-1},x_i],t_i): 1\leq i\leq n\}$, la diferencia $F(b) - F(a)$ como la suma telescópica:

$$\mathrm{F(b)} - \mathrm{F(a)} = \sum_{i=1}^{n}[F(x_i) - F(x_{i-1})]$$

por lo tanto se obtiene:

$$\left|F(b) - F(a) - \sum_{i=1}^{n} f(t_i)(x_i - x_{i-1})\right| = *$$

$$* = \left|\sum_{i=1}^{n}\{F(x_i) - F(x_{i-1}) - f(t_i)(x_i - x_{i-1})\}\right|$$

$$\leq \sum_{i=1}^{n}|\{F(x_i) - F(x_{i-1}) - f(t_i)(x_i - x_{i-1})\}|$$

$$\leq \sum_{i=1}^{n}\varepsilon(x_i - x_{i-1}) = \varepsilon\,(b-a)$$

Se concluye entonces que

$$\left|\int_a^b f(x) - \sum_{i=1}^{n} f(t_i)(x_i - x_{i-1})\right| \leq \varepsilon(b-a)$$

Esto indica que la integral puede ser aproximada por la suma de Riemann, pero seleccionando las etiquetas t_i de una manera distinta a como procede en el contexto de Riemann, es decir se ha utilizado una partición cuyos elementos provienen de

una relación de cobertura que describa de manera natural la geometría del problema es decir, partiendo de la diferenciabilidad de la integral indefinida *F*.

Cubiertas de Pierre Cousin

Una relación de cobertura β es una cubierta de Cousin de un intervalo $[a,b]$ si para cada $x\epsilon[a,b]$, existe $\delta>0$ tal que β contiene todos los pares $([c,d], x)$, para los cuales $x\epsilon[c,d] \subset [a,b]$ y $(d-c) < \delta$.

Sea β una cubierta de Cousin del intervalo $[a,b]$, entonces β es una cubierta de Cousin de cada subintervalo $[c,d]\subset[a,b]$. Esto quiere decir que si $x\epsilon[c,d]\subset[a,b]$. Como $x\epsilon[a,b]$ entonces existe $\delta>0$ tal que β contiene a todos los pares (I,x) con $x\epsilon I\subset[a,b]$ y $l(I)<\delta$. Si el par (J,x) es tal que $x\epsilon J\subset[c,d]$ con $l(J)<\delta$, entonces (J,x) debe estar en β ya que $J\subset[a,b]$.

Al considerar dos cubiertas de Cousin para comprobar que la intersección es también una cubierta de Cousin sobre el intervalo $[a,b]$, será suficiente considerar: $\delta = min\{\delta_1, \delta_2\}$.

Teorema de Cousin

Sea δ un indicador en $[a,b]$, entonces existe una partición δ - fina de $[a,b]$.

Demostración: Por el método del absurdo reducción al absurdo.

Supóngase que $[a,b]$ no tiene partición δ – fina.

Sea $c=\frac{a+b}{2}$ y considérense los subintervalos $[a,c]$ y $[c,b]$, si los dos $[a,c]$ y $[c,b]$ intervalos tuviesen una partición δ–fina entonces su unión será una partición δ–fina de $[a,b]$, es decir, alguno de los subintervalos $[a,c]$ y $[c,b]$ no tiene partición δ–fina, sin perdida de generalidad será $I'=[a_1,b_1]$, sea $c_1=\frac{a_1+b_1}{2}$ y considérese los subintervalos $[a_1,c_1]$ y $[c_1,b_1]$ y uno de los intervalos no tiene partición δ–fina, sea $I^2=[a_2,b_2]$ este subintervalo, así de esta manera, se obtiene una sucesión de intervalos compactos encajados $\{I^n\}_{n\epsilon N}$ uno con longitud $|I^n|=\frac{b-a}{2^n}$ sin partición δ–finas, luego por el teorema de los intervalos encajados se asegura que existe un único punto x en la intersección de todos los intervalos I^n sin embargo como $\delta(x)>0$ por la propiedad Arquimediana existe $p\epsilon\mathbf{N}$ tal que $|I^p|=\frac{b-a}{2^p}<\delta(x)$ por tanto $I^p\subset[x-\delta(x),\ x+\delta(x)]$.

Además el par $\{(I^p,x)\}$ es una partición formada por un solo intervalo δ–fina de I^p en contra de la elección de I^p.

Esta contradicción prueba que $[a,b]$ tiene alguna partición δ–fina, como se quería probar.

Definición de la HK-integral

Una función $f:[a,b]\rightarrow IR$ es HK integrable sobre un intervalo $[a,b]$ si existe un número A tal que para cada $\varepsilon>0$, podemos encontrar una cubierta de Cousin β de $[a,b]$ con la propiedad de que: $\left|\sum_{(I,x)\epsilon\pi} f(x)l(I)-A\right|<\varepsilon$, para cada partición π contenida en β

Generalmente, se encuentra la definición de HK-integral utilizando funciones positivas $\delta:[a,b]\rightarrow(0,\infty)$, Conocidas como gauges, si $\pi=\{([u_i,v_i],t_i):i=1,2,.....n\}$ es una partición etiquetada de $[a,b]$ entonces se dice que π es δ-fina si $t_i\epsilon[u_i,v_i]\subset[t_i-\delta(t_i),t_i+\delta(t_i)]$ de esta manera el Lema de Cousin tiene un equivalente para un gauge, es decir, si $\delta:[a,b]\rightarrow(0,\infty)$ es un gauge y si $a\leq c<d\leq b$ existe una partición δ-fina de $[c,d]$. Se dice que $f:[a,b]\rightarrow IR$ es HK- integrable si existe un $A\epsilon IR$ tal que para cada $\varepsilon>0$, existe un gauge

δ con propiedad de que si π es una partición δ-fina, entonces $\left|\sum_{(I,x)\in\pi} f(x)l(I) - A\right| < \varepsilon$ se verifica, sin embargo esta definición es equivalente a aquella en la cual se utilizan integrales superiores e inferiores.

Dada por fija una cubierta de Cousin β, la suma inferior $L(f,\beta)$ y la suma superior $S(f,\beta)$ de una función f definida sobre $[a,b]$ están dadas por

$$L(f,\beta)=inf_{\pi\subset\beta}\sum_{(I,x)\in\pi} f(x)l(I)=inf_{\pi\subset\beta}S_n(f)$$

$$S(f,\beta)=sup_{\pi\subset\beta}\sum_{(I,x)\in\pi} f(x)l(I)=sup_{\pi\subset\beta}S_n(f)$$

Donde el ínfimo y el supremo se toman sobre todas las particiones en β.

Sea f una función real definida sobre $[a,b]$, se define la integral inferior $\underline{I}$ como:

$$\underline{I} = \int_{\underline{a}}^{b} f(x)dx := sup_{\beta}L(f,\beta)$$

Donde el supremo se toma sobre todas las cubiertas de Cousin de $[a,b]$, análogamente definimos la integral superior como el ínfimo de las sumas superiores

Cuando $\underline{I}=\overline{I}$, se tiene el valor común como $\int_a^b f$ si además este valor es finito, decimos que f es integrable, se observa que $L(f,\beta) \le S(f,\beta)$.

Al considerar dos cubiertas de Cousin β_1, β_2 de $[a,b]$, tales que $\beta_1 \subset \beta_2$ cualquier partición en β_1 es una partición en β_2, pero el recíproco no es necesariamente cierto, por lo cual $L(f,\beta)\le L(f,\beta_1)$,similarmente $S(f,\beta_1)\le S(f,\beta)$, por lo cual se puede concluir que:

$L(f,\beta) \le L(f,\beta_1) \le S(f,\beta_1) \le S(f,\beta)$, si $\beta_1 \subset \beta_2$, en particular si $\beta=\beta_1 \cup \beta_2$, entonces:

$L(f,\beta_1) \le L(f,\beta) \le S(f,\beta) \le S(f,\beta_2)$, es decir, cada suma inferior es menor o igual a cada suma superior.

Si β es una cubierta de Cousin arbitraria, como: $S(f,\beta)$ es cota superior para todas las sumas inferiores, y $\underline{I}$ es la mínima cota superior de las sumas inferiores, entonces $\underline{I} \le S(f,\beta)$.

Dado que β es arbitrario, entonces $\underline{I}$ es cota inferior para todas las sumas superiores y como $\overline{I}$ es la máxima cota inferior de las sumas superiores se sigue que $\underline{I} \le \overline{I}$, por lo tanto: $L(f,\beta) \le \underline{I} \le \overline{I} \le S(f,\beta)$ para toda cubierta de Cousin β.

El Segundo Teorema Fundamental del Cálculo, IITFC

Lema. Sea F:$[a,b]\to$**R** una función derivable en t$\in[a,b]$ dado $\varepsilon > 0$, existe $\delta_\varepsilon(t) > 0$ tal que si $c,d\in[a,b]$ satisfacen: $t-\delta_\varepsilon(t) \le c \le t \le d \le t+\delta_\varepsilon(t)$ entonces:

$$|F(d) - F(c) - F'(t)\,(d - c)| \le \varepsilon\,(d - c)$$

Definición.

Sea F:$[a,b]\to$**R** una función derivable en $[a,b]$ entonces, F' es integrable en el sentido de Henstock-Kurzwiel en $[a,b]$ y además

$\int_a^x F'(t)dt = F(x) - F(a)$, para cada x$\in[a,b]$.

Prueba. Para $x = b$, sea $\varepsilon > 0$ considérese el indicador asociado a $\frac{\varepsilon}{b-a}$', por el lema anterior. Sea una partición

$$\pi = \{([x_{i-1}; x_i]; t_i)\}_{i=1}^{n}$$
$$\delta_{\frac{\varepsilon}{b-a}} - fina\ de\ [a,b]$$

de ahí se probará que $|F(b)-F(a)-S(F',\pi)| \le \varepsilon$. Como para cada i$\in\{1,2,\ldots,n\}$; x_{i-1} y x_i verifican:

$$t_i - \delta_{\frac{\varepsilon}{b-a}}(t_i) \le x_{i-1} \le t_i \le t_i + \delta_{\frac{\varepsilon}{b-a}}(t_i)$$

Por la propiedad enunciada en el lema anterior se tiene que

$$|F(x_i)-F(x_{i-1})-F'(t_i)(x_i-x_{i-1})|\leq\frac{\varepsilon(x_i-x_{i-1})}{(b-a)},$$

para cada i ϵ {*1,2,...,n*},

Para estimar $F(b) - F(a) - S(F',\pi)$ se utiliza la suma telescópica

$$F(b)-F(a)=\sum_{i=1}^{n}F(x_i)-F(x_{i-1})$$

De modo que:

$$F(b)-F(a)-S(F',\pi)=\sum_{i=1}^{n}\begin{bmatrix}F(x_i)-F(x_{i-1})\\-F'(t_i)(x_i-x_{i-1})\end{bmatrix}$$

entonces:

$$\left|F(b)-F(a)-S(F',\pi)\right|\leq\sum_{i=1}^{n}\begin{bmatrix}F(x_i)-F(x_{i-1})\\-F'(t_i)(x_i-x_{i-1})\end{bmatrix}$$

se tiene:

$$|F(b)-F(a)-S(F',\pi)|\leq\sum_{i=1}^{n}\frac{\varepsilon(x_i-x_{i-1})}{(b-a)}=\varepsilon\frac{(b-a)}{(b-a)}=\varepsilon$$

En el estudio de la integral de Riemann, el limite f de una sucesión $\{f_n\}$ de funciones Riemann integrables no es necesariamente Riemann integrable incluso el limite fuese Riemann integrable, es decir, su integral podría diferir del límite de una sucesión. El intercambio de operaciones:

$$\int_I \lim_{n\to\infty} f_n=\lim_{n\to\infty}\int_I f_n$$

Es posible si la sucesión $\{f_n\}$ converge uniformemente, pero esta condición es fuerte.

Una de las principales razones por las que la integral de Lebesgue se convertido en una herramienta central e indispensable del Análisis Matemático es por sus teoremas de convergencia la integral de Lebesgue ha remediado eficientemente el problema de asegurar la igualdad enunciada requiriendo hipótesis no tan fuertes.

El mismo tratamiento para la integral de Henstock–Kurzweil también se cumple, pero su explicación es más didáctica.

El Teorema de Convergencia Uniforme

La condición que garantiza la integrabilidad de una función limite y el intercambio citado concerniente a la noción de convergencia uniforme se sigue cumpliendo para la HK-integral.

Definición: una sucesión $\{f_k\}$ de funciones reales definidas sobre un intervalo cerrado *I* converge uniformemente sobre I a una función f si para cada ε > 0 existe $N \epsilon \mathbf{N}$ tal que $|f_k(x)-f(x)|<\varepsilon$ para todo k ≥ N y $x \epsilon I$.

Teorema Convergencia Uniforme:

Si la sucesión $\{f_n\} \subset HK(I)$ converge uniformemente a *f*, entonces $f \epsilon HK(I)$ y la igualdad de intercambio se verifica.

Demostración:

Primero se probará que $\{f_I f_n\}$ es una sucesión de Cauchy, Dado ε > 0, por la definición de convergencia uniforme de la sucesion $\{f_n\}$,existe $N \epsilon$N tal que si h, k ≥ N y $x \epsilon I$ entonces $|f_k(x)-f_h(x)|<2\varepsilon$

Por la monotonía y la linealidad de la integral, tenemos que $|\int_I f_k - \int_I f_k| \leq 2 \epsilon\, l(I)$ como ε > 0 es arbitrario la sucesion $\{\int_I f_n\}$ es la Cauchy y por lo tanto converge a un número real Aϵ**R**

Se mostrará entonces que $f \epsilon HK(I)$ y que A es su integral .Sea ε > 0 y *N* sea $k \geq N$ fijo

tal que $|\int_I f_k - A|<\varepsilon$ Ahora sea β una cubierta de Cousin, de la definicion de integrabilidad de f_k, de manera que $|S_\pi(f_k)-\int_I f_k|<\varepsilon$ siempre que $\pi \subset \beta$ tenemos que por la convergencia uniforme

$$|S_\pi(f) - S_\pi(f_k)| = \left|\sum_{(I,x)\in\pi} [f(x) - f_k(x)]l(I)\right|$$
$$\leq \sum_{(I,x)\in\pi} |f(x) - f_k(x)|\, l(I)$$
$$\leq \sum_{(I,x)\in\pi} \varepsilon l(I).$$

entonces,

$$|S_\pi(f) - A| \leq |S_\pi(f) - S_\pi(f_k)| +$$
$$+ \left|S_\pi(f_k) - \int_I f_k\right| + \left|\int f_k - A\right|$$
$$< \varepsilon(l(I) + 2)$$

Para toda $\pi \subset \beta$.por lo tanto

$$f \in HK(I)\ y\ \int_I f = A = \lim_{n\to\infty} \int_I f_n$$

3.- RESULTADOS

La relación entre las funciones de distribución y la de densidad de probabilidades, se puede enunciar desde la Integral de Riemann Stieltjes bajo la condición de que el integrando sea una función continua en todo su dominio y el integrador una función de variación acotada esto es

$$\sum_{i=1}^{n} |F(x_i) - F(x_{i-1})| < M,\ M>0$$

El estudio de los teoremas de convergencia son tratados de manera más didáctica con la integral de Henstock-Kurzweil, así evitar el tratamiento riguroso de la integral de Lebesgue, como se mostró en la prueba del Segundo Teorema Fundamental del Cálculo y el Teorema de Convergencia Uniforme sin recurrir a la teoría de la medida, es mas a partir de la integral de Henstock–Kurzweil puede definirse la teoría de la medida y la integral de Lebesgue en los Reales, como lo proponen Gordon Russell en *The integral of Lebesgue*, Denjoy, *Perron and Henstock* 1944 y Robert Bartle en *A modern theory of integration* 2001,

4.- CONCLUSIONES

La relación expresada en el Teorema Fundamental del Cálculo, muestra de manera didáctica pero formal, la relación entre la función de distribución acumulada y la función de probabilidad a través de la integral de Henstock-Kurzweil, donde las reglas desarrolladas para calcular integrales son válidas , además no existen integrales impropias en el sentido de H-K , es decir, si podemos definir la integral de Henstock-Kurzweil de una función como el límite de unas integrales de la función en su intervalos del dominio, entonces la función es inmediatamente integrable en el sentido de Henstock-Kurzweil en todo el intervalo, así se puede notar que la integral de Henstock Kurzweil es estrictamente más general que las integrales de Riemann y Lebesgue .

A diferencia de la integral de Riemann y la de Lebesgue, el Teorema Fundamental del Cálculo, garantiza que la derivada de cualquier función sobre un intervalo *I*, siempre es H-K integrable.

La enseñanza de la Teoría de la Probabilidad a nivel intermedio o superiores, debería incluir un estudio del Cálculo Diferencial e integral incluyendo la Integral de Henstock-Kurzweil para poder contar con un instrumento Matemático general y riguroso para el estudio de los teoremas límites en el análisis de la convergencia y comprender mejor los teoremas límites de la teoría de la

Probabilidad, por lo cual la recomendación de muchos autores reconocidos como Robert Bartle de Estados Unidos, Ralph Henstock de Irlanda, Jaroslav Kurzweil Republica Checa Rudolf Výrborný de Australia, Eric Schechter de Estados Unidos, Stefan Schwabick de la Republica Checa abogan por la inclusión en los libros de Cálculo, la Intgral de Henstock-Kurzweil argumentando que algunas definiciones y teoremas se pueden indicar de manera más simple y más fuertes si se utiliza la Integral de Henstock –Kurzweil.

En el estudio de la teoría de la Probabilidad no solo nos permite tratar de manera unificada la definición de integral, para la comprensión de los conceptos de las características numéricas como el valor esperado, varianza, momentos, etc. Si no también como se pudo observar en los anteriores teoremas serian de gran ayuda para la comprensión de los Teoremas límites de la teoría de la probabilidad, de esta manera el estudio del Cálculo integral debería incluir la integral de Henstock-Kurzweil en el pensum para los estudiantes de una Carrera de Ciencias.

BIBLIOGRAFÍA

Robert Bartle, (2001), "*A modern Theory of integration, Graduate studies in Mathematics*".

Javier Herrera, (2005) Tesis: "*La integral de Henstock-Kurzweil*".

Adriana Ocejo Monge, (2008) "*La integral de Henstock-Kurzweil y el Teorema fundamental del Cálculo*".

Javier Martínez Perales,(2017), "*La integral de Henstock_Kurzweil y el segundo Teorema del Cálculo*".

Luis Rincón, (2007), "*Curso intermedio de Probabilidad*", Facultad de Ciencias UNAM.

ESTUDIO DE LA COINTEGRACIÓN A TRAVÉS DE MODELOS VAR

M.Sc. Flores López, Juan Carlos

✉ *caarloslopez@gmail.com*

RESUMEN

La presente investigación tiene como componente principal el desarrollo de la aplicación de los modelos VAR y la cointegración, basado en la teoría de estas metodologías.

La cointegración basado en modelos VAR en este estudio se aplica a las series de exportación de estaño y plomo mensualmente de los años 1990.01 a 2018.07 de Bolivia.

