UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

UNIVERSIDAD NACIONAL DEL ALTIPLANO

FACULTAD DE INGENIERÍA ESTADÍSTICA E INFORMÁTICA

ESCUELA PROFESIONAL DE INGENIERÍA ESTADÍSTICA E INFORMÁTICA

"CLUSTERIZACIÓN DE LAS REGIONES DEL PERÚ, UN ANÁLISIS DE INTERDEPENDENCIA SEGÚN INDICADORES SOCIOECONÓMICOS"

TESIS

PRESENTADA POR:

Bach. GONZALO BROLYN FLORES BERMEJO

PARA OPTAR EL TÍTULO PROFESIONAL DE:

INGENIERO ESTADÍSTICO E INFORMÁTICO

PUNO – PERÚ

2020

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

DEDICATORIA

Dedico esta tesis a mi padre José Antonio Flores Marca, a mi madre Estela Bermejo Checalla y a mis seres amados, con mucha alegría y satisfacción, por haberme acompañado en cada uno de los invaluables momentos que compartimos en mi camino de formación profesional.

Gonzalo Brolyn Flores Bermejo

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

AGRADECIMIENTOS

Gracias a Dios por su inmenso amor y bendición.

Gracias a cada uno de los catedráticos, docentes y maestros que me guiaron y prepararon académicamente con sus conocimientos para lograr la presente tesis.

Gracias a mis familiares y amigos, quienes me apoyan e incentivan a perseverar en la consecución de mis objetivos.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

ÍNDICE GENERAL

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

CAPITULO III

MATERIALES Y MÉTODOS

CAPITULO IV

RESULTADOS Y DISCUSIÓN

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Área : Estadística
Tema : Análisis multivariado

FECHA DE SUSTENTACIÓN: 03 de enero del 2020

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

ÍNDICE DE FIGURAS

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

ÍNDICE DE TABLAS

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

ÍNDICE DE ACRÓNIMOS

CEPAL	COMISIÓN ECONÓMICA PARA AMÉRICA LATINA Y EL CARIBE
IDH	ÍNDICE DE DESARROLLO HUMANO
IPT	ÍNDICE DE POBREZA TOTAL
INEI	INSTITUTO NACIONAL DE ESTADÍSTICA E INFORMÁTICA
MEF	MINISTERIO DE ECONOMÍA Y FINANZAS
MIDIS	MINISTERIO DE DESARROLLO E INCLUSIÓN SOCIAL
NCHS	NATIONAL CENTER FOR HEALTH STATISTICS
OCDE	ORGANIZACIÓN PARA LA COOPERACIÓN Y EL DESARROLLO ECONÓMICO
PBI	PRODUCTO BRUTO INTERNO
PEA	POBLACIÓN ECONÓMICAMENTE ACTIVA
PIA	PRESUPUESTO INSTITUCIONAL ACTUALIZADO
PIM	PRESUPUESTO INSTITUCIONAL MODIFICADO
PNUD	PROGRAMA DE LAS NACIONES UNIDAS PARA EL DESARROLLO
VA	VALOR AGREGADO

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

RESUMEN

En el Perú, marzo de 2002, producto de acuerdos políticos se hizo una reforma constitucional que concretizaba el acuerdo de reformar el estado, pasando de un estado centralista a uno descentralizado e integrado, este gran acuerdo dio origen a tres procesos fundamentales en nuestro país: la descentralización, la regionalización y la municipalización. A pesar de la ejecución de los mencionados procesos, el Perú se destaca por ser uno de los países con mayor concentración territorial en su economía y con mayores niveles de desequilibrio en la distribución del presupuesto entre regiones. El objetivo de la investigación, es evidenciar e interpretar la estructura subyacente de agrupación homogénea de los departamentos del Perú según indicadores socioeconómicas, para ello se utilizaron variables demográficas, económicas, de educación, de salud, de vivienda, de producción, laborales y sociales, registrados y publicados en los repositorios web del INEI, MEF y MIDIS. El diseño de investigación fue no experimental y la técnica de investigación descriptiva correlacional, tomando como muestra no aleatoria el año 2018 que consta de 99 variables de 24 regiones. La metodología estadística aplicada fue la clasificación jerárquica multidimensional (Análisis Clúster) que sugiere 6 (seis) clusters: C1: La Libertad, Lambayeque, Ancash y Piura. C2: Madre de Dios y Tumbes. C3: Arequipa, Ica, Tacna y Moquegua. C4: Lima. C5: Loreto, Ucayali, Amazonas, Pasco y San Martin. C6: Cusco, Junín, Apurímac, Ayacucho, Puno, Huancavelica, Huánuco y Cajamarca. Mostrando que los retos y desafíos en las dinámicas de desarrollo económico y social superan los límites geográficos.

Palabras Clave: Análisis clúster, interdependencia, medida de distancia, método de agrupación.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

ABSTRACT

In Peru, March 2002, as a result of political agreements, a constitutional reform was made that concretized the agreement to reform the state, going from a centralist state to a decentralized and integrated one, this great agreement gave rise to three fundamental processes in our country: decentralization, regionalization and municipalization. Despite the execution of the aforementioned processes, Peru stands out for being one of the countries with the highest territorial concentration in its economy and with the highest levels of imbalance in the distribution of the budget between regions. The objective of the research is to demonstrate and interpret the underlying structure of homogeneous grouping of the departments of Peru according to socioeconomic indicators, for which demographic, economic, education, health, housing, production, labor and social variables were used, Registered and published in the web repositories of the INEI, MEF and MIDIS. The research design was non-experimental and the descriptive correlational research technique, taking the year 2018 consisting of 99 variables from 24 regions as a non-random sample. The statistical methodology applied was the multidimensional hierarchical classification (Cluster Analysis) that suggests 6 (six) clusters: C1: La Libertad, Lambayeque, Ancash and Piura. C2: Madre de Dios and Tumbes. C3: Arequipa, Ica, Tacna and Moquegua. C4: Lima. C5: Loreto, Ucayali, Amazonas, Pasco and San Martin. C6: Cusco, Junín, Apurímac, Ayacucho, Puno, Huancavelica, Huánuco and Cajamarca. Showing that the challenges in the dynamics of economic and social development go beyond geographical limits.

Keywords: Cluster analysis, interdependence, distance measurement, grouping method.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

CAPITULO I

INTRODUCCIÓN

El Perú está dividido políticamente en 24 departamentos más 01 provincia constitucional que se extienden a lo largo de tres regiones naturales: la costa, la sierra y la selva, con una población total de 31 989 256 habitantes, donde la esperanza de vida al nacer es de 76 años, 1.7% es la tasa de crecimiento anual de la población, 83**%** de personas mayores de 5 años habla castellano, este idioma convive con el quechua, aimara, asháninca y demás lenguas nativas (INEI, 2018), 930.00 soles es el salario mínimo mensual, 0.73 es el Índice de Desarrollo Humano, 17 215 700 personas son parte de la población económicamente activa y 21 657.64 es el PBI per cápita (BM, 2019).

Uno de los mayores logros del Perú en los últimos años definitivamente ha sido su índice de crecimiento económico constante. El ingreso per cápita, aumentó en más del 50% en solo diez años, cosa que no sucedía en más de tres décadas de estancamiento. El Perú alcanzó estabilidad macroeconómica y logró reducir la inflación, la deuda externa y la pobreza. Por otro lado, una expresión cultural como la gastronomía peruana gana cada día mayor terreno a nivel internacional y el país también se ha convertido en uno de los principales destinos turísticos de la región.

Estos indicadores permiten prever la expansión de la economía con grandes desafíos en materia de inclusión social y equidad de género, el Perú es un país en el que aún persisten hondas desigualdades a nivel social, los índices de desarrollo contrastan profundamente entre la capital y las provincias, así como entre las zonas urbanas y las zonas rurales. Muchos de los conflictos sociales, levantamientos y protestas de la población que viven en el interior del Perú han sido consecuencia de que éstas no se han sentido beneficiadas por las inversiones y el auge económico, donde solo 66.9 % de

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

hogares cuenta con agua potable, 4 550 personas se incorporan a la fuerza laboral por semana, 20.5 % de la población está en situación de pobreza y 3.4 % de la población está en situación de extrema pobreza (INEI, 2018).

El presente documento está dividido en cuatro capítulos a través de los cuales se expone los diferentes temas relacionados con el trabajo de tesis. Hasta aquí, en el capítulo uno se hizo la introducción, además se detalla el planteamiento del problema, la justificación, los objetivos y las hipótesis.

En el capítulo dos se presenta la revisión de la literatura, donde se expone los antecedentes que sustentan el trabajo de investigación para cada uno de los objetivos propuestos, citamos las referencias teóricas presentando el concepto de cluster, métricas de distancia, los diferentes métodos de agrupamiento, los indicadores socioeconómicos, el análisis de varianza de una vía y conceptos necesarios acerca de los gobiernos regionales.

En el capítulo tres, materiales y métodos, se describe la población, la muestra, análisis y procesamiento de los datos, la metodología seguida en la investigación y la descripción de las variables.

En el capítulo cuatro se muestra los resultados obtenidos, comparando los métodos propuestos en revisión de la literatura, de acuerdo a la metodología expuesta en el tercer capítulo.

Además, se presenta las conclusiones a las cuales se llegaron en cada uno de los objetivos, una vez finalizado el trabajo de investigación; y en recomendaciones se señala el futuro de la investigación que deriva del presente trabajo de tesis.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

1.1 PLANTEAMIENTO DEL PROBLEMA

El Perú está dividido políticamente en 24 departamentos, a nivel de país se logró un índice de crecimiento económico constante, se alcanzó estabilidad macroeconómica y se logró reducir la inflación, la deuda externa y la pobreza.

Sin embargo, persiste la desigualdad entre los 24 departamentos a nivel social y económico, los índices de desarrollo contrastan profundamente entre la capital y las regiones, así como entre las zonas urbanas y las zonas rurales. Donde solo el 66.9% de los hogares cuenta con agua potable, 20.5% de la población está en situación de pobreza y 3.4 % de la población está en situación de extrema pobreza.

No obstante, los órganos de gobierno e instituciones privadas establecen taxonomías de desarrollo que identifican grupos relativamente homogéneos y representativos de departamentos que comparten retos similares de desarrollo. Tomando en cuenta que el concepto de desarrollo es complejo y multidimensional, el problema es identificar esas taxonomías de desarrollo que puedan vislumbrar el fenómeno de desigualdad, la presente investigación interpreta esta necesidad y hace el siguiente cuestionamiento ¿Existirá una estructura subyacente de interdependencia en los indicadores socioeconómicos que hace posible la clusterización de las regiones en la República del Perú?

1.2 JUSTIFICACIÓN

Las realidades socioeconómicas de los departamentos son cada vez más diversas y heterogéneas, lo que dificulta realizar un análisis universalmente valido. Como señala Nielsen (2011), no existe un criterio de clasificación basado en la teoría del desarrollo, que sea generalmente aceptado.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Así, lograr propuestas de clasificación son imprescindibles si se desea seguir con el desarrollo del país, ya que resultan útiles para orientar las políticas de gobierno en educación, salud, producción, seguridad, vivienda, transportes y más. Para Sokal y Sneath (1963), la clasificación es uno de los procesos fundamentales de la ciencia, ya que los fenómenos deben ser ordenados para que podamos entenderlos.

En consecuencia, la presente investigación se encamino hacia el análisis de interdependencia según variables demográficas, sociales y económicas de los departamentos del Perú hasta el año 2018, para encontrar la estructura subyacente de interdependencia en los indicadores socioeconómicos que hace posible la clusterización de las regiones en la República del Perú.

1.3 OBJETIVOS DE LA INVESTIGACIÓN

1.3.1. Objetivo General

Clusterizar las regiones de la república del Perú según indicadores socioeconómicos.

1.3.2. Objetivos Específicos

- Seleccionar y evaluar las variables que identificaran los grupos para la clusterización de las regiones del Perú en grupos homogéneos según indicadores socioeconómicos.
- Elegir el criterio de agrupación de acuerdo a la medida de proximidad de casos para la clusterización de las regiones del Perú en grupos homogéneos según indicadores socioeconómicos.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

1.4 HIPÓTESIS DE LA INVESTIGACIÓN

1.4.1. Hipótesis General

El análisis Cluster evidencia una estructura subyacente de interdependencia para clusterizar las regiones de la república del Perú según indicadores socioeconómicos.

1.4.2. Hipótesis Específicos

- Las variables seleccionadas identifican grupos homogéneos de departamentos para la clusterización de las regiones del Perú según indicadores socioeconómicos.
- El criterio de agrupación de acuerdo a la medida de proximidad de casos clusteriza las regiones del Perú en grupos homogéneos según indicadores socioeconómicos.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

CAPITULO II

REVISIÓN DE LITERATURA

2.1 ANTECEDENTES DE LA INVESTIGACIÓN

Neyra (2011), propone el fomento productivo a través de la promoción de clusters territoriales en las regiones, que permitan focalizar esfuerzos en sectores con ventajas y posibilidades de ser competitivos en el escenario internacional. El trasfondo de esta política es la búsqueda del desarrollo de los mercados regionales para hacer un contrapeso económico al centro limeño. Esto se logrará a través del desarrollo e implementación de un modelo regional de articulación público-privado que, desde los propios territorios y a través de la consolidación institucional de los gobiernos descentralizados, promovida por el gobierno nacional, fortalezca la capacidad regional y local para diseñar, coordinar y articular acciones de fomento productivo y mejoramiento del entorno competitivo de los territorios. Se espera que lo anterior permita que dichos clusters capitalicen las economías de aglomeración territorial y sectorial, superen las fallas de mercado que persisten en el área de desarrollo empresarial y desarrollen una mayor capacidad de innovación.

Los obstáculos de implementar un programa de fomento de clusters productivos son principalmente socioculturales y económicos; es posible que al ser los beneficios de largo plazo y de tipo hipotético, existan obstáculos para la cooperación entre productores que pueden percibir que sus costos y riesgos son más altos que los potenciales beneficios del clúster. La propuesta requiere de un fortalecimiento de las capacidades regionales para promover la innovación productiva y la cooperación entre los distintos actores del territorio. El establecimiento de Agencias Regionales de Fomento Productivo sería una buena estrategia; estas tendrían la misión de: (i) articular la oferta nacional de recursos públicos de fomento productivo en torno a las necesidades específicas de su respectiva

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

región; (ii) planificar de forma participativa y organizar el tejido empresarial a nivel de clusters que permitan construir e implementar mejoras a la competitividad; y (iii) promover la coordinación de entidades públicas y privadas en la región, con miras a superar los problemas de entorno que afecten a los clusters regionales.

Cruz (1995), indica que el índice de pobreza se utiliza como uno de los indicadores más comunes y evidentes del fracaso en el impulso del desarrollo socioeconómico en el Perú. La tasa de pobreza, cercana al 50% de la población y con un 25% de la misma en condiciones de indigencia, es una de las más altas en América Latina. Esta tasa encubre enormes variaciones a lo largo y ancho del país, incluso dentro de los centros urbanos. Si reconocemos la pobreza, o la ausencia de desarrollo, como medida importante para evaluar la efectividad del desarrollo nacional, la política económica debe tomar en cuenta las disparidades regionales, y no referirse exclusivamente a los niveles nacionales de pobreza. Desde luego, ello implica evaluar el estado de salud y los niveles de vida por regiones. En un país que ha soportado durante las décadas recientes una enorme crisis de orden social y económico, la equidad interregional debe constituirse en meta nacional.

En el presente estudio se utiliza tanto el análisis factorial como el análisis clúster para distribuir y clasificar, de mayor a menor desarrollo, los departamentos del Perú. Los resultados muestran que la principal fuente de disparidad entre los mismos son las variables vinculadas al estado de salud. El patrón demográfico e indicadores de industrialización juegan también un rol importante. Esta investigación confirma la hipótesis del carácter dual de la sociedad peruana y muestra las agudas diferencias que existen entre grupos de departamentos de acuerdo con su nivel de industrialización y pobreza rural. El grupo Lima y Callao, más moderno e industrializado, coexiste con los grupos de departamentos de la región andina, tradicionalmente rurales. Si se considera

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

que en el Perú la igualdad interregional debe ser una de las más importantes metas nacionales, urge que la política económica incorpore medidas fiscales y compensatorias, destinadas a mejorar el estado de salud e inducir el desarrollo socioeconómico en los departamentos menos favorecidos.

Tezanos (2012), menciona que el criterio más extendido internacionalmente para medir la asistencia oficial para el desarrollo, es quizás el más sencillo, basado únicamente en un indicador de renta per cápita. Así, de acuerdo con la clasificación propuesta por el Banco Mundial, la mayoría de los países de América Latina y el Caribe (ALC) se ubican en el estrato medio de la renta mundial, lo que determina su clasificación como "países de renta media" (PRM). En efecto, como revela el análisis de diferenciación de medias realizado por Alonso (Dir. 2007, pág. 33), "desde un punto de vista estadístico, los PRM conforman un grupo específico y estadísticamente distinto del resto de los países en desarrollo".

Sin embargo, en el contexto geográfico de ALC existen diferencias notables entre los niveles de desarrollo de los países que componen el colectivo de renta media. Así, en 2010 las diferencias en términos de PIB per cápita (en paridad de poder adquisitivo) se extendieron desde los 2.914 dólares de Nicaragua, hasta los más de 15.000 dólares de Chile, Argentina, Antigua y Barbuda y San Cristóbal y Nieves. Y, en definitiva, estas abultadas diferencias en términos de ingreso enmascaran las disímiles "brechas de desarrollo" que afrontan los países de la región.

En este artículo se propone una clasificación alternativa de los PRM latinoamericanos y caribeños que trasciende al criterio tradicional de renta, basada en la técnica del análisis de conglomerados, que identifican y caracterizan tres grupos de países con perfiles socioeconómicos distintos y atiende a las principales "brechas de desarrollo"

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

(económicas, sociales y medioambientales) que limitan sus oportunidades de progreso. Esta clasificación se emplea para analizar los flujos públicos de cooperación internacional financiados por los organismos multilaterales de desarrollo y por los países donantes de la Organización para la Cooperación y el Desarrollo Económicos (OCDE) —la llamada Ayuda Oficial al Desarrollo, AOD— y recibidos por los países de ALC. La identificación de las principales brechas de desarrollo que caracterizan a cada grupo de países permite además evaluar la congruencia existente entre dichas brechas y la asignación de los recursos de AOD —tanto entre países, como entre sectores e instrumentos de cooperación— al objeto de identificar posibles "oportunidades de cooperación".

Halanoca (2017), identifica distritos con características productivas agropecuarias homogéneas con especialización agropecuaria y ve la relación que existe sobre la calidad de vida, considerando que la actividad agropecuaria en el Altiplano es una de las actividades económicas más importantes. Sin embargo, la variabilidad climática, altitud y la tenencia de tierras, recurso más importante, son las que definen en gran medida la aptitud productiva de cada distrito. Se dividieron los distritos por regiones naturales (sierra y selva) mediante imágenes satelitales, luego se agruparon variables agrícolas para reducir la dimensionalidad, posteriormente aplicando análisis de conglomerado se identificaron los distritos con similar vocación productiva. Finalmente, mediante el análisis de varianza se determinó la relación entre aptitud productiva y calidad de vida expresado por la Incidencia de Pobreza Total (IPT). El propósito fue entender la dinámica del desarrollo agropecuario expresado en términos de orientaciones, límites y posibilidades de desarrollo. Con lo cual, cada conjunto territorial sería tratado como unidades con el mismo potencial de desarrollo en lugar de que cada distrito busque alcanzar el desarrollo de manera aislada. En la región natural sierra se encontró 11 conglomerados de los cuales 2 conglomerados son atípicos (Ilave y Crucero); para la

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

región selva se identificaron 4 conglomerados; mediante el análisis de varianza se encontró que en la región natural sierra no existen diferencias significativas entre los conglomerados, sin embargo, en la región selva si se encontraron diferencias entre los conglomerados encontrados.

Pérez, Lara y Gómez (2017), investigan la evolución de la capacidad tecnología en México, a través del análisis multivariado de clúster, con base en el set de indicadores propuesto por CEPAL (2007) y recopilando los datos de diversas fuentes públicas del país para los años 2006 y 2012, con el fin de estudiar la evolución en el tiempo de dichos clusters, tratando de ver qué estados han podido mudarse a un clúster situado en posiciones más avanzadas y cuáles han retrocedido en dicho periodo. Los resultados muestran la existencia de 7 grupos de estados caracterizados por distintos niveles de capacidad tecnológica, y se detectan también 3 entidades que evolucionaron a un clúster más avanzado, tanto en lo referente a la capacidad de absorción e innovación, como en lo relativo a las capacidades tecnológicas de infraestructura.

Cárdenas y Saraiva (2016), analizan la vulnerabilidad social y la minería en el Perú, con el objetivo identificar si la minería implica mayor vulnerabilidad social. Realizaron una investigación de naturaleza cuantitativa con datos del Instituto Nacional de Estadística e Informática. Como instrumento de análisis se utilizó el programa estadístico SPSS 18.0, y por medio del análisis factorial se asociaron las variables en tres factores que denominamos: Familiar, Físico y Social. Después se utilizó la técnica de clúster, en la cual se agrupó los departamentos con características semejantes. Los resultados mostraron que el clúster 2, compuesto por una región minera que a su vez tiene presencia de minería ilegal e informal, tiene mayor impacto sobre la vulnerabilidad social, de acuerdo con el test t de Student. Se puede concluir que las regiones mineras influyen

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

en la cantidad de empleos y en la economía local, pero genera vulnerabilidad social, representada por situaciones de denuncias y robos.

Apaza (2014), implementa algoritmos genéticos para la segmentación de imágenes satelitales por conglomerados en la región Puno, con el objetivo es desarrollar una aplicación para la segmentación de imágenes basada en conglomerados, denominado Algoritmos Genéticos K-medias (AGKM). Esta aplicación fue propuesta debido al deficiente método de selección del valor de inicialización del algoritmo K-medias al tomar un número de conglomerados inicial de forma aleatoria o por cálculo de la observación visual, esto puede influir en el desempeño del algoritmo, haciendo que tenga una separación inadecuada o demore más tiempo en la búsqueda del número de conglomerados. Se implementó usando la metodología de los Algoritmos Genéticos (AGs) y K-medias, el primero tiene la finalidad de encontrar un número de conglomerados existentes en la imagen y el segundo realiza el proceso de separación. La métrica usada para evaluar la eficiencia de este algoritmo es el valor de la entropía en las imágenes, los resultados obtenidos son sometidos a una prueba estadística que nos indica que existe una ligera mejoría. Concluyendo que el AGKM ofrece una ligera mejoría con respecto al algoritmo K-medias tradicional en la segmentación de imágenes satelitales para la Región Puno.

