UNIVERSIDAD NACIONAL AUTÓNOMA
DE MÉXICO

FACULTAD DE CIENCIAS

"EL USO DEL ANÁLISIS DE CONGLOMERADOS APLICADO A LOS SISTEMAS EDÁFICOS COMO UNA APROXIMACIÓN EN LA MODELACIÓN PEDOGENÉTICA"

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

MATEMÁTICO

P R E S E N T A :

MARTÍN RAFAEL PÉREZ HERNÁNDEZ

TUTORA: DRA. ELIZABETH SOLLEIRO REBOLLEDO
ASESORA: M. EN A. P. MARIA DEL PILAR ALONSO REYES

2007

FACULTAD DE CIENCIAS

División de Estudios Profesionales

VNIVERSIDAD NACIONAL AVFNOMA DE MEXICO

ACT. MAURICIO AGUILAR GONZÁLEZ
Jefe de la División de Estudios Profesionales
Facultad de Ciencias
P r e s e n t e .

Por este medio hacemos de su conocimiento que hemos revisado el trabajo escrito titulado:

"El uso del análisis de conglomerados aplicado a los sistemas edáficos como una aproximación en la modelación pedogenética"

realizado por **Pérez Hernández Martín Rafael,** con número de cuenta **40106482-9,** quien opta por titularse en la opción de **Tesis** de la licenciatura en **Matemáticas**. Dicho trabajo cuenta con nuestro voto aprobatorio.

Tutor(a) Propietario	Dra.	Elizabeth Solleiro Rebolledo	
Asesora *Propietario*	M. en A.P.	María del Pilar Alonso Reyes	
Propietario	Dr.	Pablo Padilla Longoria	
Suplente	M. en C.	José Antonio Flores Díaz	
Suplente	Dr.	Jorge Enrique Gama Castro	

A t e n t a m e n t e
"POR MI RAZA HABLARA EL ESPÍRITU"
Ciudad Universitaria, D.F., a 20 de febrero del 2007.
EL COORDINADOR DEL COMITÉ DE TITULACIÓN
DE LA LICENCIATURA EN MATEMÁTICAS

M. EN C. AGUSTIN CONTIVEROS PINEDA

Señor sinodal: antes de firmar este documento, solicite al estudiante que le muestre la versión digital de su trabajo y verifique que la misma incluya todas las observaciones y correcciones que usted hizo sobre el mismo.

AGRADECIMIENTOS

En primer lugar agradezco a la **Universidad Nacional Autónoma de México**, pero en especial a la **facultad de ciencias**, que fue mi segunda casa durante este tiempo que realice mis estudios.

Agradezco a mis directores de tesis: **Dra. Elizabeth Solleiro Rebolledo y M. en A. P. María del Pilar Alonso Reyes** por haber aceptado dirigir mi trabajo de tesis, y pese al trabajo que tienen, me brindaron un poco de su tiempo para la revisión y corrección del mismo.

También agradezco a mis sinodales: **Dr. Pablo Padilla Longoria, Dr. Jorge Enrique Gama Castro y M. en C. José Antonio Flores Díaz**, por el tiempo que invirtieron en la revisión de mi tesis, por sus valiosos comentarios y sugerencias al contenido de ésta.

Asimismo, agradezco al **Dr. Sergey Sedov y al grupo del seminario de edafología** por las correcciones y su interés por este trabajo.

Agradezco a **INEGI** (Instituto Nacional de Estadística Geografía e Informática), la información prestada para el desarrollo de este trabajo, como también al **Biol. Jesús Solano**, quien me apoyo con la parte de la interpretación de los datos en el programa ArcView 3.3.

A mis compañeras: **Quí. Paola Molina y Mónica Romero** quienes me ayudaron a aclarar varias ideas en el presente trabajo. También al **Fis. Gerardo Centeno**, quién apoyó en el procesamiento e interpretación de los datos.

También agradezco al **Instituto de Geología de la UNAM** por las facilidades prestadas para el desarrollo de esta tesis, así como a los proyectos **PAPIIT IN112205 e IN400403**.

AGRADECIMIENTOS A MIS PADRES

Agradezco a mis padres: **Rafael Pérez López y Leticia Hernández González**, por estar cerca de mí compartiendo las experiencias más importantes de mi carrera. Porque gracias a su apoyo, he realizado una de mis mejores metas. Ustedes, que sin esperar nada, lo dieron todo. Por que nunca estuve solo. Porque siempre conté con su confianza. Por todo esto, quiero que sientan que el objetivo logrado también es suyo y que la fuerza que me ayudó a conseguirlo, fue su amor. Con cariño y adimiración.

Martín Rafael Pérez Hernández.

CONTENIDO

INTRODUCCIÓN.

LA EDAD DEL SUELO.

Determinar la edad evolutiva de los suelos es una variable muy difícil de medir, ya que son cuerpos que están en continuo cambio, sin embargo, hay algunos momentos en los que llegan, a un estado de casi equilibrio con las condiciones ambientales en las que se encuentran.

Otra problemática que se presenta para definir la edad evolutiva de los suelos, consiste en tomar en cuenta la rapidez con la que se llevan a cabo los procesos, en función del clima. También hay que considerar que en una etapa inicial evolucionan demasiado rápido, hasta que alcanzan un grado de madurez y luego su desarrollo es más lento, aunque las condiciones climáticas o cualquier otro fenómeno que se presente pueden causar que la evolución de estos se acelere, se mantenga o disminuya y en algunos casos extremos se detenga (Retallack, 1984).

Por otra parte hay muchos modelos matemáticos que han tratado de explicar la relación de los factores formadores con los procesos edafogénicos en función del tiempo (Jenny, 1941; Wilde, 1946; Ruhe, 1965, Runge, 1973; Jenny, 1980; Johnson y Rockwell, 1982; Johnson y Watson-Stegner, 1987; Phillips, 1990 y 1993; Shoji et al., 1993a). Algunos de estos modelos, aunque utilizan el lenguaje matemático, son cualitativos o completamente teóricos. Por este motivo hay la necesidad de desarrollar modelos que puedan ser aplicados a los suelos con mayor realidad y no sean tan complejos, aunque matemáticamente sean muy interesantes, pero incomprensibles para los estudiosos de ciencias de la tierra.

RESUMEN.

El capítulo 1 está conformado por el marco teórico, en el cual se presenta de una manera clara al sistema que se desea modelar, es decir, se da una explicación breve de que es el suelo, los factores que dan origen a su formación, algunas de sus propiedades o características, así como una descripción detallada de las unidades edáficas utilizadas en este trabajo. También se incluyó una parte de antecedentes, sobre algunos modelos matemáticos aplicados a las ciencias de la tierra, para conocer el alcance de los modelos aplicados a suelos.

En el capítulo 2, se da una breve introducción al análisis multivariado, en particular se hace un amplia y especial mención al uso del método de conglomerados, que es la herramienta primordial de este trabajo considerando estructuras elementales como es una distancia o disimilaridad, algunas medidas de similaridad hasta los principales algoritmos de clasificación.

El capítulo 3 corresponde a la metodología, la cual fue fundamental para consolidar este trabajo, aquí se encuentran todos los pasos que fueron utilizados desde cómo se depuraron los datos hasta la aplicación del análisis de conglomerados. También se encuentran todos los resultados obtenidos, en especial los dendogramas que aportaron gran información.

Como conclusión final este trabajo es innovador y representa el resultado de un proceso de investigación pluri-disciplinario, el cual es una primera aproximación en la modelación pedogenética, por último se incluyen anexos que consideran otros resultados que fueron aplicados como antecedentes al análisis de conglomerados.

OBJETIVO.

Lo que se pretende con este trabajo es determinar la importancia actual en la modelación del cambio climático y su influencia en la formación de diferentes unidades de suelos, en esta etapa, se han analizado algunas unidades, que pertenecen a Pachuca, Morelia, Morelos, Ciudad de México, Ciudad Altamirano y Veracruz dentro del Eje Neovolcánico.

El objetivo central del trabajo es establecer un modelo que permita relacionar unidades de suelo, sus características con el clima, particularmente la temperatura y precipitación y evaluar su influencia en el desarrollo del suelo.

JUSTIFICACIÓN.

De acuerdo al paradigma del científico ruso Dokuchaev, el suelo se forma por la acción conjunta de material parental, clima, relieve, organismos y tiempo, siendo el clima uno de los factores formadores fundamentales. Así que se puede usar al suelo como un indicador climático.

HIPÓTESIS.

El clima es considerado como un factor activo o bioclimático, según Duchaufour (1977), pues a través de la precipitación y de la temperatura, ambos parámetros actúan en conjunto y propician cambios físicos y químicos sobre los materiales de la superficie terrestre, y según Fanning y Fanning (1989), son "las fuerzas directrices que promueven los procesos que causan cambios en el suelo o en los ecosistemas durante el curso de la pedogénesis".

Por lo tanto, al modelar diversas características del suelo que dependen del clima, es posible determinar los climas del pasado, sus cambios en el presente y su dinámica futura, bajo este precepto, es posible utilizar al suelo y a las características que presenta como un indicador climático. La importancia del clima en la actualidad se debe a la creciente necesidad que existe de entender su dinámica, por los efectos directos que hay en los ecosistemas.

Los modelos actuales de cambio climático se alimentan de variables que se miden de manera directa en la atmósfera o en la hidrosfera. Sin embargo, estos pueden usarse para hacer pronósticos en el tiempo, y podrían tomar en cuenta parámetros indirectos como el suelo.

1. MARCO TEÓRICO

1.1. EL SUELO.

El suelo es un cuerpo natural que se forma en la superficie de la tierra, el cual está constituido por horizontes o capas de materia orgánica y minerales en las que existe un continuo intercambio de materia y energía; es importante su estudio por varias razones: son reguladores de los nutrientes que llegan a las plantas; son el hábitat de un sin fin de organismos; controlan la calidad de las aguas (los suelos sirven como filtros, reguladores del escurrimiento, también participa en el intercambio gaseoso con la atmósfera; gracias a su estrecha interrelación con la biota, puede ayudar a controlar la emisión de CO_2 que promueve el calentamiento global). El determinar sus propiedades físicas y químicas no sólo ayuda a clasificarlos si no a entender su función dentro de un determinado ecosistema, los procesos que dieron origen y su evolución.

El suelo se forma por la acción de cinco factores, llamados formadores, establecidos por Dockuchaev a fines del siglo XIX:

- **Material parental.-** Son materiales residuales que son depósitos de rocas expuestos a la intemperie, también pueden ser materiales transportados como minerales o fragmentos de rocas que han sido removidos de un lugar por la acción del agua, del viento o la gravedad.

- **Relieve.-**Son diversas geoformas como montañas, llanuras, taludes, valles, etc.

- **Organismos.-**Fauna y flora de diversos tamaños y tipos incluyendo al hombre.

- **Tiempo.-**Es el factor independiente, reconociéndose un tiempo cero que marca el inicio de los procesos.

- **Clima.-**El clima según Duchaufour (1977) y Fanning (1989), promueve los procesos que causan los cambios en los suelos o en los ecosistemas durante la pedogénesis.

Por definición, el clima es el estado medio de la atmósfera, el cual es determinado a partir de la recopilación de datos en un intervalo de tiempo que oscila entre 30 y 50 años. Éste depende de tres elementos esenciales: presión, humedad y temperatura; se caracteriza a través de dos parámetros: precipitación pluvial y temperatura. También se le considera como el factor medioambiental que más influye en las características del suelo, generando cambios físicos y químicos sobre la superficie terrestre.

El suelo es capaz de almacenar información sobre los ambientes del pasado, gracias a lo que se llama "*memoria del suelo*" (Targulian y Goriachkin 2004), esta memoria, integrada por diversas propiedades del suelo (materia orgánica, arcilla, profundidad entre otras) que poco cambian con el tiempo, pueden ayudar a determinar las condiciones de formación. La figura 1 explica la influencia de los factores formadores mencionados sobre los procesos que ocurren en el suelo, los cuales permiten el desarrollo de diversas propiedades. Como se ve, el análisis puede ser a la inversa y que a partir de las propiedades se establezcan los procesos y se determinen factores formadores.

El factor que interesa en el presente trabajo es el clima, ya que en la actualidad hay un creciente interés por hacer pronósticos de cambio futuro, así como hacer reconstrucciones paleoambientales. Como parte de las propiedades del suelo se seleccionaron seis que se considera, son función directa de los parámetros precipitación y temperatura: materia orgánica (MO%), capacidad de intercambio catiónico (CIC%), arcilla, pH, profundidad, así como el tipo de suelo, el cual es específico de ciertas zonas climáticas.

Figura 1. Explica la influencia de los factores formadores, sobre los procesos que ocurren en el Suelo, los cuales permiten el desarrollo de diversas propiedades.

1.2. CARACTERÍSTICAS DEL SUELO.

En el suelo, la **materia orgánica** está constituida por los compuestos de origen biológico que se presentan en los horizontes o capas sobre todo superficiales. Esta comprende tanto la flora como la fauna del suelo así como sus restos descompuestos. Es una característica muy importante del suelo y su contenido varía en función del clima (Gama et al. 1998).

La **capacidad de intercambio catiónico (CIC%)** tiene que ver con los procesos reversibles por los cuales las partículas sólidas del suelo absorben iones de la fase acuosa y al mismo tiempo, desadsorben cantidades equivalentes de otros cationes estableciendo un equilibrio entre ambas fases. Estos fenómenos se deben a las propiedades específicas del complejo coloidal del suelo que tienen cargas electrostáticas y una gran superficie. Esta propiedad es un reflejo de la fertilidad del suelo y cambia en función del tipo de minerales y contenido de materia orgánica (Gama et al. 1998).

La **arcilla** puede considerarse como material fino que, tiene un diámetro menor que 2 micras. Se trata de un material natural, que desarrolla plasticidad cuando se mezcla con agua y que, bajo este estado, puede ser deformado por presión, además constituye una clase textural de los suelos. El contenido de arcilla puede ser reflejado directamente con las condiciones de humedad.

El **pH** del suelo es logaritmo negativo de la actividad de los iones de hidrógeno en el suelo. El pH mide el grado de acidez o alcalinidad de un suelo, expresado en términos de una escala de 1 a 14. Los suelos de pH menores de 7 son ácidos y se encuentran, generalmente, en climas húmedos, mientras que los mayores de 7 reflejan condiciones más secas.

La **profundidad** total del perfil de suelo es la distancia de la superficie al límite inferior del horizonte más profundo, expresado en centímetros; se considera que los suelos de climas húmedos son profundos, en tanto los de clima seco son más someros.

1.3. UNIDADES DE SUELO.

Para el presente trabajo se utilizaron las siguientes unidades de suelos: **Leptosol, Regosol, Cambisol, Andosol, Phaeozem, Vertisol, Luvisol y Acrisol** (figura 2), las cuales se ordenaron del suelo menos evolucionado al más evolucionado y cuyas principales características se describen a continuación:

- El **Leptosol** se distingue por tener una profundidad menor a los 10 cm. Se localizan en las sierras, en laderas, barrancas, así como en lomeríos y algunos terrenos planos. Tiene características muy variables, pues pueden ser fértiles o infértiles, arenosos o arcillosos. Su susceptibilidad a la erosión depende de la zona en donde se encuentren, de la topografía y del mismo suelo. Son suelos jóvenes (no en el sentido cronológico), que pueden no estar relacionados al régimen climático.

- El **Regosol** deriva del vocablo griego "*rhegos*" que significa sabana, haciendo alusión al manto de alteración que cubre la tierra. Los Regosoles se desarrollan sobre materiales no consolidados, alterados y de textura fina. Aparecen en cualquier zona climática, son muy comunes en zonas áridas, en los trópicos secos y en las regiones montañosas.

- El **Cambisol** deriva del vocablo latino "*cambiare*" que significa cambiar, haciendo alusión al principio de diferenciación de horizontes manifestado por cambios en el color, la estructura o el lavado de carbonatos, entre otros. Los Cambisoles se desarrollan sobre materiales de alteración procedentes de un amplio abanico de rocas, entre ellos destacan los depósitos de carácter eólico, aluvial o coluvial. Aparecen sobre todas las morfologías, climas y tipos de vegetación. Permiten un amplio rango de posibles usos agrícolas. Sus principales limitaciones están asociadas a la topografía, bajo espesor, pedregosidad o bajo contenido en bases. En zonas de elevada pendiente su uso queda reducido al forestal o pascícola.

- La palabra **Andosol** deriva de dos palabras japonesas ***an*** que significa negro, ***do*** que significa suelo, se desarrollan sobre cenizas y otros materiales volcánicos ricos en elementos vítreos. Tienen altos valores en contenido de materia orgánicaT sobre un (~20%), además tienen una gran capacidad de retención de agua y mucha capacidad de cambio. Se encuentran en regiones húmedas, del ártico al trópico. Su rasgo más sobresaliente es la formación masiva de complejos amorfos humus-aluminio.