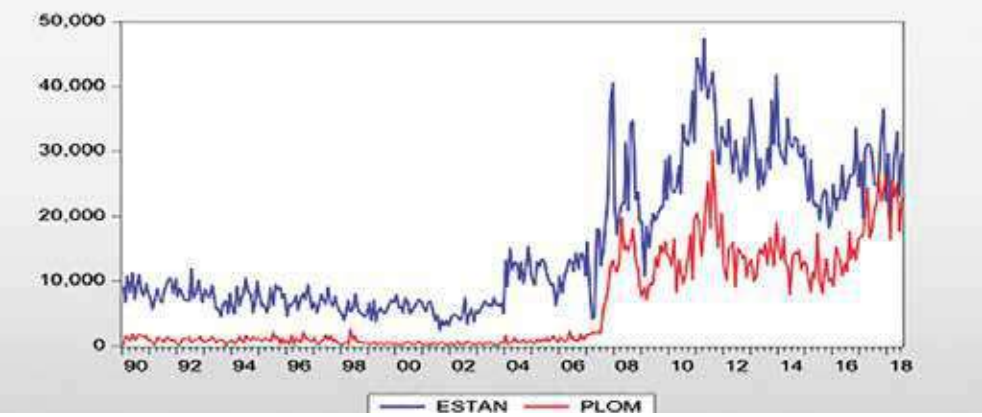
Los resultados la cointegración basados en modelos VAR, brindan un buen ajuste a la series de exportación de estaño y plomo de nuestro país, porque cumplieron con todos las condiciones teóricos que se requieren en este tipo de estudios

APLICACIÓN DE LOS MODELOS VAR Y ANÁLISIS DE INTEGRACIÓN

La exportación de minerales en miles de dólares del estaño y plomo desde 1990.01 a 2018.07 de nuestro país, es considerada en la investigación presente. La serie de datos estudiada se presentan en la Figura Nº 1

Figura Nº 1

Series de exportación en miles de dólares de Estaño y Plomo desde 1990.01 a 2018 .07



Fuente INE - Elaboración propia

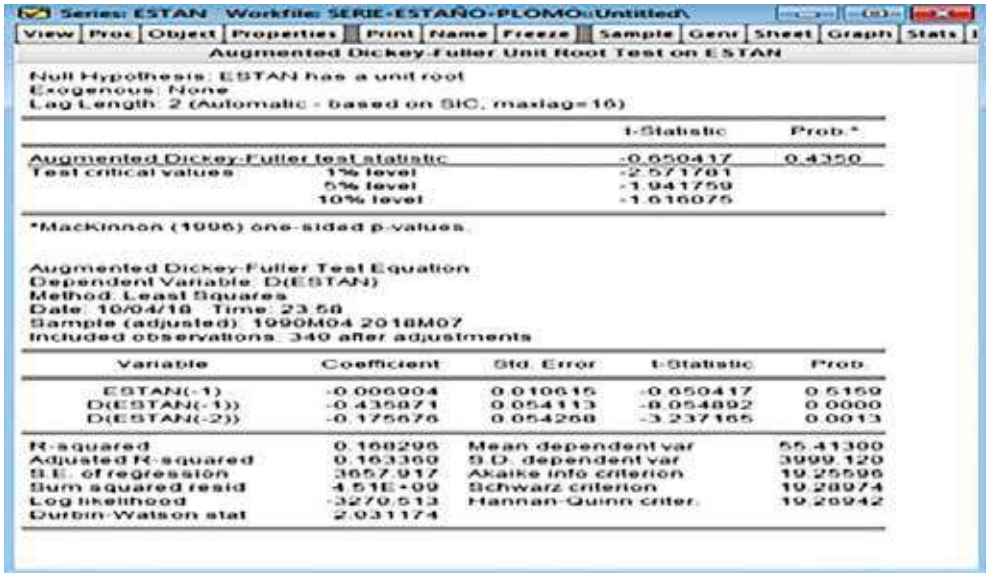
El comportamiento que presentan las series en estudio, desde 1990 al año 2006 no es de mucha fluctuación, pero a partir del año 2007 las exportaciones que tienen el estaño y el plomo es de mucha fluctuación o alta volatilidad, es así que la exportación más alta se da en el año 2011 que posteriormente se observa una baja considerable al año 2012 posteriormente se observa una subida en el año 2013 que no es el más alto y una baja notable se da en el año 2015 y desde el año 2016 al año 2018 existe una relativa estabilización pero en todos estos periodos con volatilidad.

El objetivo de esta investigación es la aplicación de los modelos VAR basados en la cointegración. Con este propósito se realiza el siguiente análisis.

En una primer instancia realizamos el test de raiz de unitaria de las series exportacion de estaño y plomo del pais, cuyos resultados son los siguientes:

Tabla Nº 1

Test de raíz unitaria de la serie exportación de Estaño



Series: ESTAN Workfile: SERIE-ESTAÑO-PLOMO::Untitled\

View | Proc | Object | Properties | Print | Name | Freeze | Sample | Genr | Sheet | Graph | Stats

Augmented Dickey-Fuller Unit Root Test on ESTAN

Null Hypothesis: ESTAN has a unit root
Exogenous: None
Lag Length: 2 (Automatic - based on SIC, maxlag=16)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-0.650417	0.4350
Test critical values:	1% level	-2.571781	
	5% level	-1.941759	
	10% level	-1.616075	

*MacKinnon (1996) one-sided p-values.

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(ESTAN)
Method: Least Squares
Date: 10/04/18 Time: 23:58
Sample (adjusted): 1990M04 2018M07
Included observations: 340 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
ESTAN(-1)	-0.006904	0.010615	-0.650417	0.5159
D(ESTAN(-1))	-0.435871	0.054113	-8.054892	0.0000
D(ESTAN(-2))	-0.175576	0.054268	-3.237165	0.0013

R-squared	0.188295	Mean dependent var	55.41300
Adjusted R-squared	0.183369	S.D. dependent var	3999.120
S.E. of regression	3657.917	Akaike info criterion	19.25596
Sum squared resid	4.51E+09	Schwarz criterion	19.28974
Log likelihood	-3270.513	Hannan-Quinn criter.	19.26942
Durbin-Watson stat	2.031174		

Fuente: Elaboración propia

De acuerdo a la Tabla Nº 1 se puede observar que la serie exportacion tiene raiz unitaria, lo que significa que es una caminata aleatoria y que no se puede tratar para analisis, por lo tanto se realiza una diferenciacion a la serie cuyo resultado se lo muestra en la Tabla Nº 2

Tabla Nº 2

Test de raíz unitaria de la serie exponencial de Estaño (diferenciado)

Series: PLOM Workfile: SERIE-ESTAÑO-PLOMO::Untitled\

View | Proc | Object | Properties | Print | Name | Freeze | Sample | Genr | Sheet | Graph | Stats | Ident

Augmented Dickey-Fuller Unit Root Test on PLOM

Null Hypothesis: PLOM has a unit root
Exogenous: None
Lag Length: 3 (Automatic - based on SIC, maxlag=16)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		0.449383	0.8108
Test critical values:	1% level	-2.571801	
	5% level	-1.941761	
	10% level	-1.616073	

*MacKinnon (1996) one-sided p-values.

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(PLOM)
Method: Least Squares
Date: 10/05/18 Time: 00:07
Sample (adjusted): 1990M05 2018M07
Included observations: 339 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
PLOM(-1)	0.005519	0.012280	0.449383	0.6534
D(PLOM(-1))	-0.572288	0.054900	-10.42426	0.0000
D(PLOM(-2))	-0.370369	0.060081	-6.164525	0.0000
D(PLOM(-3))	-0.203095	0.054853	-3.702535	0.0002

R-squared	0.253966	Mean dependent var	65.28481
Adjusted R-squared	0.247305	S.D. dependent var	2431.571
S.E. of regression	2109.581	Akaike info criterion	18.15810
Sum squared resid	1.49E+09	Schwarz criterion	18.20324
Log likelihood	-3073.797	Hannan-Quinn criter.	18.17609
Durbin-Watson stat	1.962621		

Fuente: Elaboración propia

Observando al Tabla Nº 2 con un diferenciacion se convierte en una serie estacionaria cuyo valor p-value es cero.

Tabla Nº 3

Test de raíz unitaria de la serie exportación de Plomo

Null Hypothesis: PLOM has a unit root
Exogenous: None
Lag Length: 3 (Automatic - based on SIC, maxlag=16)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		0.449383	0.8108
Test critical values:	1% level	-2.571801	
	5% level	-1.941761	
	10% level	-1.616073	

*MacKinnon (1996) one-sided p-values.

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(PLOM)
Method: Least Squares
Date: 10/05/18 Time: 00:07
Sample (adjusted): 1990M05 2018M07
Included observations: 339 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
PLOM(-1)	0.005519	0.012280	0.449383	0.6534
D(PLOM(-1))	-0.572288	0.054909	-10.42426	0.0000
D(PLOM(-2))	-0.370369	0.060881	-6.164525	0.0000
D(PLOM(-3))	-0.203095	0.054853	-3.702535	0.0002

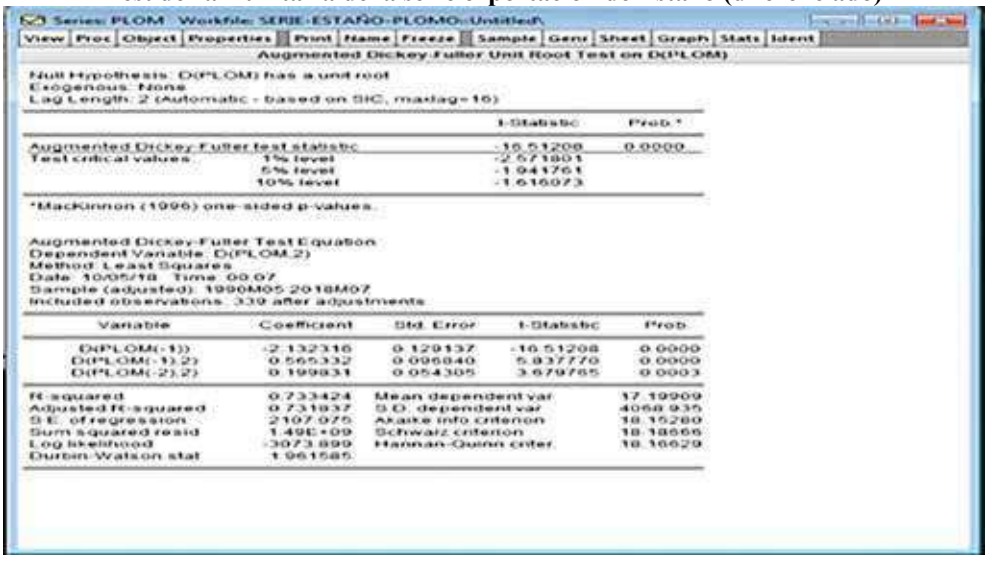
R-squared	0.253986	Mean dependent var	65.28481
Adjusted R-squared	0.247306	S.D. dependent var	2431.571
S.E. of regression	2109.581	Akaike info criterion	18.15810
Sum squared resid	1.49E+09	Schwarz criterion	18.20324
Log likelihood	-3073.797	Hannan-Quinn criter.	18.17609
Durbin-Watson stat	1.962621		

Fuente: Elaboración propia

De acuerdo a la Tabla Nº 3 se puede observar que la serie exportacion de plomo, tiene raiz unitaria, lo que significa que es una caminata aleatoria y que no se puede tratar para analisis, por lo tanto se realiza una diferenciacion a la serie cuyo resultado se lo muestra en la Tabla Nº 4. se puede ver que con una diferenciacion se convierte en una serie estacionaria.

Tabla Nº 4

Test de raíz unitaria de la serie exportación de Estaño (diferenciado)



Null Hypothesis: D(PLOM) has a unit root
Exogenous: None
Lag Length: 2 (Automatic - based on SIC, maxlag=16)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-16.51208	0.0000
Test critical values:	1% level	-2.571801	
	5% level	-1.941761	
	10% level	-1.616073	

*MacKinnon (1996) one-sided p-values.

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(PLOM,2)
Method: Least Squares
Date: 10/05/18 Time: 00:07
Sample (adjusted): 1990M05 2018M07
Included observations: 339 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
D(PLOM(-1))	-2.132316	0.129137	-16.51208	0.0000
D(PLOM(-1),2)	0.565332	0.096840	5.837770	0.0000
D(PLOM(-2),2)	0.199831	0.054305	3.679765	0.0003

R-squared	0.733424	Mean dependent var	17.19909
Adjusted R-squared	0.731937	S.D. dependent var	4058.935
S.E. of regression	2107.075	Akaike info criterion	18.15280
Sum squared resid	1.49E+09	Schwarz criterion	18.18565
Log likelihood	-3073.899	Hannan-Quinn criter.	18.16589
Durbin-Watson stat	1.961585		

Fuente: Elaboración propia

Observando al Tabla Nº 4 con un diferenciacion se convierte en una serie estacionaria.

Las series de exportacion de estaño y plomo se vuelven estacionarios a una diferenciacion. Lo que continua es determinar la regresion de la series exportacion de estaño y plomo, cuyo

resultado se muestra en la Tabla Nº 5.

Tabla Nº 5

Equation: UNTITLED Workfile: SERIE-ESTAÑO-PLOMO::Unt...

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: ESTAN
Method: Least Squares
Date: 10/05/18 Time: 00:08
Sample: 1990M01 2018M07
Included observations: 343

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	7367.403	347.3603	21.20969	0.0000
PLOM	1.310412	0.036084	36.31607	0.0000

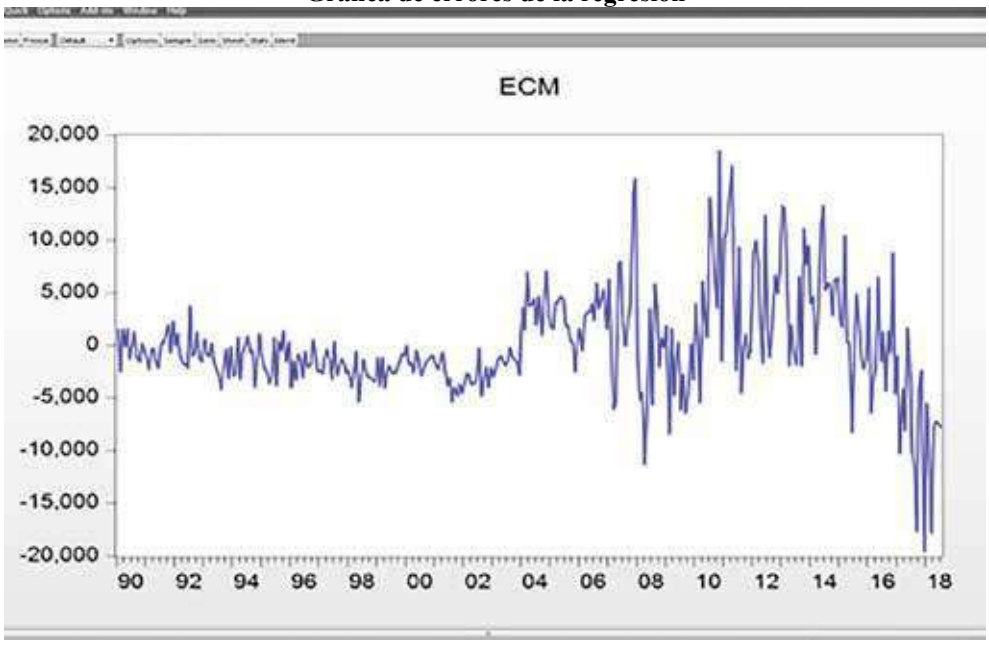
R-squared	0.794561	Mean dependent var	15480.39
Adjusted R-squared	0.793958	S.D. dependent var	10852.71
S.E. of regression	4926.244	Akaike info criterion	19.84836
Sum squared resid	8.28E+09	Schwarz criterion	19.87073
Log likelihood	-3401.993	Hannan-Quinn criter.	19.85727
F-statistic	1318.857	Durbin-Watson stat	0.929995
Prob(F-statistic)	0.000000		

Fuente: Elaboración propia

Una vez que se realiza la regresion inmediatamente se realiza el test de raiz uniataria en los errores de la regresion. Cuya grafica es la siguiente

Figura Nº 2

Gráfica de errores de la regresión



Fuente: Elaboración propia

De acuerdo a la Figura Nº 2 de errores de la regresion, aparentemente es estacionario, para verificar si e estacionario, lo que se hace es realizar la prueba de raiz unitaria de los errores cuyo resultado se presenta en la Tabla Nº 6.

Tabla Nº 6

Prueba de la raíz unitaria de los errores de la regresión

Series: ECM Workfile: SERIE-ESTAÑO-PLOMO::Untitled\

View | Proc | Object | Properties | Print | Name | Freeze | Sample | Genr | Sheet | Graph | Stats | Ident

Augmented Dickey-Fuller Unit Root Test on ECM

Null Hypothesis: ECM has a unit root
Exogenous: None
Lag Length: 2 (Automatic - based on SIC, maxlag=16)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-5.250692	0.0000
Test critical values:	1% level	-2.571781	
	5% level	-1.941759	
	10% level	-1.616075	

*MacKinnon (1996) one-sided p-values.

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(ECM)
Method: Least Squares
Date: 10/05/18 Time: 00:10
Sample (adjusted): 1990M04 2018M07
Included observations: 340 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
ECM(-1)	-0.282936	0.053886	-5.250692	0.0000
D(ECM(-1))	-0.360367	0.061671	-5.843410	0.0000
D(ECM(-2))	-0.163320	0.053940	-3.027822	0.0027

R-squared	0.300761	Mean dependent var	-27.58860
Adjusted R-squared	0.296612	S.D. dependent var	4754.333
S.E. of regression	3987.376	Akaike info criterion	19.42844
Sum squared resid	5.36E+09	Schwarz criterion	19.46222
Log likelihood	-3299.835	Hannan-Quinn criter.	19.44190
Durbin-Watson stat	2.022299		

Fuente: Elaboración propia

Al contrastar cointegracion, estamos tratando de detectar la posible existencia de relaciones de largo plazo entre las variables del modelo

Efectuar un analisis de cointegracion significa relacionar los niveles de variables como exportacion de estaño y exportacion de plomo. Es decir variables no estacionarias.

El concepto de cointegracion generaliza el concepto de correlacion en la direccion adecuada. La existencia de una tendencia estocastica comun generaria una relacion sostenible a largo plazo entre ambas variables, lo que hara que sus diferenciales reviertan a traves del tiempo, es decir que sean mean-reverting. No tiene sentido analizar relaciones entre los niveles de variables I(1) si no estan cointegradas.

Observando la Tabla Nº 6, se pude decir con certeza que los errores no tienen raiz unitaria por lo tanto cointegran los que significa series de exportacion de estaño y plomo tienen una relación sostenible de largo plazo lo que implica que se pueden realizar predicciones y analisis de funcion impulso respuesta, etc.

AJUSTE DE UN MODELO VAR

Para explicar cómo se estima un VAR, asumimos que cada ecuación contiene k valores de retardo de M (medido por las exportaciones de ESTAÑO (ESTAN)) y PLOMO (PLOM). En este caso, uno puede estimar cada una de las siguientes ecuaciones mediante OLS

$$ESTAN_{1t} = \alpha + \sum_{i=1}^{k} \beta_j ESTAN_{t-j} + \sum_{i=1}^{k} \gamma_j PLOM_{t-j} + \mu_{1t}$$

$$PLOM_t = \alpha' + \sum_{j=1}^{k} \theta_j ESTAN_{t-j} + \sum_{i=1}^{k} \gamma_j PLOM_{t-j} + \mu_{2t}$$

Son dos variables endogenas que estan en funcion de ellas mismas de sus rezagos. En este sentido estan en funcion de sus rezagos y de los rezagos de la exportacion de ESTAN Y PLOM. La variable PLOM es tambien una variable endogena que esta en funcion de ESTAN Y PLOM rezagados.

Lo que se presenta a continuacion es la estimacion del modelo VAR y los rezagos pertinentes del caso. En este caso utlizamos los informes de Akaike y Schwarz para ver el tipo de rezago que debe tener o plantearse, es decir un rezago, dos rezagos, etc. En la medida que estos son mas menores son mas robustos, por lo tanto se escoge aquel que tenga valor minimo, con esto se logra encontrar el modelo optimo.

