UNIVERSIDAD NACIONAL AGRARIA LA MOLINA

FACULTAD DE ECONOMÍA Y PLANIFICACIÓN

DEPARTAMENTO DE ESTADÍSTICA E INFORMÁTICA

TRABAJO MONOGRÁFICO

CLASIFICACIÓN DE FAMILIAS EN CAJAMARCA SEGÚN SU SITUACIÓN ECONÓMICA MEDIANTE EL ANÁLISIS DE CONGLOMERADOS

Presentado para optar el Título de Ingeniero Estadístico e Informático

JUANA MERCEDES VICENTE VASQUEZ

Modalidad de Examen Profesional

LIMA-PERÚ

2014

ÍNDICE

RESUMEN ii

I. INTRODUCCIÓN ……..……1

II. OBJETIVOS ……..……4

III. METODOLOGÍA ……..……..5

IV. RESULTADOS ……..……16

V. CONCLUSIONES ……..……24

VI. BIBLIOGRAFÍA ……..……..……26

RESUMEN

El estudio tiene como objetivo clasificar a las familias encuestadas en los distritos de San Pablo, San Luis y San Bernardino de la provincia de San Pablo en el departamento de Cajamarca según un conjunto de variables socio-económicas.

Estos datos corresponden a una investigación realizada por un grupo de personas que laboran en la Universidad del Pacifico, la encuesta fue realizada en Diciembre del 2006.

Se desea clasificar a las familias para poder brindar un mejor control en el estudio longitudinal de los proyectos a ser evaluados. Para esto, al culminar la encuesta se planteó una clasificación preliminarmente la existencia de 4 grupos de familias. Para verificar esta clasificación se utilizó el "**Análisis de Clúster"**, que es un método multivariado de clasificación.

Para el procesamiento de los datos se utilizó el programa "Minitab versión 17" y "Microsoft Excel"

I. INTRODUCCIÓN

En Diciembre del 2006 se realizó una encuesta cuyo objetivo fue hallar una descripción actualizada sobre un conjunto de familias en el departamento de Cajamarca, con el fin de medir el impacto de un determinado grupo de proyectos que se llevarían a cabo en el lapso de 2 a 3 años. El inconveniente que se presentó es que los proyectos no tenían claramente especificados la zona donde se realizará sus capacitaciones, esto conllevo a realizar una clasificación de las familias para dicha distribución de capacitaciones. La cantidad de familias encuestadas fueron de 170, de esta cantidad se descartaron 3 porqué serían familias repetidas. Estas tres encuestas fueron eliminadas porque presentaban el mismo nombre del jefe del hogar y de la esposa. Finalmente la base de datos con la que se trabajó fue de 167 familias.

Para realizar esta clasificación de familia se utilizaron las siguientes variables: miembros de hogar, miembros que trabajan, años que trabajan en la actividad principal, total de área agrícola (Has), número de viviendas que poseen, número de cuartos que posee la vivienda, área de la vivienda, número de pisos que tiene la vivienda, pago mensual de agua (S/.) pago mensual por combustible para alumbrado (S/.).

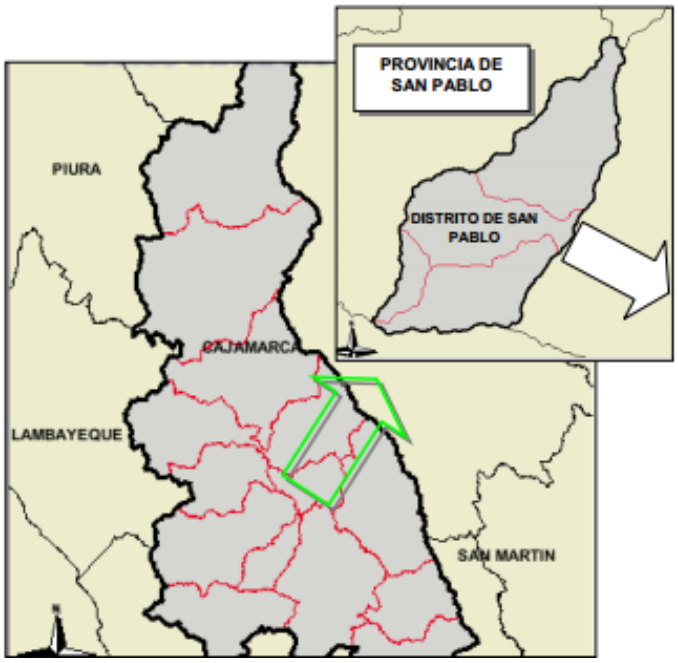
Las encuestas se realizaron en 4 días con la ayuda de 10 encuestadores a cargo de 3 investigadores responsables. Esta fue realizada en la provincia de San Pablo, abarcando 3 distritos muy cercanos; San Pablo, San Luis y San Bernardino (figura N° 01). Los encuestadores empezaron a realizar las encuestas a las 5 am hora en el cual los jefes de hogares salían a realizar sus actividades económicas. Esto conllevo a no tener una ubicación por lugar de residencia, por este motivo al realizar la

clasificación de las familias, se logró identificar las zonas adecuadas para la realización de los proyectos a implementarse en el lugar de estudio.