Vicente (2014), clasifica familias según su situación económica mediante el análisis de conglomerados, su 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 Pacífico, la encuesta fue realizada en diciembre del 2006. Se desea clasificar a las familias para poder brindar un mejor control

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

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 comprobó la existencia de 4 grupos o clúster.

Hernández (2012), propone un algoritmo de clasificación jerárquico multidimensional que redice el problema de inconsistencia, para Hernández se han propuesto diferentes alternativas para resolver problemas de tipo jerárquico, de las cuales las más destacadas son las aproximaciones locales y globales. El problema principal de los métodos locales es el de inconsistencia, este se presenta cuando se produce un error de clasificación en un cierto nodo siendo propagado a todos sus descendientes. Los clasificadores jerárquicos globales tienen el problema de producir modelos complejos y, por lo general, tienden a ser dependientes al clasificador elegido. El objetivo de este trabajo de tesis es desarrollar un nuevo método de clasificación jerárquico que tome en consideración todos los posibles caminos (ramas) en la jerarquía al momento de realizar una predicción. El método propuesto es una alternativa inspirada en la clasificación multidimensional. El método construye un clasificador multi-clase para cada nodo padre de la jerarquía. Durante la fase de clasificación, todos los clasificadores locales son aplicados simultáneamente a cada instancia, dando como resultado la clase más probable para cada clasificador. Posteriormente, se aplica uno de los tres métodos propuestos para obtener un conjunto de clases consistentes con alguna de las ramas de la jerarquía. Se desarrollaron dos extensiones al método base: La primera considera la dependencia entre los clasificadores locales aplicando el método de encadenamiento, y la segunda para clasificar a diferentes niveles de la jerarquía basados en ganancia de información. El método propuesto fue probado en tres diferentes conjuntos de datos y fue comparado con

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

los métodos del estado del arte, resultando en un desempeño predictivo similar o superior a las demás aproximaciones en todas las bases de datos.

Román (2017), implementa pruebas para una hipótesis sobre la aplicación de distancia Euclidiana para realizar agrupamientos en espacios multidimensionales. Los algoritmos de agrupamiento permiten agrupar un conjunto de datos en un conjunto de sub-clases, denominados clusters. El objetivo principal de los mismos es agrupar, en dichos clusters, instancias de datos similares entre sí. Cada instancia de datos suele ser representada en un espacio de características en donde cada característica queda presentada como una dimensión de dicho espacio. Es común, entonces, el uso de espacios de muchas dimensiones en esta representación. Una de las medidas de similitud más usadas para realizar el agrupamiento es la distancia euclidiana.

La motivación principal de este trabajo es brindar asistencia en la implementación y prueba de una hipótesis sobre el uso de la métrica de distancia euclidiana como medida de similitud en los algoritmos de agrupamiento. En la hipótesis se plantea la posibilidad de que, en espacios multidimensionales, la distancia euclidiana puede conducir a un agrupamiento erróneo en ciertas ocasiones. Es decir, puede ocurrir que se agrupen instancias en una clase cuando en realidad pertenecen a otra.

Numerosos estudios han determinado que las métricas de distancia suelen tener comportamientos erráticos en altas dimensiones. Sin embargo, no existen muchos trabajos que profundicen demasiado en esta temática debido a que el problema es naturalmente complejo. Los espacios n-dimensionales grandes (con $n > 3$) no pueden ser graficados en su totalidad, y nuestra intuición falla en ellos.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Se ha mostrado que, en algoritmos de agrupamiento particionales como K- Means, el uso de diferentes métricas de distancia puede impactar fuertemente en los resultados. Por lo tanto, la elección de la métrica debe hacerse con cuidado.

2.2 MARCO TEÓRICO

El análisis multivariante es la parte de la estadística y del análisis de datos que estudia, analiza, representa e interpreta los datos que resultan de observar más de una variable estadística sobre una muestra de individuos, donde las variables observables son homogéneas y correlacionadas, sin que alguna predomine sobre las demás. La información estadística en el análisis multivariado es de carácter multidimensional, por lo tanto, la geometría, el cálculo matricial y las distribuciones multivariante juegan un papel fundamental.

La información multivariante es una matriz de datos, pero a menudo, en el análisis multivariado la información de entrada consiste en matrices de distancias o similaridades, que miden el grado de discrepancia entre los individuos. La aplicación del análisis de interdependencia o también denominado métodos descriptivos supone la inexistencia de dependencia entre las variables explicadas y las variables explicativas, otorgando la misma consideración a todas las variables, para descubrir las interrelaciones y estructura subyacente entre ellas.

Llevar a cabo con éxito la aplicación del análisis multivariante de datos implica resolver problemas que van desde la definición del modelo hasta un diagnóstico crítico de los resultados. La aproximación a la modelización se centra en el análisis de un plan de investigación bien definido, comenzando por un modelo conceptual que detalle las relaciones a examinar. Definido el modelo, se pueden iniciar los trabajos empíricos,

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

incluyendo la selección de una técnica multivariante específica y su puesta en práctica. Finalmente, las medidas de diagnosis aseguran que el modelo no solo es válido para la muestra de datos, sino que es también generalizable.

2.2.1. Análisis de conglomerados (análisis cluster)

El término análisis clúster se utiliza para definir una serie de técnicas, fundamentalmente algoritmos, que tienen por objeto la búsqueda de grupos similares de individuos o de variables que se van agrupando en conglomerados. Dada una muestra de individuos, de cada uno de los cuales se dispone de una serie de observaciones, el análisis clúster sirve para clasificarlos en grupos lo más homogéneos posible en base a las variables observadas. Los individuos que queden clasificados en el mismo grupo serán tan similares como sea posible.

La palabra clúster, que define estas técnicas, se podría traducir por grupo, conglomerado, racimo, apiñarse, etc. El análisis clúster se usa en biología para clasificar animales y plantas, conociéndose con el nombre de taxonomía numérica. Otros nombres asignados al mismo concepto son análisis de conglomerados, análisis tipológico, clasificación automática y otros. Todos ellos pueden funcionar como sinónimos. En los paquetes estadísticos más habituales y en muchos trabajos en castellano suele aparecer el nombre de clúster análisis. Para Sokal y Sneath (1963), dos de los autores que más han influido en el desarrollo del análisis clúster, la clasificación es uno de los procesos fundamentales de la ciencia, ya que los fenómenos deben ser ordenados para que podamos entenderlos. Tanto el análisis clúster como el análisis discriminante sirven para clasificar individuos en categorías. La diferencia principal entre ellos estriba en que en el análisis

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

discriminante se conoce a priori el grupo de pertenencia, mientras que el análisis clúster sirve para ir formando grupos homogéneos de conglomerados.

El análisis clúster es un método estadístico multivariante de clasificación automática de datos. A partir de una tabla de casos-variables, trata de situar los casos (individuos) en grupos homogéneos, conglomerados o clusters, no conocidos de antemano, pero sugeridos por la propia esencia de los datos, de manera que individuos que puedan ser considerados similares sean asignados a un mismo clúster, mientras que individuos diferentes (disimilares) se localicen en clusters distintos. La diferencia esencial con el análisis discriminante estriba en que en este último es necesario especificar previamente los grupos por un camino objetivo, ajeno a la medida de las variables en los casos de la muestra. El análisis clúster define grupos tan distintos como sea posible en función de los propios datos.

El enorme campo de aplicación en numerosas disciplinas, que se inició con la clasificación de las especies biológicas, ha propiciado la diversificación de este análisis, con denominaciones específicas tales como taxonomía numérica, taximetría, nosología, nosografía, morfometría, tipología, botriología, etc. La creación de grupos basados en similaridad de casos exige una definición de este concepto, o de su complementario «distancia» entre individuos. La variedad de formas de medir diferencias multivariables o distancias entre casos proporciona diversas posibilidades de análisis. El empleo de ellas, y el de las que continuamente siguen apareciendo, así como de los algoritmos de clasificación, o diferentes reglas matemáticas para asignar los individuos a distintos grupos, depende del fenómeno estudiado y del conocimiento previo de posible agrupamiento que de él se tenga.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Puesto que la utilización del análisis clúster ya implica un desconocimiento o conocimiento incompleto de la clasificación de los datos, el investigador ha de ser consciente de la necesidad de emplear varios métodos, ninguno de ellos incuestionable, con el fin de contrastar los resultados.

Existen dos grandes tipos de análisis de clusters: Aquéllos que asignan los casos a grupos diferenciados que el propio análisis configura, sin que unos dependan de otros, se conocen como no jerárquicos, y aquéllos que configuran grupos con estructura arborescente, de forma que clusters de niveles más bajos van siendo englobados en otros de niveles superiores, se denominan jerárquicos. Los métodos no jerárquicos pueden, a su vez, producir clusters disjuntos (cada caso pertenece a un y sólo un clúster), o bien solapados (un caso puede pertenecer a más de un grupo). Estos últimos, de difícil interpretación, son poco utilizados.

Una vez finalizado un análisis de clusters, el investigador dispondrá de su colección de casos agrupada en subconjuntos jerárquicos o no jerárquicos. Podrá aplicar técnicas estadísticas comparativas convencionales siempre que lo permita la relevancia práctica de los grupos creados; así como otras pruebas multivariantes, para las que ya contará con una variable dependiente «grupo», aunque haya sido artificialmente creada.

De este modo, el horizonte de la investigación podría ampliarse, por ejemplo, con la aplicación de regresión logística y análisis discriminante con posibles nuevas variables independientes (utilizar las mismas que han servido para la confección de los grupos no sería una práctica correcta). También serían aplicables pruebas de asociación y análisis de correspondencias.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

El análisis clúster se puede utilizar para agrupar individuos (casos) y también para agrupar variables. En lo que sigue, cuando nos refiramos a grupos de individuos (o casos), debe sobreentenderse que también nos referimos a conjuntos de variables. El proceso es idéntico tanto si se agrupan individuos como variables.

Antes de iniciar un análisis clúster deben tomarse tres decisiones: selección de las variables relevantes para identificar a los grupos, elección de la medida de proximidad entre los individuos y elección del criterio para agrupar individuos en conglomerados. Es decisiva la selección de las variables que realmente sean relevantes para identificar a los grupos, de acuerdo con el objetivo que se pretenda lograr en el estudio. De lo contrario, el análisis carecerá de sentido. Para la selección de la medida de proximidad es conveniente estar familiarizado con este tipo de medidas, básicamente similitudes y distancias, ya que los conglomerados que se forman lo hacen en base a las proximidades entre variables o individuos. Puesto que los grupos que se forman en cada paso depende de la proximidad, distintas medidas de proximidad pueden dar resultados distintos para los mismos datos. Para elegir el criterio de agrupación conviene conocer, como mínimo los principales métodos de análisis clúster.

2.2.2. Distancias y similitudes

La proximidad expresa la semejanza que existe ente individuos o variables. Es decir, es el grado de asociación que existe entre ellos. Las proximidades pueden medir la distancia o la similitud (similaridad) entre individuos o variables. El valor que se obtiene en una medida de distancia es tanto mayor cuanto más alejados están los individuos o puntos entre los que se mide. En las similitudes, al contrario

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

de las distancias, el valor que se obtiene es tanto mayor cuanto más próximos están los elementos considerados. La correlación de Pearson y los coeficientes de Spearman y de Kendall son índices de similitud.

Matemáticamente se da el nombre de distancia entre dos puntos A y B, a toda medida que verifique los axiomas siguientes:

1. $d(A,B) \geq 0$ y $d(A,A) = 0$
2. $d(A,B) = d(B,A)$
3. $d(A,B) \leq d(A,C) + d(C,B)$

Un primer ejemplo de distancia es la distancia euclídea que se define como:

$$d(A,B) = \sqrt{\sum_i (A_i - B_i)^2}$$

Un segundo ejemplo de distancia es la distancia D^2 de Mahalanobis. Originariamente se utilizó para calcular la distancia entre poblaciones. La D^2 es la distancia al cuadrado entre los centroides de dos poblaciones. Recordemos que el centroide de una población es el centro de gravedad de esta población en base a un conjunto de variables (vector de las medias de las variables). El centroide es en el análisis multivariable lo que la media es en el análisis univariable. Cronológicamente la D^2 de Mahalanobis está considerada como la primera técnica de análisis multidimensional. Su autor la formuló en 1927 y se divulgó algo más tarde (Mahalanobis, 1936). Las aplicaciones de esta técnica han ido modificando su carácter original, estableciéndose una relación entre teoría y práctica, que se ha ido enriqueciendo mutuamente. Las aplicaciones prácticas han descubierto nuevos aspectos, que, generalizados en el plano formal, han dado lugar a nuevos procedimientos e interpretaciones.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Rao (1952) dio un gran impulso a esta medida con el desarrollo de proyecciones y representaciones gráficas de tipo geométrico. Las proyecciones permitieron posteriormente constatar y demostrar que la D^2 de Mahalanobis y el análisis discriminante constituyen dos aspectos del mismo proceso, en el sentido de dos formas de cálculo diferente para un mismo tipo de análisis. La D^2 de Mahalanobis permite situar poblaciones en un espacio de v dimensiones, siendo v el número de variables consideradas en el estudio. Entre las aplicaciones de la D^2 se encuentran los contrastes entre poblaciones, establecer la distancia entre dos individuos, calcular la distancia de un individuo al centroide de su grupo, etc.

Sean p poblaciones de n_1, n_2,...,n_p individuos cada una. En cada población se conocen v variables x_1, x_2,...,x_v, de modo que a cada población k le corresponde una matriz de observaciones de orden n_k x v. Se dispone por tanto de p matrices de orden n_k x v con k=1,...,p. A partir de estos datos, y en notación matricial, Mahalanobis define la distancia entre los centroides de los grupos p y q (distancia entre las poblaciones p y q) como:

$$D_{pq}^2 = (\mu_p - \mu_q)' \sum^{-1} (\mu_p - \mu_q)$$

Los vectores μ_p y μ_q son vectores columna que contienen las medias de las variables de los grupos respectivos y Σ es la matriz de varianzas covarianzas intragrupos de los grupos conjuntamente.

A partir de la D^2 se puede estimar la F de Fisher y utilizarla como prueba de contraste para dos poblaciones:

$$F = D^2 \frac{n_p n_q (n_p + n_q - v - 1)}{(n_p + n_q)(n_p + n_q - 2)v} \rightarrow F_{v, n_p + n_q - v - 1}$$

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Teniendo en cuenta que normalmente se trabaja con muestras, la fórmula que se utiliza para la D^2 es:

$$D_{pq}^2 = \left(\bar{X}_p - \bar{X}_q\right)' S^{-1} (\bar{X}_p - \bar{X}_q)$$

donde las medias poblacionales se han estimado por medias muestrales y la varianza poblacional se ha estimado por la Cuasivarianza muestral S^2.

La D^2 de Mahalanobis puede utilizarse también para medir la distancia entre dos individuos A y B de la forma siguiente:

$$D_{AB}^2 = (X_A - X_B)' \sum^{-1} (X_A - X_B)$$

La D^2 de Mahalanobis también puede utilizarse para medir la distancia de un individuo A al centroide de su grupo de la forma siguiente:

$$D^2 = (X_A - \bar{X})' \sum^{-1} (X_A - \bar{X})$$

Los individuos con mayor D^2 son los más lejanos del centro de su grupo, con lo que esta distancia puede utilizarse para detectar individuos con puntuaciones extremas (outliers).

Un ejemplo de distancia entre dos variables x e y es la distancia de Manhattan o City-block que se define como:

$$B(x, y) = \sum_i |x_i - y_i|$$

Otro ejemplo de distancia entre dos variables x e y es la distancia de Minkowski que se define como:

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

$$M(x,y) = \sum_i (|x_i - y_i|^p)^{\frac{1}{p}}$$

Un último ejemplo de distancia entre dos variables x e y es la distancia de Chebychev que se define como:

$$C(x,y) = Max|x_i - y_i|$$

Entre las medidas de similitud (similaridad) tenemos los ya conocidos coeficientes de correlación de Pearson y Sperman y los múltiples coeficientes de asociación entre variables también conocidos (lambda, tau, etc.).

Para el caso de variables cualitativas, y en general para el caso de datos binarios (o dicotómicos), que son aquéllos que sólo pueden presentar dos opciones (blanco – negro, sí – no, hombre – mujer, verdadero – falso, etc.), existen diferentes medidas de proximidad o similitud, que se verán a continuación, partiendo de una tabla de frecuencias 2x2 en la que se representa el número de elementos de la población en los que se constata la presencia o ausencia del carácter (variable cualitativa) en estudio.

Tabla 1
Distribución de frecuencias para variables cualitativas

Variable 1 → ***Variable 2*** ↓	***Presencia***	***Ausencia***
Presencia	*a*	*b*
Ausencia	*c*	*d*

Fuente: Elaboración propia

Las principales medidas son las siguientes:

Russel y Rao: $RR_{xy} = \frac{a}{a+b+c}$

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Sokal y Sneath: $SS_{xy} = \frac{2(a+d)}{2(a+d)+b+c}$

Parejas Simples: $PS_{xy} = \frac{a+d}{a+b+c+d}$

Rogers y Tanimoto: $RT_{xy} = \frac{a+d}{a+d+2(b+c)}$

Jaccard: $J_{xy} = \frac{a}{a+b+c}$

Sokal y Sneath (2): $SS2_{xy} = \frac{a}{a+2(b+c)}$

Dice y Sorensen: $D_{xy} = \frac{2a}{2a+b+c}$

Kulczynski: $K_{xy} = \frac{a}{b+c}$

Hay otro grupo de medidas denominadas medidas de similaridad para probabilidades condicionales, entre las que destacan las siguientes:

Kulczynski (medida 2): $K2_{xy} = \frac{\frac{a}{a+b}+\frac{a}{a+c}}{2}$

Sokal y Sneath (medida 4): $SS4_{xy} = \frac{\frac{a}{a+b}+\frac{a}{a+c}+\frac{d}{b+d}+\frac{d}{c+d}}{4}$

Hamann: $H_{xy} = \frac{(a+d)-(b+c)}{a+b+c+d}$

También suele considerarse un subgrupo de medidas de predicción, como la D_{xy} de Anderberg, la Y_{xy} de Yule y la Q_{xy} de Yule.

$$D_{xy} = \frac{\mathrm{m}(a,b) + \mathrm{m}(c,d) + \mathrm{m}(a,c) + \mathrm{m}(b,d) - \mathrm{m}(a+c,b+d) - \mathrm{m}\,(a+b,c+d)}{2(a+b+c+d)}$$

$$Y_{xy} = \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}} \qquad Q_{xy} = \frac{ad - bc}{ad + bc}$$

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Por último, se usan otras medidas binarias, entre las que destacan las siguientes:

Ochiai $O_{xy} = \sqrt{\frac{a}{a+b}.\frac{a}{a+c}}$

Sokal y Sneath (5) $SS5_{xy} = \frac{ad}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$

Sokal y Sneath (3) $SS3_{xy} = \frac{a+d}{b+c}$

Correlación phi $\emptyset_{xy} = \frac{ad-bc}{(a+b)(a+c)(b+c)(c+d)}$

Euclídea binaria $EB_{xy} = \sqrt{b+c}$

Diferencia de forma $DF_{xy} = \frac{(a+b+c+d)(b+c)-(b-c)^2}{(a+b+c+d)^2}$

Euclídea binaria^2 $EB_{xy}^2 = b+c$

Varianza disimilar $V_{xy} = \frac{b+c}{4(a+b+c+d)}$

Dispersión $D_{xy} = \frac{ad-bc}{(a+b+c+d)^2}$

Diferencia de tamaño $T_{xy} = \frac{(b-c)^2}{(a+b+c+d)^2}$

Lance y Wiliams $LW_{xy} = \frac{b+c}{2a+b+c}$

Diferencia de patrón $P_{xy} = \frac{bc}{(a+b+c+d)^2}$

2.2.3. Clusters jerárquicos: dendograma

Es frecuente en la investigación biológica la necesidad de clasificar los datos en grupos con estructura arborescente de dependencia, de acuerdo con diferentes niveles de jerarquía. La clasificación de especies animales o vegetales

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

constituye un buen ejemplo de este interés científico. Partiendo de tantos grupos iniciales como individuos se estudian, se trata de conseguir agrupaciones sucesivas entre ellos de forma que progresivamente se vayan integrando en clusters los cuales, a su vez, se unirán entre sí en un nivel superior formando grupos mayores que más tarde se juntarán hasta llegar al clúster final que contiene todos los casos analizados. La representación gráfica de estas etapas (Figura 1) de formación de grupos, a modo de árbol invertido, se denomina dendograma.

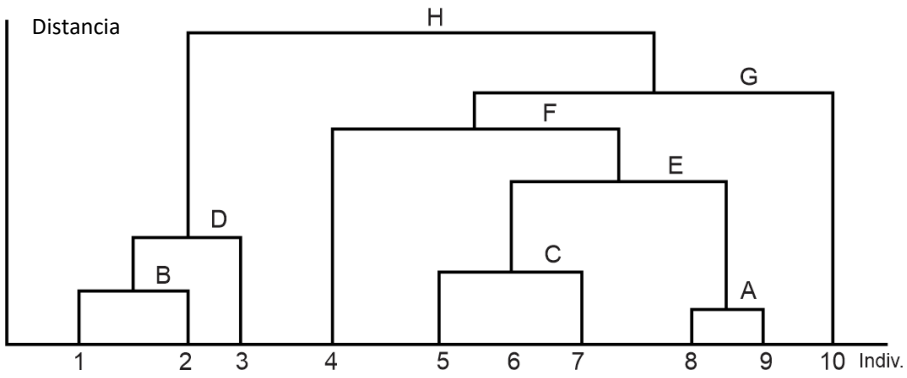

Figura 1 Dendograma

Fuente: Elaboración propia

La figura 1, que corresponde a un estudio de los individuos, muestra cómo el 8 y el 9 se agrupan en un primer clúster (A). En un nivel inmediatamente superior, se unen los individuos 1 y 2 clúster (B); y enseguida los 5,6, y 7 (C). Un paso siguiente engloba el clúster B con el individuo 3 (D); y así sucesivamente hasta que todos ellos quedan estructurados al conseguir, en el nivel más alto, el clúster total (H) que reúne los 10 casos.