- El **Phaeozem** deriva del vocablo griego "*phaios*" que significa oscuro y del ruso "zemlja" que significa tierra, haciendo alusión al color oscuro de su horizonte superficial, debido al alto contenido en materia orgánica. El material original lo constituye un amplio rango de materiales no consolidados; destacan los depósitos glaciares y el loess con predominio de los de carácter básico. Se asocian a regiones con un clima suficientemente húmedo para que exista lavado pero con una estación seca; el clima puede ir de cálido a frío y van de la zona templada a las tierras altas tropicales. El relieve es llano o suavemente ondulado y la vegetación de matorral tipo estepa o de bosque. Son suelos fértiles y soportan una gran variedad de cultivos. Sus principales limitaciones son las inundaciones y la erosión.

- El **Vertisol** proviene del vocablo latino "*vertere*" que significa verter o revolver, haciendo alusión al efecto de batido y mezcla provocado por la presencia de arcillas. El material original lo constituyen sedimentos con una elevada proporción de arcillas esmectíticas, o productos de alteración de rocas que las generen. Se encuentran en depresiones de áreas llanas o suavemente onduladas. El clima suele ser tropical, semiárido a subhúmedo o mediterráneo con estaciones contrastadas en cuanto a humedad, es decir, existe una estación seca prolongada (aproximadamente nueve meses) contrastando con un periodo de lluvias bien marcado. La vegetación suele ser de sabana, o de praderas naturales o con vegetación leñosa. Los Vertisoles se vuelven muy duros en la estación seca y muy plásticos en la húmeda. El labrado es muy difícil excepto en los cortos periodos de transición entre ambas estaciones.

- El **Luvisol** deriva del vocablo latino "*luere*" que significa lavar, haciendo alusión al lavado de arcilla de los horizontes superiores para acumularse en una zona más profunda. Los Luvisoles se desarrollan principalmente sobre una gran variedad de materiales no consolidados como depósitos glaciares, eólicos, aluviales y coluviales. Predominan en zonas llanas o con suaves pendientes de climas templados fríos o cálidos pero con una estación seca y otra húmeda, como el clima mediterráneo.

  Cuando el drenaje interno es adecuado, presentan una gran potencialidad para un gran número de cultivos a causa de su moderado estado de alteración y su, generalmente, alto grado de saturación.

- El **Acrisol** deriva del vocablo latino "*acris*" que significa muy ácido, haciendo alusión a su carácter ácido y su baja saturación en bases, provocada por su fuerte alteración. Los Acrisoles se desarrollan principalmente sobre productos de alteración de rocas ácidas, con elevados niveles de arcillas muy alteradas, las cuales pueden sufrir posteriores degradaciones. Predominan en viejas superficies con una topografía ondulada o colinada, con un clima tropical húmedo, subtropical o muy cálido. Los bosques claros son su principal forma de vegetación natural. La pobreza en nutrientes minerales, la toxicidad por aluminio, la fuerte adsorción de fosfatos y la alta suceptibilidad a la erosión, son las principales restricciones a su uso. Grandes áreas de Acrisoles se utilizan para cultivos de subsistencia, con una rotación de cultivos parcial. No son muy productivos salvo para especies de baja demanda y tolerantes a la acidez como la piña, caucho o palma de aceite.

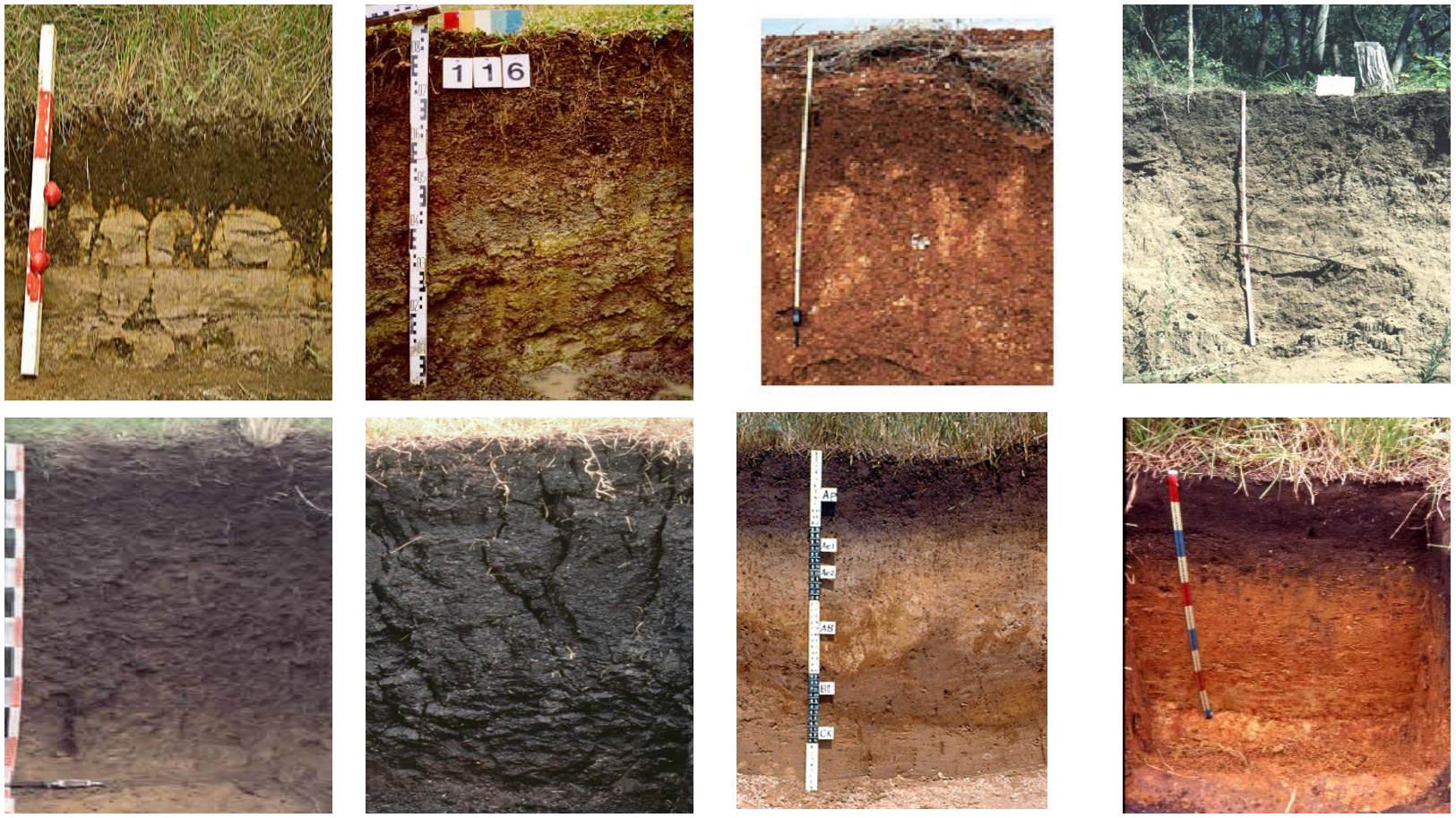


Figura 2. Fotografías de las unidades de suelo (Leptosol, Regosol, Cambisol, Andosol, Phaeozem, Vertisol, Luvisol y Acrisol) consideradas en el trabajo.

1.4. ANTECEDENTES

1.4.1. MODELOS MATEMÁTICOS APLICADOS A LAS CIENCIAS NATURALES.

Muchos modelos matemáticos han sido aplicados a diversas ciencias como son la ingeniería, la economía, la medicina, la demografía, las ciencias naturales, entre otras. Sin duda los primeros modelos aplicados a la biología han sido los que intentan describir o explicar la dinámica de poblaciones.

Quizá el modelo más antiguo de crecimiento de poblaciones es el que Leonardo de Pisa (o Fibonacci, como se le conoce desde el siglo XVIII) utilizó para describir el crecimiento de una población de conejos, documentado en su famoso libro sobre la Aritmética, ***liber abaci***, de 1202. Otro modelo es el del economista inglés Thomas R. Malthus (1766-1834), quien esbosa la hipótesis de que la tasa de crecimiento de la población de un país crece en forma proporcional a la población total, $P(t)$, de ese país en cualquier momento t. En términos matemáticos, esta hipótesis se puede expresar de la siguiente forma:

$$\frac{dP}{dt} = kP . \qquad (1)$$

Apesar de que este sencillo modelo no tiene en cuenta muchos factores (por ejemplo, la migración poblacional) que pueden influir en las poblaciones humanas, haciéndolas crecer o disminuir, predijo con mucha exactitud la población de los Estados Unidos desde 1790 hasta 1860. La ecuación diferencial (1) aún es utilizada con mucha frecuencia para modelar poblaciones de bacterias y de animales pequeños. Este modelo también se utiliza para modelar la desintegración radiactiva.

En cuanto al área de Ciencias de la Tierra, particularmente, en las Ciencias del Suelo, la modelación matemática ha sido limitada. Esto se debe al gran número de variables involucradas en la génesis del suelo, la cual depende de los cinco factores establecidos por el científico ruso V. Dokuchaev a fines del siglo XIX (clima, relieve, tiempo, material parental y organismos, en menor o mayor grado).

Algunos investigadores, en particular Jenny (1941) han tratado de demostrar, sin lograrlo, que esos factores son independientes. Sin embargo sólo el tiempo puede considerarse como independiente, pues los otros cuatro dependen, en mayor o menor grado uno del otro, del suelo mismo o de algún otro factor como puede ser la acción del hombre; también, establece la siguiente relación para explicar la génesis de un suelo de acuerdo a la acción dinámica de los factores ambientales:

$$S = f(cl, o, r, p, t)\,.$$

donde: cl = clima,
o = organismos,
r = relieve,
p = material parental,
t = tiempo.

Más recientemente, Philips (1993) señala que el modelo del factor de estado, según lo ejemplificado por la ecuación de Jenny, se puede reinterpretar como un sistema dinámico no lineal, donde el tiempo es un factor dinámico e independiente.

Ahora aplicando la teoría de sistemas dinámicos, se puede caracterizar al suelo mediante una serie de ecuaciones diferenciales ordinarias, cuyas variables se pueden medir de maneras diferentes (Philips, op cit. 1993b):

$$\frac{ds}{dt} = f_s(cl, o, r, p)\,, \qquad \frac{dr}{dt} = f_r(cl, o, s, p),$$

$$\frac{dcl}{dt} = f_{cl}(s, o, r, p), \qquad \frac{dp}{dt} = f_p(cl, o, r, s),$$

$$\frac{do}{dt} = f_o(cl, s, r, p),$$

donde: s = suelo, cl = clima, o = organismos, r = relieve y p = material parental.

Con relación a estas ecuaciones, se puede observar que el sistema no cambia o es cercano al equilibrio con respecto al tiempo (2), mostrando un comportamiento asintótico, es decir, tiende a cero.

$$\frac{dx_i}{dt} = 0 \qquad (2)$$

Mientras que el modelo en su forma más general es insoluble, puede ser solucionado con respecto a la estabilidad del sistema del suelo para una característica o un fenómeno particular. Esto sugiere que la complejidad espacial y temporal en patrones de las características del suelo y el desarrollo del mismo puedan ser inherentes en la dinámica del sistema, independientemente de la variación del medio ambiente.

Con la problemática presentada, Solleiro (1997) aplicó un modelo de regresión múltiple que permitiera modelar el desarrollo de Andosoles y que, conjugado con la información de los factores formadores, estableciera la historia genética del suelo y su tendencia evolutiva. Para ello, se seleccionaron varios Andosoles representantivos de varias etapas de desarrollo. El análisis de sus propiedades llevó a un modelo estadístico que determina su evolución, por medio de una ecuación lineal, lo que da limitaciones a su uso más extensivo.

Debido a la complejidad para la resolución de las ecuaciones, se han planteado modelos más sencillos, paramétricos. Un ejemplo es la ***metodología Provisional para la Evaluación de la Degradación de los Suelos***, que ha sido modificada para las condiciones de México (Gama et al. 1990). En dicho estudio se emplea la siguiente ecuación:

$$D = f(R', K, L, S, C, P)$$

donde :

D = pérdida de suelo por unidad de superficie
R´= indice de agresividad climática por efecto de la lluvia
K = factor de erodabilidad del suelo
L = longitud del declive
S = porcentaje de pendiente
C = factor cultivo - explotación
P = factor de prácticas de conservación

Al aplicar el modelo, los autores concluyen que la ecuación permite determinar la erodabilidad de las diferentes unidades de suelo, de una manera sencilla.

Existen otros trabajos con modelación matemática un poco más compleja como es el caso de Mc Bratney et al. (2003), en donde se hace mención a algunos métodos como la geoestadística, modelos lineales, conjuntos difusos o borrosos y redes de nervios.

Particularmente, la geoestadística es utilizada muy frecuentemente en el área de las geociencias, pues ayuda a encontrar variables distribuidas espacialmente y con ellas permitir estimaciones o simulaciones. El concepto de geoestadística, se define como la aplicación de la teoría de funciones aleatorias al reconocimiento y estimación de fenómenos naturales.

Los modelos lineales se han incluido para la predicción de atributos del suelo y para clasificarlos. Los métodos lineales para hacer agrupaciones incluyen análisis de discriminante (Hastie et al. 2001), es la técnica que ha sido aplicada en la ciencia del suelo por más de 60 años. La primera aplicación fue hecha por Cox y Martín (1937) quienes evaluaron si las características químicas daban información significativa sobre la presencia de ***azotobacter*** en suelo.

Por otro lado el uso de los conjuntos difusos, se utilizan en procesos complejos cuando no existe un modelo de solución sencillo, y en procesos no lineales; o cuando hay que introducir el conocimiento de un experto basado en conceptos imprecisos obtenidos de su experiencia; así como cuando ciertas partes del sistema a controlar (en este caso el suelo) son desconocidas o no pueden medirse de forma fiable. Asimismo cuando el ajuste de una variable puede provocar el desajuste de otras, cuando se quieran representar o trabajar con conceptos que tengan incertidumbre o imprecisión.

También las redes neuronales construyen o forman una opción de modelación matemática la cual trabaja de una manera análoga al cerebro humano. Además tienen un sistema de muchos elementos o neuronas interconectadas por los canales de comunicaciones donde los conectores llevan generalmente datos numéricos, codificados por una variedad de medios y organizados en capas.

El modelo matemático de una red neuronal abarca un sistema de funciones simples ligadas. La red consiste en un sistema de unidades de entrada, salida y unidades ocultas, que ligan las entradas a las salidas, las cuales extraen la información útil de entradas y las utilizan para predecir las salidas. Las redes de los nervios ahora se utilizan extensamente en la literatura de la ciencia de suelo, principalmente para predecir las cualidades del mismo, sobre todo las características hidráulicas (Minasny y McBratney, 2002; Chang y el Islam, 2000). Asimismo, este tipo de análisis permite predecir la probabilidad de ocurrencia de una clase de suelo en función de los factores ambientales (Zhu, 2000).

Otros modelos aplicados a las ciencias del suelo son los fractales, los cuales representan tanto teoría matemática como un método para analizar objetos naturales que no se ajustan a la geometría euclidiana. La geometría fractal fue desarrollada por Mandelbrot en 1975 en su libro la geometría fractal de la naturaleza. Los suelos son objetos que han sido caracterizados con esta teoría ya que cumplen con las propiedades de autoescalado, similitud, destacando los trabajos de Oleschko (2000).

La física de suelos es el área en la que se han desarrollado un gran número de modelos, ya que se requiere determinar los flujos de agua y aire para aplicaciones prácticas en la agricultura. Además, para hacer explícita las relaciones entre la geometría y las características hidrodinámicas del suelo, se han usado ampliamente las siguientes leyes, como la ley de Poiseuille, Laplace y algunas de sus variantes.

La ley de Poiseuille, relaciona la velocidad media del agua en un capilar cilíndrico (v), asimilado a un poro, con su radio (r) y el gradiente de energía ∇H (ver la ecuación 3):

$$V = -\frac{\rho_w g}{8\mu} \cdot r^2 \, \nabla H. \qquad (3)$$

Donde ρ_W es la densidad del agua, μ su viscosidad dinámica, g la aceleración gravitacional.