Por los datos y las experiencias dadas en otros proyectos se pudo observar que existían por lo menos 4 grupos de familias, por este motivo se aplicó un método multivariado de clasificación no jerárquico.

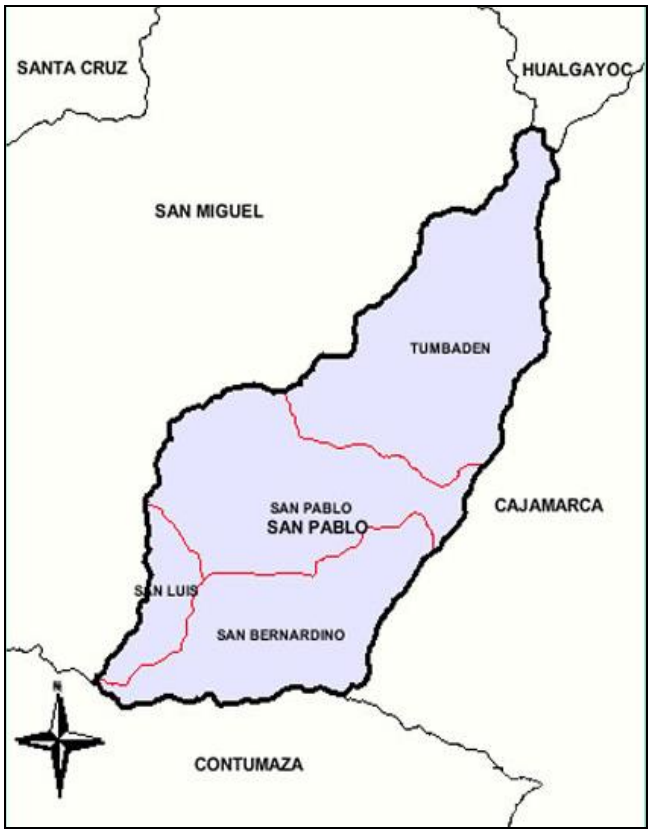
Dado que el fin del estudio es ayudar a clasificar a dichas familias para poder realizar una evaluación sobre el impacto de los proyectos que están en dicha zona, es necesario dar una adecuada recomendación para una buena distribución de las capacitaciones que brindaran en el conjunto de los proyectos.

Figura N° 1: Ubicación de la Provincia de San Pablo



Fuente: INEI

Figura N°2 Mapa de la Provincia de San Pablo



Fuente: http://www.perutoptours.com/index06sp_mapa_san_pablo.html

II. OBJETIVOS

OBJETIVO GENERAL:

Clasificar a las familias encuestadas en los distritos de San Pablo, San Luis y San Bernardino de la provincia de San Pablo en el departamento de Cajamarca según un conjunto de variables socio-económicas.

OBJETIVOS ESPECÍFICOS:

- Indicar que variables son significativas en la formación de los conglomerados.

- Aplicar la metodología del análisis multivariado de conglomerados para determinar el agrupamiento de familias en Cajamarca, según las variables Socio-económicas.

- Caracterizar a cada uno de los grupos hallados con el objetivo que ayuden a la implementación o mejoramiento de los proyectos planificados para la zona, como son las capacitaciones en diferentes ámbitos.

III. METODOLOGÍA

3.1. BASE TEÓRICA

El Análisis Cluster, conocido como Análisis de Conglomerados, es una técnica estadística multivariante que busca agrupar elementos (o variables) tratando de lograr la máxima homogeneidad en cada grupo y la mayor diferencia entre los grupos.

El Análisis Cluster tiene una importante tradición de aplicación en muchas áreas de investigación. Sin embargo, junto con los beneficios del Análisis Cluster existen algunos inconvenientes. El Análisis Cluster es una técnica descriptiva, a teórica y no inferencial.

El Análisis Cluster no tiene bases estadísticas sobre las que deducir inferencias estadísticas para una población a partir de una muestra, es un método basado en criterios geométricos y se utiliza fundamentalmente como una técnica exploratoria, descriptiva pero no explicativa.

Las soluciones no son únicas, en la medida en que la pertenencia al conglomerado para cualquier número de soluciones depende de muchos elementos del procedimiento elegido. Por otra parte, la solución cluster depende totalmente de las variables utilizadas, la adición o destrucción de variables relevantes puede tener un impacto substancial sobre la solución resultante.