Tabla Nº 7

Estimación de los parámetros del modelo VAR

File Edit Object View Proc Quick Options Add-ins

Command

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | I

Vector Autore

Vector Autoregression Estimates
Date: 10/05/18 Time: 10:29
Sample (adjusted): 1990M04 2018M07
Included observations: 340 after adjustments
Standard errors in ( ) & t-statistics in [ ]

	ESTAN	PLOM
ESTAN(-1)	0.479916	0.060213
	(0.05589)	(0.03311)
	[8.58617]	[1.81837]
ESTAN(-2)	0.209714	2.27E-05
	(0.06132)	(0.03633)
	[3.41997]	[0.00063]
ESTAN(-3)	0.127804	0.020210
	(0.05650)	(0.03347)
	[2.26193]	[0.60376]
PLOM(-1)	0.224218	0.430714
	(0.09285)	(0.05501)
	[2.41478]	[7.82988]
PLOM(-2)	0.060294	0.216225
	(0.09993)	(0.05920)
	[0.60335]	[3.65221]
PLOM(-3)	-0.049329	0.229716
	(0.09436)	(0.05590)
	[-0.52276]	[4.10917]
C	1452.399	-380.1718
	(429.233)	(254.293)
	[3.38371]	[-1.49501]
R-squared	0.893208	0.918919
Adj. R-squared	0.891284	0.917458
Sum sq. resids	4.29E+09	1.50E+09
S.E. equation	3587.983	2125.650
F-statistic	464.2032	628.9984
Log likelihood	-3261.920	-3083.926
Akaike AIC	19.22894	18.18192
Schwarz SC	19.30777	18.26075
Mean dependent	15538.24	6236.736
S.D. dependent	10881.88	7398.672
Determinant resid covariance (dof adj.)		5.48E+13
Determinant resid covariance		5.26E+13
Log likelihood		-6335.822
Akaike information criterion		37.35189
Schwarz criterion		37.50956
Number of coefficients		14

Fuente: Elaboración propia

En la Tabla Nº 7 se muestran los estimadores del modelo VAR y de acuerdo a la información obtenido a el modelo optimo que nos proporciona el Software. Con un Akaike de 19.22 que resulto ser el mas bajo. Se puede ver tambien que la variable estaño y plomo explican 89.32% y 91.89%

Por lo tanto el modelo estimado para el modelo VAR estaria dado por:

$$ESTAN_{1t} = 1452{,}399 + 0{,}479916\ ESTAN_{t-j} + 0{,}209714 ESTAN_{t-j} + 0{,}127804 ESTAN_{t-j} + 0{,}224218 PLOM_{t-j} + 0{,}020694 PLOM_{t-j} - 0{,}049329 PLOM_{t-j} + \mu_t$$

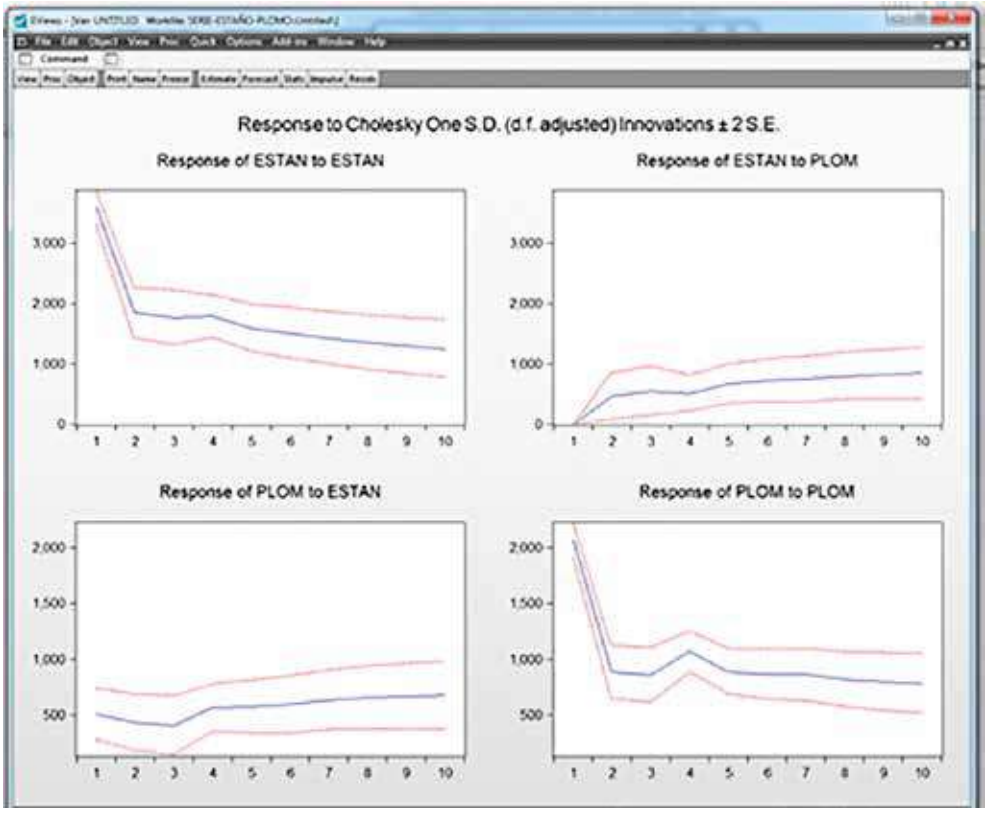
$$PLOM_t = -380{,}1718 + 0{,}020613 ESTAN_{t-j} + 2{,}27E-05 ESTAN_{t-j} + 0{,}020210 ESTAN_{t-j} + 0{,}430714 PLOM_{t-j} + 0{,}216225 PLOM_{t-j} + 0{,}229716 + \mu_t$$

FUNCIÓN IMPULSO RESPUESTA

Se hará ahora un análisis de la función impulso respuesta para los datos asociados exportación de estaño, la exportación de plomo. Los resultados se muestran en el Figura Nº 3.

Figura Nº 9

Gráfica de los resultados a un shock



Fuente: Elaboración propia

EFECTOS DE UN SHOCK DE UNA DESVIACIÓN ESTÁNDAR SOBRE LA EXPORTACIÓN DE ESTAÑO

En el Figura Nº 3 se observa que un shock de una desviación estándar sobre la perturbación asociada a la exportación de estaño produce una baja en los primero periodos para luego tratar de estabilizarse. El efecto a una desviación estándar para el plomo provoca un leve descenso para la exportación de plomo que lo mismo ocurre para el estaño

Una funcion respuesta a un impulso muestra el efecto de un cambio en los errores (innovacion)

sobre las variables endogenas del sistema. Un cambio en e1 modificara automaticamente el valor de la variable exportacion de estaño, pero no solo no solo se alterara el valor de esta varaible sino tambien en el valor de la variable exportacion de plomo, debido a la estructura dinamica del sistema.

La ordenacion en esta caso es la exportacion de estaño, exportacion de plomo, puesto como como sabemos la ordenacion no es nuetra. Una respuesta a un impulso separa los determinantes de las variables endogenas en cambios o innovaciones identificadas con variables especificas.

Tabla Nº 8

Resultados de función impulso respuesta

EViews - [Var: UNTITLED Workfile: SERIE-ESTAÑO-PLOMO::Untitled\]

File Edit Object View Proc Quick Options Add-ins Window Help

Command

View Proc Object Print Name Freeze Estimate Forecast Stats Impulse Resids

Impulse Response to Cholesky One S.D. (d.f. adjusted) Innovations

Response of ESTAN

Period	ESTAN	PLOM
1	3591.421	0.000000
2	1813.735	469.7658
3	1693.801	550.5774
4	1600.608	493.5827
5	1560.105	491.7159
6	1455.557	618.0024
7	1398.823	658.2168
8	1332.420	674.6618
9	1272.387	698.5683
10	1231.749	732.6060

Response of PLOM

Period	ESTAN	PLOM
1	507.3537	2025.821
2	418.5579	782.6517
3	321.7029	698.8565
4	291.7468	739.1969
5	538.4141	905.7331
6	514.9006	753.7264
7	518.8471	742.7266
8	541.2501	738.9320
9	587.9229	740.6399
10	599.5865	718.5942

Cholesky Ordering: ESTAN PLOM

Fuente: Elaboración propia

Las figuras adjuntas se presenta el resultado de un impulso de una vez que el error estandar de la ecuacion estimada. Vemos que en la variable exportacion de estaño, tras el impacto incial vemos una disminucion en los priemro dos periodos en forma brusca y luego en foram mas leve.

El efecto sobre la exportacion de plomo muestra un leve descenso para luego estabilizarse. Un impulso sobre exportacion plomo tiene un efecto positivo en los dos periodos y una leve estabilizacionen lo que sigue

La respuesta exportacion plomo sobre exportacion plomo se muestra una baja en los dos primeros periodos para luego tratarar de estabilizarse.

CONCLUSIONES

En el presente estudio de la cointegración a través de los modelos VAR podemos concluir de la siguiente manera:

Para estudiar varias series, es necesario tomar en cuenta la interdependencia entre ellas. Una forma de hacerlo es estimar un modelo de ecuaciones simultáneas, pero con rezagos en todas las variables.

Los Vectores Autorregresivos proporcionan una muy buena técnica para hacer pronósticos en sistemas de variables de series de tiempo interrelacionadas, donde cada variable ayuda a pronosticar a las demás variables.

Un modelo VAR es un sistema de variables que hace de cada variable endógena una función de su propio pasado y del pasado de otras variables endógenas del sistema. El estudio de las interacciones dinámicas estimadas es una de las motivaciones fundamentales delos usuarios de los modelos VAR, y, de hecho, los usos típicos de estos modelos reflejan esta motivación,

tales usos son el procesamiento de datos de las funciones impulso-respuesta y de la descomposición de la varianza del error de predicción. Las implicaciones dinámicas del modelo estimado dependerán evidentemente de la estructura de correlaciones contemporáneas reflejada en la matriz de perturbaciones.

Explicar cómo realizar esta incorporación, el cómputo de las estimaciones del VAR, de la función impulso-respuesta y de la descomposición de la varianza del error de predicción, es una parte del estudio presente.

La estimación del modelo VAR es más sencillo, ya que es posible utilizar el método de los Mínimos Cuadrados Ordinarios (MCO).

Por la experimentación que se realizó, en el presente estudio podemos decir que el modelo VAR es muy útil cuando existe evidencia de simultaneidad entre un grupo de variables, es decir series de tiempo y que sus relaciones se transmiten a lo largo de un determinado número de períodos.

Se pudo evidenciar que no todas las series cointegran, la experiencia adquirida en este estudio nos indica que se deben realizar las pruebas de raíz unitaria y la prueba de estacionariedad de la regresión de los errores y que estas deberían ser estacionarias y que caso contrario no serviría para el análisis posterior de este tipo de estudios

Los objetivos de la investigación fueron cumplidos dado que el objetivo principal de la presente investigación es desarrollar la parte teórica y la aplicación de la integración para los modelos los modelos VAR, considerando las series exportación de estaño y plomo de nuestro país.

La hipótesis de estudio planteada en la investigación se cumplió dado que la metodología de la cointegración aporta al desarrollo de los modelos vectores autoregresivos considerando las series de exportación de estaño y plomo desde 1990.01 a 2018.07.

BIBLIOGRAFÍA

Greene,W.H, "*Análisis Econométrico*". Tercera Edición. Editorial Prentice Hall. 1999.

Gujarati, D, "*Econometría*". McGraw Hill. Bogotá 1999.

Madala, G.S, "*Introducción a la Econometría*". Tercera Edición. Editorial Prentice Hall. México.

Johnston, J y Dinardo, J , "*Econometric Methods*". Cuarta Edición. Mc Graw Hill,1997.

Enders, Walter, "*Applied Econometric Time Series*". Primera Edición. Ed. John Wiley & Sons.Inc.1995

Box, G and G, Jenkins, "*Time Series Analysis*". Forecasting and control. Segunda edición. San Francisco. Holden Day. 1984.

MODELOS DE ELECCIÓN DISCRETA APLICADOS A DATOS SIMULADOS COMO APROXIMACIÓN A UN MODELO DE TRANSPORTE PARA LA CIUDAD DE LA PAZ

Lic. Paredes Alarcón, Marisol[1]

✉ *marycorreo7@yahoo.es*

RESUMEN

Este documento desarrolla modelos estadísticos de elección discreta, mediante la comparación de modelos multinomial logit y probit, a fin de diseñar un modelo para el modo de transporte elegido por los habitantes de la ciudad de La Paz, con el auxilio de datos simulados basados en fuentes provenientes de encuestas a hogares sobre los viajes. Al ser las encuestas de hogares una de las formas convencionales de recolección de información pero bastante costosas, se propicia el uso de datos simulados. Los resultados alcanzados señalan que los costos, el tiempo de espera del tramo y el tiempo de duración del tramo, son determinantes en la elección del modo de transporte para los habitantes de la ciudad de La Paz, a diferencia de las variables sociodemográficas. El modelo multinomial logit alcanzado enfatiza modos de transporte más económicos como los elegidos por la población, como el ir a pie, y también aquellos que a pesar del tiempo de espera para su abordaje, proporcionan una forma más eficiente, segura y rápida de transporte como son las nuevas modalidades de Pumakatari y Teleférico. En tanto que en el lado opuesto y con la menor probabilidad de ser elegidos se tiene al minibús y carry, seguido del micro, bus, microbús, que incitan a un mejoramiento de su capacidad técnica y propuesta económica para ofrecer un mejor servicio a la ciudadanía.

PALABRAS CLAVE

Modelos de transporte; modelo logit; modelo probit, simulación

ABSTRACT

This document develops statistical models of discrete choice, by comparing logit and probit multinomial models, in order to design a model for the mode of transport chosen by the inhabitants of the city of La Paz, with the help of related simulated data in sources from household surveys on travel. Since household surveys are one of the ways of collecting information but quite expensive, it is encouraged to use simulated data. The results obtained indicate that the costs, the waiting time of the section and the duration of the section, are decisive in the choice of the mode of transport for the inhabitants of the city of La Paz, a difference of the sociodemographic variables. The improved multinomial logistics model emphasizes more economical modes of transport such as those chosen by the population, such as going to a cake, and also those that weigh the waiting time for boarding, specifically a more efficient, safe and faster way of transport such as they are the new modalities of Pumakatari and Cable Car. While on the opposite side and with the least probability of being chosen, you have the minibus and take, followed by the micro, bus, minibus, which encourage an improvement of your technical capacity and economic proposal to offer a better service to the citizens

KEYWORDS

Transport models; logit model; probit model, simulation

1. INTRODUCCIÓN

La planificación del transporte en las grandes ciudades resulta en la actualidad un ámbito de investigación sumamente importante, particularmente en la toma de decisiones y la elaboración de políticas de transporte urbano. Una forma de abordar este problema se realiza por medio de modelos de transporte, que requieren de diversas técnicas estadísticas para su aproximación.

El transporte urbano es uno de los factores fundamentales para el desarrollo económico y social de la ciudad de La Paz y el área metropolitana, ya que más del 70% de la población depende de este medio de movilización para realizar sus actividades diarias de trabajo, educación, compras, etc. Por ello es que el interés en esta investigación se centra en el estudio de un modelo de transporte para la ciudad de La Paz.

Sin embargo, estos modelos pueden ser desarrollados tanto desde el punto de vista de la oferta como de la demanda de transporte. En el segundo caso, los modelos de demanda de transporte utilizados en la planificación se basan exclusivamente en formas convencionales de recolección de información, como son las encuestas por muestreo, efectuadas a los hogares respecto de los viajes realizados. Toda esta información estadística bastante costosa, persigue, como fin último, el diseño de un modelo de transporte para la ciudad bajo estudio.

La teoría desarrollada sobre los modelos de transporte surge en 1960 (de Dios Ortuzar y Willumsen, 2011), originalmente especificados como modelos basados en viajes, supone la realización de cuatro pasos. Este modelo trabaja sobre la hipótesis de que los usuarios realizan secuencialmente un conjunto de elecciones que caracterizan sus viajes, a base de ciertos atributos personales y del sistema de transporte. Estas elecciones dicen la relación con las decisiones de viajar (paso 1, generación de viajes) hasta un destino (paso 2, distribución de viajes) en un modo de transporte (paso 3, partición modal) y a través de una ruta determinada (paso 4, asignación). La agregación de estas decisiones individuales determina las características de operación de un sistema de transporte dado.

El requisito para el modelado de estos cuatro pasos prevé la realización de encuestas por muestreo efectuadas a los hogares, cuya ejecución supone un elevado costo. Sin embargo, en los últimos años, se han realizado investigaciones a fin de propiciar el uso de datos simulados basados en encuestas por muestreo de viajes efectuados a hogares, a fin de aprovechar otro tipo de aproximaciones para calcular modelos de transporte que permitan trabajar con escenarios hipotéticos.

Para realizar esta aproximación, se utiliza los resultados de la Encuesta municipal de movilidad intraurbana en la Región Metropolitana de La Paz publicado por el Gobierno Autónomo Municipal de La Paz (GAMLP), que se llevó a cabo de septiembre a octubre del 2014 en los macrodistritos urbanos de La Paz y El Alto y en los municipios de Palca, Mecapaca, Achocalla, Viacha, Pucarani y Laja, con un total de muestra de 1.820 hogares. Estos resultados proporcionan descripciones sociodemográficas de los hogares, que cubren atributos tales como composición del hogar, nivel socioeconómico, ocupación principal, tiempos por persona, tiempos por tramo, modo de transporte, gasto en transporte, entre otros. Esta información se utiliza como datos de entrada necesarios para aproximar los datos simulados, combinados con alguna información complementaria acerca del modo de transporte, los costos de

pasajes actualizados y la introducción del nuevo medio de transporte por teleférico.

La forma de abordar el tratamiento de la información simulada se puede realizar con modelos estadísticos de elección discreta, en la que las opciones de elección son múltiples (por ejemplo ir a pie, o tomar un micro o minibús). En este tipo de modelos de elección discreta se trabaja con los modelos Multinomial Logit (MNL), pero otra aproximación interesante es la de los modelos Multinomial Probit (MNP).

Los resultados sugieren que la construcción de modelos de elección discreta utilizando la simulación basada en encuestas por muestreo es efectiva para analizar el comportamiento de los pasajeros por medio de un modelo de transporte y puede usarse para estudiar el impacto de las políticas de gestión de la demanda en la ciudad de La Paz.

2. OBJETIVO

El objetivo general de esta investigación es desarrollar modelos estadísticos de elección discreta, mediante el Modelo Multinomial Logit (MNL), Modelo Multinomial Probit (MNP), a fin de diseñar un modelo de transporte para la ciudad de La Paz, con el auxilio de datos simulados basados en fuentes provenientes de encuestas a hogares sobre los viajes.

La pregunta de investigación intenta aproximar los patrones y características de la movilidad intraurbana en el municipio de La Paz desde el punto de vista de la demanda.

3. METODOLOGÍA

El caso del transporte en la ciudad de La Paz, se pudo analizar desde la perspectiva de trabajo con modelos de elección discreta, con un conjunto de opciones de más de dos alternativas. La elección de un individuo sobre el modo de transporte en nuestra ciudad (por ejemplo micro, taxi, minibús) es un problema de elección discreta múltiple. El enfoque que se usa para modelar elecciones discretas se basa en la teoría de la utilidad aleatoria, donde para cada una de las alternativas existentes, el individuo tiene una función de utilidad asociada, donde elige aquella alternativa que maximiza su utilidad. La función se divide en un componente determinístico y un componente aleatorio.

En este tipo de modelos de elección discreta se trabaja con los modelos Multinomial Logit (MNL), que requieren como supuesto la Independencia de Alternativas Irrelevantes (IIA), esto es independencia entre alternativas, que corresponde a establecer que los términos de error relativos a las utilidades no están correlacionadas entre sí. Debido a que este supuesto no se verifica en la mayoría de los casos, otra aproximación interesante es la de los modelos Multinomial Probit (MNP), donde la suposición es la de asumir que los errores están distribuidos de forma normal multivariante, esto puede permitir la correlación entre alternativas y las varianzas distintas entre las alternativas. Para éste último caso el inconveniente mayor es el de evaluar integrales múltiples, pero en la actualidad con la ayuda de software apropiado se pueden obtener los valores de las integrales por medio de simulaciones.

Sin embargo, analizar el comportamiento de los modelos de elección discreta para el caso del transporte en la ciudad de La Paz, requirió además comprender los determinantes más importantes y las relaciones que explican las decisiones para realizar los viajes. Por ello es que la Encuesta Municipal de movilidad intraurbana en la región metropolitana de La Paz realizada por el GAMLP recogió información de otros factores tales como variables sociodemográficas, el número de

viajes, tiempos, tramos, gastos de transporte entre otros.

3.1. PREFERENCIAS DECLARADAS Y REVELADAS

En la teoría económica desarrollada sobre el transporte, desde el punto de vista de la demanda, se la enfoca tradicionalmente en el empleo de las preferencias reveladas y las preferencias declaradas, o bien una combinación de ambas.

Datos con preferencias reveladas, significa que los datos recolectados son elecciones observadas de individuos, como en nuestro caso del modo de transporte (micro, minibús, taxi, etc.). En esta situación el individuo ya optó por un modo de transporte, y la encuesta recoge información sobre las características relacionadas a este modo de transporte que ha sido utilizado efectivamente. Datos con preferencias declaradas, significa que los individuos son enfrentados a una situación de elección, por ejemplo, la elección de un nuevo tramo de la ruta del teleférico, con nuevos precios y características en su recorrido. En resumen, se habla de preferencias reveladas cuando se observa el comportamiento real de los usuarios, por ejemplo, el medio de transporte utilizado, y de preferencias declaradas o establecidas cuando se obtienen respuestas de los individuos ante situaciones de elección hipotéticas.

Por ello, para la construcción misma del modelo tomaremos el comportamiento de las preferencias declaradas obtenidas en la encuesta, a fin de simular las respuestas para la estimación del modelo de transporte, en la que las principales variables explicativas utilizadas habitualmente son el costo del transporte, el tiempo de espera, el tiempo de duración del viaje y el nivel socioeconómico del usuario.

3.2. MODELO MULTINOMIAL LOGIT

El modelo de elección discreta más sencillo y mayormente utilizado es el logit. Su popularidad se debe al hecho de que la fórmula para las probabilidades de elección toma una forma cerrada y se interpreta más fácilmente.

De forma general el modelo logit considera que un individuo se enfrenta a J alternativas. Se define una utilidad U para cada alternativa y se supone que el individuo elige la alternativa con el más alto nivel de utilidad. La utilidad U que obtiene el tomador de decisiones de la alternativa j se descompone en:

$$\begin{cases} U_1 = \beta_1^T x_1 + \varepsilon_1 = V_1 + \varepsilon_1 \\ U_2 = \beta_2^T x_2 + \varepsilon_2 = V_2 + \varepsilon_2 \\ \vdots \\ U_j = \beta_j^T x_j + \varepsilon_j = V_j + \varepsilon_j \end{cases}$$

La alternativa k será elegida si y solo si para cualquier $j \neq k$, $U_k > U_j$, lo que nos guía a las siguientes k–1 condiciones:

$$\begin{cases} U_k - U_1 = (V_k - V_1) + (\varepsilon_k - \varepsilon_1) > 0 \\ U_k - U_2 = (V_k - V_2) + (\varepsilon_k - \varepsilon_2) > 0 \\ \vdots \\ U_k - U_j = (V_k - V_j) + (\varepsilon_k - \varepsilon_j) > 0 \end{cases}$$

Como los errores ε_j no son observados, las j–1 condiciones pueden ser reescritas en términos de los límites superiores de los j–1 errores restantes:

$$\begin{cases} \varepsilon_1 < (V_k - V_1) + \varepsilon_k \\ \varepsilon_2 < (V_k - V_2) + \varepsilon_k \\ \vdots \\ \varepsilon_j < (V_k - V_j) + \varepsilon_k \end{cases}$$

La expresión general para la probabilidad de elegir la alternativa k es:

$$(P_k \setminus \varepsilon_k) = P(U_k > U_1, \dots, U_k > U_j)$$

Note que esta probabilidad es condicional sobre el valor de ε_k. Se debe aplicar una integral para obtener el valor de la probabilidad incondicional que depende solamente de β y sobre el valor de las variables explicativas.