La ley de Laplace, relaciona el potencial de presión del agua en el suelo con el radio medio de curvatura del menisco de agua (r_c) en un poro de radio (r).

$$\psi = -\frac{2\sigma}{g r_c \rho_w} \quad . \tag{4}$$

También se obtuvo un modelo conceptual, el cual relaciona la conductividad hidráulica con la geometría del suelo, considerando un escurrimiento en dirección del eje z y realizando un corte perpendicular a dicho eje, obteniendo dos secciones de área total A_T. El flujo total de agua (Q) por una área elemental dA es $dQ = vdA$. El flujo total con respecto al área total es proporcionado por: $dq = vd\phi$, donde $dq = dQ/A_T$ y $d\phi = dA/A_T$, q es el flujo de Darcy y ϕ designa una área relativa. Si Ω representa el dominio de los poros llenos con agua entonces q se calcula como se ve en la ecuación (5):

$$q = \int_{\Omega} v d\phi = -\left(\left(\frac{\rho_w g}{8\mu} \right) \int_{\Omega} r^2 d\phi \right) \nabla H \; . \tag{5}$$

La identificación de la ecuación (6) con la ecuación (3) permite encontrar la relación entre la conductividad hidráulica y la geometría del suelo:

$$K(\Omega) = \left(\frac{\rho_w g}{8\mu} \right) \int_{\Omega} r^2 d\phi \, . \tag{6}$$

Childs y Collis (1950) atacan el problema desde un punto de vista probabilístico. Posteriormente Millington y Quirk (1961) trabajan con algunos modelos para correlacionar varias secciones de flujo y determinar el área total. Fuentes (1993) desarrolla una serie de modelos que retoman diferentes teorías de transferencia de flujos.

Con lo anterior, se puede observar que la modelación del sistema suelo es muy compleja. En principio, la génesis de suelo se enfrenta con el reto de explicar si realmente existen condiciones de equilibrio o cuasi-equilibrio y en consecuencia los cambios que registra el sistema son muy lentos, es decir, tomarlo como un sistema estático o dinámico. Richter (1987) señala claramente en su libro ***"el suelo como un reactor"***, el alcance de un modelo aplicado a los suelos, el cual sólo representará y/o describirá una parte del sistema real.

2.- ANÁLISIS DE CONGLOMERADOS (*CLUSTERS*).

2.1.- ANALISIS MULTIVARIADO.

Es el conjunto de métodos estadísticos cuya finalidad es analizar simultáneamente conjuntos de datos multivariantes dado que hay diversas variables medidas para cada individuo u objeto estudiado. Su importancia radica en que proporcionan un mejor entendimiento del fenómeno de estudio obteniendo información que los métodos estadísticos univariantes y bivariantes son incapaces de conseguir.

Dada una muestra de observaciones en un conjunto grande de variables, el *"análisis cluster"* es una técnica que agrupa a los elementos de la muestra en conglomerados, de tal forma que, respecto a la distribución de los valores de las variables, por un lado, cada conglomerado, sea lo más homogéneo posible y, por otro, sean muy distintos entre sí.

La clasificación de las especies, tal como se entiende en la actualidad, fue iniciada por C. Linneo en su famoso "Sistema Natural". Donde describió miles de especies utilizando la nomenclatura binomial, que asignaba a cada viviente el nombre latino con el género y la especie. El sistema taxonómico es una jerarquía organizada en niveles, en donde las clases disjuntas a cada nivel constituyen las llamadas taxas. Estas constituyen las categorías. Se habla así de las "especies", "géneros", "familia", "orden". La categoría género, tiene diversas taxas: los géneros que corresponden a una familia dada.

La taxonomía numérica surgió de la necesidad de ampliar los esquemas tradicionales de la sistemática de los seres vivientes. Por lo tanto, es en Biología (especialmente en Microbiología y Ecología) donde mayor aplicación ha tenido este método. Sneath y Sokal (1973) citan numerosas aplicaciones a la Biología Sistemática, haciendo amplia y especial mención a las clasificaciones de bacterias.

Posteriormente la taxonomía numérica se ha extendido a otros campos (Lingüística, Política, Economía, Psicología, Geología, Química, entre otras.), bajo el estímulo de la necesidad, cada vez mayor, de clasificar objetos. Aplicaciones, no propiamente biológicas, pueden verse en las obras editadas por Cole (1969), Hodson, Kendall et al. (1971), Romney et al. (1972), Benzecri (1976).

En general una clasificación jerárquica parte de un conjunto Ω cuyos elementos deben ser clasificados. Se trata de obtener sucesivas particiones (*"clusterings"*), organizados en diferentes niveles jerárquicos, estando cada partición formada por clases disjuntas (*"clusters"*). Los elementos de una misma clase deben ser razonablemente homogéneos (Las especies dentro de un género deben ser fenéticamente similares).

2.2.-DISTANCIAS Y DISIMILARIDADES.

Sea Ω un conjunto constituido por k objetos (especies, poblaciones geográficas, entre otras) indicado por $\{1,2,...,r,...,s,...,k\}$. Se asigna el nombre de distancia o disimilaridad entre r y s, a una medida indicada por $d(r,s)$, la cual mide el grado de semejanza entre ambos objetos (los cuales pueden ser vectores Xr y Xs), en relación a un cierto número de características cualitativas y cuantitativas. Cuanto mayor es la diferencia entre r y s, mayor es el valor positivo de $d(r,s)$. Una gran semejanza entre r y s, se refleja en un valor pequeño de $d(r,s)$.

Las propiedades que puede tener una distancia o disimilaridad, denotada por (d) son todas o algunas de las siguientes:

$i)\quad d(r,s)\geq 0$

$ii)\quad d(r,r)=0$

$iii)\quad d(r,s)=d(s,r)$ *(simetría)*

$iv)\quad d(r,s)\leq d(r,p)+d(p,s)$ *(desigualdad del triángulo)*

$v)\quad d(r,s)=0\Leftrightarrow r=s$

$vi)\quad d(r,s)\leq max\{d(r,p),d(p,s)\}$ *(desigualdad ultramétrica)*

Definición: Se dice que $d(r,s)$ *es una distancia euclidiana si para cualesquiera* $p_i, p_j \in R^m$, *en donde* R^m *es un espacio vectorial de dimensión* m, *se cumple:*

$$d(i,j)=d_2(p_i,p_j)=\left(\sum_{h=1}^{m}\left(x_{ih}-x_{jh}\right)^2\right)^{1/2}.$$

Se llamará a $d_2(p_i,p_j)$ *distancia euclidiana fundamental.*

Las anteriores distancias reciben diferentes nombres según las propiedades que verifican de acuerdo a la tabla 1:

Nombre	Propiedades
Distancia métrica:	**i), ii), iii), iv), v).**
Distancia euclidiana:	**i), ii), iii), iv).**
Disimilaridad:	**i), ii), iii).**
Distancia ultramétrica:	**i), ii), iii), vi).**

Tabla 1.- Nombres de las distancias según las propiedades que verifican.

Cuando los objetos pueden ser caracterizados por variables cuantitativas, se utiliza a la distancia de Minkowski

$$d_q(x_r, x_s) = \left(\sum_{j=1}^{n} \left| x_{rj} - x_{sj} \right|^q \right)^{1/q}.$$

que verifica las propiedades i), ii), iii) y iv), sin embargo, no son distancias euclidianas, salvo para los casos cuando $q=1$, $q=2$ y $q=\infty$, es decir:

1) Cuando $q=1$, se conoce como distancia ciudad:

$$d_1(x_r, x_s) = \sum_{j=1}^{n} \left| x_{rj} - x_{sj} \right|.$$

2) Cuando $q=2$ (distancia euclidiana), esta distancia es la más utilizada y se define de la siguiente manera:

$$d_2(x_r, x_s) = \left(\sum_{j=1}^{n} \left| x_{rj} - x_{sj} \right|^2 \right)^{1/2}.$$

3) Por último la expresión norma infinito:

$$d_\infty(x_r, x_s) = \sup_{1 \le j \le n} \left| x_{rj} - x_{sj} \right|.$$

En aplicaciones de la estadística multivariada, se caracteriza a personas, empresas, animales, etc, por medio de vectores de mediciones. Así, una persona puede estar caracterizada por: peso, estatura, etc. Hay que observar que las variables pueden ser medidas en escalas diferentes (metros, años, escalas nominales, kilogramos, etc.). La pregunta que surge de inmediato es: ¿Es correcto utilizar la distancia euclidiana? En principio, la respuesta puede ser que si las escalas de medición son distintas no deberíamos hacerlo, pues cambios en la escala producen grandes efectos en $d_2(\mathrm{Xr}, \mathrm{Xs})$.

Por ejemplo, en la tabla 2 se muestra el peso y la estatura de 3 individuos: A,B,C.

Individuos	Peso (gr)	Estatura (cm)
A	10	7
B	20	2
C	30	10

Tabla 2.- Peso y estatura de individuos A,B,C.

la distancia euclidiana entre cada individuo es:

$$d_2(A,B) = \sqrt{(10-20)^2 + (7-2)^2} = \sqrt{100+25} = \sqrt{125} = 11.18$$

$$d_2(A,C) = \sqrt{(10-30)^2 + (7-10)^2} = \sqrt{400+9} = \sqrt{409} = 20.22$$

$$d_2(B,C) = \sqrt{(20-30)^2 + (2-10)^2} = \sqrt{100+64} = \sqrt{164} = 12.80$$

Si se miden las estaturas en milímetros, se obtendrían las siguientes distancias:

$$d_2(A,B) = \sqrt{(10-20)^2 + (70-20)^2} = \sqrt{100+2500} = \sqrt{2600} = 50.9$$

$$d_2(A,C) = \sqrt{(10-30)^2 + (70-100)^2} = \sqrt{400+900} = \sqrt{1300} = 36.05$$

$$d_2(B,C) = \sqrt{(20-30)^2 + (20-100)^2} = \sqrt{100+6400} = \sqrt{6500} = 80.62$$

Por lo tanto se puede ver tabla 3.

	Estatura en (cm)	Estatura en (mm)
$d_2\,(A,B)$	11.18	50.9
$d_2\,(A,C)$	20.22	36.05
$d_2\,(B,C)$	12.80	80.62

Tabla 3.- Distancias calculadas entre los individuos A,B,C.

En este ejemplo se observa que el objeto A es el más cercano al objeto B que al objeto C, cuando la altura está expresada en cm, mientras que al expresar la altura en milímetros, el objeto A está más cercano al objeto C que al objeto B. Para solucionar este problema cada variable debe dividirse entre su desviación estándar.

$$S_j = \left[\frac{\sum_i \left(X_{ij} - \overline{X_j} \right)^2}{n-1} \right]^{\frac{1}{2}}.$$

Esto es, X_{ij} es reemplazada por X_{ij}/S_j. Aplicando esto a los datos anteriores se tiene que:

$$S_1 = \left[\frac{(10-20)^2+(20-20)^2+(30-20)^2}{2} \right]^{\frac{1}{2}} = \left[\frac{100+100}{2} \right]^{\frac{1}{2}} = \left[\frac{200}{2} \right]^{\frac{1}{2}} = 10$$

Por lo tanto los nuevos valores se muestran en la tabla 4:

$$S_2 = \left[\frac{(7-6.33)^2+(2-6.33)^2+(10-6.33)^2}{2} \right]^{\frac{1}{2}} = \left[\frac{.4444+18.777+13.444}{2} \right]^{\frac{1}{2}} = \left[\frac{32.6654}{2} \right]^{\frac{1}{2}} = 4.04$$

Individuos	Peso	Estatura
A	10/10 = 1	7/4.04 = 1.73
B	20/10 = 2	2/4.04 = 0.495
C	30/10 = 3	10/4.04 = 2.4

Tabla 4.- Valores obtenidos al dividirse entre su desviación estándar para solucionar el cambio de escala.

Otro problema que se presenta, son los efectos de la correlación entre variables. Para cubrir no sólo los efectos de cambio de escala sino también los de la correlación entre variables, se usa la **distancia de Mahalanobis**

$$d(Xr,Xs)=\left[(Xr-Xs)'\,S^{-1}(Xr-Xs)\right]^{\frac{1}{2}},$$

donde:

$$S_x=\frac{\sum_i (X_i-\overline{X})(X_i-\overline{X})'}{n-1}.$$

Ésta ha sido propuesta como una medida de distancia, invariante bajo transformaciones lineales no singulares de las variables, lo que se demuestra a continuación:

Demostración:

Sea $Y_i = AX_i+b$ una transformación lineal, entonces se tiene:

$$(Yr-Ys)=(AXr+b-(AXs+b))=(AXr+b-AXs-b)=(AXr-AXs)=A(Xr-Xs)$$

$$\therefore\ d(Yr,Ys)=\left[(Yr-Ys)'\,Sy^{-1}(Yr-Ys)\right]^{\frac{1}{2}}$$

$$=\left\{(AXr-AXs)'\left[\sum_i\frac{(AXi-A\overline{X})(AXi-A\overline{X})'}{n-1}\right]^{-1}(AXr-AXs)\right\}^{\frac{1}{2}}$$

$$=\left\{(A(Xr-Xs))'\left[\sum_i\frac{A(Xi-\overline{X})(A(Xi-\overline{X}))'}{n-1}\right]^{-1}(A(Xr-Xs))\right\}^{\frac{1}{2}}$$

$$=\left\{(Xr-Xs)'\,A'\left[\sum_i\frac{A(Xi-\overline{X})(Xi-\overline{X})'\,A'}{n-1}\right]^{-1}A(Xr-Xs)\right\}^{\frac{1}{2}}$$

$$=\left\{(Xr-Xs)'\,A'\left(A'\right)^{-1}S^{-1}A^{-1}A(Xr-Xs)\right\}^{\frac{1}{2}}$$

$$= \left\{ (Xr - Xs)' S^{-1} (Xr - Xs) \right\}^{\frac{1}{2}} = d(Xr - Xs).$$

Por lo tanto es invariante bajo cambios de escala.

Existen muchas versiones de la métrica, las cuales han sido utilizadas como una medida de disimilaridad. Por ejemplo Gower propuso la siguiente:

$$d(r,s) = \frac{1}{d} \sum_{j=1}^{d} \frac{|x_{rj} - x_{sj}|}{R_j},$$

donde R_j es el rango de la variable j. Otra medida es la de Bray-Curtis, la cual es usada ocasionalmente en Ecología, está tiene la siguiente forma:

$$d(r,s) = \frac{\sum_j |xrj - xsj|}{\sum_j xrj + \sum_j xsj} = 1 - \frac{2C}{A+B},$$

donde:

$A = \sum_j xrj$

$B = \sum_j xsj$ y $C = \sum_j \min(xrj, xsj)$. Esta medida no es una métrica si se considera el siguiente ejemplo:

sean $x_1 = (2,5,2,5,3)$, $x_2 = (3,5,2,4,3)$ y $x_3 = (9,1,1,1,1)$ entonces:

$$d_{13} = d(x_1, x_3) = \frac{|2-9| + |5-1| + |2-1| + |5-1| + |3-1|}{17+13} = \frac{7+4+1+4+2}{30} = \frac{18}{30} = 0.6$$

$$d_{12} = d(x_1, x_2) = \frac{|2-3| + |5-5| + |2-2| + |5-4| + |3-3|}{17+17} = \frac{1+0+0+1+0}{34} = \frac{2}{34} = 0.058$$

$$d_{23} = d(x_2, x_3) = \frac{|3-9| + |5-1| + |2-1| + |4-1| + |3-1|}{17+13} = \frac{6+4+1+3+2}{30} = \frac{16}{30} = 0.533$$

se puede ver claramente que $d_{13} > d_{12} + d_{23}$, por tanto no es una métrica ya que no cumple la desigualdad del triángulo.

La situación con respecto al escalamiento es difícil, pues existe falta de información en la literatura. Claramente, se debe evitar el uso de programas con cambios de escala automáticos; uno debe tener la opción de hacer un cambio de escala o no, en función de los objetivos y del tipo de trabajo a realizar.

Una medida, de disimilaridad $d(r,s)$ entre los objetos r y s no necesita ser una métrica, pues se define a ésta como una función que satisface $d(r,s) \geq 0$, $d(r,r) = 0$ y $d(r,s) = d(s,r)$. Sin embargo es posible tener $d(r,s) = 0$, con r diferente de s.

Las relaciones de orden[1] son más importantes que los valores numéricos y debe conservarse una función monótona creciente[2] de un coeficiente de disimilaridad. Por esta razón la desigualdad del triángulo[3] no es un requerimiento importante, pues una transformación monótona de una métrica no necesita satisfacerla, por ejemplo, el cuadrado de la métrica euclidiana.

1.- Definición Si S es un conjunto, un orden en S es una relación representada por el símbolo < y tiene las dos propiedades siguientes:

(i) Si $x, y \in S$, una y sólo una de las proposiciones siguientes es cierta:

$x<y$, $x=y$, $y<x$

(ii) Si $x, y, z \in S$, y si $x < y$ Y $y < z$, entonces $x < z$.

2.- Definición Sea f un número real en (a,b). Se dice que f es monótona creciente en (a,b) si $a < x < y < b$ entonces $f(x) < f(y)$.

3.- $|x+y| \leq |x| + |y|$ Esta propiedad se conoce como la "desigualdad del triángulo" y geométricamente significa que en todo triángulo un lado es igual o menor que la suma de los otros. Algebraicamente la igualdad se cumple cuando x y y son cero o números positivos y la desigualdad se da cuando x y y son números negativos.

2.3.- SIMILARIDADES.

El coeficiente de similaridad indica la relación que hay entre dos objetos dados. La similaridad entre dos objetos dados r y s puede ser una función de sus valores observados. Varias funciones han sido propuestas dependiendo del tipo de variable y del tipo de objetos.