Los algoritmos de formación de conglomerados se agrupan en dos categorías:

Algoritmos de partición: Método de dividir el conjunto de observaciones en k conglomerados (clusters), en donde k lo define inicialmente el usuario.

Algoritmos jerárquicos: Método que entrega una jerarquía de divisiones del conjunto de elementos en conglomerados.

Un método jerárquico aglomerativo parte con una situación en que cada observación forma un conglomerado y en sucesivos pasos se van uniendo, hasta que finalmente todas las situaciones están en un único conglomerado.

Un método jerárquico disociativo sigue el sentido inverso, parte de un gran conglomerado y en pasos sucesivos se va dividiendo hasta que cada observación queda en un conglomerado distinto.

El análisis de conglomerados nos va a permitir contestar a preguntas tales como:

¿Es posible identificar cuáles son las empresas en las que sería más deseable invertir?

¿Es posible identificar grupos de clientes a los que les pueda interesar un nuevo producto que una empresa va a lanzar al mercado?

¿Se pueden clasificar las bodegas de La Ribera del Duero en función de las características químicas y ópticas del vino que producen?

Los llamados métodos jerárquicos tienen por objetivo agrupar clusters para formar un nuevo o bien separar alguno ya existente para dar origen a otros dos, de tal forma que, si sucesivamente se va efectuando este proceso de aglomeración o división, se minimice alguna distancia o bien se maximice alguna medida de similitud.

Los métodos jerárquicos se subdividen en aglomerativos y disociativos. Cada una de estas categorías presenta una gran diversidad de variantes.

1. Los métodos aglomerativos, también conocidos como ascendentes, comienzan el análisis con tantos grupos como individuos haya. A partir de estas unidades iniciales se van formando grupos, de forma ascendente,

hasta que al final del proceso todos los casos tratados están englobados en un mismo conglomerado.

2. Los métodos disociativos, también llamados descendentes, constituyen el proceso inverso al anterior.

Comienzan con un conglomerado que engloba a todos los casos tratados y, a partir de este grupo inicial, a través de sucesivas divisiones, se van formando grupos cada vez más pequeños. Al final del proceso se tienen tantas agrupaciones como casos han sido tratados.

Los Métodos Jerárquicos Aglomerativos.

Algunas de las estrategias que pueden ser empleadas a la hora de unir los clusters en las diversas etapas o niveles de un procedimiento jerárquico. Ninguno de estos procedimientos proporciona una solución óptima para todos los problemas que se pueden plantear, ya que es posible llegar a distintos resultados según el método elegido. El buen criterio del investigador, el conocimiento del problema planteado y la experiencia, sugerirán el método más adecuado. De todas formas, es conveniente, siempre, usar varios procedimientos con la idea de contrastar los resultados obtenidos y sacar conclusiones, tanto como si hubiera coincidencias en los resultados obtenidos con métodos distintos como si no las hubiera.

3.2.1. Estrategia de la distancia mínima o similitud máxima.

Esta estrategia recibe en la literatura anglosajona el nombre de amalgamamiento simple (single linkage).

En este método se considera que la distancia o similitud entre dos clusters viene dada, respectivamente, por la mínima distancia (o máxima similitud) entre sus componentes.

Así, si tras efectuar la etapa K- ésima, tenemos ya formados n− K clusters, la distancia entre los clusters C_i (con n_i elementos) y C_j (con n_j elementos) sería:

$$d(C_i, C_j) = Min \quad d(C_{i1}, C_{j1}) \quad , i_1, j_1 = 1, 2, ..n - k; i_1 \neq j_1$$

$$Min \quad Min \quad d(x_i, x_m) \quad i_1, j_1 = 1, 2, \ldots, n - k; x_i$$

Mientras que la similitud, si estuviéramos empleando una medida de tal tipo, entre los dos clusters sería:

$$s(C_i, C_j) = Max \quad s(x_i, x_m) \quad , x_i \in C_i; x_m \in C_j \; i = 1, 2, .., n_i; m = 1, 2, .., n_j$$

Con ello, la estrategia seguida en el nivel K + 1 será:

1. En el caso de emplear distancias, se unirán los clusters C_i y C_j si

$$d(C_i, C_j) = Min \quad d(C_{i1}, C_{j1}) = \underset{\substack{i_1, j_1 = 1, .., n-k \\ i_1 \neq j_1}}{}$$

$$d(C_i, C_j) = Min \quad \underset{\substack{x_i \in C_{i1} \\ x_m \in C_{j1}}}{Min} (d \quad x_i, x_m \quad , l = 1, \ldots, n_i; m = 1, .., n_{j1}$$

$$d(C_i, C_j) = Min \quad d(C_{i1}, C_{j1}) \quad , i_1, j_1 = 1, 2, ..n - k; i_1 \neq j_1$$

$$Min \quad Min \quad d(x_i, x_m) \quad i_1, j_1 = 1, 2, \ldots, n - k; x_i$$

2. En el caso de emplear similitudes, se unirán los clusters Ci y Cj si

$$s \quad C_i, C_j \quad = \underset{\substack{i_1, j_1 = 1, 2, \ldots, n-K \\ i_1 \neq j_1}}{Max} \quad s(C_{i1}, C_{j1})$$

MÉTODOS JERÁRQUICOS

El Análisis Cluster Jerárquico comienza separando cada objeto en un cluster por sí mismo. En cada etapa del análisis, el criterio por el que los objetos son separados se relaja en orden a enlazar los dos conglomerados más similares hasta que todos los objetos sean agrupados en un árbol de clasificación completo.

El criterio básico para cualquier agrupación es la distancia. Los objetos que estén cerca uno del otro pertenecerían al mismo conglomerado o cluster, y los objetos que estén lejos uno del otro pertenecerán a distintos clusters. Para un conjunto de datos dado, los clusters que se construyen dependen de nuestra propia especificación de los siguientes parámetros:

- El método cluster define las reglas para la formación del cluster. Por ejemplo, cuando calculamos la distancia entre dos clusters, podemos usar el par de objetos más cercado entre clusters o el par de objeto más alejados, o un compromiso entre estos métodos.
- La medida define la fórmula para el cálculo de la distancia. Por ejemplo, la medida de distancia Euclídea calcula la distancia como una línea recta entre dos clusters. Las medidas de intervalo asumen que las variables están medidas en escala; las medidas de conteo asumen que son números discretos, y las medidas binarias asumen que toman dos valores.
- La estandarización permite igualar el efecto de las variables medidas sobre diferentes escalas.

CLASIFICACIÓN:

Asociativos o Aglomerativos: Se parte de tantos grupos como individuos hay en el estudio y se van agrupando hasta llegar a tener todos los casos en un mismo grupo.

Disociativos: Se parte de un solo grupo que contiene todos los casos y a través de sucesivas divisiones se forman grupos cada vez más pequeños.

Los métodos jerárquicos permiten construir un árbol de clasificación o dendograma.

La técnica utilizada para este análisis es la de conglomerados o llamada también análisis de Clúster. La idea básica de esta técnica es de agrupar un conjunto de observaciones en un determinado Clúster o grupo. Dentro de este análisis se encuentran los Métodos No Jerárquicos que se usan para

agrupar objetos y no variables en un conjunto de k-clúster ya predeterminados. No se tiene que especificar una matriz de distancias ni tampoco almacenar las iteraciones como si lo requieren los Métodos Jerárquicos.

Cuando no se dispone de ningún tipo de información a priori, el análisis jerárquico sería una buena opción, hacer uso de las herramientas que nos ofrece para seleccionar el número de grupos y, con esta información realizar el análisis no jerárquico que permitirá maximizar la homogeneidad dentro de cada grupo y la heterogeneidad entre unos conglomerados y otros.

El método de k-Medias es un método no jerárquico que permite asignar a cada observación el clúster que se encuentra más próximo del centroide. Generalmente se emplean la distancia euclidea.

La distancia euclidea entre dos objetos L_1 y L_2 medidos por dos variables X_1 y X_2, es:

$$d_{L_1L_2} = \sqrt{x_{11} - x_{21}^2 + x_{12} - x_{22}^2}$$

Con las variables es equivalente a:

$$d_{L_1L_2} = \sqrt{\sum_{k=1}^{p} x_{1k} - x_{2k}^2}$$

MacQueen, en 1972, emplea el término K-Medias para denotar el proceso de asignar cada individuo al cluster (de los K prefijados) con el centroide más próximo. La clave de este procedimiento radica en que el centroide se calcula a partir de los miembros del cluster tras cada asignación y no al final de cada ciclo, como ocurre en los métodos de Forgy y Jancey.

El algoritmo que propuso es el siguiente:

1. Tomar los K primeros casos como clusters unitarios.

2. Asignar cada uno de los m − K individuos restantes al cluster con el centroide más próximo. Después de cada asignación, recalcular el centroide del cluster obtenido.

3. Tras la asignación de todos los individuos en el paso segundo, tomar los centroides de los clusters existentes como puntos semilla fijos y hacer una pasada más sobre los datos asignando cada dato al punto semilla más cercano.

   El último paso es el mismo que el del metido de Forgy, excepto que la recolocación se efectúa una vez más sin esperar a que se produzca la convergencia.