Para el modelo multinomial logit aplicado a nuestro problema de transporte, se considera que tenemos i individuos que realizan su elección de transporte, y se enfrenta a j alternativas. La utilidad U que obtiene el tomador de decisiones de la alternativa j se descompone en:

$$U_{ij} = V_{ij} + \varepsilon_{ij} \; ; \forall j$$

Que consiste en una parte denominada V_{ij} que es un conjunto de variables (por ejemplo costo del transporte, tiempo de espera, etc.) conocidas, y una parte desconocida ε_{ij} que es tratada como aleatoria. El modelo multinomial logit se obtiene suponiendo que cada ε_{ij} se distribuye de forma independiente e idénticamente distribuidas.

Para ejemplificarlo, supongamos que en una situación sobre un modo de transporte elegido por un individuo entre micro, minibús y taxi, podemos asociar la utilidad como un índice de satisfacción V_j que depende de forma lineal de un costo (x) y el tiempo (y):

$$\begin{cases} V_1 = \alpha_1 + \beta x_1 + \gamma y_1 \\ V_2 = \alpha_2 + \beta x_2 + \gamma y_2 \\ V_3 = \alpha_3 + \beta x_3 + \gamma y_3 \end{cases}$$

En este caso, la probabilidad de la elección de la alternativa j aumenta con V_j. Para fines de la estimación, se debe transformar el índice de satisfacción, ya que puede tomar cualquier valor real, de tal forma que sea restringido al intervalo unitario y se pueda interpretar como una probabilidad. El modelo multinomial logit se obtiene aplicando esta transformación a los V_j. MacFadden (1974) ha demostrado que las correspondientes probabilidades están dadas por:

$$\begin{cases} P_1 = \dfrac{e^{V_1}}{e^{V_1} + e^{V_2} + e^{V_3}} \\ P_2 = \dfrac{e^{V_2}}{e^{V_1} + e^{V_2} + e^{V_3}} \\ P_3 = \dfrac{e^{V_3}}{e^{V_1} + e^{V_2} + e^{V_3}} \end{cases}$$

Donde cada una de las P_j con j=1, 2, 3 están entre cero y uno y la suma de ellas es igual a la unidad. Esta es la forma general de la función de distribución logística, y expresa la probabilidad de que un individuo escoja la alternativa j.

Esta elección discreta que realiza un individuo entre los tres medios de transporte, utiliza además el principio de maximización de la utilidad, suponiendo que el individuo, al tomar la decisión sobre un determinado medio de transporte dentro de un conjunto de alternativas disponibles, está eligiendo también alcanzar su máximo nivel de utilidad.

3.3. MODELO MULTINOMIAL PROBIT

La ventaja del MNP sobre MNL es que MNP no asume Independencia de Alternativas Irrelevantes (IIA), esto es independencia entre alternativas, que corresponde a establecer que los términos de error relativos a las utilidades no están correlacionadas entre sí. El modelo multinomial probit (MNP) se obtiene con el mismo modelo que utilizamos cuando se presentó el modelo de utilidad aleatoria. La utilidad de una alternativa aún es la suma de los dos componentes:

$$U_{ij} = V_{ij} + \varepsilon_{ij} \; ; \forall j$$

Pero la distribución conjunta de los términos de error ahora es una normal multivariante con media 0 y con una matriz de covarianza denotada con Ω.

La elección de probabilidades utilizando MNP es muy compleja. En la mayoría de las situaciones, las distribuciones normales proporcionan una representación adecuada

de los componentes aleatorios. Sin embargo, en otras ocasiones, las distribuciones son inapropiadas y pueden llevar a pronósticos incongruentes.

Debido a que se trabaja con una distribución normal multivariante, una función muy difícil de integrar, las computadoras presentan dificultades en el tiempo de cálculo o en la estimación de múltiples integrales, por ésta razón deben evaluarse numéricamente a través de simulaciones. Se utilizan varios procedimientos de simulación y pueden ser efectivos en ciertas circunstancias. Los métodos de cuadratura aproximan la integral por una función ponderada de puntos elegidos para evaluar. La cuadratura opera efectivamente cuando la dimensión de la integral es pequeña, pero no con dimensiones más altas. Se han propuesto numerosos simuladores para modelos probit; el principal simulador es el simulador de GHK, por las iniciales de los autores Geweke, Hajivassiliou y Keane que es, con mucho, el simulador de probit más utilizado y el más preciso.

3.4 ALGUNAS CARACTERÍSTICAS DEL PAQUETE MLOGIT DE R

El software R presenta entre sus versatilidades la amplia variedad de librerías, una de ellas es el paquete mlogit que permite la estimación de modelos logit y probit multinomiales con variables específicas individuales y/o alternativas.

En el modelo propuesto se trabajó con variables individuales, es decir propias de cada persona que elige un modo de transporte, como ser el nivel socioeconómico. En cambio hay otro tipo de variables denominadas alternativas que se relacionan directamente a la elección del modo de transporte de cada individuo, como ser el tiempo de espera, el costo del modo de transporte, y el tiempo del recorrido o viaje.

Cuando se trabaja con modelos logit multinomiales, uno tiene que considerar tres tipos de variables:

- Variables específicas alternativas x_{ij} con un coeficiente genérico β.
- Variables específicas individuales z_i con un coeficiente específico alternativo γ_j.
- Variables específicas alternativas w_{ij} con un coeficiente específico alternativo δ_j.

El índice de satisfacción para la alternativa j es entonces:

$$V_{ij} = \alpha_j + \beta x_{ij} + \gamma_j z_i + \delta_j w_{ij}$$

Un modelo solamente con variables específicas individuales a veces se denomina modelo logit multinomial, uno solo con variables específicas alternativas un modelo logit condicional y uno con ambos tipos de variables un modelo logit mixto. Sin embargo esto puede resultar engañoso: el modelo logit condicional también es un modelo logit para datos longitudinales en la literatura estadística y el logit mixto es uno de los nombres de un modelo logit con parámetros aleatorios. Por lo tanto, en lo que sigue, se utiliza el nombre logit multinomial para el modelo que se acaba de describir cualquiera que sea la naturaleza de las variables explicativas incluidas en el modelo.

Los coeficientes del modelo multinomial logit son estimados mediante el método de máxima verosimilitud. Bajo ciertas condiciones de regularidad, el estimador máximo verosímil es consistente y tiene una distribución normal asintótica. Actualmente se usan dos tipos de rutinas para la estimación de máxima verosimilitud. A la primera puede denominares métodos “similares a Newton”. En este caso, en cada iteración, se calcula una estimación del hessiano (matriz cuadrada de $n \times n$, de las segundas derivadas parciales),

ya sea utilizando las segundas derivadas de la función (método de Newton-Ralphson) o utilizando el producto externo del gradiente. Este enfoque es muy poderoso si la función se comporta bien, pero puede funcionar mal de otra manera y fallar después de algunas iteraciones. El segundo, actualiza en cada iteración la estimación del hessiano. A menudo es más robusto y puede tener un buen desempeño en los casos en que el primer caso no funciona.

4. RESULTADOS

Características generales del modo de transporte en La Paz

Las características generales del transporte en el municipio de La Paz para los datos simulados se determinaron principalmente con la información de la Encuesta Municipal de movilidad intraurbana. En La Paz se concentra una gran actividad económica y social, y por ello abarca no sólo al municipio en sí mismo, sino además al municipio aledaño de El Alto, por tanto el estudio del modelo de transporte también involucra este hecho.

Para los fines de este estudio, se tomó en cuenta la información proveniente del GAMLP, en la cual para la ciudad de La Paz se cuenta con 540 rutas del servicio público de transporte colectivo de pasajeros, según el siguiente detalle:

Tabla Nº 1
Rutas de servicio público por tipo de transporte. 2016

Tipo de Vehículo	Total
Bus	26
Carry	90
Micro	51
Minibus	280
Trufi	93
Total	**540**

Fuente: GAMLP

Del total de 540 rutas, se advierte la preponderancia que tienen los minibuses y carrys, que representan alrededor del 68,5% del total de rutas para el municipio de La Paz. La información simulada intenta reflejar esta estructura pero adiciona otras formas de transporte como ir a pie o elegir el teleférico.

Para fines prácticos en la elaboración del modelo de transporte, se reagruparon algunas categorías del modo de transporte de nuestra ciudad, y se trabajó con sólo 6 categorías, que son: a) A pie, b) vehículo propio, c) trufi y taxi, d) minibús y carry, e) micro junto con bus y microbús, y finalmente el f)Pumakatari agrupando además en esta categoría a la nueva modalidad de transporte por teleférico. Es bueno aclarar que no se consideró el medio de transporte del radiotaxi por tener una mínima frecuencia en los resultados de la encuesta de movilidad intraurbana del municipio de La Paz, y porque su costo de pasaje tan variable representaba un sesgo a la información simulada.

Tabla Nº 2
Porcentaje de viajes por modo de transporte en la ciudad de La Paz

Modo de Transporte	Porcentaje (%)
a) A pie	13,03%
b) Vehículo Propio	4,64%
c) Trufi y taxi	8,75%
d) Minibus, Carry	63,4%
e) Micro, Bus, Microbus	8,04%
f) Pumakatari, Teleférico	2,14%
Total	**100,0%**

Fuente: Elaboración propia en base a los datos simulados desde la encuesta de Movilidad Intraurbana del GAMLP

La cantidad de viajes por modo de transporte que se generó para el área metropolitana de La Paz y El Alto, se resume en la Tabla Nº 2. El mayor porcentaje lo representa el minibús con el 63,4%. También se puede observar que el 73,58% de los viajes se realizan en minibús, micro o Pumakatari, el 13,03% a

pie, el 4,64% en vehículo propio y solamente el 8,75% en trufi o taxi.

Otros datos relevantes de la encuesta de Movilidad Intraurbana señalan que se tienen cerca de 900 mil personas que utiliza el sistema de transporte desde y hacia la ciudad de La Paz. En general, el tiempo promedio que la población destina en cada tramo, es alrededor de 30 minutos. La información simulada intenta utilizar los valores estimados presentados por la encuesta.

El Modelo Multinomial Logit

La encuesta de Movilidad Intraurbana realizada por el GAMLP arroja un gran conjunto de variables, de las cuales se hizo una selección. De acuerdo a los trabajos realizados en varios países sobre un modelo de transporte, se han establecido como prioritarias las variables costo del tramo (Tr1cos), tiempo de espera del tramo (Tr1Esp) en minutos, la duración del tramo (Tr1Dur) en minutos. Se consideraron adicionalmente algunas variables sociodemográficas, como el nivel socioeconómico, y resultó en algún modelo relevante considerando la categoría nivel socioeconómico bajo, esto significa que este factor influye a la hora de elegir un medio de transporte como se esperaría. No obstante lo anterior, resulta mejor explicada por el costo del tramo que es la variable elegida para el modelo final.

Se tiene la elección del modo de transporte tomando en cuenta la duración del viaje, el costo del modo de transporte elegido y el tiempo de espera para tomar el medio de transporte. Las salidas utilizando el paquete mlogit del R son:

Coeficientes de la estimación:

	Estimación	Error estándar	valor z	Pr(>|z|)	
Tr1Dur	-0,0998722	0,0094165	-10,6061	< 2,2e-16	***
Tr1Cos	-0,1111034	0,0156372	-7,1051	1,203e-12	***
Tr1Esp	0,0657895	0,0053321	12,3383	< 2,2e-16	***

Significancia: 0 '*' 0,001 '**' 0,01 '*' 0,05**

Log-verosimilitud: -761,7

Todos los coeficientes son altamente significativos y tienen el signo que se esperaría. También las desviaciones estándar de los parámetros estimados son pequeñas, que significa que están bien identificadas.

Interpretando el modelo, en el caso del costo del viaje (Tr1Cos) y del tiempo que dura el viaje (Tr1Dur), ambas con signo negativo, sucede que un aumento en la probabilidad de la elección del medio de transporte disminuye con un costo más alto o con un tiempo de duración del viaje más largo. No pasa lo mismo con el tiempo de espera del modo de transporte.

A diferencia de los signos, los coeficientes no se pueden interpretar directamente, pero al dividirlos por el coeficiente de precios, obtenemos valores monetarios. Por ejemplo dividiendo el coeficiente del tiempo de duración del viaje (en minutos) sobre el coeficiente del costo del viaje (en Bs.) nos da un incremento de 0,8989 Bs. que la persona estaría dispuesta a pagar por disminuir la duración del viaje.

Las probabilidades de cada una de las alternativas del modo de transporte elegidas son:

Tabla Nº 3
Probabilidades de cada una de las alternativas

Modo de Transporte	Porcentaje (%)
a) A pie	0,29614664
b) Vehículo Propio	0,18880248
c) Trufi y taxi	0,10588330
d) Minibus, Carry	0,02315372
e) Micro, Bus, Microbus	0,10588330
f) Pumakatari, Teleférico	0,12052948

Fuente: Elaboración propia en base a los datos simulados desde la encuesta de Movilidad Intraurbana del GAMLP

Es decir de acuerdo al modelo presentado el medio de transporte con mayor probabilidad de ser elegido tomando en cuenta el costo, tiempo de espera y duracion del viaje es la opción a pie, le sigue el disponer de un vehículo propio, y después el Pumakatari y Teleférico. Son las tres primeras opciones que prefieren las personas, y en el otro extremo con la menor probabilidad de ser elegido está el minibús y carry, seguido del micro, bus, microbus. Por tanto, en la ciudad de La Paz las personas prefieren, si tuvieran la opción de elegir de acuerdo a las características del costo del transporte, del tiempo de espera para este medio y de la duración del viaje, el ir a pie, y en el lado opuesto la opción menos placentera es el minibús.

El Modelo Multinomial Probit

Para aplicar el modelo multinomial Probit, solamente se debe modificar del comando en R el argumento probit=TRUE. Para este modelo se restringió las categorías del modo de transporte a minibús, micro, vehículo propio y trufi por la importancia que cada uno de estos medios de transporte representa en la ciudad de La Paz y a fin de facilitar el procesamiento engorroso que significa trabajar con el modelo multinomial probit.

Las salidas para el modelo propuesto son:

Coeficientes de la estimación:

	Estimación	Error estándar	Valor z	Pr(>|z|)	
Trufi(intercepto)	-0,515234	0,203634	-2,5302	0,011400	*
Micro(intercepto)	-1,386419	0,333234	-4,1605	3,176e-05	***
Propio(intercepto)	-2,536006	0,874300	-2,9006	0,003724	**
Tr1Dur	-0,081552	0,020845	-3,9124	9,140e-05	***
Tr1Cos	-0,099579	0,012336	-8,0721	6,661e-16	***
Tr1Esp	0,077496	0,016744	4,6283	3,686e-06	***
Taxi.Micro	-0,404094	0,644924	-0,6266	0,530937	
Taxi.Propio	-0,139860	1,615239	-0,0866	0,930999	
Micro.Micro	0,831226	0,487264	1,7059	0,088026	.
Micro.Propio	-0,168226	1,686771	-0,0997	0,920557	
Propio.Propio	1,755431	0,746906	2,3503	0,018760	*

Significancia: 0 '*' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1**

Log-Verosimilitud: -284,39

En este caso el modelo probit propuesto tiene además de la duración del viaje, costo y tiempo de espera como coeficientes significativos del modelo, a los medios de transporte trufi, micro y vehículo propio también significativos, dejando de lado el minibús. Los errores estándar de los coeficientes también son bastante pequeños en algunos casos, y sólo en el costo del tramo menor que el del modelo multinomial logit.

5. CONCLUSIONES

En general, los resultados han sido alentadores, mostrando que la simulación de las características de la forma de transporte en la ciudad de La Paz basándose en encuestas por muestreo anteriores, arroja resultados congruentes a pesar de elegir unas pocas características del gran número de variables relevadas en la encuesta. Sin embargo, una debilidad de la simulación es la imposibilidad de modelar de mejor forma la distribución y asignación de viajes por la falta de especificidad geográfica en las características simuladas.

En cuanto a los modelos estadísticos de elección discreta, mediante la comparación de modelos multinomial logit y probit, se puede decir que los costos, el tiempo de espera del tramo y el tiempo de duración del tramo, son determinantes en la elección del modo de transporte para los habitantes de la ciudad de La Paz, a diferencia de las variables sociodemográficas.

En la ciudad de La Paz se tiene al menos seis categorías de transporte estudiadas,

de las cuales cuatro corresponden al transporte público. Las rutas de estas últimas modalidades se han ido desarrollando sin considerar estrictamente la existencia de las demás o sin una planificación adecuada. Esto lleva a que las mismas se aglomeren en diversos puntos de la ciudad, principalmente el centro y que se perjudiquen mutuamente debido a la aglomeración vehicular que provocan. Al no existir medios alternativos en la ciudad de La Paz, además de su topografía accidentada, hace que los ciudadanos sean cautivos del actual sistema de transporte, que carece de la aptitud de mejorar su capacidad técnica y financiera para ofrecer un mejor servicio a la ciudadanía, lo que repercute en los resultados de las variables estudiadas.

En el modelo multinomial logit presentado se enfatiza el medio de transporte con mayor probabilidad de ser elegido tomando en cuenta el costo, tiempo de espera y duración del viaje, como la opción a pie, le sigue el disponer de un vehículo propio, y después el Pumakatari y Teleférico. Estos últimos destacan por el hecho de presentarse como alternativas recientes de mejoramiento del transporte, pero que no logran abarcar aún un porcentaje suficiente en su uso por los habitantes de la ciudad de La Paz.

En tanto que en términos de probabilidad de elección se ubican en el lado opuesto y con la menor probabilidad de ser elegidos al minibús y carry, seguido del micro, bus, microbús entre las opciones menos placenteras.

Finalmente, los resultados encontrados permiten señalar algunas recomendaciones en relación a las políticas sobre transporte para la ciudad de La Paz. Inicialmente tomando en cuenta los datos simulados, se encuentra que los modos de transporte más económicos son los elegidos por la población, como el ir a pie, y también aquellos que a pesar del tiempo de espera para su abordaje, proporcionan una forma más eficiente, segura y rápida de transporte como son las nuevas modalidades de Pumakatari y Teleférico. Por consiguiente, resultará conveniente incentivar esta forma de traslado más seguro para que una mayor parte de la población la elija, además de diseñar rutas que logren evadir eludir de cierta forma la centralización de la ciudad circulando por áreas que evadan la alta aglomeración vehicular.

BIBLIOGRAFÍA

Anda, C., Erath, A. y Fourie, P., (2017), "Transport modeling in the age of big data". International Journal of Urban Sciences.

De Dios Ortuzar, J. y L.G. Willumsen, (2011), "Modelos de Transporte". John Wiley & Sons.

Fajardo, H.C.L., & Gomez, S.A.M., (2015), "Análisis de la elección modal de transporte público y privado en la ciudad de Popayán. Territorios", 33, pp.157-190.

Gobierno Autónomo Municipal de La Paz, (2015), "Movilidad intraurbana en la Región Metropolitana de La Paz". Secretaría Municipal de Planificación para el desarrollo. pp 39-68.

MacFadden, D., (1974), "The Measurement of Urban Travel Demand". Journal of Public Economics.

Train, K., (2003), "Discrete choice analysis methods with simulation". Cambridge: Cambridge University Press.

MEDICIÓN DEL ERROR DE MUESTREO UTILIZANDO TÉCNICAS DE CONGLOMERADOS Y GRUPOS ALEATORIOS EN UNIVERSOS AGROPECUARIOS

Lic. Pinto Ahjuacho, Jaime Tito

✉ *titojaime_pinto@yahoo.com*

RESUMEN

La estimación de parámetros del sector agropecuario, demanda el planteamiento de metodologías de muestreo que plantean estimar la varianza de los estimadores que se construyan, es importante medir la precisión, medir el error de muestreo, para lo cual acudimos a la técnica del muestreo por conglomerados y grupos aleatorios, realizando una aplicación vemos el procedimiento de determinar el error de muestreo para las variables de estudio.

PALABRAS CLAVE

Muestreo por conglomerados, Grupos Aleatorios, Error de muestreo.

ABSTRACT

The estimation of parameters of the agricultural sector, demands the approach of sampling methodologies that propose to estimate the variance of the estimators that are constructed, it is important to measure the precision, measure the sampling error, for which we turn to the technique of cluster sampling and randomized groups, making an application we see the procedure of determining the sampling error for the study variables.

KEYWORDS

Cluster sampling, Random Groups, Sampling error.