Las similaridades son consideradas generalmente como relaciones simétricas $S(r,s) = S(s,r)$. La mayoría de los coeficientes de similaridad son no negativos y son acomodados de tal manera que tengan un límite superior a 1. Aunque algunos son como el coeficiente de correlación que cumple con la condición $-1 \le S(r,s) \le 1$.

Asociada a una medida de similaridad, existe una disimilaridad $d(r,s) = 1 - S(r,s)$ la cual es simétrica y no negativa. El grado de similaridad entre dos objetos se incrementa con el incremento de $S(r,s)$ y decrece con el incremento de $d(r,s)$. Para un mismo objeto es natural tener máxima similaridad con el mismo, es decir, $S(r,r) = 1$ y $d(r,r) = 0$.

2.3.1- MEDIDAS DE SIMILARIDAD PARA VARIABLES BINARIAS

Los más simples y comunes coeficientes de similaridad son aquéllos de variables dicotómicas, donde cada variable tiene únicamente dos valores. En algunos casos éstos pueden tratarse de la ausencia o presencia de alguna cualidad, por ejemplo: alto/bajo, nuevo/usado, etc. Estos datos para dos objetos o individuos r y s pueden ser puestos de acuerdo a la tabla 5:

Tabla 5.- Medida de similaridad para los objetos r y s.

	OBJETO r			
		+	-	
OBJETO s	+	α	β	$\alpha + \beta$
	-	γ	δ	$\gamma + \delta$
		$\alpha + \gamma$	$\beta + \delta$	d

donde : $\alpha =$ número de caracteres presentes comunes

$\delta =$ número de caracteres ausentes comunes

$d = \alpha + \beta + \gamma + \delta$

Se han propuesto muchos coeficientes de similaridad de manera, que éstos combinan las cantidades α, β, γ y δ, sin embargo incluir al coeficiente δ puede ser peligroso, ya que al añadir caracteres arbitrarios no comunes, podría falsamente hacerse similares objetos o individuos que no lo son.

La asociación se mide mediante un coeficiente de similaridad S_{rs} función de α, β y γ en la mayoría de las veces:

$$S_{rs} = f(\alpha, \beta, \gamma),$$

tal que: a) es creciente en α.

b) es decreciente en β y γ.

c) es simétrica en β y γ.

Muchos autores han propuesto coeficientes de similaridad con las siguientes propiedades:

1) $\dfrac{\alpha + \beta}{d}$.

2) $\dfrac{2\alpha}{2\alpha + \beta + \gamma}$.

3) $\dfrac{\alpha}{\alpha + 2(\beta + \gamma)}$.

4) $\dfrac{\alpha}{\alpha + \beta + \gamma}$ (coeficiente de Jaccard).

5) $\dfrac{2(\alpha + \delta)}{2(\alpha + \delta) + \beta + \gamma}$.

6) $\dfrac{\alpha + \delta}{d}$ (coeficiente de Sokal).

Los coeficientes más usados en la práctica son el de Sokal y el de Jaccard (Cuadras, 1996). El primero es simplemente el cociente del total de variables en donde los dos objetos o individuos coinciden, con el número total de variables; mientras que el segundo es el mismo cociente pero ignorando a δ. El problema de incluir o no a δ, radica en que pueden hacerse falsamente similares, pero es uno quien decide en que tipo de datos incluirlos o no.

Los diferentes coeficientes de similaridad toman distintos valores para el mismo conjunto de datos. Suponiendo, por ejemplo, que dos individuos (IND1 y IND2) tienen las mismas cantidades en diez variables binarias como se presenta en la tabla 6:

VARIABLES										
	1	2	3	4	5	6	7	8	9	10
IND1	1	0	0	0	1	1	0	0	1	0
IND2	0	0	0	0	1	0	0	1	1	0

Tabla 6.- Ejemplo de dos individuos, los cuales tienen las mismas diez variables.

En este conjunto de datos α = 2, ya que sólo en las variables 5 y 9 coinciden los individuos 1 y 2, también se observa que δ = 5, pues en las variables 2, 3, 4, 7 y 10 ninguno de los dos individuos tiene la característica, así también γ = 2, ya que el individuo 1 tiene la presencia de dos variables 1 y 6 que el individuo 2 no tiene, y análogamente para β; los datos resultantes se muestran en la tabla 7.

	IND1			
		1	0	
IND2	1	α = 2	β = 1	$\alpha + \beta$ = 3
	0	γ = 2	δ = 5	$\gamma + \delta$ = 7
		$\alpha + \gamma$ = 4	$\beta + \delta$ = 6	d = 10

Tabla 7.- Información obtenida al realizar la comparación del individuo 1 vs. el individuo 2.

En el caso de datos categóricos donde las variables tienen más de dos niveles, por ejemplo, color de ojos, pueden ser tomados de la misma manera que los datos binarios, con cada nivel de la variable considerada como una variable binaria, pero, esto puede producir un gran número de caracteres no comunes. Un mejor método consiste en tomar S_{rsk} la cual toma los valores de 1 y 0 para cada variable k, en función de si los dos individuos r y s son iguales en tal variable. Tomando el promedio se tiene:

$$S_{rs} = \frac{\sum_{k=1}^{d} S_{rsk}}{d}.$$

2.3.2- MEDIDAS DE SIMILARIDAD PARA VARIABLES CUANTITATIVAS

Las variables cuantitativas pueden ser convertidas en variables binarias, y por tanto, se pueden usar los coeficientes referidos anteriormente. Profundidad, por ejemplo, podría ser transformada en debajo de 12 cm y 12 cm o más. Tal aproximación obviamente provoca una pérdida de información y por consiguiente es mucho mejor tomar medidas de similaridad que puedan ser aplicadas directamente.

Con las variables cuantitativas, una medida de similaridad entre X_r y X_s, es la correlación de los pares (X_{rj}, X_{sj}) con $j=1,...,d$, donde:

$$s_{rs} = \frac{\sum_{j=1}^{d}\left(X_{rj} - \overline{X}_r\right)\left(X_{sj} - \overline{X}_s\right)}{\left[\sum_{j=1}^{d}\left(X_{rj} - \overline{X}_r\right)^2 \sum_{j=1}^{d}\left(X_{sj} - \overline{X}_s\right)^2\right]^{1/2}},$$

además $-1 \leq s_{rs} \leq 1$. Esta medida, aparte de no estar entre 0 y 1 tiene ciertas desventajas. Por ejemplo, si $s_{rs} = 1$, esto no significa que $X_r = X_s$, sólo se refiere a que los elementos de X_r están relacionados linealmente con los de X_s. Además, el significado que se le da a $\overline{X_r}$ es la media sobre las diferentes variables del objeto r.

2.3.3- MEDIDAS DE SIMILARIDAD PARA VARIABLES DE TIPO MIXTO

El problema de los conjuntos de datos que contiene varios tipos de variables pueden ser tomados en cuenta. Un coeficiente de similaridad propuesto por Gower es usado particularmente para tales tipos de datos, el cual se define como:

$$s_{rs} = \frac{\sum_{k=1}^{d} w_{rsk}\, s_{rsk}}{\sum_{k=1}^{d} w_{rsk}}.$$

En esta fórmula, S_{rsk} es la similaridad entre r y s con respecto a la variable k y W_{rsk} es 1 o 0 dependiendo si la comparación es considerada válida con relación a la variable k. Los valores de 0 son asignados cuando la variable k no es conocida para uno o ambos individuos o cuando se excluyen las coincidencias negativas para variables binarias. Para datos categóricos S_{rsk} toma el valor 1 cuando los dos individuos tienen el mismo valor y 0 en cualquier otro caso. Para variables cuantitativas se tiene:

$$s_{rs} = 1 - \frac{|x_{rk} - x_{sk}|}{R_k},$$

donde X_{rk} y X_{sk} son los dos valores individuales de la variable k, y R_k es el rango de la variable k.

	Peso en libras	Nivel de ansiedad	Depresión	Alucinaciones	Grupo de edad
Paciente 1	120	Medio	No	No	Joven
Paciente 2	150	Moderado	Si	No	Medio
Paciente 3	110	Severo	Si	Si	Viejo
Paciente 4	145	Medio	No	Si	Viejo
Paciente 5	120	Medio	No	Si	Joven

Tabla 8.- Para ilustra el coeficiente de Gower considérese los siguientes datos, los cuales hacen referencia a cinco pacientes con problemas psiquiátricos.

En este caso, además supóngase que el investigador desea excluir las coincidencias negativas en depresión y alucinaciones para el cálculo de similaridad entre pacientes.

$$s_{12} = \frac{1\times\left(1-\frac{30}{40}\right)+1\times 0+1\times 0+0\times 1+1\times 0}{1+1+1+0+1} = 0.0625$$

$$s_{13} = \frac{1\times\left(1-\frac{10}{40}\right)+1\times 0+1\times 0+1\times 0+1\times 0}{1+1+1+1+1} = 0.150$$

$$s_{45} = \frac{1\times\left(1-\frac{25}{40}\right)+1\times 1+1\times 0+1\times 1+1\times 0}{1+1+1+1+1} = 0.475$$

Los valores para todos los pares de pacientes pueden ser puestos en la siguiente matriz de similaridad S:

$$S = \begin{pmatrix} 1.000 & & & & \\ 0.062 & 1.000 & & & \\ 0.150 & 0.200 & 1.000 & & \\ 0.344 & 0.175 & 0.425 & 1.000 & \\ 0.750 & 0.005 & 0.350 & 0.475 & 1.000 \end{pmatrix}$$

2.4.- CLASIFICACIÓN JERÁRQUICA.

En una clasificación jerárquica los grupos se van fusionando progresivamente, mientras decrece la homogeneidad entre los grupos, cada vez más amplios, que se van formando. Las técnicas de clasificación jerárquica se dividen en métodos aglomerativos y divisivos.

a) **Métodos aglomerativos:** comienzan con *n* clases o conglomerados, cada uno con un individuo u objeto, posteriormente se van uniendo hasta formar un sólo grupo con los n individuos.

b) **Métodos divisivos:** empiezan con un sólo grupo de *n* individuos y se va dividiendo el grupo hasta formar n grupos con un individuo cada uno.

Con los métodos aglomerativos o divisivos no se pueden hacer reasignaciones, pues si un algoritmo ha unido dos individuos, éstos no pueden posteriormente separarse, y cuando un algoritmo divisivo ha realizado una separación, no puede ser unido posteriormente.

Existen otros tipos de clasificaciones como las monotéticas y politéticas, las primeras están basadas en una característica única que sea muy relevante y por lo general es divisiva, pues los objetos se clasifican en los que tienen la característica y los que no la tienen, por lo que puede dar lugar a clasificaciones poco adecuadas dada la dificultad de obtener grupos lo bastante homogéneos y naturales, por ejemplo, hay

pájaros que no vuelan, mamíferos que viven en el agua, etc. La segunda se basa en un gran número de características (en general) y no exige que todos los elementos de una clase posean todas las características, sino un número suficiente para poder justificar analogías entre los miembros de una misma clase. Este tipo de clasificación es aglomerativa.

La clasificación jerárquica se representa por medio de un diagrama en dos dimensiones conocido como dendograma, el cual esquematiza la fusión o división hecha en cada paso sucesivo del análisis.

El esquema de una clasificación jerárquica se muestra en la figura siguiente:

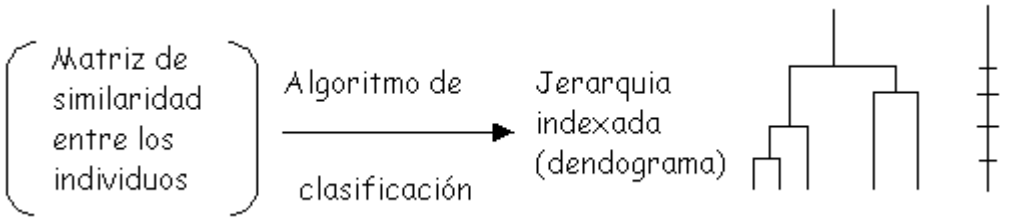


Figura 3.- Esquema de clasificación jerárquica.

Una jerarquía indexada es el resultado de una clasificación, la cual está representada gráficamente por un dendograma.

El dedograma de la figura 4 muestra la clasificación jerárquica de cinco especies hipotéticas, pero en este ejemplo (a diferencia de la taxonomía tradicional) los términos familia, género y especie tienen un significado más preciso, puesto que se habla de d-taxas o clases con distancia fenotípica d. La distancia d es el índice de la jerarquía, la cual mide el grado de homogeneidad entre las diferentes clases. Por ejemplo, la similaridad entre las especies 2 y 3 (d = 0.3) es mayor que entre las especies 4 y 5 (d = 0.5). Hay tres géneros {1}, {2,3} y {4,5}. La distancia fenotípica entre el género {1} formado por una sola especie, y el género {2,3} es 0.8. A partir de d = 0.5 se habla de familias.

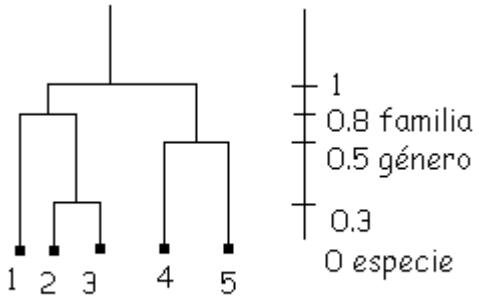


Figura 4.- Ejemplo hipotético de dendograma.

Existen dos procedimientos para dar un planteamiento matemático a la taxonomía numérica: mediante el concepto de jerarquías indexadas y en base a las propiedades de la distancia ultramétrica.

2.5.- JERARQUÍAS INDEXADAS.

En esta sección se formalizan los conceptos de taxas (género, especie, etc.), distancia fenotípica y clasificación jerárquica.

Sea $\Omega = \{1, 2, ..., n\}$ un conjunto finito. Se dice que $H \subset P(\Omega)$ es una jerarquía de partes de Ω si se verifica a los siguientes axiomas.

1) Axioma de la intersección.
Dados dos elementos de H, estos son disjuntos o uno de ellos está contenido en el otro, es decir:

$$\forall\, h, h' \in H \quad h \cap h' \in \{h, h', \varnothing\}.$$

2) Axioma de la reunión.
Todo elemento de H es el resultado de la reunión de los elementos de H que contiene o bien no contiene ningún elemento de H, es decir:

$$\forall\, h \in H, \qquad \cup \{h' \mid h' \in H,\ h' \subset h\} \in \{h, \varnothing\}.$$

Si además H contiene a Ω y a las partes formadas por un sólo elemento, es decir, $\Omega \in H$ y $\{i\} \in H \quad \forall i \in \Omega$ se dice entonces que H es una jerarquía total.

Los elementos de H (que son subconjuntos de Ω) se llaman (" conglomerados "). Si $h_1, h_2, ..., h_p$ son elementos de H tales que :

$$\Omega = h_1 + h_2 + ... + h_p$$

se dirá entonces que $\{h_1, h_2, ..., h_p\}$ es una partición de "conglomerados".

El primer axioma afirma que dos clases a un mismo nivel son disjuntas. Mientras que el segundo axioma afirma que una clase es reunión de las clases comparables de nivel inferior. Ambos axiomas reflejan la noción de que dos géneros deben ser siempre disjuntos y que todo género es reunión de las especies que los constituyen.

Por ejemplo, si $\Omega = \{1, 2, 3, 4, 5\}$ entonces :

$$H = \{ (1), (2), (3), (4), (5), (2,3), (4,5), (1,2,3), \Omega \}$$

Es una jerarquía, cuya representación gráfica es el dendograma de la figura 4.

Por último falta definir un índice que refleje la similitud entre clases (géneros, especies, etc.), que corresponda con la noción de la distancia fenotípica.

Se llama *índice* de la jerarquía H a una aplicación d, que a cada clase h hace corresponder un número real no negativo $d(h)$ tal que:

1) $d(\{i\}) = 0 \quad \forall i \in \Omega$
2) $h \subset h' \Rightarrow d(h) < d(h')$

se dice entonces que H es una jerarquía indexada.

Por ejemplo, el dendograma de la figura 4 corresponde a una jerarquía indexada. El índice d verifica lo siguiente:

$d(\{i\}) = 0$ con $i = 1, 2, ..., 6$
$d(\{2,3\}) = 0.3$
$d(\{4,5\}) = 0.5$
$d(\{1,2,3\}) = 0.8$
$d(\Omega) = 1$

El índice d se utiliza para cuantificar las diferencias entre las clases que se consideran a un mismo nivel (a nivel género, las especies 2 y 3 son más similares que las especies 4 y 5). A medida que aumenta el nivel de una clase aumenta el índice d, es decir, disminuye la similaridad entre los elementos de la clase.

Aunque d no es exactamente una similaridad sobre Ω se puede construir una disimilaridad poniendo

$$\bar{d}(i, j) = d(h) \tag{1}$$

si h es la menor clase que contiene a i, j.