Evidentemente, la decisión de todas estas agrupaciones ha de tomarse en función de la similaridad multivariante (o de su contrario distancia) proporcionada por el conjunto de variables estudiadas, ya que en cada nivel de jerarquía se unen los dos clusters más cercanos. Es, pues, importante como paso previo a un análisis

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

de clusters jerárquicos, la elección de una adecuada métrica de similaridad o disimilaridad. Se sabe que a partir de la tabla inicial de datos (n x p) es preciso calcular una matriz de distancias entre individuos (n x n). Este concepto, ya introducido anteriormente, merece aquí un tratamiento más detallado. Se han descrito numerosas formas de medir distancias multivariantes. La conocida distancia euclídea es la más sencilla y utilizada: Se usa también en el análisis de componentes principales cuyos factores son, como se sabe, muchas veces datos previos para entrar en un análisis de clusters.

Para variables cualitativas puede emplearse la distancia $\chi 2$, y, si son dicotómicas, la distancia de Jaccard. La distancia euclídea al cuadrado, la euclídea generalizada, la de bloques o Manhattan, la de Tchebycheff, la de Mahalanobis, y otras medidas de similaridad como los coeficientes de correlación de Pearson y de correlación por rangos de Kendall entre individuos, el índice de Gower, etc. dan idea de la enorme variedad de formas de enfocar el diseño de un análisis de clasificación de datos, cada una de ellas con sus ventajas e inconvenientes que, en definitiva, serán mejores o peores según las características del fenómeno estudiado y, sobre todo, de la relevancia o interpretabilidad de los grupos obtenidos. Sin embargo, las distancias más usadas son pocas y ya han sido definidas.

La segunda decisión que el investigador debe tomar es, precisamente, qué algoritmo emplear para la formación de grupos, definiendo a qué va a llamar "distancia entre clusters" para luego poder unir, a otro nivel jerárquico, aquellos clusters más próximos. Este concepto no existía en el análisis no jerárquico, puesto que allí no se unían los grupos. Es también muy variada, y continuamente ampliada, la oferta de procedimientos de agrupación.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

La aglomeración comienza con tantos grupos como individuos; cada uno de éstos constituye un cluster inicial. A medida que transcurren las etapas del proceso se van formando nuevos clusters por unión de dos individuos, de un individuo con un grupo previo, o de dos grupos anteriores entre los que exista la menor distancia.

El proceso finaliza con un único grupo (todos los individuos), pero constituido por aglomeraciones sucesivas en distintos niveles. Este es el fundamento de la agregación (ascendente); en contraposición con el proceso de disgregación (descendente), que opera de forma inversa: Parte del grupo total de individuos para llegar, tras varias etapas de partición, hasta tantos clusters como individuos. Característica importante de los métodos jerárquicos es el no permitir reasignaciones de grupos, es decir, que dos clusters (o dos individuos) que han sido unidos en un paso del proceso no pueden ya separarse en etapas sucesivas; lo que sí es posible en los métodos no jerárquicos (si bien en éstos es necesario, como se ha visto, fijar previamente el número de clusters deseado).

Se presenta a continuación una relación de los principales métodos de unión de grupos o algoritmos de clasificación jerárquica en la que, junto a su descripción, se comentan sus ventajas e inconvenientes como ayuda en la decisión del método apropiado. Suele distinguirse entre métodos aglomerativos y métodos disociativos. Entre los métodos aglomerativos tenemos los siguientes:

Método de las distancias mínimas o Enlace simple (single linkage). Considera como distancia entre dos grupos la que responde al concepto de "vecinos más cercanos" (nearest neighbor), es decir, la separación que existe entre los individuos más próximos de uno y otro grupo. Aunque presenta buenas

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

propiedades teóricas, su eficacia no ha sido la esperada en estudios de validación comparativa de métodos mediante ficheros simulados (Monte Carlo), por su tendencia al "encadenamiento" (une clusters realmente diferentes, si están próximos), y porque sólo considera la información de individuos extremos (los valores atípicos, outliers, pueden distorsionar la agrupación). Si bien no es adecuado para la obtención de grupos compactos, resulta de utilidad para clusters irregulares o elongados. El método consiste en ir agrupando los individuos que tienen menor distancia o mayor similitud, considerando como distancia entre dos clusters la distancia entre sus dos puntos más próximos.

Método de las distancias máximas o Enlace completo (complete linkage). Considera como distancia entre dos grupos la existente entre "vecinos más lejanos" (furthest neighbor), es decir, entre los individuos más separados de ambos grupos (máxima distancia que es posible encontrar entre un caso de un cluster y un caso de otro). Presenta una excesiva tendencia a producir grupos de igual diámetro, y se ve muy distorsionado ante valores atípicos moderados.

Método del promedio entre grupos o Enlace promedio (average linkage). Considera como distancia entre dos clusters, no la de los individuos más próximos ni más lejanos de ambos grupos, sino la distancia media entre todos los pares posibles de casos (uno de cada cluster). Tiende a producir clusters compactos, por lo que es muy utilizado y suele ser el método por defecto en los paquetes de software. Una variante de este método es el método de la media ponderada (average linkage within groups), en el cual se combinan los grupos de tal forma que la distancia promedio entre todos los casos en el cluster resultante sea lo más pequeña posible.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Método del centroide o Enlace centroide (centrod method). Considera como distancia entre dos grupos la existente entre sus centros de gravedad, definidos por las medias aritméticas de las variables de los individuos que componen los clusters. Es el más robusto de los métodos jerárquicos ante la presencia de casos atípicos.

Método de la mediana (median method). Considera como distancia entre dos grupos la existente entre las medianas de las variables de los individuos que componen los clusters. De este modo, los dos clusters que se combinan se ponderan de forma equivalente al método centroide, pero independientemente del número de individuos que haya en cada grupo.

Método de Ward o Enlace por mínima varianza (momento central de orden dos o pérdida de inercia mínima). Considera como distancia entre dos grupos el menor incremento de varianza residual global, o sea, si en un nivel dado existe un número de clusters de los que se deben elegir dos para una nueva fusión, se prueban todas las parejas posibles y se calcula la varianza residual global o intragrupos con cada pareja unida y todos los demás clusters. La pareja de grupos que produzca el mínimo incremento en esta varianza residual será la elegida para su unión en un nuevo nivel. En el último nivel, con todos los individuos agrupados en un sólo cluster, la varianza residual es máxima y coincide con la varianza total al ser, lógicamente, nula la varianza factorial o intergrupos (ya no hay grupos). Tiende a formar clusters esféricos o compactos, y del mismo tamaño. Requiere una distribución normal multivariante en las variables del estudio. Utiliza más información sobre el contenido de los grupos que otros métodos. Es, junto con el enlace promedio, el que ha demostrado mayor eficacia en estudios de simulación.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Siendo más precisos, en el método de Ward se calcula la media de todas las variables de cada cluster, luego se calcula la distancia euclídea al cuadrado entre cada individuo y la media de su grupo y después se suman las distancias de todos los casos. En cada paso, los clusters que se forman son aquéllos que resultan con el menor incremento en la suma total de las distancias al cuadrado intracluster. La métrica normalmente considerada en los métodos hasta aquí descritos es la euclídea o la euclídea al cuadrado. Esta última se suele usar por omisión en programas estadísticos.

Método del Enlace por densidad. Utiliza un nuevo concepto de distancia entre dos individuos (A y B): Puesto que la nube de puntos no tendrá una densidad homogénea, existirán zonas de acúmulo de casos separadas por otras menos densas, lo que puede sugerir agrupamientos naturales entre los individuos. De esta forma, para unir dos de ellos, se les puede considerar como centros de dos esferas (multidimensionales) capaces de englobar un número prefijado de casos de la nube (por ejemplo, k = 4). Se define la densidad de cada esfera como su "masa", o número de individuos que contiene (en frecuencia relativa), dividida por su volumen. Como el numerador es constante (k/n) y en el denominador figura el radio, para abarcar 4 casos muy próximos, el radio necesario será pequeño y la densidad grande. Es lógico, por tanto, definir conceptualmente una distancia como el inverso de una densidad. Al ser dos las esferas consideradas, se toma como distancia entre A y B la media de los inversos de ambas densidades. Así, usando esta distancia, se ponderan más próximos dos individuos, cada uno de ellos con k casos muy cercanos, que otros dos con k casos lejanos. En la Figura 2, el individuo A sería unido en el siguiente nivel al B, y no al C, a pesar de que existe la misma separación geométrica entre AB y AC. Esferas que no se cortan hacen no

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

considerar a sus individuos centrales como candidatos a ser unidos. Ello indicará que no están suficientemente «próximos», o que el número k de casos a abarcar debe ser mayor. Para impedir esta unión, por convenio se considera una distancia infinita cuando las esferas no se cortan; por ello, no siempre con este método se llega al final de la aglomeración a un cluster único. La sucesiva fusión de éstos se realiza mediante el método del enlace simple, pero con esta distancia definida en lugar de la euclídea. El enlace por densidad es capaz de encontrar clusters irregulares y elongados, aunque ha demostrado menos eficacia en la configuración de clusters compactos. Se han descrito diversas técnicas basadas en este algoritmo de clasificación, como por ejemplo el enlace por densidad.

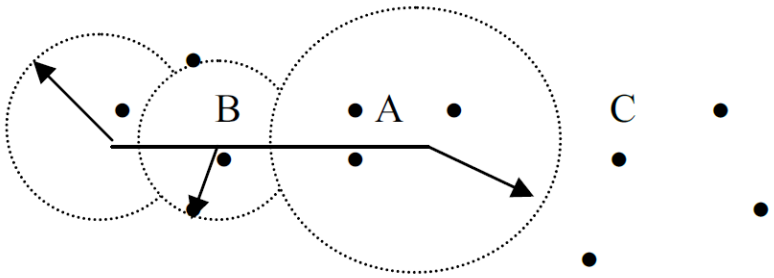

Figura 2 Enlace por densidad

Fuente: Pérez (2008)

Al aplicar cualquiera de estos métodos existe la posibilidad de que, en un determinado paso de unión, se produzcan «empates» en las distancias de dos o más individuos o grupos. En ese nivel jerárquico podría, pues, agregarse cualquiera de ellos, y se hace necesario elegir uno. Esta decisión es tomada automáticamente por los programas de acuerdo con un criterio arbitrario pero definido (por ejemplo, basado en el orden que ocupan los casos en la tabla de datos original). Si estos empates se presentan en los niveles más bajos del dendograma afectarán poco a la estructura, pero si aparecen en niveles altos podría variar

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

sustancialmente la configuración final según la decisión tomada por el programa. Una comprobación del grado de afectación podría hacerse, por ejemplo, permutando el orden de los individuos en la tabla. Altas precisiones en las medidas de las variables alejan este problema de empates.

Fórmula de Lance y Williams para la distancia entre grupos. Matemáticamente, Lance y Williams desarrollaron una fórmula general que puede ser utilizada para describir los distintos tipos de enlaces de los métodos jerárquicos aglomerativos. La fórmula de Lance y Williams para la distancia entre grupos es la siguiente:

$$D_{k(ij)} = \alpha_i D_{ki} + \alpha_j D_{kj} + \beta D_{ij} + \gamma \left| D_{ki} - D_{kj} \right|$$

donde D_{ij} es la distancia entre los grupos i y j, y α, β y γ son los tres parámetros del modelo. Se observa lo siguiente:

$$\alpha_i = \alpha_j = 1/2\,, \beta = 0\; y\; \gamma = -1/2 \Rightarrow \; enlace\; simple$$

$$\alpha_i = \alpha_j = 1/2\,, \beta = 0\; y\; \gamma = 1/2 \Rightarrow \; enlace\; completo$$

$$\alpha_i = \alpha_j = 1/2\,, \beta = -1/4\;\; y\; \gamma = 0 \Rightarrow \; método\; de\; la\; mediana$$

$$\alpha_i = \frac{n_i}{n_i + n_j}, \alpha_j = \frac{n_j}{n_i + n_j}, \beta = -\alpha_i \alpha_j \; y\; \gamma = 0 \Rightarrow \; enlace\; centroide$$

$$\alpha_i = \frac{n_i}{n_i + n_j}, \alpha_j = \frac{n_j}{n_i + n_j}, \beta = \gamma = 0 \Rightarrow \; enlace\; promedio$$

$$\alpha_i = \frac{n_k + n_i}{n_k + n_i + n_j}, \alpha_j = \frac{n_k + n_j}{n_k + n_i + n_j}, \beta = \frac{-n_k}{n_k + n_i + n_j}\; y\; \gamma = 0 \Rightarrow \; Ward$$

$$\alpha_i + \alpha_j + \beta = 1, \alpha_i = \alpha_j, \beta < 1\; y\; \gamma = 0 \Rightarrow \; método\; flexible$$

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

El último método (método flexible o cuádruple restricción) consiste en utilizar la forma de Lance y Williams variando los coeficientes según las necesidades del clasificador, pero respetando las cuatro restricciones impuestas.

Los métodos de clusters jerárquicos, por la laboriosidad de los cálculos, no resultan prácticos para procesar grandes ficheros de datos. En estos casos, puede ser aconsejable realizar un análisis previo no jerárquico, que proporcione un número preliminar razonable de clusters (en lugar de individuos) que servirán luego de partida para su posterior clasificación jerárquica.

Como resumen, los métodos jerárquicos producen resultados más ricos que los no jerárquicos. Con un solo análisis se obtiene una configuración de grupos en cada nivel de clasificación. Los mismos indicadores que en clasificación no jerárquica valoraban la adecuación del número de clusters (Criterio cúbico de clusters, Pseudo F, etc.) permiten detectar aquí el nivel jerárquico en que la separación de los grupos formados es más ostensible.

Los siete criterios que se acaban de exponer para la clasificación jerárquica aglomerativa también son utilizables en el caso de los métodos de clasificación disociativos, si bien en la práctica los que más suelen usarse son el del promedio entregrupos y el de Ward. Además de los anteriores, dentro del proceso disociativo, destacan los que se presentan muy brevemente a continuación.

En los métodos disociativos, cuando el criterio de división toma en consideración cada variable observada una a una, el método recibirá el nombre de monotético. Por el contrario, cuando se toman en cuenta todas las variables simultáneamente el método se llamarán politético. El análisis de asociación está especialmente diseñado para el caso de variables dicotómicas. Según el método

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

asociativo de Williams y Lambert se construyen tablas de contingencia 2 x 2, para cada par de variables y se calcula la ji-cuadrado para cada tabla. El criterio de participación en los grupos se basa en la variable que maximiza la ji-cuadrado.

2.2.4. Población

Las palabras población, universo o colectivo, se usan indistintamente para referirse al conjunto de todos los elementos, individuos o unidades, que presentan características comunes, susceptibles de observación, medición o experimentación y que constituye el ámbito de estudio para cualquier tipo de investigación.

La población es la totalidad de sujetos o elementos de un conjunto infinito o finito, delimitado por el investigador y su estudio se realiza mediante el censo, es decir el conteo de uno a uno de todos los elementos.

Así, de la afirmación “todas las mujeres” tenemos una declaración que se refiere a cierta unidad de análisis, de observación o de experimentación en particular -mujer-. De todo lo que existe, la afirmación se refiere a quienes comparten las características propias de una mujer. Una vez que determinamos la unidad de observación, estamos en condiciones de identificar la población que serán todas las posibles unidades de observación.

Criterios de selección de una población:

- Homogeneidad, que todos los miembros de la población tengan las mismas características según las variables que se vayan a considerar en el estudio o investigación.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

- Tiempo, se refiere al periodo de tiempo donde se ubicaría la población de interés. Determinar si el estudio es del momento o si se van a entrevistar personas de diferentes generaciones.
- Espacio, se refiere al lugar donde se ubica la población de interés.
- Cantidad, se refiere al tamaño de la población, el cual es importante porque determina o afecta el tamaño de la muestra que se va a seleccionar.

Unidad de análisis, es el objeto del cual se desea obtener información. Son los elementos, sujetos u objetos de estudio.

Dato estadístico, es toda unidad de información que se ha obtenido al realizar un estudio estadístico. Estructura a partir de la cual el investigador genera sus estudios e indagaciones.

Población accesible desconocida, conjunto de personas, elementos, cosas o valores que cumplen con las características o criterios preestablecidos y que pueden ser accesibles a alcanzar por el investigador para el estudio, pero su número no es conocido.

Población accesible conocida, conjunto de personas, elementos, cosas o valores que cumplen con las características o criterios establecidos y que es accesible por el investigador para el estudio, donde su número no es conocido.

2.2.5. Muestra

Es el sub conjunto de la población. Para que un sector de la población sea considerado como muestra, se requiere que todos los elementos de ella pertenezcan a la población. No serán muestras cuando algunos sujetos de la muestra no pertenezcan a la población.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

La selección de la muestra está íntimamente relacionada con la estimación que se hará posteriormente, de los parámetros poblacionales

El tamaño de la muestra, es el número de sujetos que se debe incluir en la muestra. Técnicamente depende de la precisión con que el investigador desea estimar el parámetro de la población en un nivel particular de confianza. No hay regla sencilla con la cual determinar esa dimensión. La estimación de la muestra requerida se obtiene con una operación algebraica, si se define con exactitud la varianza de la población, la diferencia esperada y las probabilidades deseadas de un error de tipo I y tipo II. Varios textos de estadística describen este procedimiento.

El método más seguro es usar una muestra tan grande como sea posible. Una muestra grande tiene mayores posibilidades de ser representativa de la población. Además, es probable que los datos sean más exactos y precisos, lo cual quiere decir que cuanto mayor sea la muestra, menor será el error estándar. Por lo general el error estándar de la media de una muestra es inversamente proporcional a la raíz cuadrada de n. Por ello, si se quiere duplicar la exactitud de la estimación, hay que cuadruplicar el tamaño de la muestra.

Debe recalcarse, sin embargo, que el tamaño de la muestra no basta para garantizar la exactitud. La consideración más importante al sacar una muestra es su representatividad. Una muestra puede ser grande y a pesar de ello tener algún vicio; así pues, el investigador debe reconocer que el tamaño de la muestra no contrarresta los vicios que pueda causar la aplicación de técnicas deficientes de muestreo. En la selección de la muestra la representatividad ha de ser el objetivo primordial.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

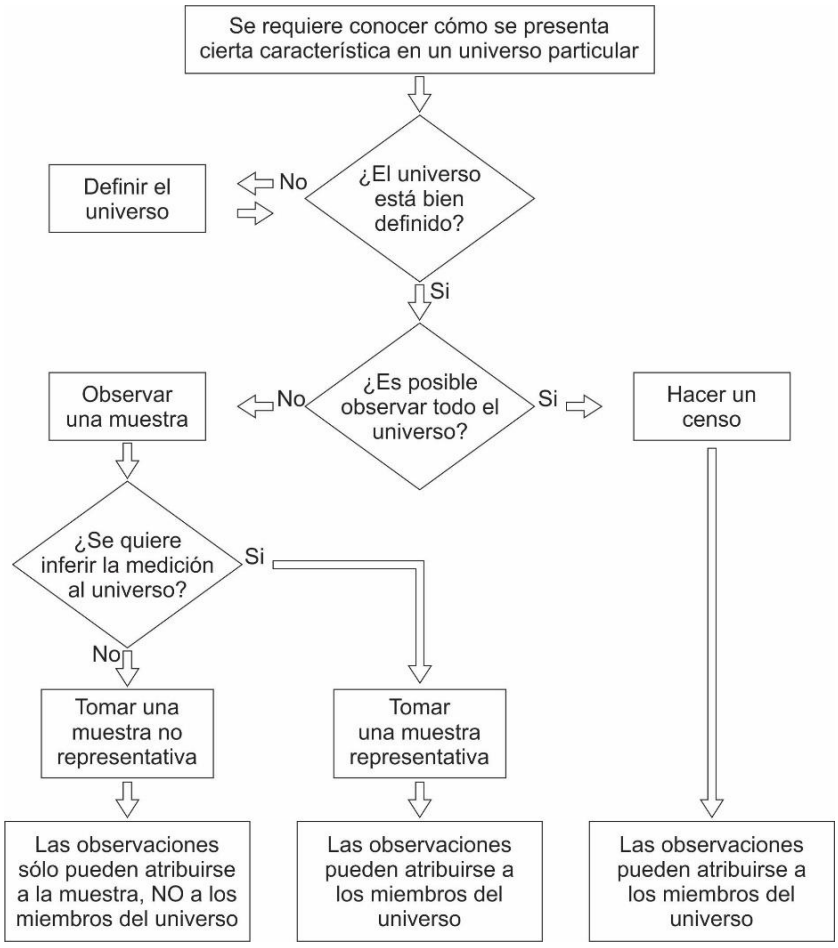

Figura 3 Muestreo

Fuente: Elaboración propia

2.2.6. Indicadores Socioeconómicos

El desarrollo y progreso están sujetos a un conjunto de procesos que tienen lugar en los distintos niveles de la estructura social y económica de un país. Estos procesos impactan en forma diferenciada en cada uno de los ciudadanos y las familias, de acuerdo a su lugar de residencia, su ubicación en la estructura productiva, el nivel de calidad de vida, el contexto institucional y las normas culturales que orientan el desarrollo de una población con características y atributos propios; de modo tal, que, para poder explicar su nivel y tendencias

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

económicas, deben vincularse, estos al contexto social de cada región y/o división político administrativa.

Las sociedades regionales presentan diferencias en sus dimensiones básicas como: demográficas, económicas, en educación, en salud, en vivienda, en producción y en convivencia dentro de su territorio, entre las áreas urbana y rural, divisiones político-administrativas, etc... Cada uno de ellos difiere en su nivel o grado de desarrollo económico, social y cultural los que llevan al mantenimiento de pautas, valores y actitudes que se asocian con diferentes tipos de comportamiento de los diversos estratos de la población.

Estas dimensiones básicas, a las que se recurre para la construcción de los diferentes grupos o estratos sociales, seleccionan algunos aspectos del nivel económico, los que, partiendo de su ubicación en la estructura productiva, definen una posición en la estructura social expresada en el grado de posesión de bienes materiales y sociales alcanzados por los miembros de ese estrato y en el reconocimiento que hacen de ello los demás. Esos aspectos socioeconómicos son, por ejemplo, la posesión o no de bienes de producción, la categoría de ocupación, el grupo de ocupación, el tipo y condiciones de la vivienda, el ingreso y diversos otros indicadores de las condiciones de existencia, como la posesión o no de ciertos bienes en el hogar, nivel educacional alcanzado, etc.