1. INTRODUCCIÓN

La medición del error de muestreo, se puede indicar que es la incertidumbre que se comete al estimar un parámetro poblacional del universo de estudio mediante el valor obtenido a partir de una muestra de ese universo, utilizándose estadígrafos, es una medida de la variabilidad que se observaría entre todas las muestras posibles si fueran seleccionadas usando el mismo diseño de muestral.

Se debe aceptar que el error muestral, puede ser debido a muchos factores, uno puede ser a causa del diseño muestral, que plantea el conocer el universo, la determinación del tamaño de la muestra, la selección de las unidades de la muestra y el procedimiento de la estimación de la variables de estudio; por lo cual se analiza la varianza del estimador, concretamente la desviación estándar del estimador, al cual se le llama error de muestreo.

El estadígrafo para medir el error de muestreo dependerá de la muestra diseñada, de la técnica de muestreo utilizada, siendo estos cálculos complejos dependiendo del diseño muestral.

Mientras más pequeños sean los márgenes de error, los resultados de las encuestas serán más exactos, por ello se debe trabajar para que las muestras sean más eficientes, para que no haya errores al publicar los resultados del

margen de error de encuestas por muestreo.

Existen técnicas de muestreo que aportan en minimizar la varianza de los estimadores en estudio de población por encuestas por muestreo, tales como el muestreo por conglomerados, métodos de estimaciones de varianza de estimadores.

En el sector agropecuario, se puede plantear algunas técnicas que muestren la estimaciones de los parámetros de estudio y la medición de los errores de muestreo, una de las técnicas es la de conglomerados que muestra que plantea la minimización de la varianza del estimador, pero buscando que sea mas mínima, se puede plantear usar la varianza del error utilizando el método de Grupos Aleatorios.

Hay razones para aplicación del muestreo por conglomerados, se ha encontrado que para muchas encuestas no se tiene una lista confiable de los elementos de la población y además sería demasiado costoso formular dicha lista, no existen listas completas y actualizadas de la gente, las viviendas, o las granjas en grandes regiones geográficas, sin embargo, a partir de los mapas de la región, se puede dividirla en unidades de área, como serian las manzanas de una ciudad y los terrenos de área, como suelen elegirse estos conglomerados porque resuelven al problema de la construcción de una lista de unidades de muestreo(Cochran, 1998).

El Muestreo por Conglomerados, sugiere y desarrolla que las unidades de muestreo se pueden agrupar en subconjuntos, denominados conglomerados, de forma tal que haya heterogeneidad entre las unidades de un mismo conglomerado y homogeneidad entre conglomerados.

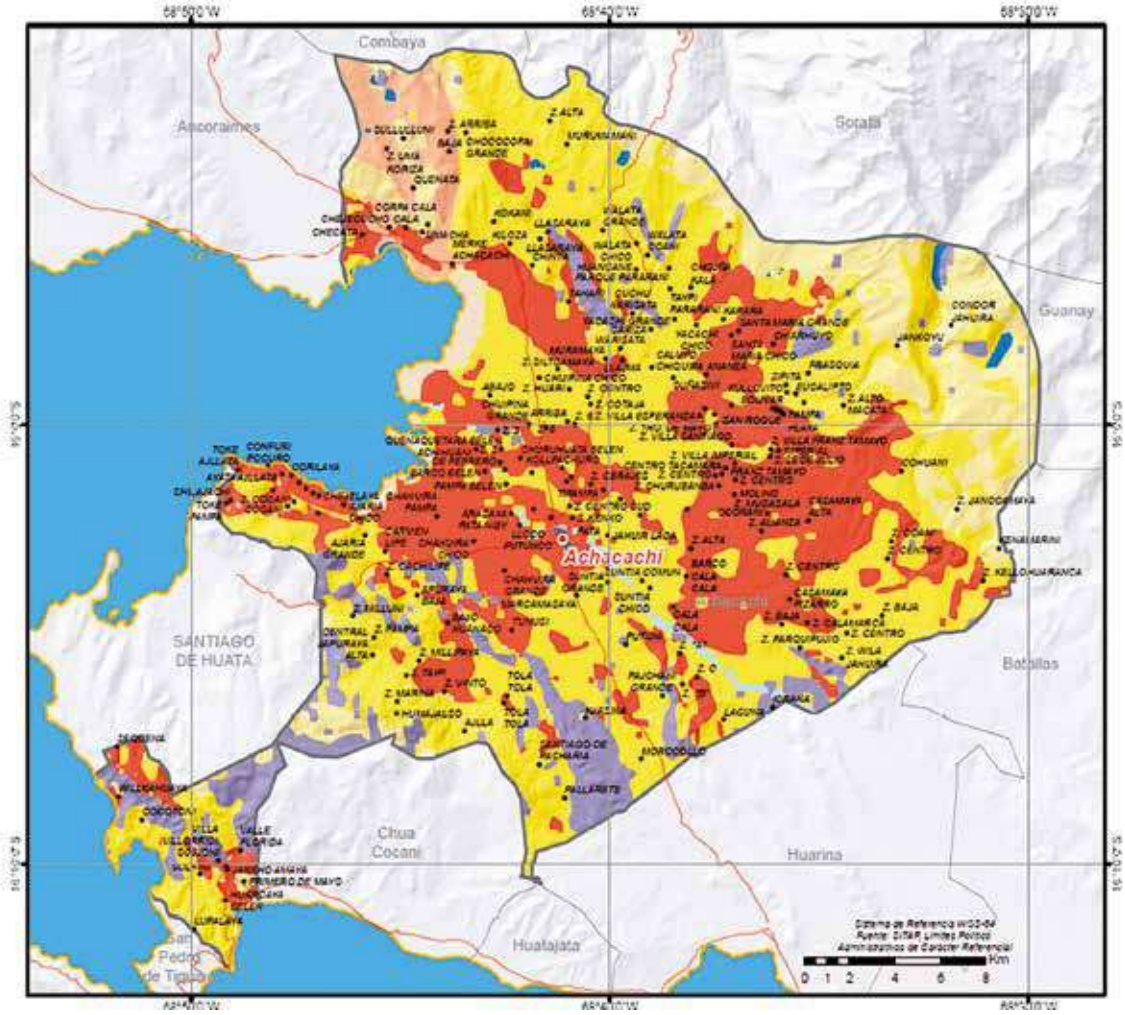
El concepto de homogeneidad entre conglomerados se refiere a que las medidas que se pueden calcular para cada conglomerado difieren poco de conglomerado en conglomerado. Al existir un patrón de conglomerados de las unidades muéstrales, se obtiene una estimación más precisa si se muestrean aleatoriamente un número determinado de conglomerados.

Para la estimación de la varianza de los estimadores, la técnica de los grupos aleatorios sugiere de una población estudio, trabajar formando grupos seleccionados en forma aleatoria y mediante una notación y procedimiento aleatorio realizar la medición de las varianzas (Miras Julio, 1976).

El objetivo general consiste en diseñar una propuesta de actividades didácticas que permita promover un uso de algunos tipos de muestreos, se plantea una aplicación de utilizar los métodos mencionados para mostrar el procedimiento de minimizar la varianza de estimadores en un universo de estudio y explicar la medición de error de muestreo.

Para conocer algunos parámetros poblacionales del sector agropecuario, se diseño una encuesta por muestreo en el municipio de Achacachi del departamento de La Paz, determinándose un tamaño de muestra y aplicar el procedimiento de estimación de la variable “Superficie cultivada de papa”. En el diseño muestral, la técnica plateada fue el muestreo por conglomerados, las unidades de primera etapa fueron los segmentos censales (Conglomerados) y unidades de segunda etapa las Unidades de Producción Agropecuaria (UPAs). El procedimiento y desarrollo de lo indicado se lo presenta a continuación.

Mapa Nº 1
DEPARTAMENTO: La Paz -PROVINCIA: Omasuyos MUNICIPIO: Achacachi



Fuente: Mapa de Referencia estadística, armada en base a Atlas de Potenciales Productivas ., SITAP –UDAPRO –Bolivia

2. MÉTODO DE MUESTREO POR CONGLOMERADOS.-

N= 498 Segmentos en el municipio de Achacachi.

n = 21 Segmentos muestra.

M=19.317 Unidades de Producción Agropecuarias (UPAs) municipio Achacachi.

VivM: Numero de viviendas Marco Muestral Agropecuario.

Mi: Número de Unidades de Producciones Agropecuarias (Listado).

UPAs MU: Número de Unidades de Producciones Agropecuarias (muestra planificada).

mi : Numero de Unidades de Producciones Agropecuarias (muestra ejecutada).

Cuadro Nº 1
Estimación de la Media de la variable "Superficie cultivada de papa" a nivel de unidades de producción agropecuarias

Segto i	*VivM*	*Mi*	*UPAs MU*	*mi* $y_i = \sum_{j=1}^{m_i} y_{ij}$	$\bar{y}_i = \frac{1}{m_i}\sum_{j=1}^{m_i} y_{ij}$	$M_i\,\bar{y}_i$	$\left(M_i\,\bar{y}_i - \overline{M}\,\bar{\bar{y}}\right)^2$
125	35	25	i	4	0,930	0,2325	5,8125	1,9176
126	14	45	8	4	0,280	0,0725	3,2625	1,3577
127	36	38	8	11	2,770	0,2518	9,5690	26,4329
128	20	32	8	3	0,570	0,1900	6,0800	2,7300
129	39	28	8	7	0,597	0,8520	2,3879	4,1607
130	52	51	8	10	3,267	0,3267	16,6617	149,6707
131	41	36	8	7	1,000	0,1428	5,1428	0,5113
132	72	72	8	6	0,147	0,0245	1,7640	7,0953
134	27	26	8	7	0,100	0,01428	0,3714	16,4535
135	75	73	8	10	0,209	0,0209	1,5257	8,4216
136	60	57	8	4	0,065	0,01625	0,92625	12,2601
137	57	57	8	7	0,209	0,02585	1,7018	7,4305
138	110	105	8	5	0,105	0,0210	2,2050	4,9403
141	70	68	8	6	0,448	0,0746	5,0773	0,4219
142	40	38	8	16	0,205	0,0128	0,4868	15,5306
143	45	45	8	6	0,136	0,0226	1,0199	11,6131
144	26	26	8	13	0,403	0,0310	0,8060	13,1167
145	50	50	8	8	0,360	0,0450	2,2500	4,7423
146	31	31	8	9	0,705	0,0783	2,4283	3,9976
147	35	33	8	9	0,955	0,1061	3,5013	0,8582
148	48	46	8	6	2,610	0,4350	20,0100	242,8080
Total							**92,99015**	**536,4706**

Fuente: Elaboración Propia

$$\hat{\bar{\bar{Y}}} = \bar{\bar{y}} = \left(\frac{N}{M\,n}\right)\sum_{i=1}^{n} M_i\,\bar{y}_i = \left(\frac{498}{19.317\,(21)}\right)92,99015 = 0,11415 =$$ Estimación de la media de la variable "Superficie cultivada de papa" a nivel de Unidades Agropecuarias Productivas.

$$\overline{M}\,\bar{\bar{y}} = 38,789\,(0,11415) = 4,4277$$

$$s_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(M_i\,\bar{y}_i - \overline{M}\,\bar{\bar{y}}\right)^2 = \frac{536,4706}{21-1} = 26,8235$$ = Varianza entre segmentos.

$$1 - f_1 = 1 - \left(\frac{n}{N}\right) = 1 - \left(\frac{21}{498}\right) = 0,9578$$ = Fracción de muestreo entre segmentos.

$$1 - f_2 = 1 - \left(\frac{m_i}{M_i}\right)$$ = Fracción de muestreo dentro del segmento i-ésimo.

$$s_{2i}^2 = \frac{1}{m_i - 1}\left(\sum_{j=1}^{mi} y_j^2 - \frac{\left(\sum_{j=1}^{mi} y_j\right)^2}{m_i} \right) = \text{Varianza dentro del segmento i-ésimo.}$$

Cuadro Nº 2
Cálculos del estimador y su varianza

Segto i	Mi	mi		$M_i^2(1-f_{2i}^2)$		S_{2i}^2	$M_i^2(1-f_{2i}^2)\frac{s_{2i}^2}{m_i}$
125	25	4	0,1600	0,8400	525,00	0,032225	4,2295
126	45	4	0,0890	0,9110	1.844,70	0,007691	3,5470
127	38	11	0,2894	0,7106	1.026,10	0,025030	2,3348
128	32	3	0,0930	0,9070	928,76	0,072300	22,3833
129	28	7	0,2500	0,7500	588,00	0,005038	0,4232
130	51	10	0,1960	0,8040	2.091,20	0,216170	45,2000
131	36	7	0,1940	0,8060	1.044,57	0,006240	0,9311
132	72	6	0,0830	0,9170	4.753,70	0,0000775	0,0614
134	26	7	0,2690	0,7310	494,16	0,00002857	0,002016
135	73	10	0,1369	0,8631	4.599,40	0,0002018	0,0928
136	57	4	0,070	0,9300	3.021,57	0,0001729	0,1306
137	57	7	0,1228	0,8772	2.850,02	0,001233	0,5020
138	105	5	0,047	0,9530	10.506,80	0,0003425	0,7197
141	68	6	0,088	0,9120	42.170,08	0,010740	7,5485
142	38	16	0,4210	0,5790	836,07	0,00005976	0,00312
143	45	6	0,1330	0,8670	1.755,67	0,00003674	0,01075
144	26	13	0,5000	0,5000	338,00	0,001462	0,03801
145	50	8	0,1600	0,8400	2.100,00	0,007078	1,8579
146	31	9	0,2900	0,7100	682,31	0,005075	0,3847
147	33	9	0,2727	0,7273	792,02	0,015980	1,4062
148	46	6	0,1304	0,8696	1.840,07	0,223830	68,6400
Total							**160,4465**

Fuente: Elaboración Propia

$$\hat{V}(\hat{\mu}) = \hat{V}(\bar{\bar{y}}) = (1-f_1)\left(\frac{1}{n\bar{M}^2}\right)s_1^2 + \frac{1}{nN\bar{M}^2}\sum_{i=1}^{n} M_i^2(1-f_{2i})\left(\frac{s_{2i}^2}{m_i}\right)$$

$$\hat{V}(\bar{\bar{y}}) = (0{,}9578)\left(\frac{1}{21(38{,}7891)^2}\right)26{,}8235 + \frac{1}{21(498)(38{,}7891)^2}160{,}4465 =$$

$$\hat{V}(\bar{\bar{y}}) = 0{,}0008131 + 0{,}00001019 = 0{,}0008232$$

$\sqrt{\hat{V}(\bar{\bar{y}})} = \sqrt{0{,}0008232} = 0{,}02869$ = Error de Muestreo de variable agropecuaria a nivel de UPAs.

Límite para el error de estimación = $1{,}96\ \sqrt{\hat{V}(\bar{\bar{y}})}$

$\bar{\bar{y}} - z\sqrt{\hat{V}(\bar{\bar{y}})} = 0{,}11415 - 1{,}96(0{,}02869) = 0{,}0579$ = Límite Inferior del estimador de la media de la variable "Superficie cultivada de papa" a nivel de UPAs en el municipio

$\bar{\bar{y}} + z\sqrt{\hat{V}(\bar{\bar{y}})} = 0{,}11415 + 1{,}96(0{,}02869) = 0{,}17038$ = Límite Superior del estimador de la media de la variable "Superficie cultivada de papa" a nivel de UPAs en el municipio.

$CV(\bar{\bar{y}}) = \frac{\sqrt{\hat{V}(\bar{\bar{y}})}}{\bar{\bar{y}}} * 100 = \frac{0{,}02869}{0{,}11415} * 100 = 25{,}13\ \%$ = Coeficiente de variación del estimador de la media.

ESTIMACIÓN DEL TOTAL DE LA VARIABLE "SUPERFICIE CULTIVADA DE PAPA" A NIVEL DE UNIDADES DE PRODUCCIÓN AGROPECUARIAS

$\hat{Y} = M\,\bar{\bar{y}} = 19.317(0{,}11415) = 2205{,}03$ = Estimador del Total de la variable a nivel UPAs en el municipio.

$V(\hat{Y}) = V(M\,\bar{\bar{y}}) = M^2 V(\bar{\bar{y}}) = (19.317)^2(0{,}0008232) = 307.174{,}18$ = Varianza del estimador del Total.

$\sqrt{\hat{V}(\hat{Y})} = \sqrt{307.174{,}18} = 554{,}23$ = Error de muestreo del estimador del Total.

$\hat{Y} - z\sqrt{\hat{V}(\hat{Y})} = 2.205{,}03 - 1{,}96(554{,}23) = 1.118{,}7$ = Límite Inferior del estimador del Total de la variable a nivel de UPAs en el municipio.

$\hat{Y} + z\sqrt{\hat{V}(\hat{Y})} = 2.205{,}03 + 1{,}96(554{,}23) = 3.291{,}3$ = Límite Superior del estimador del Total de la variable a nivel de UPAs en el municipio.

$CV(\hat{Y}) = \frac{\sqrt{\hat{V}(\hat{Y})}}{\hat{Y}} * 100 = \frac{554{,}23}{2.205{,}03} * 100 = 25{,}13\ \%$ = Coeficiente de variación del estimador del Total.

3. ESTIMACIÓN DE LA VARIANZA POR LA TÉCNICA DE GRUPOS ALEATORIOS

ESTIMACIÓN DE LA VARIANZA DEL ESTIMADOR DE LA MEDIA Y EL TOTAL, MEDIANTE LA TÉCNICA DE GRUPOS ALEATORIOS

Para la estimación de la varianza se utilizo la siguiente notación:

N = Número de Unidades de Producción Agropecuarias en el Municipio.

n= Número de Unidades de Producción Agropecuarias en la muestra.

yi= Valor observado en la j-ésima UPA

k= Número de grupos aleatorios.

$m = \frac{n}{k}$ = Número de UPAs seleccionadas aleatoriamente.

$\hat{\bar{Y}} = \frac{\sum_{i=1}^{n} y_i}{n}$ = Promedio del valor observado a nivel de la muestra de tamaño n

$\hat{\bar{Y}}_r = \frac{\sum_{i=1}^{m} y_i}{m}$ =Promedio del valor observado aleatoriamente en la submuestra m (para valores de j=1…m).

$\hat{V}_{G.A.}\left(\hat{\bar{Y}}\right) = \frac{\sum_{r=1}^{K}\left(\hat{\bar{Y}}_r - \hat{\bar{Y}}\right)^2}{k(k-1)}$ = Varianza estimada del estimador de la media (r =1, 2, …, k).

$\sqrt{\hat{V}_{G.A.}\left(\hat{\bar{Y}}\right) = \frac{\sum_{r=1}^{K}\left(\hat{\bar{Y}}_r - \hat{\bar{Y}}\right)^2}{k(k-1)}}$ = Error de muestreo del estimador de la media por el método de Grupos Aleatorizados.

$M = \sum_{i=1}^{N} M_i$ = Número de unidades de Producción Agropecuaria en el municipio de estudio.

$\hat{V}_{G.A.}\left(\hat{Y}\right) = \hat{V}\left(M\,\hat{\bar{Y}}\right) = M^2\hat{V}\left(\hat{\bar{Y}}\right) = M^2\,\frac{\sum_{r=1}^{K}\left(\hat{\bar{Y}}_r - \hat{\bar{Y}}\right)^2}{k(k-1)}$ = Varianza estimada del estimador del Total por el método de Grupos Aleatorizados. (r =1, 2, …, k).

$\sqrt{\hat{V}_{G.A.}\left(\hat{Y}\right)}$ = Error de muestreo del estimador del Total por el método de Grupos Aleatorizados.

Estimación de la Varianza del Estimador de la Media de la variable a nivel de unidades de producción agropecuarias.

n = 158 Unidades de Producción Agropecuarias en la muestra.
M = 19.317 Unidades de Producción Agropecuarias (Marco Muestral).
k = 3 grupos aleatorios.

$m = \frac{n}{k} = \frac{158}{3} = 52{,}66 \approx 53$ = Número de UPAs seleccionadas aleatoriamente

$\bar{\bar{y}} - z\sqrt{\hat{V}_{G.A}\left(\hat{\bar{Y}}\right)} = 0{,}11415 - 1{,}96(0{,}008906) = 0{,}09669$ = Límite Inferior del estimador de la media de la variable a nivel de UPAs en el municipio.

$\bar{\bar{y}} + z\sqrt{\hat{V}_{G.A.}\left(\hat{\bar{Y}}\right)} = 0{,}11415 + 1{,}96(0{,}008906) = 0{,}131605$ = Límite Superior del estimador de la media de la variable a nivel de UPAs en el municipio.

$CV(\bar{\bar{y}}) = \frac{\sqrt{\hat{V}_{G.A.}\left(\hat{\bar{Y}}\right)}}{\bar{\bar{y}}} * 100 = \frac{0{,}008906}{0{,}11415} * 100 = 7{,}80\ \%$ = Coeficiente de variación del estimador de la media.