De este modo, la disimilaridad entre dos especies del mismo género es el índice del género al que pertenecen; si son de distinto género, es el índice de la familia a la que pertenecen, etc.

Ejemplo : $\bar{d}(2,3) = 0.3,$ ya que $2, 3 \in \{2,3\}$
$\bar{d}(1,4) = 1,$ pues $1, 4 \in \Omega$

TEOREMA 1

Sea $\bar{d}(i, j)$ la disimilaridad (1) y considerando $x \geq 0$, la relación binaria en Ω

$$_iRx_j \Leftrightarrow \bar{d}(i, j) \leq x \tag{2}$$

es de equivalencia.

Demostración:

Es reflexiva, ya que si $_iRx_i \Rightarrow \bar{d}(i,i) = 0 \leq x$

Es simétrica, pues si $_iRx_j \Rightarrow \bar{d}(i, j) = d(h) = \bar{d}(j,i) \Rightarrow _jRx_i$

Es transitiva, pues si $_iRx_j \Rightarrow \bar{d}(i, j) = d(h),$ con $i, j \in h$

$$_jRx_k \Rightarrow \bar{d}(j,k) = d(h),\ \text{con}\ j, k \in h'$$

como $h \cap h' \neq \varnothing$, o $h \subset h'$ o $h' \subset h$ (axioma de la intersección). Suponiendo que $h \subset h'$ entonces $i, j, k \in h' \Rightarrow \bar{d}(i,k) \leq d(h') \leq x \Rightarrow _iRx_k$

$\therefore$ es de equivalencia.

Llamando partición ("conglomerativa") a nivel x a la partición de Ω definida por la relación R_x.

En el ejemplo anterior se tienen 5 particiones:

$$C_0 : \{1\},\{2\},\{3\},\{4\},\{5\} \quad \text{división a nivel } x = 0$$
$$C_1 : \{1\},\{2,3\},\{4\},\{5\} \quad \text{división a nivel } x = 0.3$$
$$C_2 : \{1\},\{2,3\},\{4,5\} \quad \text{división a nivel } x = 0.5$$
$$C_3 : \{1,2,3\},\{4,5\} \quad \text{división a nivel } x = 0.8$$
$$C_4 : \{1,2,3,4,5\} = \Omega \quad \text{división a nivel } x = 1$$

Las diferentes divisiones definen las especies, géneros, familias, etc. Corresponde al investigador decidir a partir de que nivel se habla de género o familia. Por ejemplo, si se establecen los géneros a partir de $x = 0.3$, dos especies i, j son del mismo género si $d(i, j) \leq 0.3$. Así, 2 y 3 serán del mismo género; pero 1, 4 y 5 definirán 3 géneros distintos. Si se establecen los géneros a partir de $x = 0.5$, entonces 4 y 5 serían del mismo género, pero 1 pertenecería a un género distinto.

Dada la relación que existe entre los conceptos de similaridad y disimilaridad, se puede también definir un índice sobre una jerarquía H que se corresponda con la noción de similaridad. Este índice es una aplicación

$$S : H \to [0,1]$$
verificando : 1) $S(\{i\}) = 1 \quad \forall i \in \Omega$
2) $h \subset h' \Rightarrow S(h) > S(h')$

En analogía con d (pero en sentido contrario), S permite cuantificar las similaridades entre las clases a un mismo nivel. Se puede definir un coeficiente de similaridad sobre Ω de la siguiente manera:

$\overline{S}(i, j) = S(h)$ **si h es la menor clase que contiene a i, j.**
La relación de equivalencia (2) adopta ahora la forma ${}_iRx_j \Leftrightarrow \overline{S}(i, j) \geq x$

2.6.- GEOMETRÍA ULTRAMÉTRICA.

Un dendograma, es una representación geométrica de una clasificación jerárquica, también es la representación gráfica de una distancia ultramétrica sobre un conjunto finito. A continuación se pueden ver los tipos de distancias y sus propiedades.

Sea $\Omega = \{1, 2,..., n\}$ un conjunto finito de objetos o individuos. Una ultramétrica u sobre Ω es una distancia que verifica:

1) $u(i,j) \geq 0$
2) $u(i,i) = 0$
3) $u(i,j) = u(j,i)$
4) $u(i,j) \leq \sup\{u(i,k), u(k,j)\}$ **(axioma ultramétrico)**

Como consecuencia inmediata u verifica la desigualdad del triángulo.

$$u(i,j) \leq \sup\{u(i,k), u(k,j)\} \leq u(i,k) + u(k,j)$$

Por ejemplo, se hará corresponder a cada par de especies i, j del dendograma anterior la distancia fenotípica d de la clase a la que pertenecen. Obteniendo la matriz de distancias

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & 0.8 & 0.8 & 1 & 1 \\ & 0 & 0.3 & 1 & 1 \\ & & 0 & 1 & 1 \\ & & & 0 & 0.5 \\ & & & & 0 \end{bmatrix} \end{array}$$

Se puede verificar que la distancia es ultramétrica, tomando $u(1,2)$ se tiene :

1) $u(1,2) = 0.8 \geq 0$
2) $u(1,1) = 0$
3) $u(1,2) = 0.8 = u(2,1)$
4) $u(1,2) = 0.8 \leq \sup\{u(1,3), u(3,2)\} = 0.8$

tomando $u(1,3)$ se tiene :

1) $u(1,3) = 0.8 \geq 0$
2) $u(1,1) = 0$
3) $u(1,3) = 0.8 = u(3,1)$
4) $u(1,3) = 0.8 \leq \sup\{u(1,4), u(4,3)\} = 1$

y de manera análoga para las demás distancias.

Se expondrá a continuación la propiedad fundamental de una ultramétrica:

TEOREMA 2

Todo triángulo en Ω es isósceles, siendo la base el lado de longitud menor, es decir, si i, j es la base

$$u(i,j) \leq u(i,k) = u(k,j)$$

Demostración :

Sea $\{i, j, k\}$ un triángulo y $u(i,j)$ el menor de sus tres lados. entonces

$$u(i,k) \leq \sup\{u(i,j), u(j,k)\} = u(j,k)$$
$$u(j,k) \leq \sup\{u(i,j), u(i,k)\} = u(i,k)$$

de donde $u(i,k) \leq u(j,k) \leq u(i,k)$

$$\therefore u(i,k) = u(j,k)$$

Con esto es posible efectuar agrupaciones en el conjunto Ω conservando la propiedad ultramétrica. Sean $i, j \in \Omega$ tal que $u(i,j) =$ mínimo; esto es i,j es el par de elementos más proximos. Considerando entonces el conjunto de n-1 elementos $\overline{\Omega} = \{1, \ldots, \{i,j\}, \ldots\}$ y definiendo una distancia en $\overline{\Omega}$ como sigue :

$$\bar{u}(k, \{i,j\}) = u(k,i) = u(k,j) \qquad \forall\ k \neq i, j$$
$$\bar{u}(k,m) = u(k,m) \qquad \forall\ k, m \neq i, j$$

entonces $\bar{u}$ es una ultramétrica en $\overline{\Omega}$.

Agrupando seguidamente los elementos más próximos de $\overline{\Omega}$, se podría definir análogamente una distancia sobre el conjunto de n-2 elementos restantes. En general se verifica :

TEOREMA 3

Sea $\Omega = h_1 + h_2 + \ldots + h_p$ **una partición de** Ω **y sea** u **una ultramétrica sobre** $\{h_1, h_2, \ldots, h_p\}$. **Si** h_i, h_j **son tales que :**

$$u(h_i, h_j) = \text{mínimo}$$

la distancia $\bar{u}$ **definida por :**

$$\bar{u}(h_k, h_i \cup h_j) = u(h_k, h_i) = u(h_k, h_j) \quad k \neq i, j$$

$$\bar{u}(h_k, h_m) = u(h_k, h_m) \quad k, m \neq i, j$$

es una ultramétrica sobre el conjunto $\{h_1, \ldots, h_i \cup h_j, \ldots, h_p\}$

Demostración :

$u(h_k, h_i) = u(h_k, h_j)$ **es válido porque** h_i, h_j, h_k **es un triángulo que cumple las condiciones del teorema 2.**

Considerando el triángulo $h_a, h_b, h_i \cup h_j$ **se verifica :**

$$\bar{u}(h_a, h_b) = u(h_a, h_b) \leq \sup\{u(h_a, h_i), u(h_b, h_i)\} = \sup\{\bar{u}(h_a, h_i \cup h_j), \bar{u}(h_b, h_i \cup h_j)\}$$

y $\bar{u}(h_a, h_i \cup h_j) = u(h_a, h_i) \leq \sup\{u(h_a, h_b), u(h_i, h_b)\} = \sup\{\bar{u}(h_a, h_b), \bar{u}(h_i \cup h_j, h_b)\}$

por lo que : $\bar{u}(h_a, h_b) \leq \sup\{\bar{u}(h_a, h_i \cup h_j), \bar{u}(h_b, h_i \cup h_j)\}$

y $\bar{u}(h_a, h_i \cup h_j) \leq \sup\{\bar{u}(h_a, h_b), \bar{u}(h_i \cup h_j, h_b)\}$

de manera análoga para $\bar{u}(h_b, h_i \cup h_j)$ **Q.E.D**

Si existen tres clases h_i, h_j, h_k **tales que :**

$$u(h_i, h_j) = u(h_i, h_k) = u(h_j, h_k) = \text{mínimo}$$

se agruparían para formar $\{h_1, \ldots, h_i \cup h_j \cup h_k, \ldots, h_p\}$ **y se definiría la ultramétrica** $\bar{u}$ **de forma parecida. Se procedería de la misma forma para más de tres clases.**

2.7.- ALGORITMO FUNDAMENTAL DE CLASIFICACIÓN.

Una jerarquía indexada H está formada por una sucesión de divisiones "conglomerativas" $C_0, C_1,..., C_m$ de Ω, a niveles respectivamente mayores.

Cuando se define una ultramétrica en Ω es posible construir tales divisiones mediante el algoritmo fundamental que consta de los siguientes pasos:

1) $C_0 = \{1\},\{2\},...,\{n\}$

2) Sea $C_{r-1} : h_1,..., h_p$ la división en el paso r y u una distancia ultramétrica sobre las clases de C_{r-1}. Agrupando las clases h_i, h_j tales que $u(h_i, h_j) =$ mínimo.

3) Formando la división $C_r : h_1,..., h_i \cup h_j,..., h_p$ y definiendo una distancia ultramétrica $\bar{u}$ sobre las clases de C de acuerdo con el teorema 3, es decir :
$$\bar{u}(h_k, h_i \cup h_j) = u(h_k, h_i) = u(h_k, h_j)$$

4) Repitiendo los pasos 2) y 3) las veces necesarias hasta llegar a la partición $C_m = \Omega$.

Por construcción, el resultado de este algoritmo es una jerarquía indexada H de índice, $d(h_i \cup h_j) = u(h_i, h_j)$ si h_i, h_j son las clases más próximas en la partición C_{r-1}.

Ejemplo:

Se aplicará el algoritmo fundamental a partir de la siguiente matriz.

división

$$C_0 : \{1\},\{2\},\{3\},\{4\},\{5\} \quad d(\{i\}) = 0$$

distancias

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & 0.8 & 0.8 & 1 & 1 \\ & 0 & 0.3 & 1 & 1 \\ & & 0 & 1 & 1 \\ & & & 0 & 0.5 \\ & & & & 0 \end{bmatrix} \end{array}$$

$$C_1 : \{1\},\{2,3\},\{4\},\{5\} \quad d(\{2,3\}) = 0.3 \quad \begin{matrix} & 1 & (2,3) & 4 & 5 \\ 1 & 0 & 0.8 & 1 & 1 \\ (2,3) & & 0 & 1 & 1 \\ 4 & & & 0 & 0.5 \\ 5 & & & & 0 \end{matrix}$$

$$C_2 : \{1\},\{2,3\},\{4,5\} \quad d(\{4,5\}) = 0.5 \quad \begin{matrix} & 1 & (2,3) & (4,5) \\ 1 & 0 & 0.8 & 1 \\ (2,3) & & 0 & 1 \\ (4,5) & & & 0 \end{matrix}$$

$$C_3 : \{1,2,3\},\{4,5\} \quad d(\{1,2,3\}) = 0.8 \quad \begin{matrix} & (1,2,3) & (4,5) \\ (1,2,3) & 0 & 1 \\ (4,5) & & 0 \end{matrix}$$

$$C_4 : \{1,2,3,4,5\} = \Omega \quad d(\Omega) = 1 \quad \begin{matrix} & \Omega \\ \Omega & 0 \end{matrix}$$

El teorema 4 establece la relación existente entre ultramétrica y jerarquía indexada.

TEOREMA 4

Una distancia ultramétrica u sobre un conjunto finito Ω define una jerarquía indexada sobre Ω. Recíprocamente, una jerarquía indexada sobre Ω define una distancia ultramétrica.

Demostración:

La primera parte del teorema es una consecuencia del algoritmo fundamental de clasificación.

Sea H una jerarquía indexada. Se define $u(i, j) = d(h)$ si h es la menor clase que contiene a i, j; la distancia u verifica :

1) $u(i, j) \geq 0$
2) $u(i, i) = d(\{i\}) = 0 \quad \forall\, i \in \Omega$
3) $u(i, j) = u(j, i)$
4) $u(i, j) \leq \sup\{u(i, k), u(k, j)\}$

Sean $h_{i,j}, h_{i,k}, h_{j,k}$ tales que $u(i, j) = d(h_{i,j})$, etc. Como $h_{i,k} \cap h_{j,k} \neq \varnothing$ entonces por el axioma de la intersección $h_{i,k} \subset h_{j,k}$ o bien $h_{j,k} \subset h_{i,k}$. En el primer caso se tiene $i, j, k \in h_{j,k} \therefore\ u(i, j) \leq u(j, k)$ luego $u(i, j) \leq \sup\{u(i, k), u(k, j)\}$. En el segundo caso se tendrá $i, j, k \in h_{i,k} \therefore\ u(i, j) \leq u(i, k)$ por lo que $u(i, j) \leq \sup\{u(i, k), u(k, j)\}$. Q.E.D

Una propiedad interesante de la ultramétrica es la varianza monótona. Considerando la preordenación asociada a una ultramétrica u y sea $f(u) = \hat{u}$ una transformación monótona, es decir:

$$u(i, j) \leq u(i', j') \Rightarrow \hat{u}(i, j) \leq \hat{u}(i', j')$$

satisface :

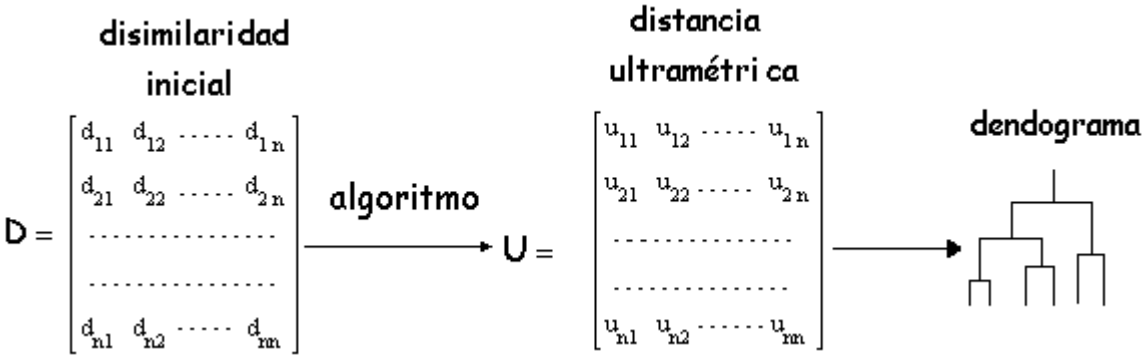
a) La preordenación asociada a $\hat{u}$ sigue siendo la misma.
b) $\hat{u}$ es también una ultramétrica.
c) u y $\hat{u}$ definen una misma jerarquía H, diferenciándose solamente en el índice de la jerarquía. Se dice que u y $\hat{u}$ son ultramétricas equivalentes.

2.8.- PRINCIPALES ALGORITMOS DE CLASIFICACIÓN.

El algoritmo fundamental de clasificación conduce a una clasificación jerárquica si la similaridad d entre los objetos de Ω cumple con las condiciones de una ultramétrica. Desafortunadamente, en un problema real la disimilaridad no es, en general ultramétrica.

Un algoritmo de clasificación consiste en transformar razonablemente la disimilaridad inicial para convertirla en ultramétrica, y con esto poder construir la jerarquía indexada.

Una representación de lo que sucede se muestra a continuación:



Algoritmo de clasificación:

1) Se inicia el algoritmo con la partición $C_0 : \{1\}, \{2\}, \ldots, \{n\}$

2) Sea $C_{r-1} : h_1, h_2, \ldots, h_p$ la partición en el paso r y d es una disimilaridad sobre las clases C_{r-1}, se agrupan las clases h_i, h_j; tales que, $d(h_i, h_j) =$ mínimo.