Los indicadores socioeconómicos revelan el estado o situación de la unidad de estudio. Constituyen criterios objetivos para clasificar o dividir a las unidades de estudio. Suelen combinarse para determinar la clase social. Sirven también para explicar el comportamiento del individuo en determinado medio.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

2.2.7. ANOVA de una vía para datos independientes

El ANOVA de una vía, ANOVA con un factor o modelo factorial de un solo factor es el tipo de análisis que se emplea cuando los datos no están pareados y se quiere estudiar si existen diferencias significativas entre las medias de una variable aleatoria continua en los diferentes niveles de otra variable cualitativa o factor. Es una extensión de los *t-test independientes* para más de dos grupos.

Las hipótesis contrastadas en un ANOVA de un factor son:

- H_0: No hay diferencias entre las medias de los diferentes grupos: $\mu_1=\mu_2...=\mu_k=\mu$
- H_a: Al menos un par de medias son significativamente distintas la una de la otra.

Otra forma de plantear las hipótesis de un ANOVA es la siguiente. Si se considera μ como el valor esperado para una observación cualquiera de la población (la media de todas las observaciones sin tener en cuenta los diferentes niveles), y α_i el efecto introducido por el nivel *i*. La media de un determinado nivel (μ_i) se puede definir como:

$$\mu_i = \mu + \alpha_i$$

- H_0: Ningún nivel introduce un efecto sobre la media total: $\alpha_1=\alpha_2=...\alpha_k=0$
- H_a: Al menos un nivel introduce un efecto que desplaza su media: Algún $\alpha i\neq 0$

Como se ha mencionado anteriormente, la diferencia entre medias se detecta a través del estudio de la varianza entre grupos y dentro de grupos. Para

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

lograrlo, el ANOVA requiere de una descomposición de la varianza basada en la siguiente idea:

Variabilidad total = variabilidad debida a los diferentes niveles del factor + variabilidad residual variabilidad debida a los diferentes niveles del factor + variabilidad residual

Lo que es equivalente a:

variabilidad explicada por el factor + variabilidad no explicada por el factor variabilidad explicada por el factor + variabilidad no explicada por el factor

Lo que es equivalente a:

(varianza entre niveles) + (varianza dentro de los niveles)

Para poder calcular las diferentes varianzas en primer lugar se tienen que obtener las Sumas de Cuadrados (SS o Sc):

- **Suma de Cuadrados Total o Total Sum of Squares (TSS):** mide la variabilidad total de los datos, se define como la suma de los cuadrados de las diferencias de cada observación respecto a la media general de todas las observaciones. Los grados de libertad de la suma de cuadrados totales es igual al número total de observaciones menos uno (N-1).
- **Suma de cuadrados del factor o Sum of Squares due to Treatment (SST):** mide la variabilidad en los datos asociada al efecto del factor sobre la media (la diferencia de las medias entre los diferentes niveles o grupos). Se obtiene como la suma de los cuadrados de las desviaciones de la media de cada proveedor respecto de la media general, ponderando cada diferencia al cuadrado por el número de observaciones de cada grupo. Los

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

grados de libertad correspondientes son igual al número niveles del factor menos uno (k-1).

- **Suma de cuadrados residual/error o Sum of Squares of Errors (SSE):** mide la variabilidad dentro de cada nivel, es decir, la variabilidad que no es debida a variable cualitativa o factor. Se calcula como la suma de los cuadrados de las desviaciones de cada observación respecto a la media del nivel al que pertenece. Los grados de libertad asignados a la suma de cuadrados residual equivale la diferencia entre los grados de libertad totales y los grados de libertad del factor, o lo que es lo mismo (N-k). En estadística se emplea el termino error o residual ya que se considera que esta es la variabilidad que muestran los datos debido a los errores de medida. Desde el punto de vista biológico tiene más sentido llamarlo *Suma de cuadrados dentro de grupos* ya que se sabe que la variabilidad observada no solo se debe a errores de medida, si no a los muchos factores que no se controlan y que afectan a los procesos biológicos.

$$TSS=SSE+SST$$

Una vez descompuesta la suma de cuadrados se puede obtener la descomposición de la varianza dividiendo la Suma de Cuadrados entre los respectivos grados de libertad. De forma estricta, al cociente entre la Suma de Cuadrados y sus correspondientes grados de libertad se le denomina *Cuadrados Medios o Mean Sum of Squares* y pueden ser empleado como estimador de la varianza:

ANOVA se define como análisis de varianza, pero en un sentido estricto, se trata de un análisis de la Suma de Cuadrados Medios.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

- $\hat{S}_T^2 = \frac{TSS}{N-1}$ = Cuadrados Medios Totales = Cuasivarianza Total (varianza muestral total)
- $\hat{S}_t^2 = \frac{SST}{k-1}$ = Cuadrados Medios del Factor = Intervarianza (varianza entre las medias de los distintos niveles)
- $\hat{S}_E^2 = \frac{SSE}{N-k}$ = Cuadrados Medios del Error = Intravarianza (varianza dentro de los niveles, conocida como varianza residual o de error)

Una vez descompuesta la estimación de la varianza, se obtiene el estadístico F_{ratio} dividiendo la intervarianza entre la intravarianza:

$$F_{ratio} = \frac{\text{Cuadrados Medios del Factor}}{\text{Cuadrados Medios del Error}} = \frac{\hat{S}_t^2}{\hat{S}_E^2} = \frac{intervarianza}{intravarianza} \sim F_{k-1,N-k}$$

Dado que por definición el estadístico F_{ratio} sigue una distribución *F Fisher-Snedecor* con k−1 y N−t grados de libertad, se puede conocer la probabilidad de obtener valores iguales o más extremos que los observados.

2.2.8. Gobiernos Regionales del Perú

Los Gobiernos Regionales del Perú son instituciones públicas encargadas de la administración superior de cada uno de los departamentos. Son considerados personas jurídicas de derecho público con autonomía política, económica y administrativa en los asuntos de su competencia. Los gobiernos regionales del Perú se componen de dos órganos: un Consejo Regional y un Gobernador Regional.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Según el ordenamiento jurídico peruano, la gestión de los Gobiernos Regionales corresponde al gobierno a nivel regional. Este nivel de gobierno fue introducido en la legislación peruana con la puesta en vigencia de la Constitución del 79, pero inició su activación en la forma que en la actualidad lleva a partir de los años 2000, al modificarse la constitución para añadirla. En el proceso que se contempla en ella y en el orden jurídico peruano, todos los departamentos del país han de integrarse para conformar regiones mediante referéndum hasta que la totalidad del territorio se encuentre regionalizado, salvo la Provincia Constitucional del Callao y la Provincia de Lima, con autonomía regional por ser capital del país, tiene su propia Municipalidad en funciones de Gobierno Regional como de Municipalidad Provincial.

De acuerdo con la Ley Orgánica de Gobiernos Regionales, las responsabilidades de los gobiernos regionales incluyen el desarrollo de la planificación regional, ejecución de proyectos de inversión pública, promoción de las actividades económicas y administración de la propiedad pública.

De acuerdo con el Artículo 191° de la Constitución Política del Perú, Ley de Reforma de los Artículos 191º, 194° y 203° de la Constitución Política del Perú; la estructura orgánica básica de los gobiernos regionales está conformada de la siguiente manera:

- **Gobernador Regional**, constituye el órgano ejecutivo y sus funciones incluyen proponer y ejecutar el presupuesto, designar a los oficiales de gobierno, promulgar decretos y resoluciones, ejecutar planes y programas regionales y administrar las propiedades y rentas regionales. Ley N° 27867, Ley Orgánica de Gobiernos Regionales. Artículo Nº 21.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

- **Consejo Regional**, debate y vota sobre el presupuesto sugerido por el presidente regional, también supervisa a todos los oficiales de gobierno y puede deponer de su cargo al presidente, su vicepresidente y a cualquier miembro del consejo. Ley N° 27867, Ley Orgánica de Gobiernos Regionales, Artículo Nº 15.

- **Consejo de Coordinación Regional,** está integrado por los alcaldes provinciales y representantes de la sociedad civil y tiene un papel consultivo en los asuntos de planeamiento y presupuesto, no tiene poderes ejecutivos ni legislativos. Ley N° 27867, Ley Orgánica de Gobiernos Regionales, Artículo N.º 11B.

El Gobernador Regional y el Concejo Regional sirven por un periodo de cuatro años. Ley N° 27867, Ley Orgánica de Gobiernos Regionales, Artículo Nº 11. De acuerdo con la Ley N° 30305 - Ley de Reforma Constitucional, el Gobernador Regional y el Vicegobernador Regional no pueden ser reelectos inmediatamente, aunque transcurrido un período como mínimo, pueden volver a postular.

Los Gobiernos Regionales del Perú son:

- Gobierno Regional de Amazonas
- Gobierno Regional de Áncash
- Gobierno Regional de Apurímac
- Gobierno Regional de Arequipa
- Gobierno Regional de Ayacucho
- Gobierno Regional de Cajamarca
- Gobierno Regional del Callao

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

- Gobierno Regional del Cusco
- Gobierno Regional de Huancavelica
- Gobierno Regional de Huánuco
- Gobierno Regional de Ica
- Gobierno Regional de Junín
- Gobierno Regional de Lambayeque
- Gobierno Regional de La Libertad
- Gobierno Regional de Lima
- Gobierno Regional de Loreto
- Gobierno Regional de Madre de Dios
- Gobierno Regional de Moquegua
- Gobierno Regional de Pasco
- Gobierno Regional de Piura
- Gobierno Regional de Puno
- Gobierno Regional de San Martín
- Gobierno Regional de Tacna
- Gobierno Regional de Tumbes
- Gobierno Regional de Ucayali

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Figura 4 Gobiernos Regionales del Perú

Fuente: Choque (2014)

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

CAPITULO III

MATERIALES Y MÉTODOS

3.1 POBLACIÓN

Para el presente proyecto la población está constituida por datos del año 2018 de las 24 regiones geopolíticas demarcadas en el territorio de la República del Perú.

3.2 MUESTRA

La muestra está conformada por muestreo no aleatorio tomando el año 2018 que consta de 99 variables de 24 regiones.

3.3 TÉCNICA E INSTRUMENTOS DE RECOGIDA DE DATOS

Los datos se obtuvieron de diferentes repositorios públicos:

- Gobierno del Perú
- Instituto Nacional de Estática e Informática
- Ministerio de Economía y Finanzas
- Ministerio de Desarrollo e Inclusión Social
- Agencia de Promoción de la Inversión Privada
- Instituto Peruano de Economía
- Banco Mundial
- Programa de las Naciones unidas para el desarrollo
- Comisión Económica para América Latina y el Caribe

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

3.4 ANÁLISIS Y PROCESAMIENTO DE DATOS

Durante el procesamiento y análisis de datos se usaron las siguientes técnicas:

- Para seleccionar y evaluar las variables que identificaran los grupos para la clusterización de las regiones del Perú en grupos homogéneos según indicadores socioeconómicos se usó la técnica descriptiva conocida como análisis exploratorio de datos.
- Para elegir el criterio de agrupación de acuerdo a la medida de proximidad de casos se usó el análisis cluster.

3.5 METODOLOGÍA DE LA INVESTIGACIÓN

Para llevar a cabo con éxito la aplicación de cualquier técnica de análisis multivariante deben resolverse asuntos que van desde el problema de definición del modelo hasta un diagnóstico crítico de los resultados. La aproximación a la modelización se centra en el análisis de un plan de investigación bien definido, comenzando por un modelo conceptual que detalle las relaciones a examinar. Definido el modelo, se pueden iniciar los trabajos empíricos, incluyendo la selección de una técnica multivariante específica y su puesta en práctica. Después de haber obtenido resultados significativos, el asunto central es la interpretación. Finalmente, las medidas de diagnosis aseguran que el modelo es válido.

En el primer paso en la práctica del análisis multivariante se definió el problema de investigación, objetivos y técnica multivariante conveniente. Se vio en primer lugar el problema en términos conceptuales, definiendo los conceptos e identificando las relaciones fundamentales a investigar. Como se propuso una técnica de interdependencia se determinaron las dimensiones de la estructura o similitud. Con los objetivos y el

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

modelo conceptual especificados, se eligió la técnica multivariante de análisis de conglomerados.

En el segundo paso se desarrolló el proyecto de análisis poniendo en práctica la técnica del análisis de conglomerados. Con el modelo conceptual establecido, se especificó un plan de análisis que dirijo el conjunto de supuestos que subyacen en la aplicación de la técnica. Estos supuestos son los tipos de variables permitidas y métodos de estimación.

En el tercer paso se evaluó los supuestos básicos del análisis de conglomerados. Una vez recogidos los datos, se evaluó si se cumplen los supuestos subyacentes como los supuestos de correlación entre las variables.

En el cuarto paso se estimó el modelo multivariante y la valoración del ajuste del modelo. Una vez satisfechos todos los supuestos del modelo, se procedió a su estimación efectiva realizando a continuación una valoración global del ajuste del modelo (parámetros significativos individual y globalmente).

En el quinto paso se interpretó los valores obtenidos. Una vez estimado el modelo fue necesario interpretar los resultados de acuerdo a los valores teóricos posibles. Esta interpretación condujo a la re-especificación del modelo y a su nueva estimación hasta que la interpretación de los resultados se ajustó coherentemente a los valores teóricos.

En el sexto paso se validó el modelo multivariante. Una vez estimado el modelo fue necesario aceptar los resultados con el grado de fiabilidad más alto posible mediante la aplicación de contrastes específicos.

Las fases anteriores pueden esquematizarse de la siguiente manera:

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

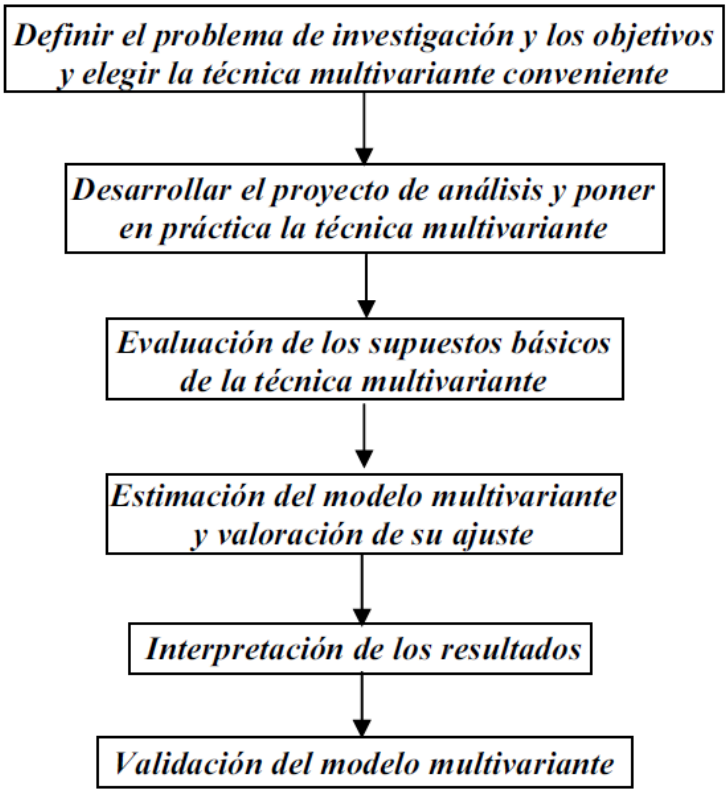

Figura 5 Fases del análisis multivariante

Fuente: Pérez (2008)

3.6 DESCRIPCIÓN DE LAS VARIABLES

Se consideraron variables económicas, sociales, demográficas y de hogares, estandarizados por el Instituto Nacional de Estadística e Informática para los 24 departamentos de la república del Perú hasta el año 2018, las cuales se detallan a continuación:

Esperanza de vida, medido en años.

Tasa global de fecundidad, medido por el promedio de hijos/as por mujer.

Mortalidad en la niñez, medido por el número de muertes de niños de hasta 5 años por cada 1,000 nacidos.

Crecimiento poblacional, medido en porcentaje.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Nacimientos, medido en número de personas.

Defunciones, medido en número de personas.

Población electoral, medido en número de personas.

Población inmigrante, medido en número de personas.

Población emigrante, medido en número de personas.

Población censada, medido en número de personas.

Población urbana censada, medido en número de personas.

Población rural censada, medido en número de personas.

Tasa de Analfabetismo, medido en porcentaje.

Tasa de Analfabetismo masculino, medido en porcentaje.

Tasa de Analfabetismo femenino, medido en porcentaje.

Tasa de Asistencia escolar (inicial), medido en porcentaje.

Tasa de Asistencia escolar (primaria), medido en porcentaje.

Tasa de Asistencia escolar (secundaria), medido en porcentaje.

Nivel de educación de la Población de 15 y más años (Primaria), medido en porcentaje.

Nivel de educación de la Población de 15 y más años (Secundaria), medido en porcentaje.

Nivel de educación de la Población de 15 y más años (Sup. No Universitaria), medido en porcentaje.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Nivel de educación de la Población de 15 y más años (Sup. Universitaria), medido en porcentaje.

Rendimiento en lectura de los estudiantes de 2do grado de primaria y secundaria con rendimiento satisfactorio en comprensión lectora, medido en porcentaje.

Rendimiento en matemáticas de los estudiantes de 2do grado de primaria y secundaria con rendimiento satisfactorio en matemáticas, medido en porcentaje.

Colegios con acceso a internet (escuelas de primarias y secundarias), medido en porcentaje.

Cobertura de algún Seguro de Salud, medido en porcentaje.

Tasa de desnutrición crónica en menores de 5 años, medido en porcentaje.

Niños con anemia (De 6 a 35 meses de edad), medido en porcentaje.

Desnutrición crónica de niños menores de 5 años que tienen una longitud o talla menor a la esperada para su edad y sexo según el patrón NCHS, medido en porcentaje.

Morbilidad de la población que padece problemas de salud crónicos y no crónicos, medido en porcentaje.

Cobertura del personal médico por cada 10,000 habitantes, medido en número de personas.

Cobertura hospitalaria por cada 100,000 habitantes, medido en número de hospitales.

Partos institucionales, medido en porcentaje.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Población en edad de trabajar (De 14 y más años de edad), medido en miles de personas.

PEA, medido en miles de personas.

PEA ocupada, medido en miles de personas.

PEA Ocupada en Agricultura, pesca, minería, medido en miles de personas.

PEA Ocupada en Manufactura, medido en miles de personas.

PEA ocupada en empresas de 1-10 trabajadores, medido en porcentaje.

PEA desempleada, medido en miles de personas.

PEA ocupada adecuadamente empleada, medido en porcentaje.

PEA ocupada con educación superior, medido en porcentaje.

Creación de empleo formal, medido en promedio móvil tres años de la variación anual.

Empleo informal de la PEA ocupada, medido en porcentaje.

Desempleo juvenil urbano de la PEA, medido en porcentaje.

VA Bruto, medido en porcentaje.

Manufactura, medido en porcentaje.

Construcción, medido en porcentaje.

Agricultura, Ganadería, Caza y Silvicultura, medido en porcentaje.

Electricidad, Gas y Agua, medido en porcentaje.

Telecomunicaciones y Servicios De Información, medido en porcentaje.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Administración Publica y Defensa, medido en porcentaje.

Alojamiento y Restaurantes, medido en porcentaje.

Transporte, Almacén, Correo y Mensajería, medido en porcentaje.

Comercio, medido en porcentaje.

Pesca y Acuicultura, medido en porcentaje.

Extracción De Petróleo, Gas y Minerales, medido en porcentaje.

Otros Servicios, medido en porcentaje.

Hogares con acceso a agua potable, medido en porcentaje.

Hogares con acceso a desagüe, medido en porcentaje.

Hogares con acceso a alumbrado eléctrico, medido en porcentaje.

Hogares con acceso a TV Cable, medido en porcentaje.

Hogares con acceso a telefonía fija, medido en porcentaje.

Hogares con acceso a telefonía móvil, medido en porcentaje.

Hogares con acceso a Internet, medido en porcentaje.

Hogares que utilizan gas para cocinar, medido en porcentaje.

Hogares que disponen de alumbrado eléctrico por red pública, medido en porcentaje.

Precio de la electricidad, medido en centavos de US$/kW.h.

Hogares que se abastecen de agua mediante la red pública, medido en porcentaje.

Continuidad de la provisión de agua, medido en número de horas al día.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Hogares que residen en viviendas con red pública de alcantarillado, medido en porcentaje.

Hogares que cuentan con el servicio de internet, medido en porcentaje.

Hogares con al menos un miembro que tiene teléfono celular, medido en porcentaje.

Densidad del transporte aéreo, medido en número de movimientos de pasajeros vía aérea (entrada y salida) por cada 1,000 habitantes.

Ingreso promedio mensual, medido en nuevos soles.

Pobreza, medido en porcentaje.

Pobreza extrema, medido en porcentaje.

Producto Bruto Interno real, medido en millones de soles.

Producto Bruto Interno real per cápita, medido en nuevos soles.

Stock de capital por trabajador, medido en nuevos soles.

Presupuesto público per cápita, medido en nuevos soles.

Gasto real por hogar mensual, medido en nuevos soles.

Incremento del gasto real por hogar, medido en promedio móvil tres años de la variación anual.

Disponibilidad de servicios financieros, medido en número de agentes bancarios, oficinas o cajeros automáticos por cada 100,000 habitantes adultos.

Acceso al crédito, medido en porcentaje del número de deudores entre el total de habitantes adultos.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Brecha de género en ingresos laborales, medido en porcentaje de ingresos laborales femeninos necesarios para alcanzar los ingresos masculinos.

Ejecución de la inversión pública, medido en porcentaje del gasto devengado del PIM de inversión (incluye gobierno local, regional y nacional).

Variación del índice de precios al consumidor, medido en porcentaje.

PBI Agricultura, Ganadería, Caza y Silvicultura, medido en porcentaje.

PBI Pesca y Acuicultura, medido en porcentaje.

PBI Extracción De Petróleo, Gas y Minerales, medido en porcentaje.

PBI Manufactura, medido en porcentaje.

PBI Electricidad, Gas y Agua, medido en porcentaje.

PBI construcción, medido en porcentaje.

PBI Comercio, medido en porcentaje.

PBI Transporte, Almacén, Correo y Mensajería, medido en porcentaje.

PBI Alojamiento y Restaurantes, medido en porcentaje.

PBI Telecomunicación y Servicios de Información, medido en porcentaje.

PBI Administración Publica y Defensa, medido en porcentaje.

PBI Otros Servicios, medido en porcentaje.

Percepción de la gestión pública, medido por el porcentaje de la población adulta que considera que la gestión pública del Gobierno Regional es buena o muy buena.