Cuadro Nº 3
Estimación de la varianza por grupos aleatorios

		GRUPOS ALEATORIOS			RESULTADOS
		1	2	3	
	$\sum_{i=1}^{m} y_i$	5,44	4,244	5,189	
	m	53	53	52	158
	$\hat{\bar{Y}}_r$	0,102641	0,08007	0,09978	
	$\hat{\bar{Y}}$				0,101778
	$\left(\hat{\bar{Y}}_r - \hat{\bar{Y}}\right)^2$	0,000000744	0,0004712	0,000003992	0,0004759
	$K(k\text{-}1)$				3(3-1)
Media	$\hat{V}_{G.A.}\left(\hat{\bar{Y}}\right)$				0,00007932
	$\sqrt{\hat{V}_{G.A.}\left(\hat{\bar{Y}}\right)}$				**0,008906**
Total	M				19.317
	$\hat{V}_{G.A.}\left(\hat{Y}\right)$				29.597,97
	$\sqrt{\hat{V}_{G.A.}\left(\hat{Y}\right)}$				**172,04**

Fuente: Elaboración Propia

Estimación de la Varianza del Estimador del Total de la variable a nivel de unidades de producción agropecuarias.

$\hat{Y} = M\,\hat{\bar{y}} = 19.317(0{,}11415) = 2205{,}03$ = Estimador del Total de la variable a nivel UPAs en el municipio.

$\hat{V}_{G.A.}\left(\hat{Y}\right) = \hat{V}\left(M\,\hat{\bar{y}}\right) = M^2\hat{V}\left(\hat{\bar{Y}}\right) = (19.317)^2(0{,}00007932) = 29.597{,}97$ =Varianza del estimador del Total.

$\sqrt{\hat{V}_{G.A.}\left(\hat{Y}\right)} = \sqrt{29.597{,}97} = 172{,}04$ = Error de muestreo del estimador del Total.

$\hat{Y} - z\sqrt{\hat{V}_{G.A.}\left(\hat{Y}\right)} = 2.205{,}03 - 1{,}96(172{,}04) = 1.867{,}83$ =Límite Inferior del estimador del Total de la variable a nivel de UPAs en el municipio.

$\hat{Y} + z\sqrt{\hat{V}_{G.A.}\left(\hat{Y}\right)} = 2.205{,}03 + 1{,}96(172{,}04) = 2.542{,}22$ =Límite Superior del estimador del Total de la variable a nivel de UPAs en el municipio.

$CV(\hat{Y}) = \dfrac{\sqrt{\hat{V}_{G.A.}\left(\hat{Y}\right)}}{\hat{Y}} * 100 = \dfrac{172{,}04}{2.205{,}03} * 100 = 7{,}80\,\%$ =Coeficiente de variación del estimador del Total.

La precisión del estimador de ambos métodos se puede ver en el siguiente cuadro:

Cuadro Nº 4
Error de muestreo de la media

	Muestreo por Conglomerados	Técnica de Grupos Aleatorios
Estimador	0,11415	
Error de muestreo	0,02869	0,008906
Límite Inferior (N.C. 95%)	0,0579	0,09669
Límite Superior (N.C. 95%)	0,17038	0,131605
Coeficiente de Variación (%)	25,13	7,80

Fuente: Elaboración Propia

Cuadro Nº 5
Error de muestreo del Total

	Muestreo por Conglomerados	Técnica de Grupos Aleatorios
Estimador	2.205	
Error de muestreo	554,23	172,04
Límite Inferior (N.C. 95%)	1.118,7	1.867,83
Límite Superior (N.C. 95%)	3.291,3	2.542,22
Coeficiente de Variación (%)	25,13	7,80

Fuente: Elaboración Propia

4. CONCLUSIONES.-

El muestreo por conglomerados, propone un tratamiento de la información que se puede acomodar al sector agropecuario, donde se trabaja para validar la homogeneidad entre conglomerados y permite la comparación de varianzas entre conglomerados.

Esta técnica de muestreo, facilita al investigador que pueda asignar sus recursos limitados a los pocos conglomerados o áreas seleccionadas aleatoriamente cuando se usan muestras por conglomerados.

El muestreo por conglomerados, muestra un proceso controlable que permite conocer los estimadores de la media y el total como también sus varianzas.

En el proceso de cálculo se puede ver la heterogeneidad dentro del grupo o conglomerado que es fundamental para un buen diseño del muestreo por conglomerados, que muestra que los elementos dentro de cada grupo debe ser tan heterogéneos como la población objetivo.

La técnica de Grupos Aleatorios que estima la varianza de estimadores, muestra una minimización de las varianzas, siendo un buen medidor del error de muestreo.

La utilización de diferentes técnicas de muestreo que minimizan la varianza de los estimadores son buenas herramientas para resolver problemas de estimación en el sector agropecuario.

BIBLIOGRAFÍA

Cochran William G.,(1998), "Técnicas de muestreo", Decima Cuarta Edición;, JOHN WILEY & SONS, México.

Lohr Sharon L.,(1999), "Muestreo: Diseño y Análisis", Duxbury Press, USA.

Woodruff, R. S., (1971) "A simple method for approximating the variance of a complicated estimate". Journal of the American Statistical Association 66:411-414.

Binder D.A., (1983) "On the variances of asymptotically normal estimators from complex surveys International Statistical" Review 51:279-292.

Miras Amor Julio, (1976), "Estimación de errores de muestreo", INE España.

ANALISIS DE CONGLOMERADOS

Dr. Cs. Gustavo Ruiz Aranibar[1]

✉ *ruizaranibargustavo@gmail.com.bo*

RESUMEN

El análisis de conglomerados implica agrupar objetos, sujetos o variables, con características similares en grupos. La semejanza o disimilitud de los objetos se mide por un índice particular de asociación. Se consideran los tipos de métodos que agrupan variables basadas en la estructura de correlación de variables.

En algunos estudios geológicos es conveniente agrupar muestras similares en las que se han realizado muchas mediciones y medir el grado de similitud entre los grupos. Utilizando el coeficiente de correlación o la función de distancia, la matriz resultante suele ser demasiado grande para la interpretación directa. El análisis de conglomerados, es una técnica desarrollada por sicólogos, es un método de búsqueda de relaciones en una gran matriz simétrica. Las variables o grupos de variables especificados pueden usarse entonces para agrupar las muestras por función de distancia.

PALABRAS CLAVE

Coeficiente de correlación, dendograma, distancia, taxonomía, similitud.

ABSTRACT

Cluster analysis involves grouping objects, subjects or variables, with similar characteristics into groups. Similarity or dissimilarity of objects is measured by a particular index of association. Types of methods that cluster variables based on correlation structure of variables.

In some geologic studies it is desirable to group together similar samples on which many measurements have been made, and to measure the degree of similarity between the groups. Using either a coefficient, the matching coefficient, or the distance function, the resulting matrix is usually too large for direct interpretation. Cluster analysis, a technique developed by psychologists, is a method of searching for relationships in a large symmetrical matrix. Specified variables or groups of variables can then be used in clustering the samples by distance function.

KEYWORDS

Correlation coefficient, dendogram, distance, taxonomy, similarity.

1. INTRODUCCIÓN.

El Análisis de Conglomerados (AC) también conocido como Cluster Analysis o Taxonomía Numérica, es una técnica estadística multivariable, cuya finalidad es dividir un conjunto de objetos en grupos de forma que los objetos en un mismo grupo sean muy similares entre sí (cohesión interna del grupo) y de los objetos de conglomerados diferentes sean distintos (aislamiento externo del grupo); éste permite agrupar los elementos o variables de un archivo de datos en función del parecido o similitud existente entre ellos, buscando agrupar elementos (o variables) y tratando de lograr la máxima homogeneidad entre los grupos y la mayor diferencia entre los ellos, es una técnica descriptiva, teórica y no inferencial.

[1] Se agradece a la UAGRM por la beca otorgada con fondos del IDH, para cursar y culminar exitosamente el Doctorado en Ciencias de la Educación Superior. Especializado en Estadística e Informática

El AC permite clasificar las unidades de análisis en grupos homogéneos de tal manera que las unidades pertenecientes a uno de los grupos o conglomerados serán lo más parecidas entre sí, aunque muy diferentes respecto a los otros grupos o dicho de otra manera es la tarea de agrupar un conjunto de objetos de tal manera que los miembros del mismo grupo sean más similares, en algún sentido u otro, siendo la tarea principal de la minería de datos exploratorios, técnica común en el análisis de datos estadísticos.

2. OBJETIVOS DEL ANÁLISIS DE CONGLOMERADOS.

El principal objetivo es agrupar objetos (personas, empresas, productos, etc.) en conglomerados, de forma que cada objeto es muy parecido a los que hay en el conglomerado con respecto a algún criterio de selección predeterminado.

Se expone las dos decisiones sobre las que se apoya esta técnica de análisis:

1. Elección de una medida de proximidad entre los individuos.

2. Elección de un criterio a partir del cual agrupar a los individuos o unidades de análisis (secciones censales, países, ciudades,etc.) en los conglomerados.

Para lo cual se debe:

- Plantear el problema a resolver por un AC.

  En el planteamiento del problema de conglomerados se debe considerar las variables en la que se basara el agrupamiento. Haciendo notar que la inclusión de una o más variables irrelevantes distorsiona una solución de agrupamiento, que podría ser útil o no.

  En esencia, el conjunto de las variables elegidas debe describir la semejanza entre los objetos en términos relevantes para el problema de investigación; en la investigación exploratoria, el investigador debe valerse de su juicio e intuición, se aconseja utilizar para los conglomerados que se utilicen un número de muestras mayores a 100.

- Establecer medidas de semejanza y de distancia entre los objetos a clasificar en función del tipo de datos analizados.

- Analizar algunos de los métodos de clasificación propuestos en la literatura, debido a la existencia de diferentes métodos tales como los jerárquicos aglomerativos, el algoritmo de las k-medias, y otros, que permiten determinar el número de grupos.

- Interpretar los resultados obtenidos.

  Como técnica de agrupación de variables, el AC es similar al análisis factorial; pero, mientras que la factorización es más bien poco flexible en algunos de sus supuestos (linealidad, normalidad, variables cuantitativas, etc.) y siempre estima de la misma manera la matriz de distancias, la aglomeración es menos restrictiva en sus supuestos (no exige linealidad, ni simetría, permite variables categóricas, etc.) y admite varios métodos de estimación de la matriz de distancias.

3. CONCEPTOS GENERALES DEL AC.

La taxonomía es la ciencia de la clasificación de los seres, elementos de una de las ciencias naturales, las describe, denomina y clasifica ordenadamente atendiendo a sus afinidades y relaciones.

El AC tuvo su origen cuando se utilizó en antropología por Driver y Kroeber en 1932 e introducido a la sicología por Zubin en 1938 y Robert Tryon en 1939, fue utilizado por Cattell en 1943 para la clasificación de la personalidad sicológica basada en teoría de rasgos.

El AC es una técnica usada para clasificar objetos o casos en grupos relativamente homogéneos llamados conglomerados. Los objetos de cada conglomerado tienden a ser similares entre sí y diferentes de los objetos de otros conglomerados. Como técnica de agrupación de casos, el AC es similar al análisis discriminante. Sin embargo, mientras que el análisis discriminante efectúa la clasificación tomando como referencia un criterio o variable dependiente (los grupos de clasificación), el AC permite detectar el número óptimo de grupos y su composición únicamente a partir de la similitud existente entre los casos; además, el AC no asume ninguna distribución específica para las variables.

Tanto el AC como el análisis discriminante se interesan en la clasificación. Sin embargo, el análisis discriminante requiere de un conocimiento previo del conglomerado o la pertenencia al grupo de cada objeto o caso incluido, para desarrollar la regla de clasificación.

En el AC, todo un conjunto de relaciones interdependientes, no distingue entre variables dependientes e independientes, sino que examina las relaciones interdependientes entre el conjunto completo de variables. Los objetos en un grupo son relativamente similares en términos de estas variables y diferentes de los objetos de otros grupos. Cuando se usa de esta manera, el AC es la contrapartida del análisis factorial, ya que no reduce el número de variables sino de objetos, a los que agrupa en un número mucho menor de conglomerados.

El método jerárquico es idóneo para determinar el número óptimo de conglomerados existente en los datos y el contenido de los mismos. El método de K medias permite procesar un número ilimitado de casos, pero sólo permite utilizar un método de aglomeración, y es este método, el que se describirá y aplicará en el presente trabajo.

Existe un notable contraste entre el AC con el análisis de varianza, la regresión, el análisis discriminante y el análisis factorial; los cuales se fundamentan en un razonamiento estadístico amplio. Aunque muchos de los procedimientos de conglomeración tienen propiedades estadísticas importantes, debe reconocerse fundamentalmente su sencillez.

Los siguientes estadísticos y conceptos se asocian con el AC.

- Calendario de aglomeración: este programa brinda información sobre objetos o casos que se combinan en cada etapa del proceso de conglomeración jerárquica.

- Centroide del conglomerado: es la media de los valores de las variables de todos los objetos o casos de un conglomerado particular.

- Centros del conglomerado: son los puntos de partida en la conglomeración no jerárquica. Los conglomerados se construyen en torno a estos centros.

- Pertenencia al conglomerado: indica el conglomerado al que corresponde cada objeto o caso.

- Dendograma: conocido como gráfica de árbol, es un medio gráfico para presentar los resultados de la conglomeración. Las líneas verticales representan conglomerados que están unidos. La posición de la línea en la escala, indica las distancias en las que se unen los conglomerados. El diagrama de árbol muestra los grupos que se forman al crear conglomerados de observaciones en cada paso y sus niveles de similitud. El nivel de similitud se mide en el eje vertical (alternativamente se puede mostrar el nivel de distancia) y las diferentes observaciones se especifican en el eje horizontal, como se observará en su aplicación.

Distancias entre los centros de los conglomerados: estas distancias indican cuán separados están los pares individuales de conglomerados. Los que están muy separados son distintos y, por lo tanto, son deseables. El objetivo principal del AC es definir la estructura de los datos colocando las observaciones más parecidas de los datos, en conglomerados de acuerdo a las distancias obtenidas de la matriz de distancia.

4. ANÁLISIS DE CONGLOMERADOS.

La clasificación es la colocación de objetos en grupos más o menos homogéneos, de tal manera que se revela la relación entre grupos. Este es el punto fuerte en especial de los taxonomistas, que intentan deducir el linaje de las criaturas vivas de sus características y semejanzas. La taxonomía es altamente subjetiva y depende de las habilidades de los taxonomistas, desarrollados a través de años de experiencia. En este sentido, el campo es análogo en muchos aspectos a la geología. En geología, un grupo de investigadores se han vuelto insatisfechos con la subjetividad y buscan nuevas técnicas de clasificación que incorporen las capacidades masivas de manejo de información creando una base de datos en la computadora. Estos trabajadores se llaman taxonomistas numéricos y son responsables de muchos de los avances en la clasificación numérica. En el AC se tiene que:

a. Los conglomerados resultantes deben mostrar un alto grado de homogeneidad interna (dentro del conglomerado) y un alto grado de heterogeneidad externa (entre conglomerados).

b. Gráficamente, los objetos dentro de los conglomerados estarán muy próximos, y los diferentes conglomerados muy alejados.

c. El AC permite la inclusión de múltiples variables para llevar a cabo la agrupación de objetos.

5. FUNDAMENTOS TEÓRICOS PARA EL AC

En la actualidad, la taxonomía numérica es el centro de una controversia entre los biólogos, al igual que el acrimonioso debate entre los sicólogos que se arremolinaron alrededor del análisis factorial en los años 1930 y 1940. Como en esa disputa, las técnicas de la taxonomía numérica han sido exageradamente promovidas por

algunos practicantes. Además, afirmaron que una taxonomía numéricamente derivada representaría mejor la filogenia de un grupo de organismos que cualquier otro tipo de clasificación. Esto, por supuesto, no puede ser demostrado. En la actualidad, los fundamentos teóricos del AC son incompletos, se sabe poco de las propiedades estadísticas de los métodos taxonómicos numéricos y no se dispone de pruebas de significación. Muchos de los métodos de taxonomía numérica son importantes en la investigación geológica, especialmente en la clasificación de los invertebrados fósiles y el estudio de la paleoecología.

Si se tiene una colección de objetos que desea organizar en una clasificación jerárquica como en biología, estos objetos se denominan "unidades taxonómicas operativas". En cada objeto, se realiza una serie de medidas que constituyen el conjunto de datos.

Teniéndose n objetos, con medidas de m características, el conjunto de datos forma una matriz de observaciones. A continuación, se calculará una medida de semejanza o similitud entre cada par de objetos. Se han utilizado los coeficientes de semejanza, como ser el coeficiente de correlación , y la distancia euclidiana estandarizada que es el coeficiente de distancia, se observara que, por definición, los elementos de la diagonal principal de esta matriz son nulos.

Las fórmulas que se utilizan para el AC son:

$$X_{ij} \quad \forall\ i = 1,\ldots,n \ \ y \ \ j = 1,\ldots,m$$

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

$$S = \sqrt{\frac{\sum_{i=1}^{n} X_i^2}{n} - \left(\frac{\sum_{i=1}^{n} X_i}{n}\right)^2}$$

$$Z_i = \frac{X_i - \overline{X}}{S}$$

$$d_{ij} = \sqrt{\frac{\sum_{k=1}^{m} (X_{ik} - X_{jk})^2}{m}}$$

$$r_{ij} = \frac{\operatorname{cov}(x, y)}{s_x s_y}$$

$$r_{ij} = \frac{n\sum XY - \sum X \sum Y}{\sqrt{n\sum X^2 - (\sum X)^2} * \sqrt{n\sum Y^2 - (\sum Y)^2}}$$

$$-1 \le r_{ij} \le 1$$

Dónde:

n = número de observaciones.

m = número de variables de cada observación.

X_{nm} = matriz de observaciones.

X_{mn} = matriz transpuesta de observaciones.

X_{ik} = k-ésima medida en el objeto i.

X_{jk} = k-ésima medida sobre el objeto j.

S = desviación estándar.

Z_i = variable normalizada.

d_{ij} = distancia entre el objeto i y el objeto j.

r_{ij} = coeficiente de correlación entre dos columnas i y j.

r_{mm} = matriz de coeficientes de correlación.

6. MEDIDAS DE DISTANCIA.

Las distancias entre los centros de los conglomerados, indican que cuánto más separados están los pares individuales de conglomerados y son distintos, por lo tanto, son deseables. En la medición de las distancias

se utilizan diferentes medidas, especialmente medidas para variables cuantitativas, las más utilizadas son:

1. Distancia euclídeana y distancia euclídea al cuadrado
2. Distancia métrica de Chebychev
3. Distancia de Manhattan
4. Distancia de Minkowski
5. Distancia de Mahalanobis

Todas estas distancias no son invariantes a cambios de escala por lo que se aconseja estandarizar los datos si las unidades de medida de las variables no son comparables.

7. MÉTODOS DE CLASIFICACIÓN.

Entre los muchos tipos de métodos que existen cabe destacar los siguientes:

- Repartición: tienen un número de grupos, g fijado de antemano, como objetivo y agrupa los objetos para obtener los g grupos. Comienzan con una solución inicial y los objetos se reagrupan de acuerdo con algún criterio de optimalidad.
- Métodos tipo Q: son similares al análisis factorial y utilizan como información la matriz XX' utilizando las variables como objetos y los objetos como variables.
- Procedimientos de localización de modas: agrupan los objetos en torno a modas con el fin de obtener zonas de gran densidad de objetos separados unos de otros por zonas de poca densidad.
- Métodos que permiten solapamiento: permiten que los grupos tengan elementos en común.
- Método de Ward, tiene tendencia a formar conglomerados más compactos y de igual tamaño.

EL Método jerárquico se caracteriza porque en cada paso del algoritmo sólo un objeto cambia de grupo y los grupos están anidados en los de pasos anteriores. Si un objeto ha sido asignado a un grupo ya no cambia más de grupo, este método es utilizado en el ejemplo de aplicación. Es un método aglomerativo, que comienza con n conglomerados de un objeto. En cada paso del algoritmo se recalculan las distancias entre los grupos existentes y se unen los dos grupos más similares o menos disimilares. El algoritmo acaba con un conglomerado conteniendo todos los elementos. Para determinar qué grupos se unen o se dividen, se utiliza una función objetivo o criterio que, en el caso de los métodos aglomerativos recibe el nombre de enlace.

8. INTERPRETACIÓN DE RESULTADOS

Una distancia baja indica que los dos objetos son similares o están muy cerca o juntos, una gran distancia indica disimilitud. Comúnmente, la matriz de datos originales se estandariza antes de calcular las mediciones de distancias, entonces estos nuevos valores tienen un promedio nulo y una desviación estándar la unidad, esto asegura que cada variable es ponderada igualmente, de lo contrario, la distancia será influenciada más fuertemente por la variable que tiene la mayor magnitud.

Se han desarrollado varias técnicas de agrupamiento, en este trabajo se desarrollará una técnica simple de agrupamiento, llamado el método ponderado par-grupo con promedios aritméticos, luego se señalará algunas modificaciones útiles a este esquema.