3) Se forma la partición de p-1 clases en $C_r : h_1, \ldots, h_i \cup h_j, \ldots, h_p$. Sin embargo ahora no se puede definir una ultramétrica $\overline{d}$ sobre las clases de C_r (aplicando el teorema 3), pues en general no será cierto que $d(h_i, h_k) = d(h_j, h_k)$, es decir, h_i, h_j, h_k no es un triángulo isósceles. Se debe definir la distancia de $h_i \cup h_j$ a h_k como una función de $d(h_i, h_k)$ y $d(h_j, h_k)$.

$$\overline{d}(h_k, h_i \cup h_j) = f\{d(h_i, h_k), d(h_j, h_k)\} \qquad (3)$$

mientras que $\overline{d}(h_k, h_m)$ permanece inalterada para las demás clases. Si para algún h_k se verifica que $d(h_i, h_k) = d(h_j, h_k)$, es conveniente que la función f verifique la condición:

$$f\{d(h_i, h_k), d(h_j, h_k)\} = d(h_i, h_k) = d(h_j, h_k) \qquad (4)$$

El propósito de la transformación (3), es modificar todo triángulo con base h_i, h_j para convertirlo en isósceles.

Algunos algoritmos hacen depender $\bar{d}$ también de $d(h_i, h_j)$, es decir:

$$\bar{d}(h_k, h_i \cup h_j) = f\{d(h_i, h_k), d(h_j, h_k), d(h_i, h_j)\}.$$

4) Se repiten los pasos 2) y 3) las veces necesarias hasta llegar a Ω. Los algoritmos de clasificación diferirán en la forma de definir $\bar{d}$ al pasar de la partición C_{r-1} a la partición C_r por fusión de las clases más próximas h_i, h_j.

A continuación se presentarán los métodos más utilizados, así como la manera en que estos son definidos con mayor frecuencia, incluyendo el método de Lance y Williams el cuál permitirá ver cómo a partir de éste se generan los demás métodos y con este mismo ver si se cumple la condición (4).

2.8.1.- MÉTODO DE LIGA SIMPLE (VECINO MÁS CERCANO).

Sean h_1 y h_2 dos conglomerados, la distancia entre ellos se define como la menor disimilaridad entre un miembro de h_1 y uno de h_2.

$$d_{(h_1)(h_2)} = d(h_1, h_2) = min\{d_{r,s} \mid r \in h_1, s \in h_2\}$$

Se ejemplificará el proceso de fusión a partir de la siguiente matriz de disimilaridades:

$$C_0 : \{1\},\{2\},\{3\},\{4\},\{5\} \qquad a_0 = 0$$

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & 7 & 1 & 9 & 8 \\ & 0 & 6 & 3 & 5 \\ & & 0 & 8 & 7 \\ & & & 0 & 4 \\ & & & & 0 \end{bmatrix} \end{array}$$

La menor de las distancias es $d(1,3) = 1 = a_1$, por lo que se unen 1 y 3 en un mismo conglomerado y se definen las distancias a este nuevo conglomerado $\{1,3\}$, es decir:

$$C_1 : \{1,3\},\{2\},\{4\},\{5\} \qquad a_1 = 1$$

$$d_{2(1,3)} = min\{d_{21}, d_{23}\} = min\{7,6\} = 6$$

$$d_{4(1,3)} = min\{d_{41}, d_{43}\} = min\{9,8\} = 8$$

$$d_{5(1,3)} = min\{d_{51}, d_{53}\} = min\{8,7\} = 7$$

$$\begin{array}{c} \\ \{1,3\} \\ 2 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccc} \{1,3\} & 2 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & 6 & 8 & 7 \\ & 0 & 3 & 5 \\ & & 0 & 4 \\ & & & 0 \end{bmatrix} \end{array}$$

Estas distancias forman el primer renglón, y las demás distancias permanecen igual.

La menor de las distancias en esta nueva matriz es $d_{2(4,5)} = 3 = a_3$, por lo que 2 se une al conglomera do que tiene al 4 y al 5, para formar el conglomerado $\{2,4,5\}$ y posteriormente se definen las distancias a este conglomerado.

$$C_2 : \{1,3\},\{2,4\},\{5\} \qquad a_2 = 3$$

$$d_{(1,3)(2,4)} = min\{d_{(1,3)2}, d_{(1,3)4}\} = min\{6,8\} = 6$$

$$d_{5(2,4)} = min\{d_{52}, d_{54}\} = min\{5,4\} = 4$$

$$\begin{array}{c} \\ \{1,3\} \\ \{2,4\} \\ 5 \end{array} \begin{array}{c} \begin{array}{ccc} \{1,3\} & \{2,4\} & 5 \end{array} \\ \begin{bmatrix} 0 & 6 & 7 \\ & 0 & 4 \\ & & 0 \end{bmatrix} \end{array}$$

La menor de las distancias en esta nueva matriz es $d(2,4) = 3 = a_2$, por lo que se unen 2 y 4 en un mismo conglomerado y se definen las distancias a este nuevo conglomerado $\{2, 4\}$, es decir :

$$C_3 : \{1,3\},\{2,4,5\} \qquad a_3 = 4$$

$$d_{(1,3)(2,4,5)} = min\{d_{(2,4)(1,3)}, d_{5(1,3)}\} = min\{6,7\} = 6$$

$$\begin{array}{c} \\ \{1,3\} \\ \{2,4,5\} \end{array} \begin{array}{c} \begin{array}{cc} \{1,3\} & \{2,4,5\} \end{array} \\ \begin{bmatrix} 0 & 6 \\ & 0 \end{bmatrix} \end{array}$$

$$C_4 : \{1,2,3,4,5\} = \Omega \qquad a_4 = 6$$

$$\begin{array}{c} \\ \Omega \end{array} \begin{array}{c} \begin{array}{c} \Omega \end{array} \\ \begin{bmatrix} 0 \end{bmatrix} \end{array}$$

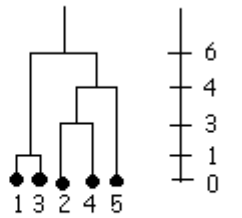
El dendograma y la matriz ultramétrica resultante (ver figura 5) se obtienen de los valores a_k; por ejemplo:

Los objetos 1 y 2 son unidos en el mismo conglomerado cuando la distancia entre los conglomerados a los que pertenecen son menores, es decir, en este caso cuando a = 6, por lo que la distancia de 1 a 2 es 6.

Los objetos 2 y 5 son unidos en el mismo conglomerado cuando la distancia entre los conglomerados a los que pertenecen son menores, es decir, en este caso cuando a = 4, por lo que la distancia de 2 a 5 es 4, resultando la matriz de la figura 5:

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & 6 & 1 & 6 & 6 \\ & 0 & 6 & 3 & 4 \\ & & 0 & 6 & 6 \\ & & & 0 & 4 \\ & & & & 0 \end{bmatrix} \end{array}$$

(Ultramétrica resultante).



(dendograma correspondiente al método de liga simple).

Figura 5.- Dendograma y matriz ultramétrica correspondientes al método de liga simple.

2.8.2.- MÉTODO DE LIGA COMPLETA (VECINO MÁS LEJANO).

Este método es contrario al método de liga simple, ya que se toma la mayor disimilaridad entre los miembros de *h1* contra los de *h2*.

$$d_{(h_1)(h_2)} = d(h_1, h_2) = max\{d_{r,s} \mid r \in h_1, s \in h_2\}$$

Para ejemplificar este método, se iniciará con una matriz de disimilaridades en la cual existe empate y ver con esto la diferencia de ultramétricas resultantes al tomar una u otra.

(primera solución)

$$C_0 : \{1\},\{2\},\{3\},\{4\},\{5\} \qquad a_0 = 0 \qquad \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & 1 & 1 & 2 & 5 \\ & 0 & 2 & 3 & 4 \\ & & 0 & 7 & 8 \\ & & & 0 & 6 \\ & & & & 0 \end{bmatrix} \end{array}$$

La menor distancia en esta matriz es $d\,(1,2) = d\,(1,3) = 1$. Si se toma a $d\,(1,2) = 1 = a_1$ primero, entonces 1 y 2 forman un conglomerado y se definen las distancias a este conglomerado como sigue:

$$C_1 : \{1,2\},\{3\},\{4\},\{5\} \qquad a_1 = 1$$

$$\begin{array}{c} \\ \{1,2\} \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccc} \{1,2\} & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & 2 & 3 & 5 \\ & 0 & 7 & 8 \\ & & 0 & 6 \\ & & & 0 \end{bmatrix} \end{array}$$

$$d_{3(1,2)} = max\{d_{3(1)}, d_{3(2)}\} = max\{1,2\} = 2$$
$$d_{4(1,2)} = max\{d_{4(1)}, d_{4(2)}\} = max\{2,3\} = 3$$
$$d_{5(1,2)} = max\{d_{5(1)}, d_{5(2)}\} = max\{5,4\} = 5$$

Estas distancias forman el primer renglón, y las demás distancias permanecen igual.

La menor distancia en esta nueva matriz es $d_{3(1,2)} = 2 = a_2$, por lo que 3 se une al conglomerado $\{1, 2\}$ para formar el conglomerado $\{1, 2, 3\}$ y tomando las distancias respectivas a este, se tiene:

$$C_2 : \{1,2,3\},\{4\},\{5\} \qquad a_2 = 2$$

$$\begin{array}{c} \\ \{1,2,3\} \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccc} \{1,2,3\} & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & 7 & 8 \\ & 0 & 6 \\ & & 0 \end{bmatrix} \end{array}$$

$$d_{4(1,2,3)} = max\{d_{4(1,2)}, d_{4(3)}\} = max\{3,7\} = 7$$
$$d_{5(1,2,3)} = max\{d_{5(1,2)}, d_{5(3)}\} = max\{5,8\} = 8$$

Siguiendo este procedimiento de forma análoga se obtiene:

$$C_3 : \{1,2,3\},\{4,5\} \qquad a_3 = 6$$

$$\begin{array}{c} \\ \{1,2,3\} \\ \{4,5\} \end{array} \begin{array}{c} \begin{array}{cc} \{1,2,3\} & \{4,5\} \end{array} \\ \begin{bmatrix} 0 & 8 \\ & 0 \end{bmatrix} \end{array}$$

$$d_{(1,2,3)(4,5)} = max\{d_{(1,2,3)4}, d_{(1,2,3)5}\} = max\{7,8\} = 8$$

$$C_4 : \{1,2,3,4,5\} \qquad a_4 = 8$$

$$\begin{array}{c} \\ \Omega \end{array} \begin{array}{c} \begin{array}{c} \Omega \end{array} \\ \begin{bmatrix} 0 \end{bmatrix} \end{array}$$

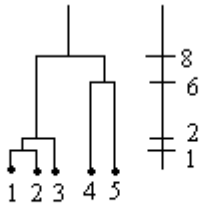


Figura 6.- Dendograma correspondiente al método del máximo.

El dendograma y la matriz ultramétrica son formados con los valores de a_k como se expuso anteriormente.

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & 1 & 2 & 8 & 8 \\ & 0 & 2 & 8 & 8 \\ & & 0 & 8 & 8 \\ & & & 0 & 6 \\ & & & & 0 \end{bmatrix} \end{array}$$

(ultramétrica resultante).

Ahora tomando a $d_{1(3)} = 1$ como el mínimo en el primer paso.

(segunda solución).

$$C_0 : \{1\},\{2\},\{3\},\{4\},\{5\} \qquad a_0 = 0 \qquad \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & 1 & 1 & 2 & 5 \\ & 0 & 2 & 3 & 4 \\ & & 0 & 7 & 8 \\ & & & 0 & 6 \\ & & & & 0 \end{bmatrix} \end{array}$$

$$C_1 : \{1,3\},\{2\},\{4\},\{5\} \qquad a_1 = 1 = d_{1(3)} \qquad \begin{array}{c} \\ \{1,3\} \\ 2 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccc} \{1,3\} & 2 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & 2 & 7 & 8 \\ & 0 & 3 & 4 \\ & & 0 & 6 \\ & & & 0 \end{bmatrix} \end{array}$$

$$d_{2(1,3)} = max\{d_{2(1)}, d_{2(3)}\} = max\{1,2\} = 2$$
$$d_{4(1,3)} = max\{d_{4(1)}, d_{4(3)}\} = max\{2,7\} = 7$$
$$d_{5(1,3)} = max\{d_{5(1)}, d_{5(3)}\} = max\{5,8\} = 8$$

$$C_2 : \{1,2,3\},\{4\},\{5\} \qquad a_2 = 2$$
$$d_{4(1,2,3)} = max\{d_{4(1,3)}, d_{4(2)}\} = max\{7,3\} = 7$$
$$d_{5(1,2,3)} = max\{d_{5(1,3)}, d_{5(2)}\} = max\{8,4\} = 8$$

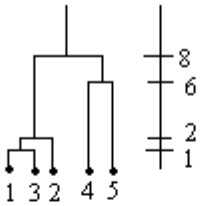
$$\begin{array}{c} \\ \{1,2,3\} \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccc} \{1,2,3\} & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & 7 & 8 \\ & 0 & 6 \\ & & 0 \end{bmatrix} \end{array}$$

$$C_3 : \{1,2,3\},\{4,5\} \qquad a_3 = 6$$
$$d_{(1,2,3)(4,5)} = max\{d_{4(1,2,3)}, d_{5(1,2,3)}\} = max\{7,8\} = 8$$

$$\begin{array}{c} \\ \{1,2,3\} \\ \{4,5\} \end{array} \begin{array}{c} \begin{array}{cc} \{1,2,3\} & \{4,5\} \end{array} \\ \begin{bmatrix} 0 & 8 \\ & 0 \end{bmatrix} \end{array}$$

$$C_4 : \{1,2,3,4,5\} = \Omega \qquad a_4 = 8$$

$$\begin{array}{c} \\ \Omega \end{array} \begin{array}{c} \begin{array}{c} \Omega \end{array} \\ \begin{bmatrix} 0 \end{bmatrix} \end{array}$$



(dendograma de la segunda solución del método del máximo).

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & 2 & 1 & 8 & 8 \\ & 0 & 2 & 8 & 8 \\ & & 0 & 8 & 8 \\ & & & 0 & 6 \\ & & & & 0 \end{bmatrix} \end{array}$$

(ultramétrica resultante).

Observaciones:

1) Si se considera la clase de ultramétricas u (i, j) inferiores a d (i, j), es decir,

$$u(i, j) \leq d(i, j) \qquad \forall i, j \in \Omega,$$

entonces la jerarquía indexada por el método del mínimo tiene como ultramétrica asociada u_1 el elemento máximo de esta clase, además u_1 es único y representa la aproximación a la disimilaridad d.

2) Si se considera $u(i, j)$ tal que:

$$u(i, j) \geq d(i, j) \qquad \forall i, j \in \Omega,$$

la jerarquía indexada construida por el método del máximo tiene como ultramétrica asociada u_2 un elemento minimal de esta clase.

u_2 no es necesariamente única y representa la aproximación por exceso a la disimilaridad d. Si toda las $d(i, j)$ con $i \neq j$ son diferentes, entonces u_2 es única.

3) La ultramétrica inferior máxima u_1 y la ultramétrica superior minimal u_2 que resultan de los dos métodos mínimo y máximo verifican:

$$u_1(i, j) \leq d(i, j) \leq u_2(i, j),$$

$u_1 = u_2 \Leftrightarrow d$ es también ultramétrica.

2.8.3.- MÉTODO DEL CENTROIDE.

La distancia entre dos conglomerados se define como la distancia entre los centroides de cada conglomerado. Si

$$\bar{x}_1 = \sum_{r \in h_1} \frac{x_r}{n_1},$$

es el centroide de los n_1 miembros de h_1 y $\bar{x}_2$ se define igual pero en h_2, entonces :

$$d_{(h_1)(h_2)} = d(h_1, h_2) = p(\bar{x}_1, \bar{x}_2),$$

donde p es una medida de proximidad como la correlación, el cuadrado de la distancia euclídia $\|\bar{x}_1 - \bar{x}_2\|^2$ u otras disimilaridades.

Se puede empezar con una matriz de proximidades con elementos p (x_r, x_s) y, en cada paso, los dos conglomerados más cercanos son unidos y reemplazados por el nuevo centroide.

$$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$$

2.8.4.- MÉTODO DE WARD.