   Notemos que, usando los K primeros individuos como puntos semilla, este método tiene la virtud de ser el menos caro de todos los métodos discutidos. El cómputo total de operaciones desde la configuración inicial hasta la final involucra sólo K(2m−K) cálculos de distancias, (K −1)(2m−K) comparaciones de distancias y m − K cálculos de centroides.

   Hay que comentar que el conjunto de clusters construido en el paso segundo del algoritmo depende de la secuencia en la que los individuos han sido procesados. MacQueen (1967) efectuó algunos estudios preliminares en este sentido; su experiencia indicó que la ordenación de los datos tiene solamente un efecto marginal cuando los clusters están bien separados.

   El algoritmo utilizado por esta técnica es el de MacQueen(cuyo procedimiento teórico es:

1. Se toma al azar k clúster iniciales.
2. Para el conjunto de observaciones, se vuelve a calcular las distancias a los centroides de los clúster y se resignan a los que estén más próximos. Se vuelven a recalcular los centroides de los k clúster después de las reasignaciones de los elementos.

3. Se repiten los dos pasos anteriores hasta que no se produzca ninguna reasignación, es decir, hasta que los elementos se estabilicen en algún grupo.

   Usualmente, se especifican k centroides iniciales y se procede al paso (2) y, en la práctica, se observan la mayor parte de reasignaciones en las primeras iteraciones.

   A continuación se describirá un ejemplo que muestre la forma de aplicar el algoritmo de MacQueen a través del método de k-medias:

   En la columna inicial del Cuadro N° 01, se identifica la posición del dato de la variable en análisis, en la segunda columna se encuentra el valor del mismo. Luego se eligen inicialmente 2 centroides, ubicados en ese caso particular en las posiciones 2 y 7. En la columna con etiqueta distancia 1 se registra la distancia de cada dato al primer centroide. De igual forma, en la siguiente columna se registra la distancia de cada dato al siguiente centroide. Luego se escogen las distancias mínimas, y en la última columna de la tabla se realiza la asignación de elementos a cada uno de los grupos o clúster.

Cuadro N° 1: Ejemplo de algoritmo de K-medias

Número	Variable	Distancia 1	Distancia 2	Mínima d	Clúster
1	9	1	2	1	1
2	10	0	3	0	1
3	4	6	3	3	2
4	5	5	2	2	2
5	9	1	2	1	1
6	3	7	4	4	2
7	7	3	0	0	2
8	25	15	18	15	1
9	8	2	1	1	2
10	0	10	7	7	2

Luego se vuelven a estimar los centros, como el promedio de las distancias dentro de cada conglomerado, de la siguiente forma:

Mínima distancia del clúster 1/ Número de Ítems del clúster 1.

Luego:

Clúster 1: 17/4=4.25.

Clúster 2: 17/6=2.83.

Entonces los nuevos centroides son: 4.25, 2.83 respectivamente.

Luego se calcula la distancia de cada elemento a los nuevos centros. Este proceso se repite iterativamente hasta un número de veces propuesto por el usuario o hasta que no varié la configuración dentro de los grupos.

3.2. DESCRIPCIÓN DE LAS VARIABLES

En el cuadro N° 2 se presentan los códigos de cada una de las variables consideradas en el presente estudio.

Cuadro N° 2: Codificación de variables

Códigos	Descripción
MH	Número de miembros del hogar
MT	Número de miembros que trabajan
AÑOSACT	Años trabajando en la actividad principal
AREAAGRI	Total de área agrícola medido en hectáreas
VIVIPO	Número de viviendas que posee la familia
CUARVIVI	Numero de cuartos que posee la vivienda
AREAVI	Total del área de la vivienda (m^2)
PIVI	Número de pisos que tiene la vivienda
AGUA	Pago mensual por consumo de agua (S/.)
COMBALUM	Pago mensual por combustible para alumbrado (S/.)

3.3. ANÁLISIS ESTADÍSTICO

El análisis estadístico que se aplicó a este conjunto de datos comprende primeramente un análisis exploratorio y posteriormente a un análisis multivariado.

Estos análisis fueron procesados con el programa estadístico MINITAB versión 17.

3.4 ANÁLISIS EXPLORATORIO

En el análisis exploratorio se evaluó el comportamiento de cada una de las variables obteniéndose sus estadísticas descriptivas.

ANÁLISIS MULTIVARIADO

Se tiene una matriz de datos en una tabla de (nxp) donde n representa a las familias encuestadas, que en este caso son 167, y el valor de p es igual a 10, las cuales son las variables de estudio. Se estandarizaron las variables, para evitar la influencia de no deseada de las variables que presentan diferentes unidades de medida. Se aplicó la técnica de clúster no jerárquico con el método de k-medias, con un k=4.