Conflictos sociales, medido por el número de conflictos sociales activos, latentes y resueltos.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Criminalidad, medido por el número de denuncias de delitos por cada 1,000 habitantes.

Homicidios, medido por el número de homicidios por cada 100,000 habitantes.

Presencia policial, medido por el número de habitantes por efectivo policial.

Resolución expedientes judiciales, medido por el número de expedientes resueltos de la carga judicial (pendientes + ingresantes).

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

CAPITULO IV

RESULTADOS Y DISCUSIÓN

4.1. CONSTRUCCIÓN DE LOS CONGLOMERADOS

La Clusterización se desarrolló analizando a los 24 departamentos del Perú (unidades de estudio), donde cada uno de ellos fueron procesados con los diferentes análisis y métodos para el logro de los objetivos de la investigación.

Tabla 2
Resumen de procesamiento de casos[(a)]

Válido		Perdidos		Total	
N	Por centaje	N	Por centaje	N	Por centaje
24	100,0%	0	0,0 %	24	100,0%

a. Distancia euclídea utilizada

Fuente: Elaboración propia

Para comprender el fenómeno de estudio, primero se analizaron las variables de manera descriptiva. (ver Anexo A. Análisis Exploratorio), se observó la dispersión de las variables en diagramas de cajas donde el caso Lima es el departamento con variables atípicas, el caso Lima se localiza en la parte superior de la distribución, indica que se trata de distribuciones asimétricas positivas, lo cual se corrigió sustituyendo los datos originales por su raíz cuadrada para transformar a las variables con datos atípicos. Se logró reducir la asimetría de las variables y, por lo tanto, la influencia que puedan tener los datos atípicos.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Antes de proceder con el análisis cluster en sí, fue necesario comprobar hasta qué punto los datos cumplen con los supuestos como la ausencia de correlación entre las variables y que las variables estén medidas en la misma escala, para lo cual estas fueron estandarizadas o tipificadas.

Para el análisis cluster, el dendrograma, o árbol jerárquico, es una representación gráfica del resultado del proceso de agrupamiento en forma de árbol. El gráfico representa los conglomerados de los departamentos mediante segmentos horizontales y las etapas de fusión mediante segmentos verticales. La distancia entre las etapas de fusión es proporcional a la distancia a la que se unen los departamentos en cada etapa (en una escala estandarizada de 25 puntos). El dendrograma es, pues, una herramienta útil para evaluar la homogeneidad de los conglomerados y decidir el número óptimo de grupos.

4.1.1. Primera conglomeración: medida de distancia euclídea y método de agrupación entre grupos

Para la primera conglomeración donde usamos la medida de distancia euclídea y el método de agrupación entre grupos, a una jerarquía de 6 puntos (de 25 puntos), el dendrograma diferencia 5 clusters de departamentos con perfiles socio-demográficos y económicos distintos. Así, los 24 departamentos se agrupan en los siguientes 5 conglomerados, de acuerdo con su nivel de industrialización, índice de pobreza, nivel de educación, datos demográficos, calidad de salud, servicios públicos y la convivencia social.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

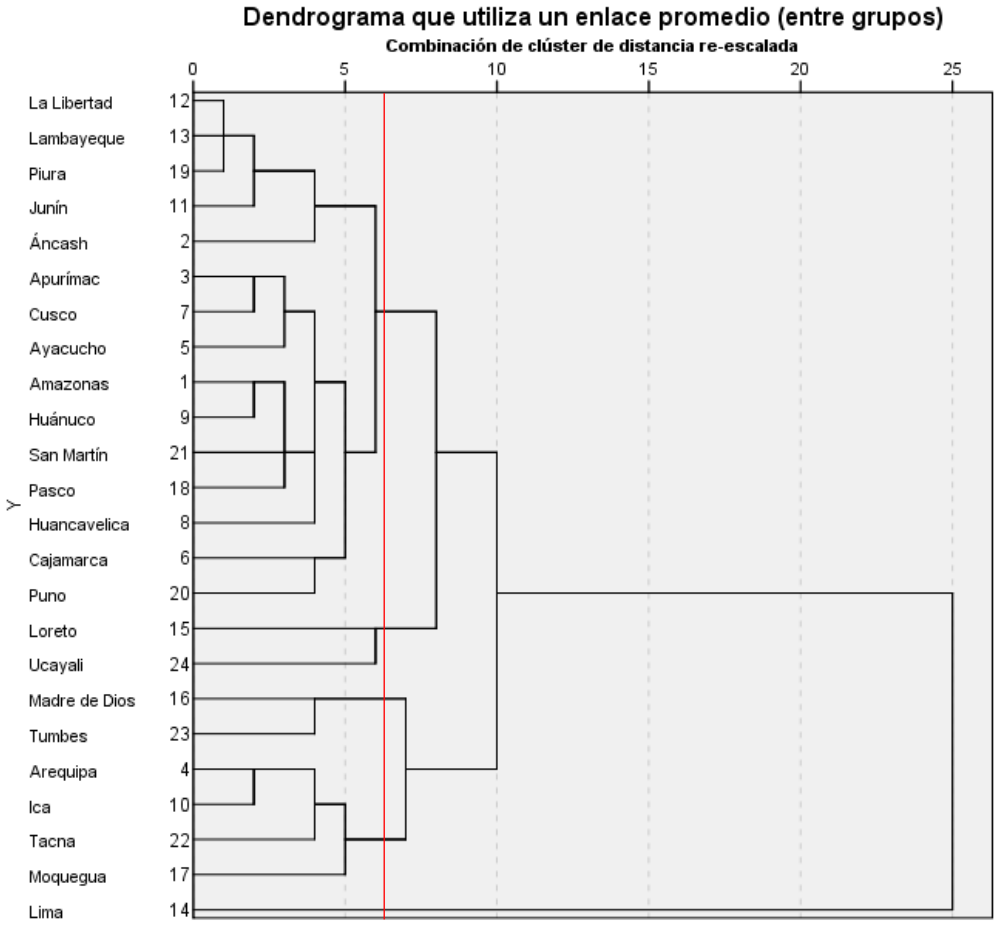

Figura 6 Dendograma para la medida de distancia euclídea y método de agrupación entre grupos.

Fuente: Elaboración propia

C1: La Libertad, Lambayeque, Piura, Junín, Ancash, Apurímac, Cusco, Ayacucho, Amazonas, Huánuco, San Martin, Pasco, Huancavelica, Cajamarca y Puno.

C2: Loreto y Ucayali.

C3: Madre de Dios y Tumbes.

C4: Arequipa, Ica, Tacna y Moquegua.

C5: Lima.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

4.1.2. Segunda conglomeración: medida de distancia euclídea y método de agrupación dentro de grupos

Para la segunda conglomeración donde usamos la medida de distancia euclídea y el método de agrupación dentro de grupos, a una jerarquía de 12 puntos (de 25 puntos), el dendrograma diferencia 6 clusters de departamentos con perfiles socio-demográficos y económicos distintos. Así, los 24 departamentos se agrupan en los siguientes 6 conglomerados, de acuerdo con su nivel de industrialización, índice de pobreza, nivel de educación, datos demográficos, calidad de salud, servicios públicos y la convivencia social.

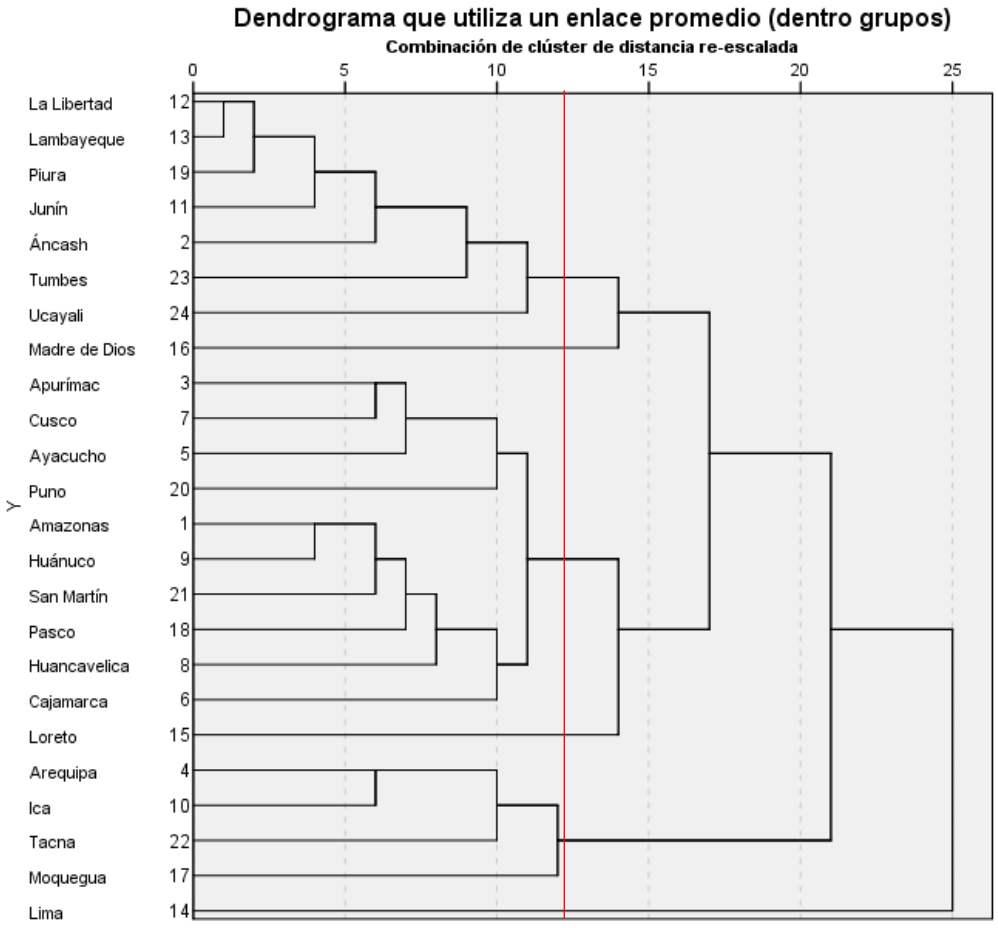

Figura 7 Dendograma para la medida de distancia euclídea y método de agrupación dentro de grupos.

Fuente: Elaboración propia

C1: La Libertad, Lambayeque, Piura, Junín, Ancash, Tumbes y Ucayali.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

C2: Madre de Dios.

C3: Apurímac, Cusco, Ayacucho, Puno, Amazonas, Huánuco, San Martin, Pasco, Huancavelica y Cajamarca.

C4: Loreto.

C5: Arequipa, Ica, Tacna y Moquegua.

C6: Lima.

4.1.3. Tercera conglomeración: medida de distancia euclídea y método de agrupación enlace completo

Para la tercera conglomeración donde usamos la medida de distancia euclídea y el método de agrupación enlace completo, a una jerarquía de 6 puntos (de 25 puntos), el dendrograma diferencia 7 clusters de departamentos con perfiles socio-demográficos y económicos distintos. Así, los 24 departamentos se agrupan en los siguientes 7 conglomerados, de acuerdo con su nivel de industrialización, índice de pobreza, nivel de educación, datos demográficos, calidad de salud, servicios públicos y la convivencia social.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

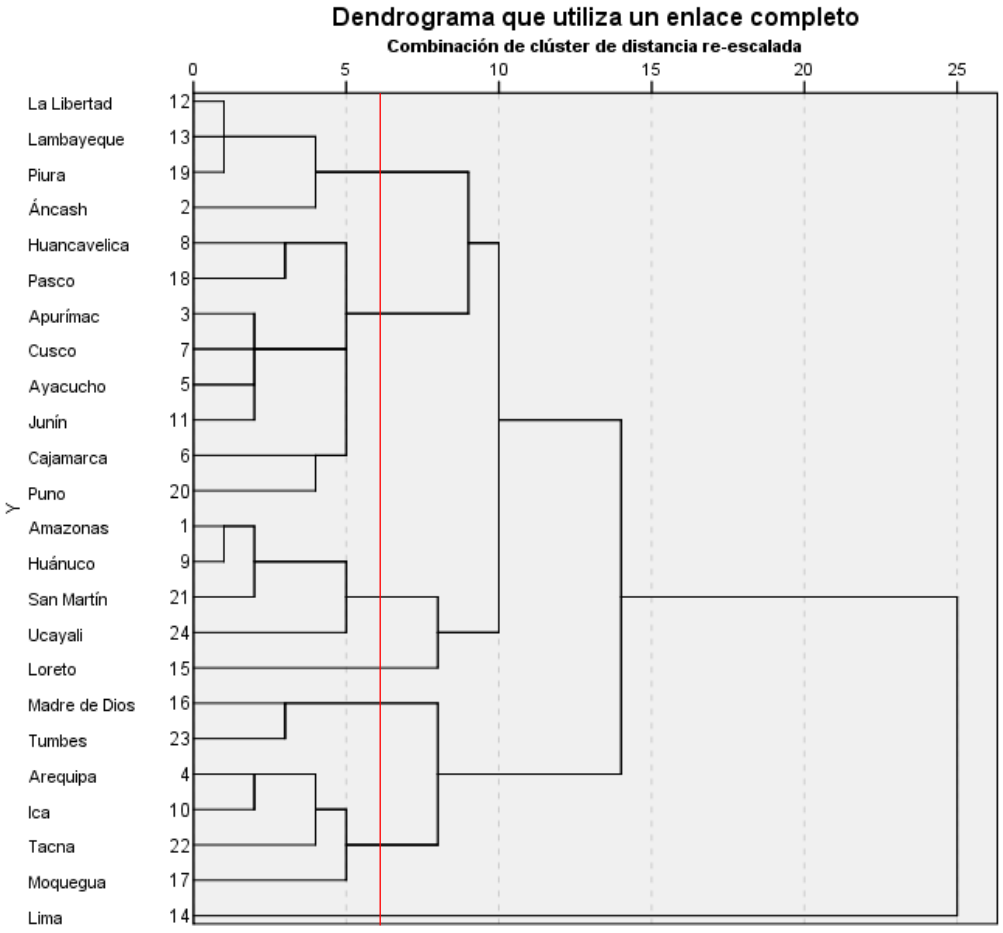

Figura 8 Dendograma para la medida de distancia euclídea y método de agrupación enlace completo.

Fuente: Elaboración propia

C1: La Libertad, Lambayeque y Ancash.

C2: Huancavelica, Pasco, Apurímac, Cusco, Ayacucho, Junín, Cajamarca y Puno.

C3: Amazonas, Huánuco, San Martin y Ucayali.

C4: Loreto.

C5: Madre de Dios y Tumbes.

C6: Arequipa, Ica, Tacna y Moquegua.

C7: Lima.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

4.1.4. Cuarta conglomeración: medida de distancia euclídea y método de agrupación enlace de Ward

Para la cuarta conglomeración donde usamos la medida de distancia euclídea y el método de agrupación enlace de Ward, a una jerarquía de 3 puntos (de 25 puntos), el dendrograma diferencia 6 clusters de departamentos con perfiles socio-demográficos y económicos distintos. Así, los 24 departamentos se agrupan en los siguientes 6 conglomerados, de acuerdo con su nivel de industrialización, índice de pobreza, nivel de educación, datos demográficos, calidad de salud, servicios públicos y la convivencia social.

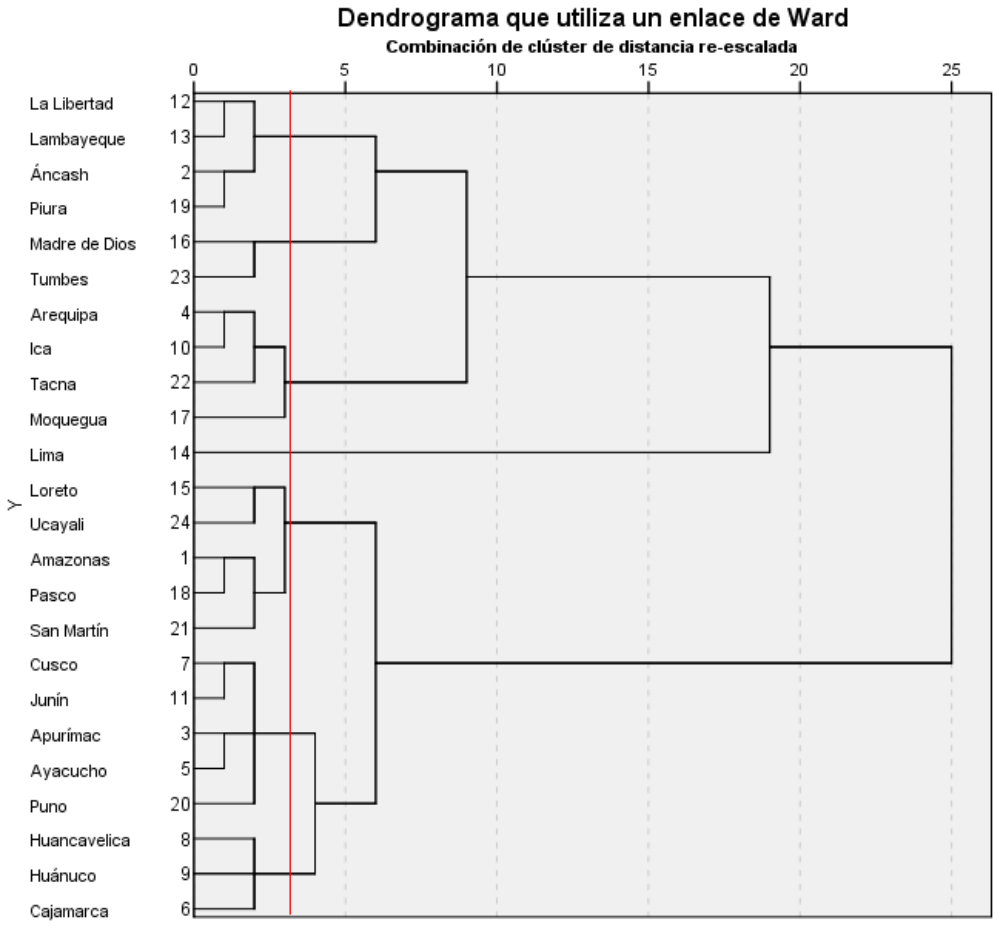

Figura 9 Dendograma para la medida de distancia euclídea y método de agrupación enlace de Ward.

Fuente: Elaboración propia

C1: La Libertad, Lambayeque, Ancash y Piura.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

C2: Madre de Dios y Tumbes.

C3: Arequipa, Ica, Tacna y Moquegua.

C4: Lima.

C5: Loreto, Ucayali, Amazonas, Pasco y San Martin.

C6: Cusco, Junín, Apurímac, Ayacucho, Puno, Huancavelica, Huánuco y Cajamarca.

4.1.5. Quinta conglomeración: medida de similitud coseno de vectores y método de agrupación entre grupos

Para la quinta conglomeración donde usamos la medida de similitud coseno de vectores y el método de agrupación entre grupos, a una jerarquía de 14 puntos (de 25 puntos), el dendrograma diferencia 5 clusters de departamentos con perfiles socio-demográficos y económicos distintos. Así, los 24 departamentos se agrupan en los siguientes 5 conglomerados, de acuerdo con su nivel de industrialización, índice de pobreza, nivel de educación, datos demográficos, calidad de salud, servicios públicos y la convivencia social.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

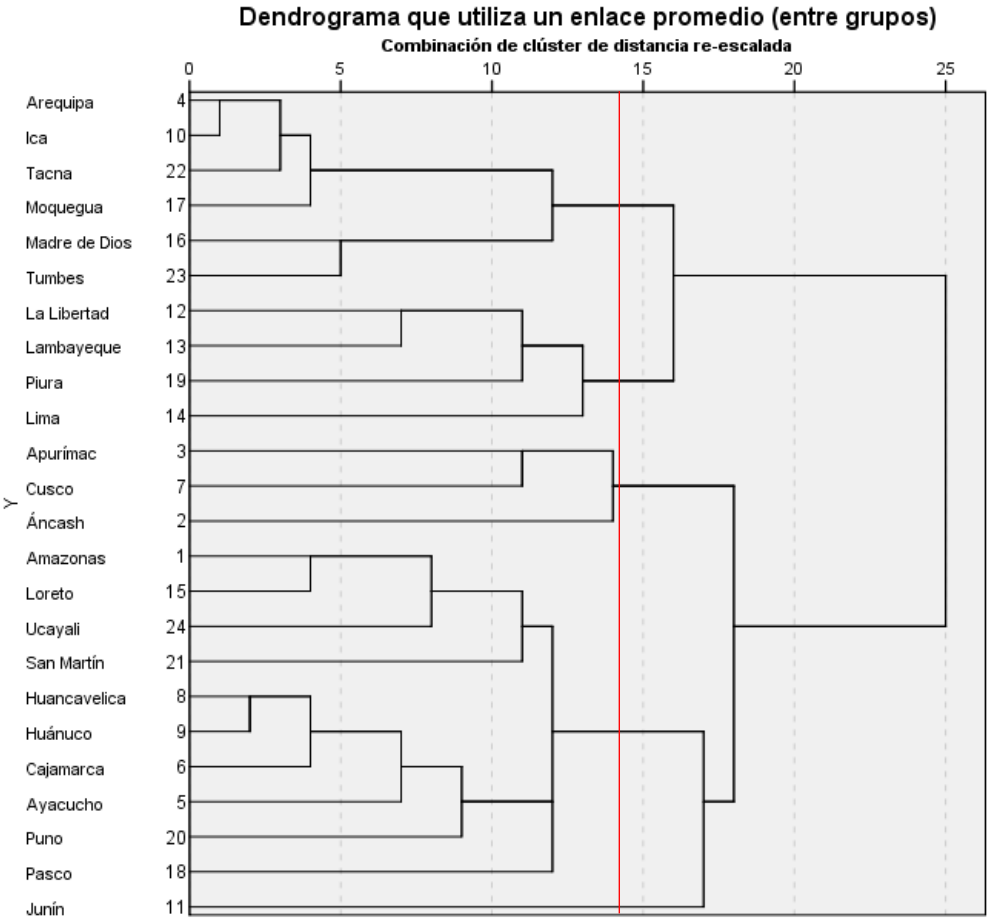

Figura 10 Dendograma para la medida de similitud coseno de vectores y método de agrupación entre grupos.

Fuente: Elaboración propia

C1: Arequipa, Ica, Tacna, Moquegua, Madre de Dios y Tumbes.

C2: La Libertad, Lambayeque, Piura y Lima.

C3: Apurímac, Cusco y Ancash.