9. CONSTRUCCIÓN DE UN DENDOGRAMA.

El dendograma ó gráfico en forma de árbol,

es una herramienta visual para ayudar a decidir el número de conglomerados que podrían representar mejor la estructura de los datos. Se utiliza el dendograma para observar cómo se forman los conglomerados en cada paso y para evaluar los niveles de similitud (o distancia) de los conglomerados que se forman.

La decisión acerca de la agrupación final, se la conoce como cortar el dendograma. Cortar el dendograma es similar a trazar una línea vertical a lo largo del dendograma para especificar la agrupación final, si el dendograma está orientado horizontalmente mediante una línea horizontal sucede lo contrario, también se pueden comparar agrupaciones finales en los dendogramas para determinar cuál de ellas tiene más sentido para los datos. En la determinación del número final de conglomerados a formar o regla de parada, no hay un procedimiento determinado, decidiéndolo el investigador en la fase de interpretación de los datos.

Para una comprensión en la construcción de un dendograma, teniéndose una matriz de correlaciones, que es simétrica de los coeficientes de similitud, entre seis objetos supuestos, identificados como A,B,..,F, para las filas y para las columnas, como se muestra a continuación:

Tabla Nº 1
Matriz de correlaciones de seis variables

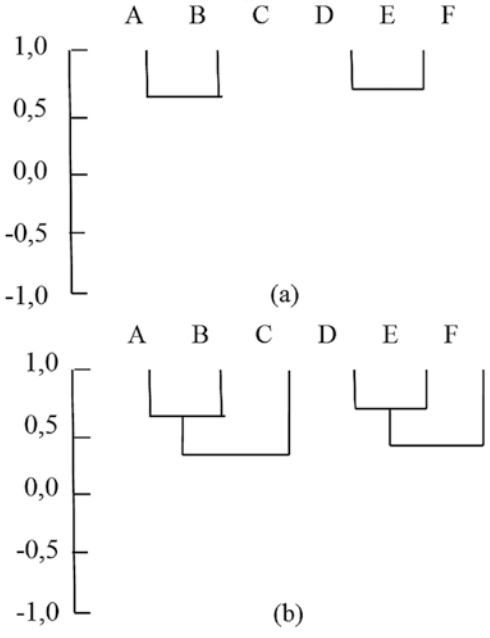
A	B	C	D	E	F
1,00	(0,57)	0,12	-0,65	-0,62	-0,39
(0,57)	1,00	(0,46)	-0,79	-0,72	-0,72
0,12	0,46	1,00	-0,58	-0,61	-0,52
-0,65	-0,79	-0,58	1,00	(0,66)	(0,41)
-0,62	-0,72	-0,61	(0,66)	1,00	0,40
-0,39	-0,72	-0,52	0,41	0,40	1,00

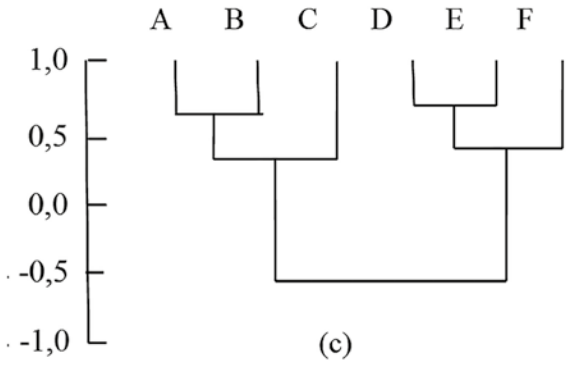
Fuente: Obtienido de Davis C. John. Statistics and Data Analysis in Geology.

El primer paso en la agrupación por el método de grupo de pares, es encontrar mutuamente las correlaciones más altas en la matriz para formar los centros de los conglomerados. La más alta correlación en cada columna de la matriz de la tabla 1, se muestra entre paréntesis. Los objetos A y B forman parejas mutuamente altas, porque A se asemeja mucho a B y B se asemeja más a A. Sin embargo, C y B no forman una pareja mutuamente alta, porque aunque C se asemeja mucho a B, B se parece más a A que a C. Para calificarlo como un par mutuamente alto, el coeficiente r_{ij} debe ser el coeficiente más alto en sus respectivas columnas.

Se puede indicar la semejanza entre los pares mutuamente altos en un diagrama como el de la Fig. 1. En la gráfica (a). el objeto A está conectado a B a un nivel de 0,57, indicando el grado de su similitud mutua. De la misma manera, D y E están conectados. Este es el primer paso en la construcción de un dendograma, o diagrama de árbol, que es la forma más común de mostrar los resultados de la agrupación.

Figura Nº 1
(a) Dendograma con grupos iníciales.
(b) Conexión de los objetos restantes a los grupos
(c) Conexión final de dos grupos, completando el dendograma





Fuente: Elaboración del autor

A continuación, la similitud de la matriz debe recalcularse, tratando a los elementos agrupados como un solo elemento. Hay varios métodos para hacer esto. La técnica simple que se está considerando, las nuevas correlaciones entre todos los grupos y los objetos no agrupados se recalculan mediante el cálculo aritmético simple. Así, a nueva correlación entre el grupo AB y el objeto C es igual a la suma de las correlaciones de los elementos comunes a AB y C, dividido entre 2. La tabla 2, contiene los resultados de estos nuevos cálculos.

Tabla Nº 2
Matriz de correlaciones promedio entre dos conglomerados y dos variables

AB	C	DE	F
1,00	(0,29)	-0,70	-0,55
(0,29)	1,00	-0,59	-0,52
-0,70	-0,59	1,00	(0,41)
-0,55	-0,52	(0,41)	1,00

Fuente: Elaboración del autor

El procedimiento de agrupamiento se repite; los pares mutuamente altos son buscados y agrupados. En este ciclo, el objeto C se une al grupo AB y el objeto F se une al grupo de (Fig. 1. b). Entonces el proceso continúa hasta que todos los racimos se unan. La matriz final de similitudes será una matriz de 2x2 entre los dos agrupamientos restantes, como se muestra en la tabla 3. Esto indica que el grupo ABC tiene una semejanza de -0,59 con el grupo DEF. Así el dendrograma puede ser completado. (Fig. 1. c)

Tabla Nº 3
Matriz de correlaciones promedio entre dos grupos

ABC	DEF
1,00	-0,59
0,59	1,00

Fuente: Elaboración del autor

Las características esenciales de este método de análisis de conglomerados pueden resumirse en la forma siguiente:

a. El coeficiente de correlación se utiliza como medida de similitud.
b. Las similitudes más altas se agrupan en primer lugar.
c. Dos objetos sólo pueden conectarse si tienen correlaciones mutuamente más altas entre sí.
d. Después de que se agrupan dos objetos, se promedian sus correlaciones con todos los demás objetos.

Una modificación obvia de este esquema es incorporar alguna otra medida de similitud. Aunque se han propuesto muchas medidas, sólo dos se utilizan ampliamente; el coeficiente de correlación y el coeficiente de distancia.

Como era de esperar, una distancia baja indica que los dos objetos son similares, o "estar cerca o juntos", ya que una gran distancia indica disimilitud. Comúnmente, la matriz de datos originales se estandarizan antes de calcular mediciones de distancia. Esto asegura que cada variable es ponderada igualmente, de lo contrario, la distancia será influenciada más fuertemente por la variable que tiene la mayor magnitud. Así por ejemplo, se puede medir tres ejes perpendiculares sobre una colección de muestras. Si se mide dos de los

ejes en centímetros y el tercero en milímetros, el tercer eje tendrá proporcionalmente diez veces la influencia sobre el coeficiente de distancia de las otras dos variables.

Se han desarrollado varias técnicas de agrupamiento: una consideración de todas las posibles variaciones y sus méritos relativos que están fuera del alcance de este trabajo, recomendándose el texto Benzécri J. P. señalada en la bibliografía para mayor conocimiento. Se describe la técnica de agrupamiento, llamado el método ponderado par-grupo con promedios aritméticos.

10. CONSIDERACIONES ENTRE EL ANÁLISIS DE CONGLOMERADOS, EL ANÁLISIS DE COMPONENTES PRINCIPALES Y EL ANÁLISIS FACTORIAL.

El AC es una técnica estadística clasificadora, pero, en realidad, es una técnica que, como el ACP o como el AF, pretenden representar una realidad en la que no se consigue visualizar, una realidad cuya representación original es multidimensional y es imposible que la podamos ver en su estado puro.

En realidad tanto el ACP, el AF y el AC son técnicas que tratan de representar una nube de puntos originales situada en un espacio de tantas dimensiones que es imposible visualizar. Y cada una de ellas, también, pueden ser usadas como métodos clasificatorios, como métodos para crear subpoblaciones, subgrupos o subsubgrupos.

Las diferencias fundamentales entre ellas es la forma de presentación que utilizan y la forma de resolver el problema de no visualización de la nube de puntos originales. El ACP y el AF construyen una nube de puntos de la misma naturaleza pero en menor número de dimensiones perdiendo una parte de la información original. En cambio, el AC crea una representación distinta a la de la nube de puntos. Crea otro tipo de representación, cambia la forma, no lo hace mediante una nube de puntos, lo hace mediante un dendograma, pero cada una de estas opciones tiene sus ventajas y sus desventajas.

El ACP y el AF respetan el tipo de representación de una nube de puntos, pero al reducir dimensiones se pierde información y esto es un problema, especialmente si la pérdida es importante. El AC respeta la nube de puntos originales, no reduce dimensiones y no se pierde información, pero sí, se cambia el mecanismo de representación. Esta se representa mediante un dendograma. Se puede decir que en el ACP y el AF se hace una representación figurativa y en el AC se hace una representación abstracta.

En el AC se define una noción de distancia entre puntos, se necesita elegir una distancia, una medida que cuantifique distancias entre los individuos dentro de la nube de puntos originales, y aquí aparece el primer problema del AC, porque existen muchas distancias propuestas.

En el AC la distancia euclídea es la más utilizada, que calcula la distancia en línea recta entre los puntos en el espacio o en el hiperespacio de la nube de puntos originales, siendo esta distancia en realidad una aplicación del teorema de Pitágoras.

La distancia Mahalanobis es de mucho prestigio en estadística, se trata de una distancia que toma en cuenta las distancias que hay entre cada una de las variables y las relativiza respecto a la dispersión que tiene cada una de estas variables originales.

Estos temas puede el lector profundizar, consultando la bibliografía señalada en las revistas Varianza Nº 10, 12 y 15, del IETA, Carrera de Estadística - FCPN de la UMSA.

11. TRABAJO COMPUTACIONAL

El AC de un pequeño conjunto de datos es relativamente simple, se vuelve arduo cuando (n) el tamaño de la muestra es grande, como también (m) el número de variables. Además, las rutinas gráficas para construir dendogramas se vuelven muy complejas, por estas razones, se desarrollo el programa computacional utilizado (8), que permite calcular en base a la matriz de observaciones, la matriz estandarizada, la matriz de distancias, la agrupación media ponderada de grupos por pares y la construcción del dendograma; también, se tiene la opción de realizar estos cálculos determinando la matriz de correlación, haciéndose notar que los resultados de la agrupación media ponderada de grupos por pares y la construcción del dendograma son diferentes en estos dos casos. El dendograma resultante con el programa computacional es un diagrama de árbol que muestra los grupos que se forman al crear conglomerados de variables en cada paso y sus niveles de similitud.

11. CONCLUSIONES

Una característica de este trabajo, es que necesita poca explicación. Las muestras observadas con sus respectivas variables constituyen la matriz de observaciones, las cuales son estandarizadas, cuando las variables medidas no son directamente comparables, cuyos valores son independientes en las unidades empleadas, son adimensionales, se caracterizan por tener una media igual a cero y una desviación estándar igual a la unidad, de esta manera se permite comparar los datos procedentes de diferentes muestras o poblaciones. El coeficiente de correlación indica mayor similitud a valores absolutos altos, mientras que el coeficiente de distancia indica mayor similitud a la menor distancia. Por lo tanto, las correlaciones deben estar vinculadas o conectadas a un valor alto, y los coeficientes de distancia deben estar unidos a valores bajos.

La matriz de distancia parece ser menos susceptible a cambios drásticos entre los diferentes métodos de agrupamiento. Sin embargo, no hay pruebas estadísticas disponibles para este método de agrupación, ni se ha desarrollado ninguna teoría estadística y aplicada.

En el AC, se puede conglomerar variables que permitan identificar grupos homogéneos, en este caso, las unidades usadas para el análisis son las variables y las medidas de distancia se calculan para todos los pares de variables de acuerdo a los valores de la matriz de coeficientes de correlación, cuyos valores se usan como medida de semejanza (lo opuesto a la distancia) entre las variables.

La conglomeración jerárquica de variables ayuda a identificar variables que hacen una contribución única de datos, la conglomeración también puede usarse para reducir el número de variables en el conglomerado, llamada componente del conglomerado; a menudo se reemplaza un conjunto grande de variables con un conjunto de componentes de conglomerados con poca pérdida de información. Se conglomeran las variables, porque es más sencillo interpretar los componentes conglomerados que en el análisis de componentes principales. El gran beneficio del AC es que proporciona una forma de clasificar los objetos que es relativamente simple y directo, y presenta los

resultados de una manera familiar y fácil de entender.

Luego de obtener los resultados, se concluye que los métodos multivariantes del AC y el análisis factorial, ayudan a reducir la información proporcionada. De esta manera facilitar la toma de decisiones en diferentes estudios, permitiendo analizar fácilmente una serie de variables agrupándolas para poder simplificar los estudios de mercado, investigación de productos, publicidad, estudios sobre precios, etc.

12. APLICACIONES DEL ANÁLISIS DE CONGLOMERADOS

- Segmentación del mercado: por ejemplo, puede agruparse a los consumidores según los beneficios que buscan en la compra de un producto. Cada conglomerado estaría formado por consumidores que son relativamente homogéneos en términos de los beneficios que buscan. Este procedimiento se conoce como segmentación por beneficios.

- Entender la conducta de los compradores: el AC puede usarse para identificar grupos homogéneos de compradores. Luego se examina por separado la conducta de compras de cada grupo. El AC también se ha empleado para identificar las estrategias que usan los compradores de automóviles cuando buscan información externa.

- Identificar oportunidades de nuevos productos: al agrupar marcas y productos, es posible determinar conjuntos competitivos dentro del mercado. Las marcas del mismo conglomerado compiten mucho más entre sí que con las marcas de otros conglomerados. Una empresa puede comparar sus ofertas actuales con las de sus competidores para identificar posibles oportunidades de productos nuevos.

- Elegir mercados de prueba: al agrupar ciudades en conglomerados homogéneos, es posible elegir ciudades comparables para probar diversas estrategias de marketing y poder clasificar a los consumidores.

- Reducir los datos: el AC es útil como herramienta general de reducción de datos, para desarrollar conglomerados o subgrupos de datos que sean más fáciles de manejar que las observaciones individuales. El análisis multivariado posterior no se realiza en las observaciones individuales, sino en los conglomerados. Por ejemplo, para describir las diferencias en la conducta de uso del producto por parte de los consumidores, primero se dividiría a éstos en conglomerados luego, las diferencias entre los grupos se examinaría con el análisis discriminante múltiple.

En Geología, Minas y Metalurgia: se usa el AC para clasificar los minerales por su tamaño, pureza, explotación, etc., para formar grupos de pixels en imágenes digitalizadas enviadas por un satélite desde un planeta a la tierra para identificar los terrenos.

En Odontología, Medicina y Bioquímica: se usa el AC para clasificar las diferentes clases de dientes, seres vivos con los mismos síntomas y características patológicas, remedios, tipos de enfermedades, compuestos químicos, etc.

En Agronomía: para comparar los productos agrícolas ya sea de una misma o de diferentes

especies, tanto de semillas como del producto cosechado, rendimiento de producción, etc.

En la taxonomía: el AC se utiliza para agrupar especies naturales.

Problema. En base a la fuente de datos obtenida para fines ilustrativos, se tienen los datos del análisis petrográfico de veinte muestras mineralógicas con ocho variables cada una, con contenido de óxidos en rocas ígneas, se desea obtener la matriz estandarizada, la matriz de distancias (o también la matriz de correlación), la agrupación media ponderada de grupos por pares, el dendograma tanto para las muestras mineralógicas como para las variables.

Datos:

Cuadro Nº 1

Matriz de observaciones de 20 muestras con 8 variables

Nombre de la muestra	Nº de muestra	SiO_2	Al_2O_3	Fe_2O_3	FeO	MgO	CaO	Na_2O	K_2O
		1	2	3	4	5	6	7	8
Sienita	1	61,7	15,1	2	2,3	3,7	4,6	4,4	4,5
Sienita	2	58,3	17,9	3,2	1,7	1,5	3,7	5,9	5,3
Sienita	3	51,2	17,6	3,5	4,3	3,2	4,5	5,7	4,4
Monzonita	4	54,4	14,3	3,3	4,1	6,1	7,7	3,4	4,2
Diorita	5	58	15,7	0,7	2,8	5	10,9	3	3,2
Diorita	6	46,9	15,9	2,9	10	7	9,6	2,7	0,7
Diorita	7	58	*17,3*	2,2	3,8	2,2	4,3	4,3	4,1
Cuarzo diorita	8	55,5	16,5	2,7	4,6	6,7	6,7	3,2	2,5
Gabro	9	55,4	15,3	2,7	5,5	5,8	9,9	2,9	1,5
Gabro	10	55,9	13,5	2,7	5,9	6,5	8,9	2,4	1,7
Norita	11	47,2	14,5	1,6	13,8	5,2	8,1	3,1	1,2
Norita	12	48,2	18,3	1,3	6,1	10,8	9,4	1,3	0,7
Hiperesteno gabro	13	44,8	18,8	2,2	4,7	11,3	14,6	0,9	0,1
Hiperesteno gabro	14	47	14,1	0,8	15	16	2,3	0,4	1,7
Sienita	15	59,8	17,3	3,6	1,6	1,2	3,8	5	5,1
Cuarzo sienita	16	66,2	16,2	2	0,2	0,8	1,3	6,5	5,8
Sienita alterada	17	50	9,9	3,5	5	11,9	8,3	2,4	5
Monzonita	18	57,4	18,5	3,7	2,1	1,7	6,8	4,5	3,7
Monzonita	19	59,8	15,8	3,8	3,3	2,2	3,9	3	4,4
Diabase	20	52,2	18,2	3,3	4,4	4,7	6,5	4,6	1,9

Fuente: Elaboración Propia

Solución:

Cuadro Nº 2
Matriz de observaciones estandarizadas

	1	2	3	4	5	6	7	8
1	1,10	0,91	0,72	0,36	0,52	0,60	1,12	1,24
2	1,04	1,08	1,15	0,26	0,21	0,48	1,50	1,46
3	0,91	1,06	1,26	0,67	0,45	0,58	1,45	1,21
4	0,97	0,86	1,19	0,64	0,86	1,00	0,87	1,16
5	1,03	0,95	0,25	0,43	0,70	1,42	0,76	0,88
6	0,84	0,96	1,04	1,55	0,98	1,25	0,69	0,19
7	1,03	1,04	0,79	0,59	0,31	0,56	1,09	1,13
8	0,99	0,99	0,61	0,71	0,94	0,87	0,81	0,69
9	0,99	0,92	0,97	0,85	0,81	1,29	0,74	0,41
10	1,00	0,81	0,97	0,92	0,91	1,16	0,61	0,47
11	0,84	0,87	0,58	2,14	0,73	1,05	0,79	0,33
12	0,86	1,10	0,47	0,95	1,52	1,22	0,33	0,19
13	0,80	1,13	0,79	0,73	1,59	1,90	0,23	0,03
14	0,84	0,85	0,29	2,33	2,25	0,30	0,10	0,47
15	1,07	1,04	1,30	0,25	0,17	0,49	1,27	1,40
16	1,18	0,98	0,72	0,03	0,11	0,17	1,65	1,60
17	0,89	0,60	1,26	0,78	1,67	1,00	0,61	1,38
18	1,02	1,12	1,33	0,33	0,24	0,88	1,15	1,02
19	1,07	0,95	1,37	0,51	0,31	0,51	0,76	1,21
20	0,93	1,10	1,19	0,68	0,66	0,84	1,17	0,52

Fuente: Elaboración Propia

Cuadro Nº 3
Matriz de distancias

	1	2	3	4	5	6	7	8	9	10
1	0,000	0,258	0,265	0,288	0,385	0,664	0,132	0,318	0,463	0,463
2	0,258	0,000	0,201	0,415	0,597	0,805	0,257	0,531	0,622	0,638
3	0,265	0,201	0,000	0,303	0,552	0,632	0,222	0,423	0,489	0,502
4	0,288	0,415	0,303	0,000	0,389	0,490	0,305	0,274	0,306	0,295
5	0,385	0,597	0,552	0,389	0,000	0,559	0,416	0,275	0,344	0,367
6	0,664	0,805	0,632	0,490	0,559	0,000	0,613	0,406	0,273	0,263
7	0,132	0,257	0,222	0,305	0,416	0,613	0,000	0,320	0,439	0,445
8	0,318	0,531	0,423	0,274	0,275	0,406	0,320	0,000	0,230	0,217
9	0,463	0,622	0,489	0,306	0,344	0,273	0,439	0,230	0,000	0,088
10	0,463	0,638	0,502	0,295	0,367	0,263	0,445	0,217	0,088	0,000
11	0,747	0,889	0,724	0,649	0,661	0,297	0,678	0,534	0,489	0,472