IV. RESULTADOS

4.1. ANÁLISIS EXPLORATORIO

En el cuadro N° 3 se presenta las estadísticas descriptivas halladas para cada una de las variables.

Cuadro N° 3: Estadísticas Descriptivas de las Variables

Códigos	n	Media	Desviación Estándar	Mínimo	Máximo
MH	167	4.407	1.831	1	11
MT	167	1.6347	0.9075	1	6
AÑOSACT	167	25.76	16.04	0	66
AREAAGRI	167	2.005	1.415	0	11.5
VIVIPO	167	1.1976	0.5623	1	4
CUARVIVI	167	3.641	1.857	1	10
AREAVI	167	74.46	35.62	12	200
PIVI	167	1.6527	0.514	1	3
AGUA	167	1.991	1.364	0	13
COMBALUM	167	6.849	5.933	0	35

El promedio de miembros que tiene un hogar es de 4 personas, con un mínimo de una persona y un máximo de 11 personas. De estos miembros de hogares el promedio de miembros que trabajan es de 2 personas por familia. El promedio de los años que se dedican a la actividad principal es de 26 años. El promedio de área agrícola sembrada cosechada que posee las familias es de 2 hectáreas.

Con respecto a los pagos el promedio por el pago de agua es de 2 nuevos soles mensuales y un pago de 7 nuevos soles en combustible para alumbrado.

En cuanto a su vivienda el promedio de viviendas que poseen la familia es de 1.2 viviendas, el promedio de área construida es de 74.46 metros cuadrados, el

número promedio de cuartos en la vivienda es de 4 habitaciones y el número promedio de pisos es de 2.

Así mismo, se realizó los gráficos de caja de algunas de las variables:

Grafico N° 01: Área de vivienda

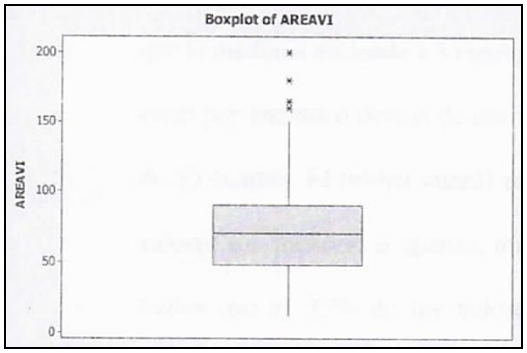


Grafico N° 02: Número de cuartos de la vivienda

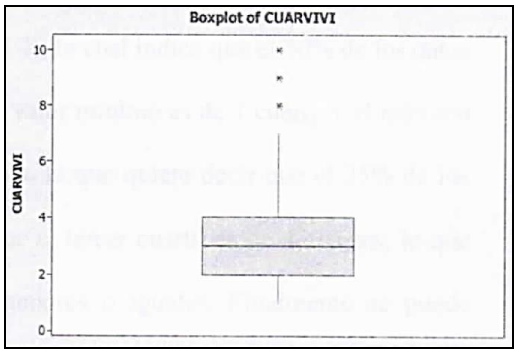


Grafico N° 03: Total de área agrícola

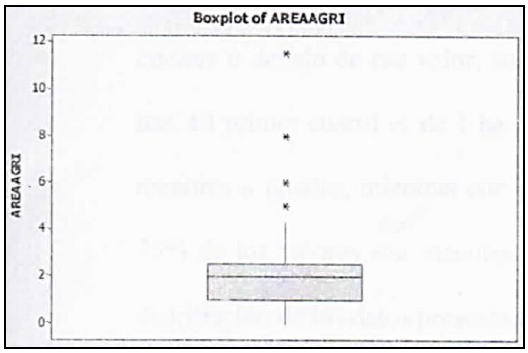
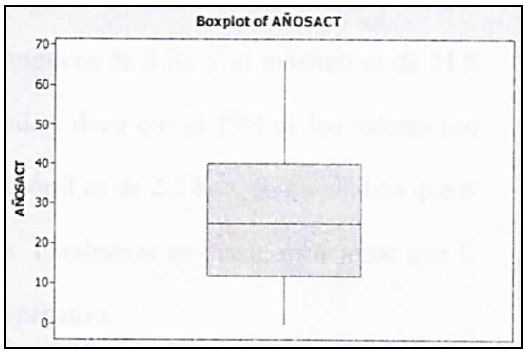