C4: Amazonas, Loreto, Ucayali, San Martin, Huancavelica, Huánuco, Cajamarca, Ayacucho, Puno y Pasco.

C5: Junín.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

4.1.6. Sexta conglomeración: medida de similitud coseno de vectores y método de agrupación dentro de grupos

Para la sexta conglomeración donde usamos la medida de similitud coseno de vectores y el método de agrupación dentro de grupos, a una jerarquía de 15 puntos (de 25 puntos), el dendrograma diferencia 4 clusters de departamentos con perfiles socio-demográficos y económicos distintos. Así, los 24 departamentos se agrupan en los siguientes 4 conglomerados, de acuerdo con su nivel de industrialización, índice de pobreza, nivel de educación, datos demográficos, calidad de salud, servicios públicos y la convivencia social.

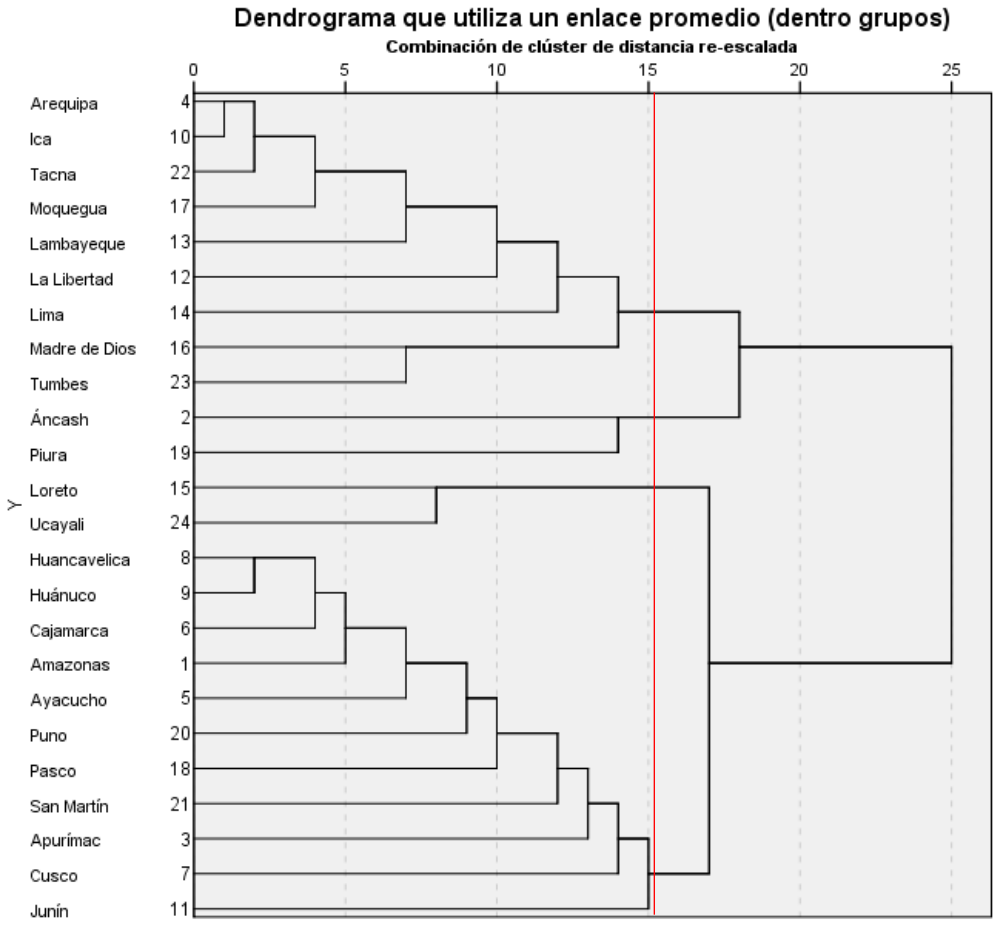

Figura 11 Dendograma para la medida de similitud coseno de vectores y método de agrupación dentro de grupos.

Fuente: Elaboración propia

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

C1: Arequipa, Ica, Tacna, Moquegua, Lambayeque, La Libertad, Lima, Madre de Dios y Tumbes.

C2: Ancash y Piura.

C3: Loreto y Ucayali.

C4: Huancavelica, Huánuco, Cajamarca, Amazonas, Ayacucho, Puno, Pasco, San Martin, Apurímac, Cusco y Junín.

4.1.7. Séptima conglomeración: medida de similitud coseno de vectores y método de agrupación enlace completo.

Para la séptima conglomeración donde usamos la medida de similitud coseno de vectores y el método de agrupación enlace completo, a una jerarquía de 13 puntos (de 25 puntos), el dendrograma diferencia 5 clusters de departamentos con perfiles socio-demográficos y económicos distintos. Así, los 24 departamentos se agrupan en los siguientes 5 conglomerados, de acuerdo con su nivel de industrialización, índice de pobreza, nivel de educación, datos demográficos, calidad de salud, servicios públicos y la convivencia social.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

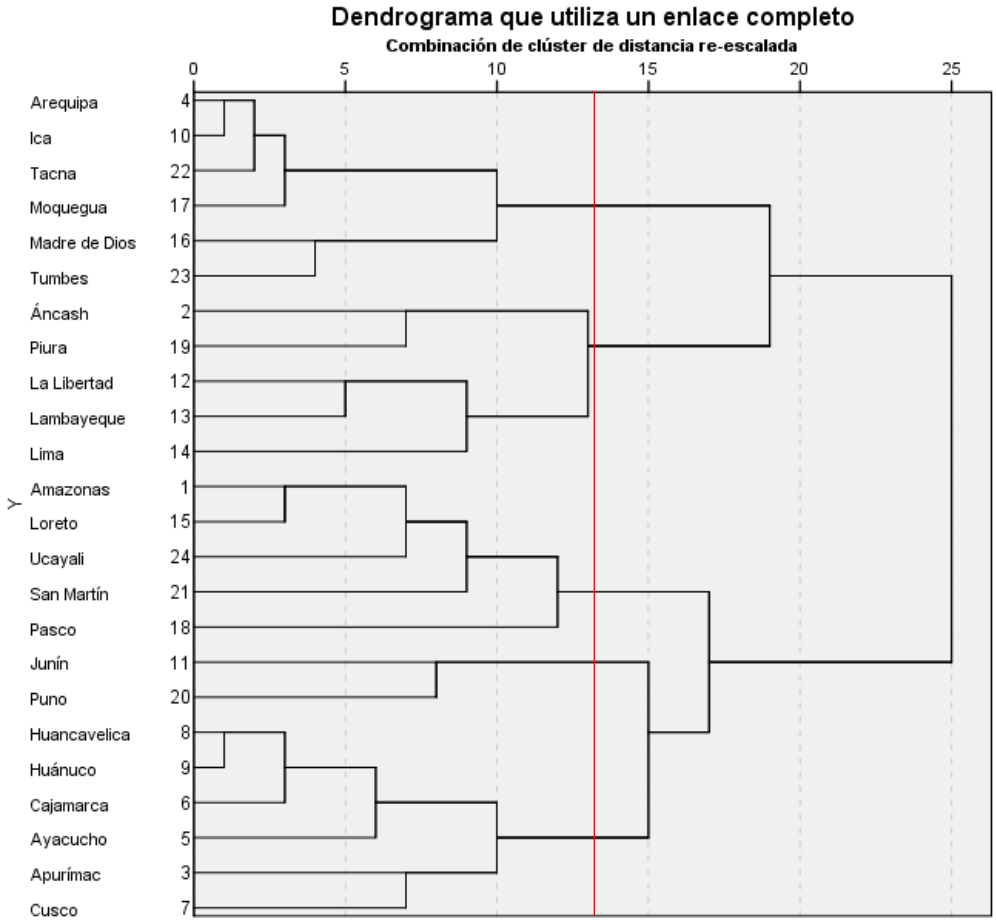

Figura 12 Dendograma para la medida de similitud coseno de vectores y método de agrupación enlace completo.

Fuente: Elaboración propia

C1: Arequipa, Ica, Tacna, Moquegua, Madre de Dios y Tumbes.

C2: Ancash, Piura, La Libertad, Lambayeque y Lima.

C3: Amazonas, Loreto, Ucayali, San Martin y Pasco.

C4: Junín y Puno.

C5: Huancavelica, Huánuco, Cajamarca, Ayacucho, Apurímac y Cusco.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

4.2. CLUSTERIZACIÓN POR MEDIDA DE DISTANCIA EUCLÍDEA Y MÉTODO DE AGRUPACIÓN ENLACE DE WARD

El resultado de los diferentes clusters es producto del análisis multivariado, en concreto del análisis clúster jerárquico, la selección de las variables está respaldada por la evaluación de los supuestos de normalidad (ver Anexo B. Pruebas de Normalidad).

La medida de Distancia Euclídea y el método de agrupación Enlace de Ward se eligieron para analizar e identificar la estructura subyacente de clusterización de los 24 departamentos del Perú, en grupos homogéneos según indicadores socioeconómicos (ver Anexo C. Matriz de Proximidades).

La Tabla 3. Informe de medias, proporciona una visión especifica de la conglomeración, donde podemos verificar la disimilaridad entre los 6 clusters. Los valores medios de las variables de cada cluster (centroide) sirven de base para definir el perfil de cada uno de los clusters.

Tabla 3
Informe de Medias de los 6 conglomerados

Ward Method	1	2	3	4	5	6
Esperanza de vida	72,2800	75,7750	72,4625	76,6000	79,9000	74,1500
Tasa Global de Fecundidad (hijos x mujer)	3,1800	2,5725	2,6000	2,1500	1,9000	2,8000
Mortalidad en la niñez	32,8000	19,0000	30,1250	18,0000	15,0000	21,0000
Crecimiento poblacional	0,7280	0,9050	-0,3150	1,7275	1,5800	2,5050
Nacimientos	13357,0000	29076,0000	18391,8750	12136,5000	156984,0000	3975,5000
Defunciones	2250,0000	7637,0000	4756,6250	3148,7500	43735,0000	668,5000
Población electoral	427234,6000	1131533,0000	758919,7500	532112,5000	7283679,0000	133786,5000
Población inmigrante	95578,2000	169681,2500	77789,3750	175361,7500	2985643,0000	52002,0000
Población emigrante	148141,2000	324789,7500	377914,7500	98326,7500	462442,0000	27164,5000

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Población censada	565359,8000	1478917,0000	881986,8750	684422,5000	9485405,0000	182966,5000
Población urbana censada	376159,0000	1133309,2500	468399,8750	626009,2500	9324796,0000	163667,5000
Población rural censada	189200,8000	345607,7500	413587,0000	58413,2500	160609,0000	19299,0000
Tasa de Analfabetismo	7,2000	7,4500	10,7375	3,5250	2,0000	4,0000
Tasa de Analfabetismo masculino	4,5000	4,2750	4,8000	1,3750	1,0000	3,4500
Tasa de Analfabetismo femenino	10,4200	10,6250	16,7750	5,8500	2,9000	4,8000
Tasa de Asistencia escolar (secundaria)	77,2800	80,5500	84,0000	90,4500	88,4000	85,2000
Nivel de educación de la Población de 15 y más años (Primaria)	30,8800	25,3750	28,9375	15,3000	10,1000	21,4000
Nivel de educación de la Población de 15 y más años (Sup. No Universitaria)	10,8400	12,7250	10,0625	17,0500	16,0000	16,5000
Nivel de educación de la Población de 15 y más años (Sup. Universitaria)	11,1600	13,8750	12,8125	19,8000	24,2000	12,2000
Asistencia escolar inicial	84,5880	87,2150	88,1338	91,6775	88,3900	87,7000
Rendimiento en lectura	21,6740	29,3350	23,6475	45,5425	39,7900	26,2400
Rendimiento en matemáticas	17,4340	22,5375	21,5963	36,8775	26,1500	16,9650
Colegios con acceso a internet	24,5620	53,6750	36,2850	65,0750	78,9200	54,9450

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Cobertura de algún Seguro de Salud	82,8200	77,2250	81,6125	65,4750	73,7000	76,7000
Tasa de desnutrición crónica en menores de 5 años	19,0400	14,5250	21,4875	4,9500	5,1000	7,7500
Niños con anemia (De 6 a 35 meses de edad)	56,1000	41,1250	51,5750	37,2250	33,3000	52,0500
Desnutrición crónica	14,8800	10,0250	15,6250	2,9750	3,2000	4,9500
Cobertura del personal médico	7,0140	16,0550	9,0550	27,5475	36,0100	10,6250
Partos institucionales	80,6200	90,4000	91,5125	98,1000	98,5000	96,8500
Población en edad de trabajar (De 14 y más años de edad)	451,9400	1150,3250	731,8250	511,9750	7899,5000	146,5500
PEA	337,4400	805,2250	615,0125	355,7000	5032,2000	110,6000
PEA ocupada	329,2400	781,0750	598,8000	342,0500	4694,3000	107,1000
PEA Ocupada en Agricultura, pesca, minería	132,8000	218,7500	285,7000	63,9250	62,0000	21,4500
PEA Ocupada en Manufactura	17,9000	72,9500	38,3375	31,2250	656,5000	6,9000
PEA ocupada en empresas de 1-10 trabajadores	79,7600	73,2300	83,7625	68,2500	59,8000	78,2000
PEA desempleada	8,2000	21,6000	16,1875	13,6500	337,9000	3,5000
Empleo adecuado	41,6780	47,6150	33,0150	62,6525	64,2600	63,7300
Educación de la fuerza laboral	21,4180	27,7500	20,8238	39,8900	42,5400	27,8550
Empleo informal	82,8960	77,4550	86,6913	67,4725	55,9500	79,5350
Desempleo juvenil urbano	10,1900	11,8450	13,7363	11,8125	16,4800	6,8500

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Administ. Publica y Defensa	4,7000	4,6250	5,4000	2,7500	4,3000	3,8000
Comercio	2,0800	3,2250	2,1375	2,2500	2,7000	2,8000
Hogares con acceso a agua potable	37,5800	61,4250	43,3625	77,8750	95,0000	68,9500
Hogares con acceso a desagüe	52,2600	75,3500	56,5500	87,3500	94,8000	63,5500
Hogares con acceso a alumbrado eléctrico	86,2200	95,6250	90,7875	96,2750	99,6000	94,4500
hogares con acceso a TV Cable	37,6200	32,3000	16,4500	34,0500	58,9000	51,5000
Hogares con acceso a telefonía fija	5,9200	17,5500	6,5875	17,0250	46,6000	9,4000
Hogares con acceso a telefonía móvil	84,9800	90,9000	86,9500	93,7000	94,0000	94,0500
Hogares con acceso a Internet	13,1000	21,3500	10,0375	34,8500	52,9000	17,2500
Hogares que utilizan gas para cocinar	23,8480	39,0075	25,6425	70,2700	63,0200	48,9600
Cobertura de electricidad	84,4800	95,8025	89,0488	95,9675	99,3400	94,2050
Cobertura de agua	73,2380	89,2825	83,8975	93,0750	94,4500	83,9400
Cobertura de desagüe	43,5240	69,8550	48,2100	81,2775	89,7200	54,2450
Hogares con internet	10,7860	21,2250	8,3300	31,8375	45,5700	16,5550
Hogares con al menos un celular	82,8700	89,7475	84,0825	92,6575	92,8800	91,8950
Ingreso promedio mensual	1096,2200	1126,8000	937,1125	1488,1500	1921,1000	1465,0500
Pobreza	28,6300	21,7875	32,6938	10,8750	13,3500	8,4000
Pobreza extrema	6,1000	3,4000	7,3375	0,7000	0,7000	0,7000
Producto Bruto Interno real	5234,4100	16818,6600	8918,7263	13692,6175	215457,8400	2442,4300
Producto Bruto Interno real per cápita	9430,5460	11224,1200	8242,5750	26754,8400	19860,4900	13850,4400

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Stock de capital por trabajador	17546,1800	21358,3325	14419,0150	49413,9900	37933,1400	23563,2000
Presupuesto público per cápita	4285,6600	3054,2275	4213,3463	4923,4875	7602,0300	5594,1600
Gasto real por hogar mensual	1326,7540	1525,6800	1019,5338	1574,0575	2423,6200	1731,6700
Acceso al crédito	20,1780	32,1250	22,3313	37,7225	46,8800	33,4300
PBI Agricultura, Ganadería, Caza y Silvicultura	3,0400	5,5000	3,6750	4,2250	15,3000	0,6000
PBI Pesca y Acuicultura	0,3200	13,5750	0,3625	2,6250	26,4000	2,0000
PBI Manufactura	0,5800	3,9250	0,7625	3,4250	60,9000	0,3000
PBI Electricidad, Gas y Agua	0,6000	2,6500	3,0250	2,2250	53,0000	0,2000
PBI Construcción	1,2000	3,8750	2,6250	4,4750	38,3000	0,6500
PBI Comercio	1,2800	3,8250	1,6125	2,5000	54,2000	0,7000
PBI Transporte, Almacén, Correo y Mensajería	0,7000	3,7250	1,5000	2,8750	57,4000	0,4000
PBI Alojamiento y Restaurantes	0,7800	2,8250	1,4125	1,7000	65,9000	0,3000
PBI Telecom. Y Servicios de Información	0,7000	3,2000	1,1375	1,9250	66,1000	0,3500
PBI Administ. Publica y Defensa	1,5200	3,8000	2,4625	2,0250	48,2000	0,6000
PBI Otros Servicios	0,9000	2,9750	1,4625	1,9250	63,6000	0,3500
Percepción de la gestión pública	44,2360	31,0100	31,5100	29,0000	31,8000	26,5250
Criminalidad	6,2040	8,6925	4,9488	10,1375	12,7600	18,8450
Homicidios	9,2040	7,4250	6,6113	5,9600	6,6600	19,7750

Fuente: Elaboración propia.

4.3. VALIDACIÓN DE LA CLUSTERIZACIÓN

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Para validar la clusterización (análisis de interdependencia) de las regiones del Perú según indicadores socioeconómicos, se planteó la siguiente hipótesis:

$H_0 : \mu_{C1} = \mu_{C2} = \mu_{C3} = \mu_{C4} = \mu_{C5} = \mu_{C6}$ Las medias de los clusters son iguales.

$H_a : \mu_{Ci} \neq \mu_{Cj} \quad para\ algun\ i \neq j$ No todas las medias de los clusters son iguales.

De esta manera, para la presente investigación de clusterización de las regiones del Perú en grupos homogéneos según indicadores socioeconómicos, usando la medida de distancia euclídea y el método de agrupación enlace de Ward, se planteó:

H_0 (hipótesis nula), Las medias de los clusters son iguales o la clusterización, según indicadores socioeconómicos, no es determinante para evidenciar la estructura subyacente de agrupación homogénea de los departamentos del Perú.

Ha (hipótesis alterna), No todas las medias de los cluster son iguales o la clusterización es determinante para evidenciar la estructura subyacente de agrupación homogénea de los departamentos del Perú.

En la Tabla 4 Análisis de Varianza, podemos ver la tabla anova para cada una de las 78 variables respecto a los 6 clusters agrupados usando la medida de distancia Euclídea y el método de agrupación enlace de Ward, y en particular el p-valor el cual nos permitirá obtener el resultado del contraste, si utilizamos como nivel de error del contraste el 5%, también denominado alfa, comprobamos que el p valor, de las 78 variables involucradas en la clusterización, son menores que el nivel de significación utilizado y por lo tanto, los datos presentas suficiente evidencia como para rechazar la hipótesis nula, las medias son iguales para todos los clusters, por lo tanto aceptamos la hipótesis alterna, la clusterización es determinante para evidencias la estructura subyacente de agrupación homogénea de los 24 departamentos del Perú.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Tabla 4

Análisis de Varianza

ANOVA

		Suma de cuadrados	gl	Media cuadrática	F	Sig.
Esperanza de vida	Entre grupos	107,874	5	21,575	17,621	,000
	Dentro de grupos	22,039	18	1,224		
	Total	129,913	23			
Tasa Global de Fecundidad (hijos x mujer)	Entre grupos	3,045	5	,609	7,095	,001
	Dentro de grupos	1,545	18	,086		
	Total	4,590	23			
Mortalidad en la niñez	Entre grupos	982,158	5	196,432	6,576	,001
	Dentro de grupos	537,675	18	29,871		
	Total	1519,833	23			
Crecimiento poblacional	Entre grupos	19,841	5	3,968	4,456	,008
	Dentro de grupos	16,030	18	,891		
	Total	35,871	23			
Nacimientos	Entre grupos	19927257903,250	5	3985451580,650	70,822	,000
	Dentro de grupos	1012937716,375	18	56274317,576		
	Total	20940195619,625	23			
Defunciones	Entre grupos	1604947262,708	5	320989452,542	100,187	,000
	Dentro de grupos	57670421,125	18	3203912,285		
	Total	1662617683,833	23			
Población electoral	Entre grupos	43930518404591,760	5	8786103680918,352	106,049	,000
	Dentro de grupos	1491288196398,200	18	82849344244,344		
	Total	45421806600989,960	23			
Población inmigrante	Entre grupos	7959016104935,450	5	1591803220987,090	250,667	,000
	Dentro de grupos	114304804456,175	18	6350266914,232		
	Total	8073320909391,625	23			
Población emigrante	Entre grupos	441579800606,325	5	88315960121,265	10,784	,000
	Dentro de grupos	147410893158,300	18	8189494064,350		
	Total	588990693764,625	23			
Población censada	Entre grupos	74906915773004,660	5	14981383154600,932	95,871	,000
	Dentro de grupos	2812803498481,175	18	156266861026,732		
	Total	77719719271485,830	23			
Población urbana censada	Entre grupos	75420407379334,750	5	15084081475866,950	145,123	,000
	Dentro de grupos	1870915918836,875	18	103939773268,715		
	Total	77291323298171,620	23			
Población rural censada	Entre grupos	530186204210,658	5	106037240842,132	5,029	,005
	Dentro de grupos	379534798130,300	18	21085266562,794		
	Total	909721002340,958	23			
Tasa de Analfabetismo	Entre grupos	201,529	5	40,306	10,467	,000
	Dentro de grupos	69,316	18	3,851		
	Total	270,845	23			