Usando una idea propuesta por Ward para el caso univariado , Wishart propuso unir los dos conglomerados que minimicen $I_{(k_1)(k_2)}$, esto es, el incremento del total del cuadrado de las sumas de las distancias dentro de los conglomerados, donde $\bar{x}$ es el centroide de k_1 y k_2.

$$I_{(k_1)(k_2)} = \sum_{r \in k_1 \cup k_2} \|x_r - \bar{x}\|^2 - \left\{ \sum_{r \in k_1} \|x_r - \bar{x}_1\|^2 + \sum_{r \in k_2} \|x_r - \bar{x}_2\|^2 \right\}$$

$$= \sum_{\alpha=1}^{2} n_\alpha \|\bar{x}_\alpha - \bar{x}\|^2 = \frac{n_1 n_2}{n_1 + n_2} \|\bar{x}_1 - \bar{x}_2\|^2$$

En particular los objetos r y s $\quad I_{(r)(s)} = \frac{1}{2}\|\bar{x}_r - \bar{x}_s\|^2 = \frac{1}{2} d_{rs}^2$

Comenzando con $D = \left[d_{rs}^2\right]$ se puede definir la distancia entre dos conglomerados como :

$$d_{(k_1)(k_2)} = 2 I_{(k_1)(k_2)}$$

Este método ha sido propuesto por distintos autores bajo los nombres de conglomerados de varianza mínima, método de suma de cuadrados, método de Ward y otros.

2.8.5.- MÉTODO DE LA MEDIANA.

Este método es parecido al del centroide, sólo que ahora se define $\bar{x}$ de la siguiente forma:

$$\bar{x} = \frac{1}{2}(\bar{x}_1 + \bar{x}_2).$$

Este método fue introducido para cubrir una desventaja del método del centroide, la cual es que si la medida o el tamaño de dos grupos para ser unidos son muy diferentes, entonces el centroide del nuevo grupo podría caer muy cercano al grupo más grande o incluso dentro de él. Las propiedades o características del grupo pequeño son por tanto perdidas virtualmente. La estrategia puede ser hecha de tal manera que se asuman los grupos de igual tamaño, por lo que el nuevo centroide caerá en medio de los dos centroides de cada uno de los grupos.

2.8.6.- MÉTODO DEL PROMEDIO ENTRE GRUPOS (U.P.G.M.A).

La distancia entre h_1 y h_2 se define como el promedio de las $n_1 n_2$ disimilaridades entre todos los pares.

$$d_{(h_1)(h_2)} = d(h_1, h_2) = \frac{1}{n_1 n_2} \sum_{r \in h_1} \sum_{s \in h_2} d_{rs}.$$

2.8.7.- MÉTODO FLEXIBLE DE LANCE Y WILLIAMS.

Lance y Williams mostraron que el siguiente método da lugar a los métodos anteriores:

$$\begin{aligned} d_{(h_k)(h_i \cup h_j)} &= \alpha_i d_{(h_k)(h_i)} + \alpha_j d_{(h_k)(h_j)} + \beta\, d_{(h_i)(h_j)} + \gamma \left| d_{(h_k)(h_i)} - d_{(h_k)(h_j)} \right| \\ &= \alpha_i d(h_k, h_i) + \alpha_j d(h_k, h_j) + \beta\, d(h_i, h_j) + \gamma \left| d(h_k, h_i) - d(h_k, h_j) \right|. \end{aligned}$$

Este método incluye como caso particular a los métodos anteriores y permite observar como es que se cumple la condición (4) expuesta anteriormente, es decir:

$$\overline{d}\left(h_k, h_i \cup h_j\right) = f\left[d\left(h_i, h_k\right), d\left(h_j, h_k\right)\right] = d\left(h_i, h_k\right) = d\left(h_j, h_k\right).$$

por ejemplo, para $\alpha_i = \alpha_j = \frac{1}{2}, \beta = 0$ y $\gamma = -\frac{1}{2}$ se obtiene el método del mínimo

$$d\left(h_k, h_i \cup h_j\right) = \frac{1}{2} d\left(h_k, h_i\right) + \frac{1}{2} d\left(h_k, h_j\right) - \frac{1}{2}\left|d\left(h_k, h_i\right) - d\left(h_k, h_j\right)\right|$$

Si $d\left(h_k, h_i\right) > d\left(h_k, h_j\right) \Rightarrow \left|d\left(h_k, h_i\right) - d\left(h_k, h_j\right)\right| = d\left(h_k, h_i\right) - d\left(h_k, h_j\right)$

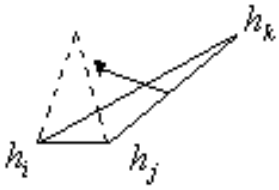
$$\therefore d\left(h_k, h_i \cup h_j\right) = \frac{1}{2} d\left(h_k, h_i\right) + \frac{1}{2} d\left(h_k, h_j\right) - \frac{1}{2} d\left(h_k, h_i\right) + \frac{1}{2} d\left(h_k, h_j\right) = d\left(h_k, h_j\right).$$

Si $d\left(h_k, h_i\right) < d\left(h_k, h_j\right) \Rightarrow \left|d\left(h_k, h_i\right) - d\left(h_k, h_j\right)\right| = d\left(h_k, h_j\right) - d\left(h_k, h_i\right)$

$$\therefore d\left(h_k, h_i \cup h_j\right) = \frac{1}{2} d\left(h_k, h_i\right) + \frac{1}{2} d\left(h_k, h_j\right) + \frac{1}{2} d\left(h_k, h_i\right) - \frac{1}{2} d\left(h_k, h_j\right) = d\left(h_k, h_i\right).$$

$$\therefore \overline{d}\left(h_k, h_i \cup h_j\right) = min\left\{d\left(h_k, h_i\right), d\left(h_k, h_j\right)\right\}.$$

Lo que geométricamente significa, es que el triángulo se deforma hasta obtener dos lados iguales que coinciden con el menor de los lados que no son base.



El método del máximo verifica de manera análoga la condición (4) con $\alpha_i = \alpha_j = \frac{1}{2}$, $\beta = 0$ y $\gamma = \frac{1}{2}$.

$$d\left(h_k, h_i \cup h_j\right) = \frac{1}{2} d\left(h_k, h_i\right) + \frac{1}{2} d\left(h_k, h_j\right) + \frac{1}{2}\left|d\left(h_k, h_i\right) - d\left(h_k, h_j\right)\right|$$

Si $d\left(h_k, h_i\right) > d\left(h_k, h_j\right) \Rightarrow \left|d\left(h_k, h_i\right) - d\left(h_k, h_j\right)\right| = d\left(h_k, h_i\right) - d\left(h_k, h_j\right)$

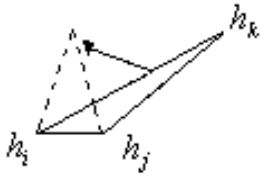
$$\therefore d\left(h_k, h_i \cup h_j\right) = \frac{1}{2} d\left(h_k, h_i\right) + \frac{1}{2} d\left(h_k, h_j\right) + \frac{1}{2} d\left(h_k, h_i\right) - \frac{1}{2} d\left(h_k, h_j\right) = d\left(h_k, h_i\right).$$

Si $d\left(h_k, h_i\right) < d\left(h_k, h_j\right) \Rightarrow \left|d\left(h_k, h_i\right) - d\left(h_k, h_j\right)\right| = d\left(h_k, h_j\right) - d\left(h_k, h_i\right)$

$$\therefore d\left(h_k, h_i \cup h_j\right) = \frac{1}{2} d\left(h_k, h_i\right) + \frac{1}{2} d\left(h_k, h_j\right) - \frac{1}{2} d\left(h_k, h_i\right) + \frac{1}{2} d\left(h_k, h_j\right) = d\left(h_k, h_j\right).$$

$$\therefore \overline{d}\left(h_k, h_i \cup h_j\right) = max\left\{d\left(h_k, h_i\right), d\left(h_k, h_j\right)\right\}.$$

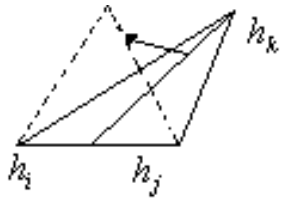
El triángulo se deforma pero ahora tomando el lado mayor que no es base.



Para el método de la mediana $\alpha_i = \alpha_j = \frac{1}{2}, \beta = -\frac{1}{4}$ y $\gamma = 0$ definiendo

$$\overline{d}\left(h_k, h_i \cup h_j\right) = \frac{1}{2} d\left(h_k, h_i\right) + \frac{1}{2} d\left(h_k, h_j\right) - \frac{1}{4} d\left(h_i, h_j\right)$$

en este caso se deforma el triángulo de modo que los dos lados iguales coinciden con la mediana del mismo.



este método no verifica la condición (4) y puede presentar inversiones, es decir, que el triángulo isósceles que resulte tenga la base mayor que los lados iguales.

Lance y Williams sugirieron usar el método flexible con las siguientes restricciones : $\alpha_i + \alpha_j + \beta = 1, \alpha_i = \alpha_j, \beta < 1$ y $\gamma = 0$ y un número pequeño para β tal como $\beta = -0.25$.

2.9.- COMPARACIÓN DE MÉTODOS.

2.9.1.- MONOTONÍA.

Si el proceso de fusión es representado por un dendograma con la propiedad de incremento de los valores a_k, entonces se requiere $a_k = mind_{(h_1)(h_2)}$ en el paso k. Lo que pasa en el paso k y $k+1$ puede ser descrito en términos de cuatro conglomerados h_1, h_2, h_3, h_4. Suponiendo que en el paso k se forma $h_1 \cup h_2$, esto es:

$$a_k = d(h_1, h_2) < min\{d(h_1, h_3), d(h_2, h_3), d(h_3, h_4)\}$$

y en el paso $k+1$ se une: $1)\ h_3 \cup h_4$ o $2)\ h_1 \cup h_2 \cup h_3$.

En el caso 1) se tiene $d_{k+1} = d(h_3, h_4) > d(h_1, h_2) = a_k$

para el caso 2) se consideran los métodos de liga simple y liga completa, teniendo:

$$d(h_3, h_1 \cup h_2) = a_{k+1} = min\{d(h_3, h_1), d(h_3, h_2)\} \Rightarrow a_{k+1} > a_k$$

$$d(h_3, h_1 \cup h_2) = a_{k+1} = max\{d(h_3, h_1), d(h_3, h_2)\} \Rightarrow a_{k+1} > a_k$$

para los otros métodos tenemos $(\gamma = 0, ... \alpha_i, \alpha_j > 0)$ se encuentra que:

$$d(h_3, h_1 \cup h_2) > (\alpha_i + \alpha_j + \beta) d(h_1, h_2)$$

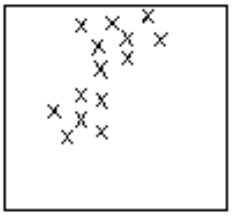
por lo que se requiere que $\alpha_i + \alpha_j + \beta > 1$ para la monotonía. Por lo que todos los métodos anteriores cumplen con la monotonía, excepto el método del centroide y el de la media en los que puede suceder que $a_{k+1} < a_k$. Lo que quiere decir que el triángulo isósceles puede presentar inversiones, es decir, que el triángulo resultante tenga la base mayor que los lados iguales.

2.9.2.- PROPIEDADES ESPACIALES.

Las proximidades iniciales pueden ser consideradas como espacios definidos con propiedades conocidas. Sin embargo, por la forma de los conglomerados, esto no significa que dentro de ellos se definan espacios con las propiedades originales.

Un algoritmo es espacio contractivo si la ultramétrica asociada a la clasificación jerárquica tiende a aproximar a los objetos respecto a sus disimilaridades iniciales. Si tiende a alejarlos, se dice entonces que es dilatante. Si por el contrario, no cambia de forma apreciable las disimilaridades, el algoritmo es conservativo. El método del mínimo y el del máximo son, respectivamente, espacio contractivo y espacio dilatante. En general los métodos basados en medias son espacios conservativos.

Algunos algoritmos espacio contractivos, tal como el método del mínimo tienden a unir en un solo conglomerado, grupos que están relativamente separados por la presencia de algún punto entre ellos. Este problema es conocido como cadena y la representación gráfica es la siguiente.



Los diferentes algoritmos de clasificación están representados en la tabla 1 como se muestra a continuación:

flexible	UPGMA	centroide	mediana	Ward	máximo	mínimo	método
α	$\frac{n_i}{n_i + n_j}$	$\frac{n_i}{n_i + n_j}$	$\frac{1}{2}$	$\frac{n_i + n_k}{n_i + n_j + n_k}$	$\frac{1}{2}$	$\frac{1}{2}$	α_i
α	$\frac{n_j}{n_i + n_j}$	$\frac{n_j}{n_i + n_j}$	$\frac{1}{2}$	$\frac{n_j + n_k}{n_i + n_j + n_k}$	$\frac{1}{2}$	$\frac{1}{2}$	α_j
$\beta < 1$	**0**	$\frac{-n_i n_j}{(n_i + n_j)^2}$	$-\frac{1}{4}$	$\frac{-n_k}{n_i + n_j + n_k}$	**0**	**0**	β
0	**0**	**0**	**0**	**0**	$\frac{1}{2}$	$-\frac{1}{2}$	γ
contractivo	**conservativo**	**conservativo**	**conservativo**		**dilatante**	**contractivo**	**espacio de distorcion**
si	**si**	**no**	**no**	**si**	**si**	**si**	**monotonía**
no	**no**	**si**	**si**	**no**	**no**	**no**	**inversiones**
si	**si**	**no**	**no**	**si**	**si**	**si**	**condición (4)**

Tabla 1.- Correspondiente a las propiedades de los métodos.

2.9.3.- MEDIDAS DEL GRADO DE DISTORSIÓN.

Las técnicas de clasificación jerárquica imponen una estructura jerárquica en los datos, por lo que es necesario ver si al aplicar el algoritmo de los datos originales cambian considerablemente, y con esto ver si existe una buena o mala clasificación.

El método más comúnmente usado es el llamado coeficiente de correlación cofenético. Este es simplemente la correlación que existe entre los pares de distancias (d_{ij}, u_{ij}) donde d_{ij} = disimilaridad inicial y u_{ij} = ultramétrica correspondiente.

Para ilustrar el uso del coeficiente de correlación cofenético, se tomarán las matrices de disimilaridad original y la ultramétrica correspondiente, asociada al método del máximo correspondiente a la primera solución.

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array}\begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & 1 & 1 & 2 & 5 \\ & 0 & 2 & 3 & 4 \\ & & 0 & 7 & 8 \\ & & & 0 & 6 \\ & & & & 0 \end{bmatrix} \end{array}$$

(disimilaridad original).

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array}\begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & 1 & 2 & 8 & 8 \\ & 0 & 2 & 8 & 8 \\ & & 0 & 8 & 8 \\ & & & 0 & 6 \\ & & & & 0 \end{bmatrix} \end{array}$$

(ultramétrica resultante).

$$d_{ij} = 1,1,2,5,2,3,4,7,8,6 \quad \textbf{con } \bar{d}_i = 3.9$$

$$u_{ij} = 1,2,8,8,2,8,8,8,8,6 \quad \textbf{con } \bar{u}_i = 5.9$$

$$\text{y } r_u = \rho = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}}$$

entonces

$$r_u = \frac{\sum (d_{ij} - \bar{d}_i)(u_{ij} - \bar{u}_i)}{\sqrt{\sum (d_{ij} - \bar{d}_i)^2}\sqrt{\sum (u_{ij} - \bar{u}_i)^2}} = \frac{(1-3.9)(1-5.9)+(1-3.9)(2-5.9)+\ldots+(6-3.9)(6-5.9)}{\sqrt{(1-3.9)^2+\ldots+(6-3.9)^2}\sqrt{(1-5.9)^2+\ldots+(6-5.9)^2}}$$

$$= \frac{14.21+11.31-3.99+2.31+7.41-1.89+0.21+6.51+8.61+0.21}{\sqrt{(-2.9)^2+(-2.9)^2+\ldots+(2.1)^2}\sqrt{(-4.9)^2+(-3.9)^2+\ldots+(0.1)^2}} = \frac{44.9}{56.66} = 0.79244617$$

r_u verifica $0 \le r_u \le 1$. Por tanto, cuando r_u es próximo a 1 existe una clara estructura jerárquica entre los objetos de Ω, y cuando r_u es pequeño indica una distorción notable al aplicar el algoritmo. Jardine y Sibson propusieron otra medida de distorción , la cual se define como :

$$\lambda_\alpha = \frac{\left(\sum_{ij} \left| d_{ij} - u_{ij} \right|^{\frac{1}{\alpha}} \right)^\alpha}{\left(\sum_{ij} d_{ij} \right)^\alpha}.$$

Dando diferentes valores a α, se tiene una medida de ajuste entre d_{ij} y u_{ij} que satisface : $0 \le \lambda_\alpha \le 1$. En este caso una buena clasificación ocurre cuando λ_α es próxima a *cero*.

3.-METODOLOGIA.

3.1.-LA MUESTRA.