Grafico N° 04: Años que trabaja en la actividad principal



Con respectos a los gráficos de caja se puede mencionar que:

a) Para el área de la vivienda (Grafico N° 1), se ha determinado que la mediana asciende a 70 m^2 (cuartil 2), lo cual indica que el 50% de las familias están por encima o debajo de ese valor, su valor mínimo es de 12 m^2 y el máximo 200 m^2. El primer cuartil es 48 m^2, o que quiere decir que el 25% de los valores son menores o iguales, mientras que el tercer cuartil es de 900 m^2, lo que indica que el 75% de los valores son menores o iguales. Finalmente se puede mencionar que la distribución de los datos presenta asimetría positiva.
b) Con respecto al número de cuartos de la vivienda (Grafico N° 2), se ha determinado que la mediana asciende a 3 cuartos (cuartil 2), lo cual indica que el 50% de las familias están por encima o debajo de ese valor; su valor mínimo es de 1 cuarto y el máximo de 10 cuartos. El primer cuartil es 2 cuartos, lo que quiere decir que el 25% de las familias los valores son menores o iguales, mientras que el tercer cuartil es de 4 cuartos, lo que indica que el 75% de las familias los valores son menores o iguales. Finalmente se puede mencionar que la distribución de los datos presenta asimetría positiva.
c) En cuanto al total del área agrícola (Grafico N° 3), se ha determinado que la mediana ascienda a 2 has (cuartil 2), lo cual indica que el 50% de las familias están por encima o debajo de ese valor; su valor mínimo es de 0 ha. Y el máximo es de 11.5 has. El primer cuartil es de 1 ha., lo que quiere decir que el 25% de las familias los valores son menores o iguales, mientras que el tercer cuartil es de 2.5 has., lo que indica que el 75% de las familias los valores son menores o iguales. Finalmente se puede mencionar que la distribución de las familias presenta asimetría negativa.
d) Con respecto a la cantidad de años que trabaja en la actividad principal (Grafico N° 04), se ha determinado que la mediana asciende a 25 años (cuartil 2), lo cual indica que el 50% de las familias están por encima o debajo de ese valor; su valor mínimo es de 0 años y el máximo de 66 años. El primer cuartil es de 12 años, lo que quiere decir que el 25% de las familias los valores son menores o iguales, mientras que el tercer cuartil es de 40 años, lo que indica que el 75% de

las familias los valores son menores o iguales. Finalmente se puede mencionar que la distribución de las familias presenta asimetría positiva.

4.2. ANÁLISIS MULTIVARIADO

Las variables fueron estandarizadas, el número de conglomerados que se consideró es de 4. La aplicación del análisis multivariado de conglomerados no jerárquico, permitió saber si esta clasificación preliminar tiene sustento. Utilizando el método k – medias se logró la siguiente información.

Cuadro N° 4: Resultados del Método Cluster

	Número of Observations	Within Cluster Sum of Squares	Average Distance From Centroid	Maximun Distance From Centroid
Cluster1	23	215.435	2.858	6.145
Cluster2	32	225.018	2.551	4.752
Cluster3	24	269.830	3.207	5.631
Cluster4	88	455.835	2.194	4.706

Del cuadro anterior se puede mencionar como ha sido distribuidos los clúster identificados, además de la distancia promedio al centroide y la máxima distancia al centroide evaluado luego de evaluar todas las iteraciones del algoritmo de las k-medias.

Así mismo, se puede mencionar que el Minitab ha permitido determinar la distancia entre el clúster y los demás clúster centroide, tal como se puede evidenciar en el siguiente cuadro.

Cuadro N° 5: Distancias entre clúster

	Cluster1	Cluster2	Cluster3	Cluster4
Cluster1	0.0000	3.1653	3.5374	2.8595
Cluster2	3.1653	0.0000	3.9685	2.3721
Cluster3	3.5374	3.9685	0.0000	3.3503
Cluster4	2.8595	2.3721	3.3503	0.0000

En el cuadro N° 6 se presentan las familias agrupadas por conglomerado que dieron como resultado de las corridas el Minitab.

También se evaluó con 3 grupos y con 5 grupos pero se encontró una gran dispersión en la distribución de familias.

Cuadro N° 6: familias agrupadas por conglomerados.

Grupos	Total	Familias
I	23	1, 3, 5, 7, 8, 9, 17, 44, 45, 47, 65, 66, 68, 85, 94, 100, 101, 143, 147, 151, 152, 155, 156
II	32	2, 12, 20, 27, 33, 54, 55, 58, 60, 63, 70, 71, 76, 80, 83, 91, 92, 97, 126, 129, 133, 134, 145, 146, 148, 149, 154, 157, 158, 159, 160, 167
III	24	18, 19, 34, 35, 36, 37, 43, 59, 73, 78, 79, 108, 109, 110, 112, 114, 116, 120, 136, 162, 163, 164, 165, 166
IV	88	4, 6, 10, 11, 13, 14, 15, 16, 21, 22, 23, 24, 26, 28, 29, 30, 31, 32, 38, 39, 40, 41, 42, 46, 48, 49, 50, 51, 52, 53, 56, 57, 61, 62, 64, 67, 69, 72, 74, 75, 77, 81, 82, 84, 86, 87, 88, 89, 90, 93, 95, 96, 98, 99, 102, 103, 104, 105, 106, 107, 111, 113, 115, 117, 118, 119, 121, 122, 123, 124, 125, 127, 128, 130, 131, 132, 135, 137, 138, 139, 140, 141, 142, 144, 150, 153, 161

4.3. CARACTERIZACIÓN DE LOS GRUPOS FORMADOS:

El primer grupo esta agrupado por 23 familias, el segundo por 32, el tercero por 24 y el cuarto por 88 familias, cuyas caracterizaciones fueron evaluadas a través del siguiente cuadro.