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Tasa de Analfabetismo masculino	Entre grupos	42,958	5	8,592	7,464	,001
	Dentro de grupos	20,720	18	1,151		
	Total	63,678	23			
Tasa de Analfabetismo femenino	Entre grupos	517,609	5	103,522	10,686	,000
	Dentro de grupos	174,380	18	9,688		
	Total	691,990	23			
Tasa de Asistencia escolar (secundaria)	Entre grupos	452,925	5	90,585	7,250	,001
	Dentro de grupos	224,888	18	12,494		
	Total	677,813	23			
Nivel de educación de la Población de 15 y más años (Primaria)	Entre grupos	921,662	5	184,332	6,899	,001
	Dentro de grupos	480,954	18	26,720		
	Total	1402,616	23			
Nivel de educación de la Población de 15 y más años (Sup. No Universitaria)	Entre grupos	188,245	5	37,649	9,546	,000
	Dentro de grupos	70,988	18	3,944		
	Total	259,233	23			
Nivel de educación de la Población de 15 y más años (Sup. Universitaria)	Entre grupos	295,445	5	59,089	9,718	,000
	Dentro de grupos	109,448	18	6,080		
	Total	404,893	23			
Asistencia escolar inicial	Entre grupos	114,352	5	22,870	5,527	,003
	Dentro de grupos	74,489	18	4,138		
	Total	188,841	23			
Rendimiento en lectura	Entre grupos	1722,417	5	344,483	12,257	,000
	Dentro de grupos	505,883	18	28,105		
	Total	2228,300	23			
Rendimiento en matemáticas	Entre grupos	1023,282	5	204,656	6,385	,001
	Dentro de grupos	576,986	18	32,055		
	Total	1600,268	23			
Colegios con acceso a internet	Entre grupos	5957,186	5	1191,437	12,398	,000
	Dentro de grupos	1729,735	18	96,096		
	Total	7686,921	23			
Cobertura de algún Seguro de Salud	Entre grupos	870,238	5	174,048	3,557	,021
	Dentro de grupos	880,752	18	48,931		
	Total	1750,990	23			
Tasa de desnutrición crónica en menores de 5 años	Entre grupos	1025,050	5	205,010	7,598	,001
	Dentro de grupos	485,683	18	26,982		
	Total	1510,733	23			
Niños con anemia (De 6 a 35 meses de edad)	Entre grupos	1329,921	5	265,984	4,216	,010
	Dentro de grupos	1135,675	18	63,093		
	Total	2465,596	23			

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Desnutrición crónica	Entre grupos	641,837	5	128,367	8,406	,000
	Dentro de grupos	274,883	18	15,271		
	Total	916,720	23			
Cobertura del personal médico	Entre grupos	1697,476	5	339,495	10,617	,000
	Dentro de grupos	575,567	18	31,976		
	Total	2273,043	23			
Partos institucionales	Entre grupos	868,317	5	173,663	4,018	,013
	Dentro de grupos	777,922	18	43,218		
	Total	1646,238	23			
Población en edad de trabajar (De 14 y más años de edad)	Entre grupos	52134704,243	5	10426940,849	113,019	,000
	Dentro de grupos	1660650,607	18	92258,367		
	Total	53795354,850	23			
PEA	Entre grupos	20692591,685	5	4138518,337	96,790	,000
	Dentro de grupos	769640,828	18	42757,824		
	Total	21462232,513	23			
PEA ocupada	Entre grupos	17926952,911	5	3585390,582	88,491	,000
	Dentro de grupos	729305,249	18	40516,958		
	Total	18656258,160	23			
PEA Ocupada en Agricultura, pesca, minería	Entre grupos	223862,071	5	44772,414	6,877	,001
	Dentro de grupos	117193,702	18	6510,761		
	Total	341055,773	23			
PEA Ocupada en Manufactura	Entre grupos	377972,684	5	75594,537	188,434	,000
	Dentro de grupos	7221,096	18	401,172		
	Total	385193,780	23			
PEA ocupada en empresas de 1-10 trabajadores	Entre grupos	1065,381	5	213,076	14,830	,000
	Dentro de grupos	258,624	18	14,368		
	Total	1324,004	23			
PEA desempleada	Entre grupos	101292,381	5	20258,476	395,117	,000
	Dentro de grupos	922,899	18	51,272		
	Total	102215,280	23			
Empleo adecuado	Entre grupos	3523,999	5	704,800	15,689	,000
	Dentro de grupos	808,639	18	44,924		
	Total	4332,638	23			
Educación de la fuerza laboral	Entre grupos	1369,509	5	273,902	18,107	,000
	Dentro de grupos	272,288	18	15,127		
	Total	1641,797	23			
Empleo informal	Entre grupos	1620,200	5	324,040	24,105	,000
	Dentro de grupos	241,976	18	13,443		
	Total	1862,176	23			
Desempleo juvenil urbano	Entre grupos	113,619	5	22,724	3,030	,037
	Dentro de grupos	134,978	18	7,499		
	Total	248,597	23			
	Entre grupos	20,012	5	4,002	6,138	,002

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Administ. Publica y Defensa	Dentro de grupos	11,738	18	,652		
	Total	31,750	23			
Comercio	Entre grupos	4,285	5	,857	3,618	,019
	Dentro de grupos	4,264	18	,237		
	Total	8,550	23			
Hogares con acceso a agua potable	Entre grupos	6847,123	5	1369,425	5,939	,002
	Dentro de grupos	4150,347	18	230,575		
	Total	10997,470	23			
Hogares con acceso a desagüe	Entre grupos	4672,543	5	934,509	9,646	,000
	Dentro de grupos	1743,857	18	96,881		
	Total	6416,400	23			
Hogares con acceso a alumbrado eléctrico	Entre grupos	372,973	5	74,595	6,637	,001
	Dentro de grupos	202,317	18	11,240		
	Total	575,290	23			
hogares con acceso a TV Cable	Entre grupos	3575,247	5	715,049	9,510	,000
	Dentro de grupos	1353,398	18	75,189		
	Total	4928,645	23			
Hogares con acceso a telefonía fija	Entre grupos	1853,919	5	370,784	17,591	,000
	Dentro de grupos	379,394	18	21,077		
	Total	2233,313	23			
Hogares con acceso a telefonía móvil	Entre grupos	292,185	5	58,437	4,360	,009
	Dentro de grupos	241,273	18	13,404		
	Total	533,458	23			
Hogares con acceso a Internet	Entre grupos	2998,810	5	599,762	17,199	,000
	Dentro de grupos	627,704	18	34,872		
	Total	3626,513	23			
Hogares que utilizan gas para cocinar	Entre grupos	7254,241	5	1450,848	13,318	,000
	Dentro de grupos	1960,828	18	108,935		
	Total	9215,069	23			
Cobertura de electricidad	Entre grupos	522,755	5	104,551	8,279	,000
	Dentro de grupos	227,303	18	12,628		
	Total	750,058	23			
Cobertura de agua	Entre grupos	1122,255	5	224,451	3,279	,028
	Dentro de grupos	1232,027	18	68,446		
	Total	2354,283	23			
Cobertura de desagüe	Entre grupos	5570,203	5	1114,041	14,219	,000
	Dentro de grupos	1410,287	18	78,349		
	Total	6980,489	23			
Hogares con internet	Entre grupos	2562,559	5	512,512	19,794	,000
	Dentro de grupos	466,068	18	25,893		
	Total	3028,627	23			
Hogares con al menos un celular	Entre grupos	392,929	5	78,586	6,417	,001
	Dentro de grupos	220,449	18	12,247		

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

	Total	613,378	23			
Ingreso promedio mensual	Entre grupos	1609892,967	5	321978,593	15,652	,000
	Dentro de grupos	370287,252	18	20571,514		
	Total	1980180,218	23			
Pobreza	Entre grupos	2016,027	5	403,205	7,040	,001
	Dentro de grupos	1030,890	18	57,272		
	Total	3046,916	23			
Pobreza extrema	Entre grupos	183,128	5	36,626	3,811	,016
	Dentro de grupos	172,984	18	9,610		
	Total	356,111	23			
Producto Bruto Interno real	Entre grupos	41025421895,837	5	8205084379,167	286,367	,000
	Dentro de grupos	515741968,529	18	28652331,585		
	Total	41541163864,366	23			
Producto Bruto Interno real per cápita	Entre grupos	1062666499,401	5	212533299,880	4,560	,007
	Dentro de grupos	838920999,142	18	46606722,175		
	Total	1901587498,543	23			
Stock de capital por trabajador	Entre grupos	3747585007,598	5	749517001,520	5,438	,003
	Dentro de grupos	2480945294,854	18	137830294,159		
	Total	6228530302,452	23			
Presupuesto público per cápita	Entre grupos	21787362,450	5	4357472,490	3,730	,017
	Dentro de grupos	21030144,761	18	1168341,376		
	Total	42817507,211	23			
Gasto real por hogar mensual	Entre grupos	2625592,357	5	525118,471	14,462	,000
	Dentro de grupos	653597,997	18	36311,000		
	Total	3279190,354	23			
Acceso al crédito	Entre grupos	1424,572	5	284,914	7,962	,000
	Dentro de grupos	644,146	18	35,786		
	Total	2068,719	23			
PBI Agricultura, Ganadería, Caza y Silvicultura	Entre grupos	164,799	5	32,960	4,945	,005
	Dentro de grupos	119,974	18	6,665		
	Total	284,773	23			
PBI Pesca y Acuicultura	Entre grupos	1057,038	5	211,408	9,897	,000
	Dentro de grupos	384,502	18	21,361		
	Total	1441,540	23			
PBI Manufactura	Entre grupos	3408,034	5	681,607	436,900	,000
	Dentro de grupos	28,082	18	1,560		
	Total	3436,116	23			
PBI Electricidad, Gas y Agua	Entre grupos	2514,477	5	502,895	113,903	,000
	Dentro de grupos	79,473	18	4,415		
	Total	2593,950	23			
PBI Construcción	Entre grupos	1253,558	5	250,712	112,665	,000
	Dentro de grupos	40,055	18	2,225		
	Total	1293,613	23			

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

PBI Comercio	Entre grupos	2632,801	5	526,560	311,121	,000
	Dentro de grupos	30,464	18	1,692		
	Total	2663,265	23			
PBI Transporte, Almacén, Correo y Mensajería	Entre grupos	2986,595	5	597,319	355,606	,000
	Dentro de grupos	30,235	18	1,680		
	Total	3016,830	23			
PBI Alojamiento y Restaurantes	Entre grupos	3990,474	5	798,095	495,297	,000
	Dentro de grupos	29,004	18	1,611		
	Total	4019,478	23			
PBI Telecom. Y Servicios de Información	Entre grupos	4022,205	5	804,441	867,517	,000
	Dentro de grupos	16,691	18	,927		
	Total	4038,896	23			
PBI Administ. Publica y Defensa	Entre grupos	2041,519	5	408,304	479,910	,000
	Dentro de grupos	15,314	18	,851		
	Total	2056,833	23			
PBI Otros Servicios	Entre grupos	3699,091	5	739,818	873,873	,000
	Dentro de grupos	15,239	18	,847		
	Total	3714,330	23			
Percepción de la gestión pública	Entre grupos	812,555	5	162,511	4,771	,006
	Dentro de grupos	613,102	18	34,061		
	Total	1425,657	23			
Criminalidad	Entre grupos	367,889	5	73,578	15,305	,000
	Dentro de grupos	86,532	18	4,807		
	Total	454,421	23			
Homicidios	Entre grupos	317,895	5	63,579	9,296	,000
	Dentro de grupos	123,108	18	6,839		
	Total	441,002	23			

Fuente: Elaboración propia.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

4.4. DISCUSIÓN

A partir de los hallazgos, aceptamos la existencia de una estructura subyacente para Clusterizar las regiones de la república del Perú según indicadores socioeconómicos. Estos resultados guardan relación con lo que sostiene Cruz (1995), la sociedad peruana muestra diferencias entre los clusters de departamentos. Los resultados muestran que la principal fuente de disparidad entre los mismos son las variables vinculadas al estado de salud, el patrón demográfico e indicadores de industrialización juegan también un rol importante. Esta investigación muestra las agudas diferencias que existen entre grupos de departamentos de acuerdo con su nivel de industrialización, índice de pobreza, nivel de educación, datos demográficos, calidad de salud, servicios públicos y la convivencia social.

En lo que respecta a la metodología basada en la técnica de análisis de conglomerados, esta es usada por Tezanos (2012) que identifica a tres grupos de países en América Latina y el Caribe (ALC) con perfiles socioeconómicos distintos, propone una clasificación alternativa que trasciende al criterio tradicional de renta, basada en la técnica de análisis de conglomerados, atendiendo a las principales brechas de desarrollo "económicas, sociales y medioambientales".

La técnica de análisis de conglomerados también es usada por Halanoca (2017) que identifica distritos de la región Puno con similar vocación productiva y la relación que existe sobre la calidad de vida, considerando la variabilidad climática, altitud territorial, tenencia de tierras, aptitud productiva y microrregiones naturales, dando como resultado, conjuntos territoriales con dinámicas de desarrollo agropecuario disimilares.

En la presente investigación de clusterización de las regiones del Perú, según indicadores socioeconómicos, la Figura 11. República del Perú en conglomerados,

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

muestra los clusters distribuidos en el territorio nacional; los 6 clusters están conformados de la siguiente manera:

Clúster 1: La Libertad, Lambayeque, Ancash y Piura.

Clúster 2: Madre de Dios y Tumbes.

Clúster 3: Arequipa, Ica, Tacna y Moquegua.

Clúster 4: Lima.

Clúster 5: Loreto, Ucayali, Amazonas, Pasco y San Martin.

Clúster 6: Cusco, Junín, Apurímac, Ayacucho, Puno, Huancavelica, Huánuco y Cajamarca.

Figura 13 República del Perú en conglomerados.

Fuente: Elaboración propia.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

V. CONCLUSIONES

PRIMERA: Las variables seleccionadas fueron 78 indicadores socioeconómicos, obtenidos mediante pruebas de correlación y análisis de varianza, entre ellas, variables demográficas, económicas, de educación, de salud, de vivienda, de producción, laborales y sociales.

SEGUNDA: Para determinar la clusterización de los departamentos del Perú, los resultados y la esencia de los datos sugieren que existe una estructura subyacente de clusterización en grupos homogéneos de departamentos según sus variables socioeconómicas, estos clusters de departamentos internamente poseen características similares, pero difieren entre sí de manera sustancial, así podemos especificar una conglomeración con 6 clusters de departamentos, análoga a la ordenación geográfica, que proporciona una herramienta susceptible a la modulación de distancia para generar más o menos clusters, permitiendo a los diseñadores de políticas económicas en el Perú mejor perspectiva en la distribución del gasto corriente y de inversión, pues el modelo de clasificación fue determinado robusto y confiable.

TERCERA: Los criterios utilizados para lograr la clusterización de los 24 departamentos del Perú fueron el método de Ward como criterio de agrupación y la medida de distancia Euclídea como medida de proximidad.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

VI. RECOMENDACIONES

PRIMERA: Utilizar métodos multivariantes, para evidenciar estructuras subyacentes de conglomeración, clusters lo suficientemente heterogéneos entre si y homogéneos entre sus miembros, los cuales construyeron un modelos robustos y confiables.

SEGUNDA: Implementar métodos de automatización para la clusterización y el análisis del desarrollo de las regiones de la república del Perú, aunque la decisión del número de clusters dependerá relativamente de los humanos, pues las necesidades y requerimientos para definir este, son complejos y varían en el tiempo según las realidades de cada región.

TERCERA: Métodos como machine learning con aprendizaje no supervisado podrían dar como resultado modelos de clasificación en tiempo real y agregando data acumulada en el tiempo se podrían establecer rutas de desarrollo departamental.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

VII. REFERENCIAS BIBLIOGRÁFICAS

Abella Alberto, Ortiz Marta, De Pablos Carmen (2014). “Meloda, a metric to assess open data reuse / Meloda, métrica para evaluar la reutilización de datos abiertos”. El profesional de la información, vol. 23, n. 6, pp. 582-588.

Aguilar Giovanna (1998). “Crecimiento y desarrollo regional en el Perú”. Pontificia universidad católica del Perú, Departamento de economía.

Apaza Meliton (2014). “Implementación de algoritmos genéticos para la segmentación de imágenes satelitales por conglomerados de la región Puno - 2013”. Universidad Nacional del Altiplano Puno.

Arriaza Manuel (2010) “Guía práctica de análisis de datos”. Instituto de investigación y formación agraria y pesquera. Fondo Europeo de desarrollo regional, Junta de Andalucía. [ISBN 84-611-1661-5].

Cárdenas Jhiannina, Saraiva María (2016). “Vulnerabilidad social y la minería en el Perú: un análisis comparativo”.

Choque Roger, Choque Carlos (2014). “Descentralización, territorio y Gobernabilidad”

Censos Nacionales 2017: XII de población, VII de vivienda y III de comunidades indígenas. Resultados definitivos – censos2017.inei.gob.pe

Contraloría general de la República (2014). “Estudio del proceso de descentralización en el Perú”. Programa de las Naciones Unidas para el desarrollo.

Cruz Amparo (1995). “Clasificación de los departamentos en el Perú por análisis factorial y de acumulación”.

Cuadras C. (2019). “Nuevos métodos de análisis multivariante”. CMC Editions, 08023 Barcelona, España.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Halanoca Nestor (2017). “Identificación de zonas homogéneas con especialización agropecuaria y niveles de desarrollo en el departamento de Puno”. Universidad Nacional del Altiplano Puno.

Hernández Julio (2012). “Clasificación Jerárquica Multidimensional”. Instituto Nacional de Astrofísica, Óptica y Electrónica. Puebla, México.

Instituto Nacional de Estadística e Informática. Sistemas de consulta – inei.gob.pe

Neyra Gonzalo (2011). “Propuesta para la política nacional de desarrollo regional en el Perú”. Universidad Católica de Santa María, Consorcio de Investigación Económica y Social.

Pacompia Alvaro (2017). “Organización de datos multidimensionales en un sistema de recomendaciones basado en data clustering e inteligencia de enjambres”. Universidad Nacional del Altiplano Puno.

Pérez Carla, Lara Graciela, Gómez Denise (2017). “Evolución de la capacidad tecnológica en México. Aplicación del análisis estadístico multivariante de clúster”.

Pérez Cesar (2008). “Técnicas de análisis multivariante de datos”. Pearson education S.A., Madrid. [ISBN 978-84-205-4104-4].

Román Alfonso (2017). “Implementación de pruebas para una hipótesis sobre la aplicación de distancia Euclidiana para realizar agrupamientos en espacios multidimensionales”. Universidad Nacional del Centro. Buenos Aires, Argentina.

Tezanos Sergio (2012). “Conglomerados de desarrollo en América Latina y el Caribe: Una aplicación al análisis de la distribución de la asistencia oficial para el desarrollo”. Naciones Unidad, CEPAL.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Valero Pedro (2011). “Análisis Cluster”. Metodología de las CC del Comp-Universitat de València.

Vicente Mercedes (2014). “Clasificación de las familias en Cajamarca según su situación económica mediante el análisis de conglomerados”. Universidad Nacional Agraria la Molina.

Yamane, T. (1967). Statistics, an introductory analysis. Harper and Row, Nueva York.

Zapata Gilberto (2011). “Evaluación de la Política de descentralización y regionalización en el Perú: caso gobierno regional de Junín”. Universidad Nacional de Ingeniería.

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

ANEXOS

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

A. ANÁLISIS EXPLORATORIO

Estadísticos descriptivos				
Variables	N	Media	Mediana	Desv. Desviación
	Válido			
Superficie	24	52618,7475	35889,5100	71722,67183
Esperanza de vida	24	74,1167	73,6000	2,37664
Tasa Global de Fecundidad (hijos x mujer)	24	2,6288	2,5450	0,44674
Mortalidad en la niñez	24	25,4167	26,5000	8,12894
Crecimiento poblacional	24	0,7600	0,8250	1,24884
Nacimientos	24	22654,3750	18699,5000	30173,55279
Defunciones	24	5729,9167	3839,0000	8502,21885
Población electoral	24	933890,2083	646599,5000	1405297,53311
Población inmigrante	24	232084,3750	104338,0000	592464,30490
Población emigrante	24	248885,8750	259003,5000	160025,90738
Población censada	24	1182807,9167	832073,0000	1838237,80540
Población urbana censada	24	929891,6250	514573,5000	1833164,56362
Población rural censada	24	252916,2917	233958,0000	198879,57929
Tasa de Analfabetismo	24	7,3250	7,3000	3,43160
Tasa de Analfabetismo masculino	24	3,8083	4,2000	1,66392
Tasa de Analfabetismo femenino	24	11,0292	10,3500	5,48512
Tasa de Asistencia escolar (primaria)	24	91,5458	92,1000	1,97065
Tasa de Asistencia escolar (secundaria)	24	83,3833	84,9500	5,42864
Nivel de educación de la Población de 15 y más años (Primaria)	24	25,0625	24,8000	7,80918
Nivel de educación de la Población de 15 y más años (Secundaria)	24	43,2083	43,5000	4,27031
Nivel de educación de la Población de 15 y más años (Sup. No Universitaria)	24	12,6167	12,3000	3,35723
Nivel de educación de la Población de 15 y más años (Sup. Universitaria)	24	14,2333	14,3500	4,19572
Asistencia escolar inicial	24	87,8071	87,5650	2,86539
Rendimiento en lectura	24	28,7221	26,7450	9,84290
Rendimiento en matemáticas	24	23,2367	22,1250	8,34127

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Colegios con acceso a internet	24	44,8708	39,8350	18,28152
Cobertura de algún Seguro de Salud	24	77,7042	77,3500	8,72525
Tasa de desnutrición crónica en menores de 5 años	24	15,2333	16,0000	8,10457
Niños con anemia (De 6 a 35 meses de edad)	24	47,6625	45,7500	10,35374
Desnutrición crónica	24	11,0208	11,7000	6,31327
Morbilidad	24	67,7292	66,2000	7,81189
Cobertura del personal médico	24	14,1325	10,2150	9,94123
Cobertura hospitalaria	24	2,2517	2,0200	1,03788
Partos institucionales	24	90,8917	93,4000	8,46024
Población en edad de trabajar (De 14 y más años de edad)	24	956,5042	621,5000	1529,35557
PEA	24	687,6833	500,1000	965,99200
PEA ocupada	24	659,9000	488,6500	900,63403
PEA Ocupada en Agricultura, pesca, minería	24	174,3833	163,8500	121,77238
PEA Ocupada en Manufactura	24	61,8000	33,3500	129,41235
PEA ocupada en empresas de 1-10 trabajadores	24	77,1258	77,2500	7,58719
PEA desempleada	24	27,3500	11,9000	66,66440
Empleo adecuado	24	46,0542	43,0950	13,72500
Educación de la fuerza laboral	24	26,7704	26,5200	8,44881
Creación de empleo formal	24	0,4288	0,2600	1,33480
Empleo informal	24	79,2808	81,0300	8,99801
Desempleo juvenil urbano	24	11,9021	11,5400	3,28764
VA Bruto	24	3,2542	3,5000	3,98824
Manufactura	24	6,8417	6,0500	5,43682
Construcción	24	5,4250	5,2500	9,50278
Agricultura, Ganadería, Caza y Silvicultura	24	7,9458	6,5500	8,82378
Electricidad, Gas y Agua	24	5,5625	5,3000	16,82544
Telecomunicaciones y Serv. De Información	24	5,3542	4,8500	3,25963
Administ. Publica y Defensa	24	4,5042	4,5500	1,17491
Alojamiento y Restaurantes	24	3,5667	3,6500	0,90634
Transporte, Almacén, Correo y Mensajería	24	4,2292	4,1500	1,56357