12	0,671	0,880	0,754	0,546	0,465	0,376	0,673	0,360	0,358	0,332
13	0,819	0,998	0,877	0,641	0,559	0,474	0,827	0,541	0,434	0,440
14	1,052	1,236	1,086	0,947	0,992	0,716	1,037	0,817	0,873	0,811
15	0,262	0,100	0,211	0,381	0,592	0,780	0,251	0,519	0,592	0,604
16	0,332	0,229	0,396	0,581	0,672	0,959	0,365	0,635	0,772	0,781
17	0,553	0,691	0,581	0,332	0,570	0,580	0,601	0,459	0,491	0,443
18	0,281	0,255	0,224	0,291	0,483	0,635	0,248	0,418	0,429	0,459
19	0,280	0,305	0,267	0,279	0,541	0,641	0,241	0,422	0,472	0,462
20	0,351	0,434	0,290	0,278	0,446	0,420	0,308	0,269	0,257	0,288

Fuente: Elaboración Propia

Cuadro Nº 3
Matriz de distancias continuación

	11	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
1	0,747	0,671	0,819	1,052	0,262	0,332	0,553	0,281	0,280	0,351
2	0,889	0,880	0,998	1,236	0,100	0,229	0,691	0,255	0,305	0,434
3	0,724	0,754	0,877	1,086	0,211	0,396	0,581	0,224	0,267	0,290
4	0,649	0,546	0,641	0,947	0,381	0,581	0,332	0,291	0,279	0,278
5	0,661	0,465	0,559	0,992	0,592	0,672	0,570	0,483	0,541	0,446
6	0,297	0,376	0,474	0,716	0,780	0,959	0,580	0,635	0,641	0,420
7	0,678	0,673	0,827	1,037	0,251	0,365	0,601	0,248	0,241	0,308
8	0,534	0,360	0,541	0,817	0,519	0,635	0,459	0,418	0,422	0,269
9	0,489	0,358	0,434	0,873	0,592	0,772	0,491	0,429	0,472	0,257
10	0,472	0,332	0,440	0,811	0,604	0,781	0,443	0,459	0,462	0,288
11	0,000	0,544	0,704	0,659	0,880	1,007	0,744	0,777	0,758	0,591
12	0,544	0,000	0,286	0,663	0,860	0,989	0,552	0,729	0,730	0,534
13	0,704	0,286	0,000	0,872	0,970	1,136	0,629	0,804	0,847	0,640
14	0,659	0,663	0,872	0,000	1,226	1,316	0,825	1,171	1,079	0,971
15	0,880	0,860	0,970	1,226	0,000	0,293	0,666	0,205	0,223	0,414
16	1,007	0,989	1,136	1,316	0,293	0,000	0,827	0,451	0,469	0,598
17	0,744	0,552	0,629	0,825	0,666	0,827	0,000	0,612	0,556	0,546
18	0,777	0,729	0,804	1,171	0,205	0,451	0,612	0,000	0,222	0,270
19	0,758	0,730	0,847	1,079	0,223	0,469	0,556	0,222	0,000	0,350
20	0,591	0,534	0,640	0,971	0,414	0,598	0,546	0,270	0,350	0,000

Fuente: Elaboración Propia

Agrupación media ponderada de grupos por pares:

Columna 1 y 2 = observaciones combinadas, dentro el grupo

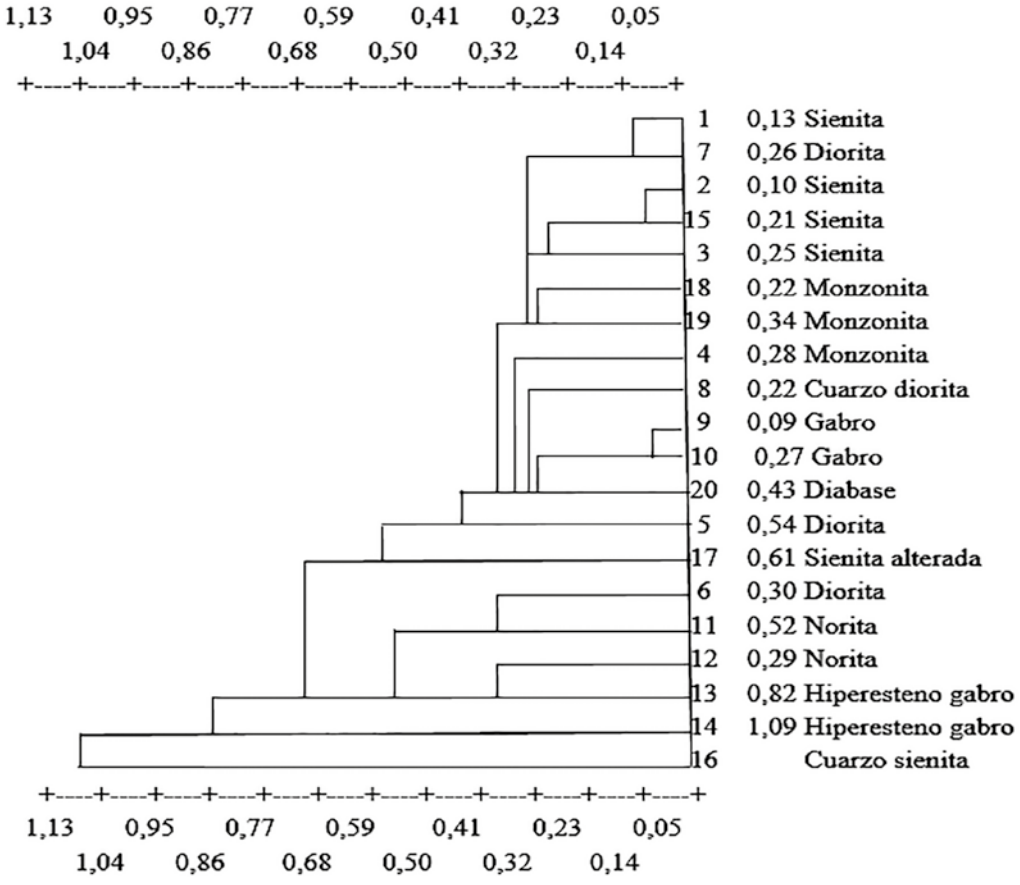
Columna 3 = nivel de correlación del agrupamiento.

Cuadro Nº 4
Agrupación media ponderada de grupos por pares

	1	2	3
1	1	7	0,1323
2	2	15	0,1003
3	9	10	0,0876
4	12	13	0,2862
5	2	3	0,2060
6	8	9	0,2235
7	18	19	0,2219
8	2	18	0,2462
9	6	11	0,2969
10	8	20	0,2709
11	1	2	0,2564
12	4	8	0,2825
13	1	4	0,3441
14	1	5	0,4283
15	1	17	0,5374
16	6	12	0,5244
17	1	6	0,6130
18	1	14	0,8201
19	1	16	1,0899

Fuente: Elaboración Propia

Figura Nº 3
Dendograma para las muestras mineralógicas



Fuente: Elaboración Propia

En este dendograma se observa los grupos que se forman entre las muestras de minerales, cuanto más cerca se encuentran, mayor es la similitud y cuanto más alejadas estén no se tendrá una similitud entre ellas, a pesar de que todas tienen cierta afinidad por el contenido de óxidos en cada muestra, lo cual conlleva a formar grupos afines. Para mayor comprensión de este dendograma, se debe leer de derecha a izquierda.

Trazando una línea vertical de corte imaginaria al nivel de similitud de aproximadamente 0,10; se tienen tres conglomerados, el primer conglomerado (extremo derecho) se compone de dos observaciones (las observaciones de filas 1 y 7 de la matriz de datos), el segundo conglomerado a la derecha se compones de dos observaciones 2 y 17 y el tercer conglomerado también se compone de dos observaciones 9 y 10.

Si se traza la línea vertical de corte imaginaria al nivel de similitud de aproximadamente 0,40 más a la izquierda, se tendrá cuatro conglomerados, y así sucesivamente. De manera que si se corta el dendograma mucho más a la izquierda, habrá menos conglomerados finales. Los tamaños de los conglomerados deben ser significativos, así en el dendograma resultante se observa tres conglomerados, generados como resultado de 6, 6, y 7 elementos, no teniendo sentido formar un conglomerado con un solo caso, como ocurre con los dos últimos.

En cuanto a la decisión sobre el número de conglomerados, no hay reglas exactas, ni rápidas, pero si existen algunos lineamientos:

a. Las consideraciones, conceptuales o prácticas pueden sugerir cierto número de conglomerados.

b. En los procedimientos de conglomeración jerárquica, puede usarse como criterio las distancias en las que se combinan los conglomerados.

Utilizando la matriz de correlación, se obtendrá el dendograma en función de las variables, luego de estandarizar la información, los cálculos y el gráfico, son los siguientes:

Cuadro Nº 5
Matriz de coeficientes de correlación

	1	2	3	4	5	6	7	8
1	1,000	0,075	0,176	-0,744	-0,759	-0,554	0,691	0,747
2	0,075	1,000	0,045	-0,332	-0,385	-0,009	0,302	-0,128
3	0,176	0,045	1,000	-0,413	-0,437	-0,149	0,434	0,412
4	-0,744	-0,332	-0,413	1,000	0,648	0,156	-0,633	-0,648
5	-0,759	-0,385	-0,437	0,648	1,000	0,410	-0,872	-0,593
6	-0,554	-0,009	-0,149	0,156	0,410	1,000	-0,579	-0,683
7	0,691	0,302	0,434	-0,633	-0,872	-0,579	1,000	0,732
8	0,747	-0,128	0,412	-0,648	-0,593	-0,683	0,732	1,000

Fuente: Elaboración Propia

Cuadro Nº 6
Matriz de distancias

	1	2	3	4	5	6	7	8
1	0,000	0,155	0,344	0,681	0,664	0,487	0,355	0,425
2	0,155	0,000	0,363	0,656	0,644	0,444	0,395	0,522
3	0,344	0,363	0,000	0,793	0,783	0,577	0,403	0,463
4	0,681	0,656	0,793	0,000	0,478	0,664	0,898	0,961
5	0,664	0,644	0,783	0,478	0,000	0,551	0,941	0,929
6	0,487	0,444	0,577	0,664	0,551	0,000	0,730	0,820
7	0,355	0,395	0,403	0,898	0,941	0,730	0,000	0,332
8	0,425	0,522	0,463	0,961	0,929	0,820	0,332	0,000

Fuente: Elaboración Propia

Agrupación media ponderada de grupos por pares:

Columna 1 y 2 = observaciones combinadas, dentro del grupo

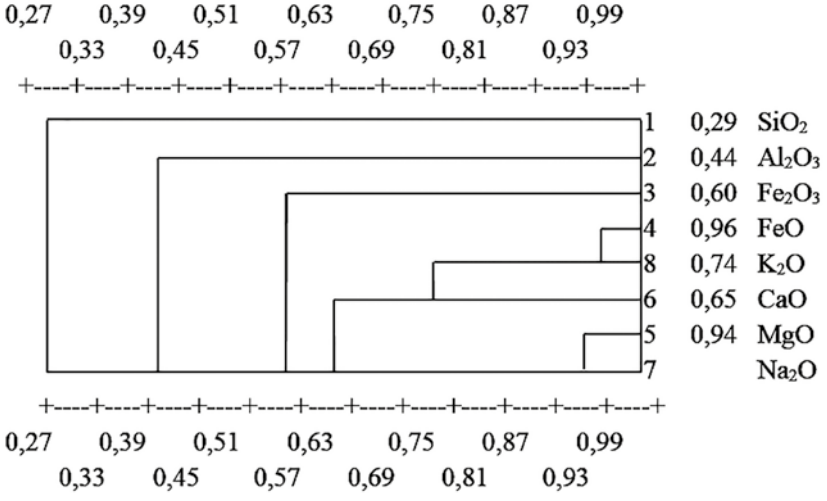
Columna 3 = nivel de correlación del agrupamiento

Cuadro Nº 7
Agrupación media ponderada de grupos por pares

	1	2	3
1	4	8	0,9607
2	5	7	0,9412
3	4	6	0,7420
4	4	5	0,6500
5	3	4	0,5978
6	2	3	0,4405
7	1	2	0,2921

Fuente: Elaboración Propia

Figura Nº 4
Dendograma para las variables



Fuente: Elaboración Propia

Los valores a lo largo del eje X representan las similitudes.

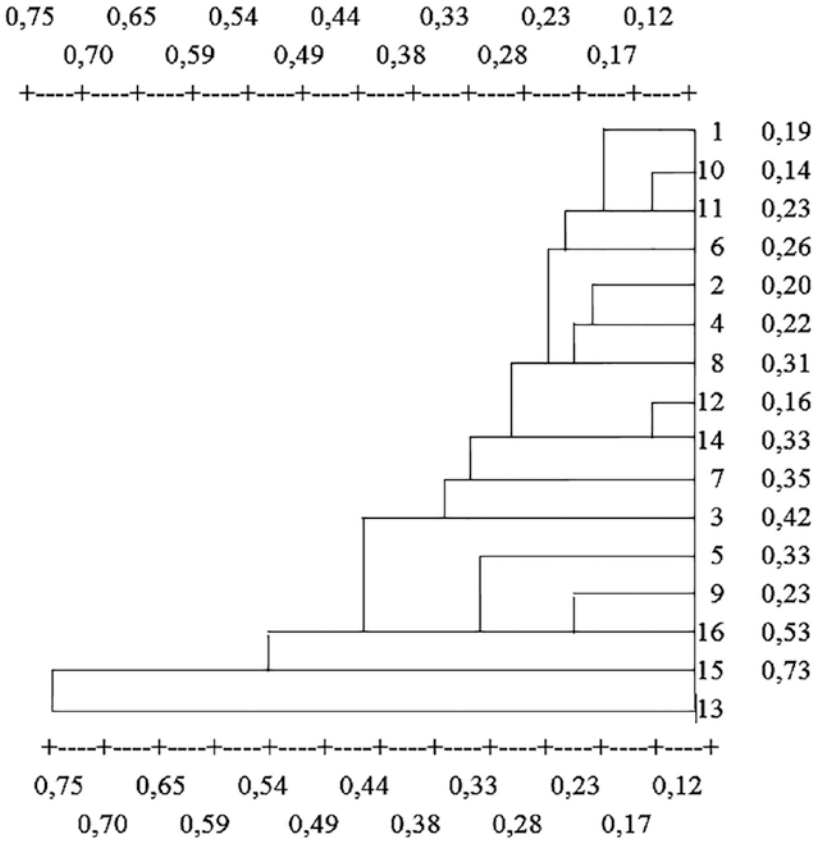
En este dendograma se observa los grupos que se forman entre las variables de los óxidos.

Problema. En base a la fuente de datos obtenida de la Sección de Bioestadística del S.S.U. de la UMSA para fines ilustrativos, se tienen algunos datos del análisis de sangre de diez y seis muestras con quince variables cada una, de pacientes que tienen diabetes.

La matriz de datos observados es de 16 por 15, y siendo los cálculos semejantes a los del problema anterior debido a que se utilizo el mismo programa computacional, solo se presentará los dendogramas respectivos de muestras y variables como se demuestra a continuación:

Figura Nº 5

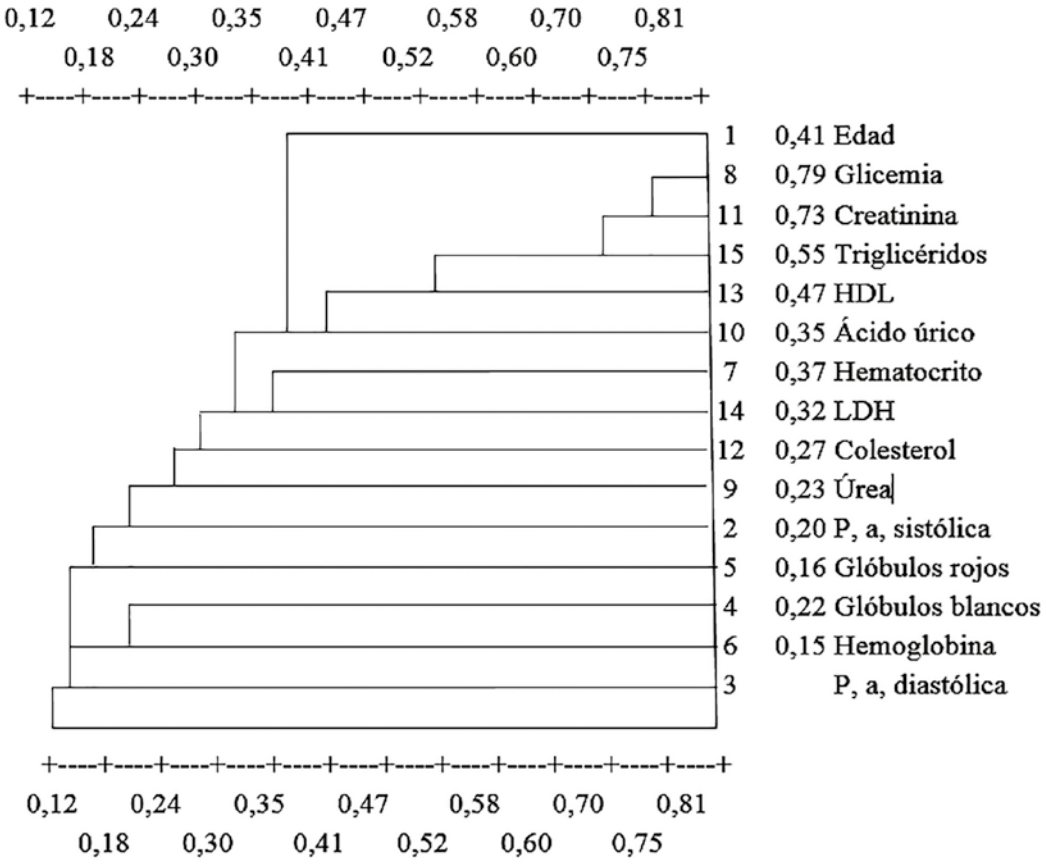
Dendograma para las variables



Fuente: Elaboración Propia

Figura Nº 6

Dendograma para las muestras de enfermos con diabetes



Fuente: Elaboración Propia

Colaboración.

Ing. Guillermo Salinas Reguerin
Ingeniero Comercial. Universidad Loyola
La Paz – Bolivia (2004)
Unidad de Bioestadística, SSU, UMSA

BIBLIOGRAFÍA

1. BENZÉCRI J. P. & Collaborateurs. L´Analyse Des Donnees, La Taxinomia. Ed. Dunod, Paris, Francia, 1973, pp. 615 – XIII.

2. DAGNELIE Pierre, Théorie et Méthodes Statistiques. (volume 1). Les Presses Agronomiques de Gembloux, Gembloux, Belgica, 1973 (2da. Edición), pp. 378 – IX.

3. DAGNELIE Pierre, Analyse Statistique á Plusieurs Variables. Les Presses Agronomiques de Gembloux, Gembloux, Belgica, 1975 (2da. Edición), pp. 362 – XIV.

4. DAVIS C. John. Statistics and Data Analysis in Geology. John Wiley & Sons, Inc., Toronto, Canada,1973, pp. 550 – VII.

5. KLOCKMANN F. y RAMDOHR P., Tratado de Mineralogia (Versión en aleman).

Editorial Gustavo Gili, Barcelona, España, 1961, pp. 736 – IX.

6. KRUMBEIN W. C. and GRAYBILL A. Franklin, An Introduction to Statistical Models in Geology. McGraw-Hill Book Company, Toronto,Canada, 1965, pp. 475 – XV.

7. RUIZ Aranibar Gustavo. Factores que Inciden en el Rendimiento Académico y Evaluación Docente. U.A.G.R.M. Santa Cruz – Bolivia. 2010, pp. 190 – IX.

8. RUIZ Aranibar Gustavo[2] . Libreria Científica de Programas Informáticos, La Paz -Bolivia.

" Todo trabajo de investigación es el resultado del esfuerzo que se realiza con: orden, voluntad, paciencia y estudio constante, para dar a conocer a nuestros semejantes, sabiendo que el final de los trabajos de investigación que se realizan, son el comienzo de las investigaciones que otros proseguirán en forma más profunda"

Dr. Cs Gustavo Ruiz Aranibar

[2] Calle 20 y Av. Ballivian, Nº 8035, Calacoto, La Paz – Bolivia, Tel. 591-22772162 Cel, 67111778
gustavoruiz432@hotmail.com.bo ruizaranibargustavo@gmail.com.bo Blog: Gustavo Ruiz Aranibar

Calle 27 de Cota Cota
Bloque F.C.P.N. - Primer Piso

La Paz - Bolivia