Cuadro N° 7: Medias de Variables por grupos

Grupos	MH	MT	AÑOSACT	AREAAGRI	VIVIPO	CUARVIVI	AREAVI	PIVI	AGUA	COMBALUM
Total	4.407	1.635	25.760	2.005	1.198	3.641	74.460	1.653	1.991	6.849
I	3.925	1.227	20.670	1.368	1.024	3.015	64.135	1.645	0.939	8.193
II	3.566	1.199	22.726	1.379	1.052	3.045	60.443	1.605	1.438	6.871
III	3.999	1.236	19.847	1.386	1.027	3.017	65.490	1.626	0.960	8.425
IV	3.670	1.261	22.337	1.562	1.089	3.028	63.603	1.597	1.319	5.616

Con respecto al grupo I se puede mencionar que todas las familias han sido clasificadas por las siguientes características:

El grupo I: posee la característica de la mayor cantidad de pisos de vivienda por conglomerado individual y es el más cercano al valor medio total.

El grupo II: Posee las características de tener más años trabajando en la actividad principal, más cuartos dentro de la vivienda y más pago mensual por el consumo de agua. Además de tener valores medios más cercanos a los valores medios de las variables especificadas.

El grupo III: posee las características de tener más miembros en el hogar y mayor área en la vivienda. Además de tener valores medios más cercanos a los valores medios de las variables especificadas.

El grupo IV: Posee las características de tener más miembros del hogar trabajando, así como mayor área agrícola y mayor cantidad de viviendas por individuo. Además de tener valores más cercanos a los valores medios de las variables especificadas.

V. CONCLUSIONES

1. La metodología de análisis clúster permitió una clasificación de las familias encuestadas en Cajamarca.

2. Se comprobaron la existencia de 4 grupos o clúster que están definidas en el Cuadro N° 6.

3. El clúster o grupo I posee la característica de la mayor cantidad de pisos vivienda por clúster individual y es el más cercano al valor medio total. El clúster o grupo II posee las características de tener más años de trabajo en la actividad principal, más cuartos dentro de la vivienda y más pago mensual por el consumo de agua. Además de tener valores medios más cercanos a los valores medios de las variables especificadas y el clúster o grupo III posee las características de tener más miembros en el hogar y mayor área en la vivienda. Además de tener valores medios más cercanos a los valores medios de las variables especificadas.

4. El clúster o grupo IV posee las características de tener más miembros del hogar trabajando, así como mayor área agrícola y mayor cantidad de viviendas por individuo. Además de tener valores medios más cercanos a los valores medios de las variables especificadas.

VI. RECOMENDACIONES

1. Se recomienda utilizar métodos opcionales de clasificación como Análisis de Clasificación Bayesiana.

2. Se recomienda realizar la detección y eliminación de valores atípicos, con la finalidad de que no perturben el análisis.

3. Para corroborar esta clasificación se recomienda continuar con un análisis discriminante.

VII. BIBLIOGRAFÍA

1. De la Fuente Santiago Fernandez.2011. Análisis Conglomerados. Facultad de ciencias Económicas y administrativas. UAM. México.

2. Hair – Anderson – Tatham – Blanck. 1999 “Análisis Multivariante”. Prentice Hall. Quinta Edición. México.

3. Luque Martínez Teodoro, 2000. “Técnicas de análisis de datos en investigación de Mercados”. España.

4. Macqueen J. Some methods for classification and analysis of multivariate observations. University of California, Los Angeles.

5. Peña Daniel, 2002. “Análisis de datos Multivariante”. España.

6. Uriel Jiménez E. “Análisis Multivariante Aplicado”. Editorial Thomson. 2005.

7. Universidad Nacional Agraria La Molina. V Curso de Actualización para Bachilleres de Estadística 2007. “Técnicas Multivariadas Avanzadas”. Departamento de Estadística e Informática.

8. Webpages.ull.es/users/aramos/CLUSTERS.ppt

9. http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/AMult/tema5am.pdf

10. www.uam.es/departamentos/economicas/econapli/se03/cluster.doc