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Comercio	24	2,4042	2,4500	0,60969
Pesca y Acuicultura	24	28,9125	0,0000	90,64037
Extrac. De Petróleo, Gas y Minerales	24	-0,0375	-0,4000	14,91182
Otros Servicios	24	4,3750	4,4000	0,67903
Hogares con acceso a agua potable	24	55,2042	53,1500	21,86666
Hogares con acceso a desagüe	24	66,1000	64,7000	16,70251
Hogares con acceso a alumbrado eléctrico	24	92,2292	93,7500	5,00126
hogares con acceso a TV Cable	24	31,1250	29,8000	14,63861
Hogares con acceso a telefonía fija	24	11,9167	9,5000	9,85396
Hogares con acceso a telefonía móvil	24	89,2083	89,3000	4,81600
Hogares con acceso a Internet	24	19,0833	15,8000	12,55685
Hogares que utilizan gas para cocinar	24	38,4346	34,3950	20,01637
Cobertura de electricidad	24	91,2342	91,7650	5,71062
Precio de la electricidad	24	12,1438	12,2150	3,04097
Cobertura de agua	24	84,5471	88,0400	10,11732
Continuidad de la provisión de agua	24	16,0579	17,1900	6,05468
Cobertura de desagüe	24	58,5850	57,1200	17,42124
Hogares con internet	24	17,1458	13,0000	11,47517
Hogares con al menos un celular	24	87,2208	87,4150	5,16416
Densidad del transporte aéreo	24	579,6825	310,8600	662,58880
Ingreso promedio mensual	24	1178,7083	1121,0000	293,41914
Pobreza	24	23,5625	24,6000	11,50976
Pobreza extrema	24	4,4875	3,6500	3,93486
Producto Bruto Interno real	24	18329,5696	8510,1800	42498,67715
Producto Bruto Interno real per cápita	24	13023,7725	10439,1950	9092,72882
Stock de capital por trabajador	24	23801,3271	19497,7950	16456,17408
Presupuesto público per cápita	24	4409,8450	3997,2450	1364,41590
Gasto real por hogar mensual	24	1378,1646	1442,6050	377,58905
Incremento del gasto real por hogar	24	-0,3213	-0,0350	2,17277
Disponibilidad de servicios financieros	24	483,0000	445,0000	235,51092

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Acceso al crédito	24	28,0279	29,0550	9,48390
Brecha de género en ingresos laborales	24	33,3963	34,8800	7,79604
Ejecución de la inversión pública	24	67,7754	69,0300	7,76710
Variacion del índice de precios al consumidor	24	3,0667	3,1300	0,63238
PBI Agricultura, Ganadería, Caza y Silvicultura	24	4,1667	3,5000	3,51873
PBI Pesca y Acuicultura	24	4,1542	0,4500	7,91679
PBI Extrac. De Petróleo, Gas y Minerales	24	4,1708	2,7000	4,34356
PBI Manufactura	24	4,1625	0,9500	12,22278
PBI Electricidad, Gas y Agua	24	4,1708	1,6000	10,61981
PBI Construcción	24	4,1667	2,0500	7,49960
PBI Comercio	24	4,1750	1,6000	10,76077
PBI Transporte, Almacén, Correo y Mensajería	24	4,1708	1,4500	11,45279
PBI Alojamiento y Restaurantes	24	4,1583	1,0500	13,21968
PBI Telecom. Y Servicios de Información	24	4,1625	1,1500	13,25157
PBI Administ. Publica y Defensa	24	4,1667	2,2000	9,45661
PBI Otros Servicios	24	4,1708	1,3500	12,70797
Percepción de la gestión pública	24	33,2563	31,7150	7,87306
Conflictos sociales	24	8,8750	4,5000	7,78690
Criminalidad	24	8,1825	7,8550	4,44494
Homicidios	24	8,2775	7,1050	4,37881
Presencia policial	24	718,8300	683,0350	186,23679
Resolución expedientes judiciales	24	42,6521	41,9200	7,91362

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

B. PRUEBAS DE NORMALIDAD

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
Superficie	0,269	24	0,000	0,517	24	0,000
Esperanza de vida	0,169	24	0,076	0,934	24	0,118
Tasa Global de Fecundidad (hijos x mujer)	0,178	24	0,047	0,928	24	0,087
Crecimiento poblacional	0,086	24	,200*	0,972	24	0,708
Nacimientos	0,331	24	0,000	0,498	24	0,000
Defunciones	0,327	24	0,000	0,483	24	0,000
Población electoral	0,336	24	0,000	0,454	24	0,000
Población inmigrante	0,408	24	0,000	0,321	24	0,000
Población emigrante	0,094	24	,200*	0,962	24	0,478
Tasa de Analfabetismo	0,128	24	,200*	0,959	24	0,412
Tasa de Analfabetismo masculino	0,129	24	,200*	0,941	24	0,175
Tasa de Analfabetismo femenino	0,107	24	,200*	0,956	24	0,356
Tasa de Asistencia escolar (primaria)	0,152	24	0,156	0,948	24	0,240
Tasa de Asistencia escolar (secundaria)	0,185	24	0,033	0,941	24	0,169
Nivel de educación de la Población de 15 y más años (Primaria)	0,130	24	,200*	0,978	24	0,857
Nivel de educación de la Población de 15 y más años (Secundaria)	0,147	24	0,193	0,962	24	0,488
Nivel de educación de la Población de 15 y más años (Sup. No Universitaria)	0,123	24	,200*	0,970	24	0,658
Nivel de educación de la Población de 15 y más años (Sup. Universitaria)	0,109	24	,200*	0,963	24	0,496
Cobertura de algún Seguro de Salud	0,118	24	,200*	0,962	24	0,484
Tasa de desnutrición crónica en menores de 5 años	0,108	24	,200*	0,954	24	0,329
Niños con anemia (De 6 a 35 meses de edad)	0,139	24	,200*	0,941	24	0,167
Población en edad de trabajar (De 14 y más años de edad)	0,339	24	0,000	0,435	24	0,000
PEA	0,329	24	0,000	0,471	24	0,000
PEA ocupada	0,321	24	0,000	0,484	24	0,000
PEA Ocupada en Agricultura, pesca, minería	0,103	24	,200*	0,934	24	0,122
PEA Ocupada en Manufactura	0,371	24	0,000	0,381	24	0,000
PEA ocupada en empresas de 1-10 trabajadores	0,110	24	,200*	0,963	24	0,492

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

PEA desempleada	0,441	24	0,000	0,313	24	0,000
Ingreso promedio mensual	0,182	24	0,038	0,935	24	0,123
Hogares con acceso a agua potable	0,097	24	,200*	0,962	24	0,471
Hogares con acceso a desagüe	0,098	24	,200*	0,939	24	0,153
Hogares con acceso a alumbrado eléctrico	0,150	24	0,175	0,936	24	0,130
hogares con acceso a TV Cable	0,118	24	,200*	0,966	24	0,559
Hogares con acceso a telefonía fija	0,188	24	0,029	0,826	24	0,001
Hogares con acceso a telefonía móvil	0,125	24	,200*	0,904	24	0,027
Hogares con acceso a Internet	0,159	24	0,122	0,916	24	0,048
Hogares que utilizan gas para cocinar	0,142	24	,200*	0,920	24	0,059
Pobreza	0,188	24	0,029	0,913	24	0,041
Pobreza extrema	0,180	24	0,042	0,830	24	0,001
Nacimientos registrados	0,302	24	0,000	0,482	24	0,000
Servicio del sistema financiero	0,347	24	0,000	0,428	24	0,000
Población censada	0,332	24	0,000	0,452	24	0,000
Población urbana censada	0,342	24	0,000	0,394	24	0,000
Población rural censada	0,119	24	,200*	0,901	24	0,023
Producto Bruto Interno real	0,408	24	0,000	0,336	24	0,000
Producto Bruto Interno real per cápita	0,221	24	0,004	0,704	24	0,000
Stock de capital por trabajador	0,198	24	0,015	0,745	24	0,000
Presupuesto público per cápita	0,195	24	0,019	0,902	24	0,024
Gasto real por hogar mensual	0,137	24	,200*	0,923	24	0,068
Incremento del gasto real por hogar	0,137	24	,200*	0,937	24	0,139
Disponibilidad de servicios financieros	0,119	24	,200*	0,956	24	0,368
Acceso al crédito	0,116	24	,200*	0,980	24	0,896
Cobertura de electricidad	0,147	24	0,191	0,935	24	0,124
Precio de la electricidad	0,132	24	,200*	0,957	24	0,388
Cobertura de agua	0,177	24	0,050	0,857	24	0,003
Continuidad de la provisión de agua	0,129	24	,200*	0,936	24	0,136
Cobertura de desagüe	0,171	24	0,067	0,912	24	0,038

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Hogares con internet	0,175	24	0,056	0,908	24	0,031
Hogares con al menos un celular	0,084	24	,200*	0,937	24	0,142
Densidad del transporte aéreo	0,192	24	0,022	0,816	24	0,001
Mortalidad en la niñez	0,168	24	0,077	0,921	24	0,060
Desnutrición crónica	0,101	24	,200*	0,970	24	0,661
Morbilidad	0,141	24	,200*	0,970	24	0,661
Cobertura del personal médico	0,202	24	0,013	0,818	24	0,001
Cobertura hospitalaria	0,159	24	0,119	0,865	24	0,004
Partos institucionales	0,160	24	0,113	0,848	24	0,002
Acceso a seguro de salud	0,105	24	,200*	0,987	24	0,984
Analfabetismo	0,134	24	,200*	0,936	24	0,131
Asistencia escolar inicial	0,120	24	,200*	0,961	24	0,464
Asistencia escolar primaria y secundaria	0,096	24	,200*	0,979	24	0,881
Población con secundaria a más	0,098	24	,200*	0,953	24	0,316
Rendimiento en lectura	0,154	24	0,147	0,932	24	0,106
Rendimiento en matemáticas	0,155	24	0,140	0,914	24	0,043
Colegios con acceso a internet	0,133	24	,200*	0,929	24	0,094
Nivel de ingresos por trabajo	0,140	24	,200*	0,914	24	0,043
Brecha de género en ingresos laborales	0,133	24	,200*	0,949	24	0,258
Empleo adecuado	0,123	24	,200*	0,943	24	0,186
Educación de la fuerza laboral	0,133	24	,200*	0,921	24	0,063
Creación de empleo formal	0,160	24	0,112	0,933	24	0,112
Empleo informal	0,175	24	0,054	0,913	24	0,042
Desempleo juvenil urbano	0,118	24	,200*	0,972	24	0,715
Ejecución de la inversión pública	0,154	24	0,143	0,954	24	0,322
Percepción de la gestión pública	0,135	24	,200*	0,965	24	0,544
Conflictos sociales	0,234	24	0,001	0,854	24	0,003
Criminalidad	0,116	24	,200*	0,921	24	0,061
Homicidios	0,168	24	0,077	0,838	24	0,001
Presencia policial	0,134	24	,200*	0,955	24	0,355

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

Resolución expedientes judiciales	0,095	24	,200*	0,975	24	0,798
Variación del índice de precios al consumidor	0,110	24	,200*	0,973	24	0,730
VA Bruto	0,185	24	0,033	0,890	24	0,013
Manufactura	0,262	24	0,000	0,671	24	0,000
Construcción	0,076	24	,200*	0,955	24	0,345
Agricultura, Ganadería, Caza y Silvicultura	0,202	24	0,012	0,778	24	0,000
Electricidad, Gas y Agua	0,225	24	0,003	0,758	24	0,000
Telecomunicaciones y Serv. De Información	0,097	24	,200*	0,978	24	0,859
Administ. Publica y Defensa	0,137	24	,200*	0,914	24	0,044
Alojamiento y Restaurantes	0,139	24	,200*	0,954	24	0,329
Transporte, Almacén, Correo y Mensajería	0,108	24	,200*	0,952	24	0,300
Comercio	0,123	24	,200*	0,967	24	0,585
Pesca y Acuicultura	0,395	24	0,000	0,398	24	0,000
Extrac. De Petróleo, Gas y Minerales	0,273	24	0,000	0,785	24	0,000
Otros Servicios	0,098	24	,200*	0,952	24	0,296
PBI Porcentual Total	0,395	24	0,000	0,330	24	0,000
PBI Agricultura, Ganadería, Caza y Silvicultura	0,155	24	0,139	0,860	24	0,003
PBI Pesca y Acuicultura	0,347	24	0,000	0,577	24	0,000
PBI Extrac. De Petróleo, Gas y Minerales	0,190	24	0,026	0,836	24	0,001
PBI Manufactura	0,399	24	0,000	0,321	24	0,000
PBI Electricidad, Gas y Agua	0,431	24	0,000	0,351	24	0,000
PBI Construcción	0,320	24	0,000	0,411	24	0,000
PBI Comercio	0,413	24	0,000	0,323	24	0,000
PBI Transporte, Almacén, Correo y Mensajería	0,416	24	0,000	0,321	24	0,000
PBI Alojamiento y Restaurantes	0,424	24	0,000	0,286	24	0,000
PBI Telecom. Y Servicios de Información	0,442	24	0,000	0,283	24	0,000
PBI Administ. Publica y Defensa	0,440	24	0,000	0,317	24	0,000
PBI Otros Servicios	0,457	24	0,000	0,279	24	0,000
*. Esto es un límite inferior de la significación verdadera.						
a. Corrección de significación de Lilliefors						

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

C. ANÁLISIS DE CONGLOMERADOS

Matriz de proximidades (1/4)						
Distancia euclídea al cuadrado						
Caso	1:Amazonas	2:Áncash	3:Apurímac	4:Arequipa	5:Ayacucho	6:Cajamarca
1:Amazonas	0,000	145,047	105,554	268,600	71,938	102,296
2:Áncash	145,047	0,000	94,760	159,047	119,710	150,466
3:Apurímac	105,554	94,760	0,000	187,040	77,230	104,056
4:Arequipa	268,600	159,047	187,040	0,000	228,387	304,625
5:Ayacucho	71,938	119,710	77,230	228,387	0,000	101,216
6:Cajamarca	102,296	150,466	104,056	304,625	101,216	0,000
7:Cusco	114,214	95,493	69,374	135,259	78,983	118,845
8:Huancavelica	87,195	180,642	101,549	339,705	97,338	96,099
9:Huánuco	64,542	146,398	85,750	247,261	68,155	92,220
10:Ica	238,129	149,509	184,769	66,383	199,232	313,113
11:Junín	101,543	90,231	68,955	124,784	71,997	81,710
12:La Libertad	128,541	86,325	121,548	90,144	126,047	160,275
13:Lambayeque	151,779	101,654	116,988	82,162	133,795	194,763
14:Lima	850,725	675,474	756,723	515,849	797,180	842,955
15:Loreto	91,095	208,142	191,344	349,855	139,236	159,296
16:Madre de Dios	169,712	223,659	167,962	179,090	205,438	299,200
17:Moquegua	240,031	207,600	206,109	102,335	218,257	324,854
18:Pasco	59,138	118,348	89,275	229,297	100,741	120,342
19:Piura	96,169	63,778	90,812	139,255	94,071	116,705
20:Puno	102,605	147,835	109,844	264,642	87,982	81,607
21:San Martín	68,656	138,317	75,525	187,157	83,225	117,279
22:Tacna	253,484	216,714	233,652	96,496	221,063	353,173

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

23:Tumbes	181,725	132,541	142,856	118,057	167,785	266,266
24:Ucayali	89,701	185,884	142,703	215,604	116,595	178,300

Matriz de proximidades (2/4)

Distancia euclídea al cuadrado

Caso	7:Cusco	8:Huancavelica	9:Huánuco	10:Ica	11:Junín	12:La Libertad
1:Amazonas	114,214	87,195	64,542	238,129	101,543	128,541
2:Áncash	95,493	180,642	146,398	149,509	90,231	86,325
3:Apurímac	69,374	101,549	85,750	184,769	68,955	121,548
4:Arequipa	135,259	339,705	247,261	66,383	124,784	90,144
5:Ayacucho	78,983	97,338	68,155	199,232	71,997	126,047
6:Cajamarca	118,845	96,099	92,220	313,113	81,710	160,275
7:Cusco	0,000	115,807	91,624	171,712	68,777	102,541
8:Huancavelica	115,807	0,000	65,208	324,864	124,309	201,765
9:Huánuco	91,624	65,208	0,000	234,232	100,762	126,771
10:Ica	171,712	324,864	234,232	0,000	122,412	77,453
11:Junín	68,777	124,309	100,762	122,412	0,000	64,212
12:La Libertad	102,541	201,765	126,771	77,453	64,212	0,000
13:Lambayeque	120,321	218,051	141,490	70,612	78,532	46,378
14:Lima	679,654	954,649	820,777	572,714	655,530	541,333
15:Loreto	188,776	170,821	164,110	295,577	164,535	216,200
16:Madre de Dios	132,846	258,652	174,561	170,198	179,070	166,529
17:Moquegua	184,592	330,967	266,244	108,694	166,089	159,389
18:Pasco	106,614	84,511	88,468	184,273	82,032	112,428
19:Piura	75,874	157,466	107,298	127,977	60,234	56,297
20:Puno	98,811	106,694	102,733	229,546	76,371	136,435
21:San Martín	89,983	110,013	77,898	171,184	64,391	90,299
22:Tacna	190,684	348,495	266,170	97,769	158,702	138,756
23:Tumbes	132,375	249,080	168,269	89,276	127,097	106,467

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

24:Ucayali	137,718	160,549	109,099	188,714	96,289	134,315

Matriz de proximidades (3/4)						
Distancia euclídea al cuadrado						
Caso	13:Lambayeque	14:Lima	15:Loreto	16:Madre de Dios	17:Moquegua	18:Pasco
1:Amazonas	151,779	850,725	91,095	169,712	240,031	59,138
2:Áncash	101,654	675,474	208,142	223,659	207,600	118,348
3:Apurímac	116,988	756,723	191,344	167,962	206,109	89,275
4:Arequipa	82,162	515,849	349,855	179,090	102,335	229,297
5:Ayacucho	133,795	797,180	139,236	205,438	218,257	100,741
6:Cajamarca	194,763	842,955	159,296	299,200	324,854	120,342
7:Cusco	120,321	679,654	188,776	132,846	184,592	106,614
8:Huancavelica	218,051	954,649	170,821	258,652	330,967	84,511
9:Huánuco	141,490	820,777	164,110	174,561	266,244	88,468
10:Ica	70,612	572,714	295,577	170,198	108,694	184,273
11:Junín	78,532	655,530	164,535	179,070	166,089	82,032
12:La Libertad	46,378	541,333	216,200	166,529	159,389	112,428
13:Lambayeque	0,000	581,489	232,846	157,121	135,773	133,072
14:Lima	581,489	0,000	885,598	775,971	673,613	831,712
15:Loreto	232,846	885,598	0,000	252,908	335,488	152,005
16:Madre de Dios	157,121	775,971	252,908	0,000	180,582	176,800
17:Moquegua	135,773	673,613	335,488	180,582	0,000	191,108
18:Pasco	133,072	831,712	152,005	176,800	191,108	0,000
19:Piura	56,567	595,372	156,155	170,374	181,082	100,325
20:Puno	164,972	802,881	136,011	222,786	270,802	97,578

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

21:San Martín	105,383	745,332	147,166	123,300	204,300	72,098
22:Tacna	125,352	673,573	363,275	194,472	118,209	226,522
23:Tumbes	68,298	697,806	263,409	93,659	136,072	142,285
24:Ucayali	127,240	781,078	118,603	124,405	240,986	119,844

Matriz de proximidades (4/4)

Distancia euclídea al cuadrado

Caso	19:Piura	20:Puno	21:San Martín	22:Tacna	23:Tumbes	24:Ucayali
1:Amazonas	96,169	102,605	68,656	253,484	181,725	89,701
2:Áncash	63,778	147,835	138,317	216,714	132,541	185,884
3:Apurímac	90,812	109,844	75,525	233,652	142,856	142,703
4:Arequipa	139,255	264,642	187,157	96,496	118,057	215,604
5:Ayacucho	94,071	87,982	83,225	221,063	167,785	116,595
6:Cajamarca	116,705	81,607	117,279	353,173	266,266	178,300
7:Cusco	75,874	98,811	89,983	190,684	132,375	137,718
8:Huancavelica	157,466	106,694	110,013	348,495	249,080	160,549
9:Huánuco	107,298	102,733	77,898	266,170	168,269	109,099
10:Ica	127,977	229,546	171,184	97,769	89,276	188,714
11:Junín	60,234	76,371	64,391	158,702	127,097	96,289
12:La Libertad	56,297	136,435	90,299	138,756	106,467	134,315
13:Lambayeque	56,567	164,972	105,383	125,352	68,298	127,240
14:Lima	595,372	802,881	745,332	673,573	697,806	781,078
15:Loreto	156,155	136,011	147,166	363,275	263,409	118,603
16:Madre de Dios	170,374	222,786	123,300	194,472	93,659	124,405
17:Moquegua	181,082	270,802	204,300	118,209	136,072	240,986
18:Pasco	100,325	97,578	72,098	226,522	142,285	119,844
19:Piura	0,000	115,088	82,393	183,738	107,762	107,925
20:Puno	115,088	0,000	113,259	270,165	205,270	124,881

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis

UNIVERSIDAD
NACIONAL DEL ALTIPLANO
Repositorio Institucional

21:San Martín	82,393	113,259	0,000	206,863	122,718	87,937
22:Tacna	183,738	270,165	206,863	0,000	158,830	225,802
23:Tumbes	107,762	205,270	122,718	158,830	0,000	143,017
24:Ucayali	107,925	124,881	87,937	225,802	143,017	0,000

repositorio.unap.edu.pe
No olvide citar adecuadamente esta tesis