Se utilizaron datos proporcionados por INEGI (Instituto Nacional de Estadística Geografía e Informática), el cual cuenta con información semidetallada de las unidades principales que se encuentran en el país. De dichos datos, se seleccionaron ciertas características como son: materia orgánica (M.O%), pH, arcilla (%), capacidad de intercambio catiónico (CIC%), precipitación Y temperatura, para las siguientes localidades: Morelia, Pachuca, Cuernavaca, Ciudad Altamirano, Ciudad de México y Veracruz. Estas propiedades se consideran que varían de acuerdo a las condiciones climáticas y, por lo mismo, satisfacen muy bien los requisitos para ser considerados en la modelación.

Primero se trabajó con el software de Sistemas de Información Geográfica Arc View este programa, permitió obtener la información de las propiedades del suelo por medio de mapas de climas y cartas edafológicas, luego se construyó una tabla con esa información en Excel en donde se organizaron los datos de acuerdo a las localidades muestreadas, obteniendo un total de 219 datos.

Posteriormente se graficaron en 2 dimensiones (en Excel) con las variables precipitación y temperatura para las siguientes unidades: Andosol, Vertisol, Cambisol, Leptosol, Luvisol, Phaeozem, Regosol y Acrisol. También se realizaron las gráficas de la relación temperatura vs. precipitación haciendo las combinaciones Andosol+Acrisol, Andosol+Vertisol, Luvisol+Vertisol, Vertisol+Phaeozem+Andosol, Vertisol+Phaeozem, y la gráfica que contiene a todas las unidades de suelos, con la finalidad de observar la distribución de cada uno de los suelos.

3.2.-REGRESION LOGISTICA.

El tamaño de la muestra que se consideró en esta etapa fue de 219 datos, los cuales fueron proporcionados por el Instituto Nacional de Estadística, Geografía e Informática (INEGI), estos datos tienen información acerca de la precipitación, temperatura y el tipo de suelo. Respecto al tipo de suelo se consideraron 9 unidades edáficas, correspondientes a los estados de Pachuca, Morelia, Morelos, Ciudad de México, Ciudad Altamirano y Veracruz.

El tipo de suelo es una variable de tipo categórica, por tanto, se tuvo que dar un valor a cada unidad de suelo, el cual se asignó de manera arbitraria, de acuerdo a la tabla 8.

UNIDADES DE SUELO		
Andosol = 1	**Litosol = 4**	**Phaeozem = 7**
Luvisol = 2	**Cambisol = 5**	**Planosol = 8**
Acrisol = 3	**Regosol = 6**	**Vertisol = 9**

Tabla 8.- Muestra el orden en que fueron asignados los valores a la variable tipo de suelo.

La regresión logística se usa con frecuencia para modelar la probabilidad de que una unidad experimental caiga en un grupo particular, en función a la información medida en la propia unidad. Estos modelos se pueden usar con fines de discriminación.

Luego, se utilizó el software Stata 7.0; en donde se introdujeron las variables independientes que son: precipitación y temperatura y la variable respuesta o también conocida como variable dependiente, en este caso, el tipo de suelo, y se aplicó el **modelo logístico**.

El software toma como categoría de referencia a la especie o en este caso a la unidad de suelo con la mayor cantidad de datos, que en este caso es el Andosol, aunque se puede elegir cualquier otra unidad de suelo como categoría de referencia.

Esta regresión indica la probabilidad de ocurrencia de cada unidad de suelo de acuerdo a las variables independientes que se proponen.

3.3.-APLICACIÓN DEL ANALISIS DE CONGLOMERADOS.

En esta segunda etapa se analizó la base de datos proporcionada por INEGI, con un tamaño de muestra de 16,062 datos. Se tomaron los estados del centro de México, donde el material parental es volcánico; dichos estados son los siguientes: Pachuca, Morelia, Morelos, Ciudad de México, Ciudad Altamirano y Veracruz, sin embargo, el tamaño de la muestra disminuyo a 219 datos (ver tabla 1 en Anexos), pues algunos de los datos de la muestra total pertenecen a otros estados que no se están considerando en este trabajo, ya que no pertenecen a la zona del Eje Neovolcánico,

también dentro de los datos que se están considerando hubo datos que se depuraron ya que en algunos de los casos la información que contenían era errónea y en otros la información no estaba disponible y para no afectar el análisis que se realizó posteriormente se decidió no considerarlos.

Posteriormente se graficaron las unidades de suelo, con sus valores de precipitación y temperatura, para ver su distribución espacial, después se agregaron las variables siguientes: capacidad de intercambio cationico (CIC%), arcilla(%), materia orgánica (M.O%) y pH, y también se hicieron las gráficas en tres dimensiones tomando como variables independientes a la temperatura y la precipitación y como dependientes a las propiedades del suelo (ver figura 2 en anexos), con el objetivo de evaluar sus tendencias.

A los datos anteriores se les realizó un Análisis de Regresión Lineal en el paquete estadístico Stata para ver si los datos se comportaban de manera lineal.

El siguiente paso fue aplicar el análisis de conglomerados o de "*Cluster*", para ello se utilizo el paquete estadístico Statistica 6.0, y también se agregó una nueva variable al análisis (profundidad).

Lo que se pretende con este nuevo análisis es determinar la influencia del clima en la formación de diferentes unidades de suelos. En el caso de este análisis a cada unidad se le asignó un valor del 1 al 8, ordenándolas de acuerdo a su grado de desarrollo: menor-Leptosol, mayor-Acrisol (ver tabla 9).

1.- Leptosol	5.- Phaeozem
2.- Regosol	6.- Vertisol
3.- Cambisol	7.- Luvisol
4.- Andosol	8.- Acrisol

Tabla 9.- Este orden sigue un criterio teórico, ya que se tomó en cuenta del suelo de menor desarrollo (Leptosol) al de mayor desarrollo (Acrisol).

En el Análisis de Conglomerados se consideraron 106 datos, pues al agregar las variables tipo de suelo y profundidad, algunos datos no corresponden a las unidades de suelo mencionadas anteriormente, mientras que había datos que no tenían la

información de la profundidad, por tanto no se consideraron en el análisis de conglomerados. También se utilizaron las variables precipitación, temperatura, arcilla (%), pH, capacidad de intercambio catiónico (CIC %) y materia orgánica (M.O %).

Se realizó un análisis de conglomerado, tomando el método de la liga simple y el método de la liga completa para las variables precipitación, temperatura, arcilla (%), pH, capacidad de intercambio catiónico (CIC %), profundidad y materia orgánica (M.O %).

Como se sabe en el análisis de conglomerados se utiliza la información de una serie de variables para cada sujeto u objeto y conforme a ellas mide la similitud. Una vez medida la similitud se integran en grupos homogéneos internamente y diferentes entre sí.

Posteriormente se aplicó el método de liga simple, el cual consiste en unir los grupos considerando la menor de las distancias existentes entre los miembros más cercanos de distintos grupos. Este método se aplicó de tres maneras diferentes, en la primera se consideró a la temperatura con las variables siguientes: capacidad de intercambio catiónico (CIC%), arcilla (%), materia orgánica (M.O%), pH, tipo de suelo y profundidad (ver figura 5 en resultados), en la segunda se tomó a la precipitación y a las propiedades del suelo (ver figura 6 en resultados), sin considerar a la temperatura y por último en la tercera se incluyo la precipitación, la temperatura y las demás variables (ver figura 7 en resultados).

En el software Statistica 6.0 se modificó la variable tipo de suelo, como es una variable de tipo categórica quise saber que cambios existen en los conglomerados al asignarle un valor numérico mayor, tal y como sigue:

100.- Leptosol	500.- Phaeozem
200.- Regosol	600.- Vertisol
300.- Cambisol	700.- Luvisol
400.- Andosol	800.- Acrisol

Tabla 10.- Recordemos que este orden sigue un criterio teórico, ya que se tomó en cuenta del suelo de menor desarrollo (Leptosol) al de mayor desarrollo (Acrisol), pero ahora considerando a la variable tipo de suelo de 100 en 100.

3.4.-RESULTADOS Y DISCUSIÓN.

3.4.1.-MODELO LOGISTICO.

El modelo resultante fue el siguiente:

$$P(y=2)=\frac{e^{x\beta(2)}}{1+e^{x\beta(2)}+e^{x\beta(3)}+...+e^{x\beta(9)}}$$

donde:

$$e^{x\beta(2)}=e^{-0.125-0.002\times precipitación+0.1007\times temperatura}=0.9731$$
$$e^{x\beta(3)}=e^{-0.507-0.001\times precipitación+0.3705\times temperatura}=0.8714$$
$$e^{x\beta(4)}=e^{-0.096-0.0071\times precipitación+0.4188\times temperatura}=1.3701$$
$$e^{x\beta(5)}=e^{-6.0797-0.0039\times precipitación+0.585\times temperatura}=0.0040$$
$$e^{x\beta(6)}=e^{-1.9436-0.0056\times precipitación+0.4795\times temperatura}=0.2299$$
$$e^{x\beta(7)}=e^{2.2061-0.0089\times precipitación+0.4545\times temperatura}=14.1781$$
$$e^{x\beta(8)}=e^{1.4732-0.0071\times precipitación+0.2893\times temperatura}=5.7857$$
$$e^{x\beta(9)}=e^{1.5256-0.0087\times precipitación+0.4577\times temperatura}=7.2037$$
$$1+e^{x\beta(2)}+e^{x\beta(3)}+...+e^{x\beta(9)}=1+\sum_{j=2}^{9}e^{x\beta(j)}=31.6164$$

y la precipitación y la temperatura igualadas a 1.

Con el análisis realizado anteriormente, este modelo tiene la posibilidad de manifestar cada conducta según cambien las variables independientes, en este caso la unidad de suelo que tiene la mayor probabilidad de ocurrencia es el Phaeozem, ya que tiene una probabilidad de 0.4484, la cual es mayor con respecto a las otras unidades de suelo (ver tabla 11).

	PREC	1			
	TEMP	1			
e^xß(2)=	0.973166589			P(Y=2)=	0.030780403
e^xß(3)=	0.871438486			P(Y=3)=	0.027562832
e^xß(4)=	1.370122292			P(Y=4)=	0.043335762
e^xß(5)=	0.004092497			P(Y=5)=	0.000129442
e^xß(6)=	0.229994473			P(Y=6)=	0.007274523
e^xß(7)=	14.17812098	Phaeozem	→	P(Y=7)=	0.44844149
e^xß(8)=	5.785761584			P(Y=8)=	0.182998547
e^xß(9)=	7.203737582			P(Y=9)=	0.22784788
	30.61643448				
1-> ANDOSOL					
2-> LUVISOL					
3-> ACRISOL					
4-> LITOSOL					
5-> CAMBISOL					
6-> REGOSOL					
7-> FEOZEM					
8-> PLANOSOL					
9-> VERTISOL					

Tabla 11.- Datos obtenidos del modelo logístico para las siguientes unidades de suelo (1.Andosol, 2.Luvisol, 3.Acrisol, 4.Litosol, 5.Cambisol, 6.Regosol, 7.Phaeozem, 8.Planosol y 9.Vertisol), y considerando a la precipitación y la temperatura con valor de 1.

Como se observa el resultado de este análisis arrojó resultados significativos, pero este método no cumple con el objetivo, el cual consiste en establecer un modelo que relacione unidades de suelo y sus propiedades con el clima.

3.4.2.-ANALISIS DE REGRESION.

En el análisis de regresión lineal, al obtener los resultados se observó que ninguno de ellos mostraban tendencias lineales. Casi todos los casos se podían ajustar a curvas polinomiales dificultando su aplicación práctica (ver tabla 12).

. regresión arcilla_ precipit temper

fuente	SS	df	MS		
				número de obs. =	219
				F(2, 216) =	2.18
modelo	1452.10064	2	726.050318	Prob > F	= 0.1152
residuo	71839.0866	216	332.588364	R-cuadrada	= 0.0198
				Adj R- cuadrada	= 0.0107
Total	73291.1872	218	336.198106	raíz MSE	= 18.237

arcilla	coeficiente	Std. Err.	t	P>|t|	[95%	intervalo de confianza]
precipit |	-.0027007	.0039051	-0.69	0.490	-.0103977	.0049963
temperat |	.5857858	.2973157	1.97	0.050	-.0002256	1.171797
_cons |	11.3829	6.225968	1.83	0.069	-.8885273	23.65433

regresión cic_ precipit temperat

fuente	SS	df	MS		
				número de obs. =	219
				F(2, 216) =	1.51
modelo	652.849662	2	326.424831	Prob > F	= 0.2232
residuo	46686.7174	216	216.14221	R-cuadrada	= 0.0138
				Adj R- cuadrada	= 0.0047
Total	47339.567	218	217.153977	raíz MSE	= 14.702

cic	coeficiente	Std. Err.	t	P>|t|	[95%	intervalo de confianza]
precipit	.0050925	.0031481	1.62	0.107	-.0011124	.0112975
temperat	.1531191	.2396812	0.64	0.524	-.3192945	.6255326
_cons	7.702348	5.019069	1.53	0.126	-2.190274	17.59497

regresión ph precipit temperat

fuente	SS	df	MS	número de obs. =	219
				F(2, 216) =	3.76
modelo	79.6743712	2	39.8371856	Prob > F =	0.0249
residuo	2290.78188	216	10.6054717	R-cuadrada =	0.0336
				Adj R- cuadrada =	0.0247
Total	2370.45625	218	10.8736525	raíz MSE =	3.2566

ph	coeficiente	Std. Err.	t	P>|t|	[95%	intervalo de confianza]
precipit	.0015543	.0006973	2.23	0.027	.0001799	.0029288
temperat	-.0844339	.053092	-1.59	0.113	-.1890787	.0202109
_cons	1.667412	1.111779	1.50	0.135	-.5239122	3.858736

regresión mo precipit temperat

fuente	SS	df	MS	número de obs. =	219
				F(2, 216) =	2.20
modelo	40.4037842	2	20.2018921	Prob > F =	0.1132
residuo	1982.54728	216	9.1784596	R-cuadrada =	0.0200
				Adj R- cuadrada =	0.0109
Total	2022.95107	218	9.27959206	raíz MSE =	3.0296

mo	coeficiente	Std. Err.	t	P>|t|	[95%	intervalo de confianza]
precipit	-.0007059	.0006487	-1.09	0.278	-.0019845	.0005728
temperat	-.0887186	.0493912	-1.80	0.074	-.186069	.0086317
_cons	6.124951	1.03428	5.92	0.000	4.086377	8.163525

Tabla 12.- Análisis de Regresión aplicado en Stata 7.0 (tomando a la precipitación y temperatura como variables independientes y a las propiedades del suelo como dependientes), donde se puede observar que no existen tendencias lineales, pues el valor de r-cuadrada es muy pequeño.

3.4.3.-ANALISIS DE CONGLOMERADOS.

Con respecto a los conglomerados se observó que no hay ningún cambio en los dendogramas, además se ve claramente que la variable que más influye sobre todo el conjunto de variables es la precipitación (ver figuras 3 y 4).

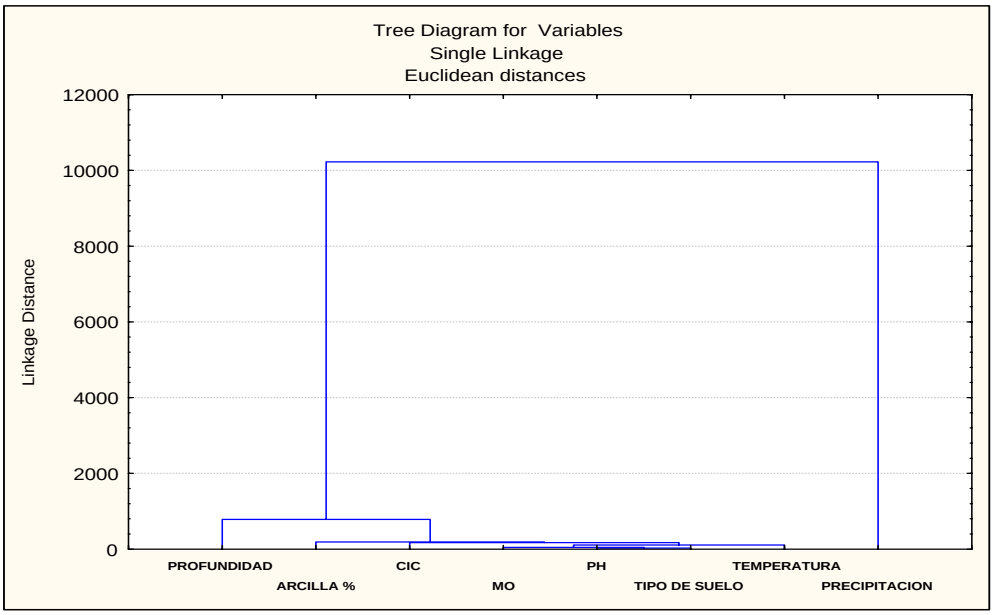


Figura 3.- Método de liga simple, aplicado a las variables (profundidad, arcilla %, cic %, pH, materia orgánica, tipo de suelo, temperatura y precipitación), donde se muestra que la variable precipitación está influenciando a las demás variables.

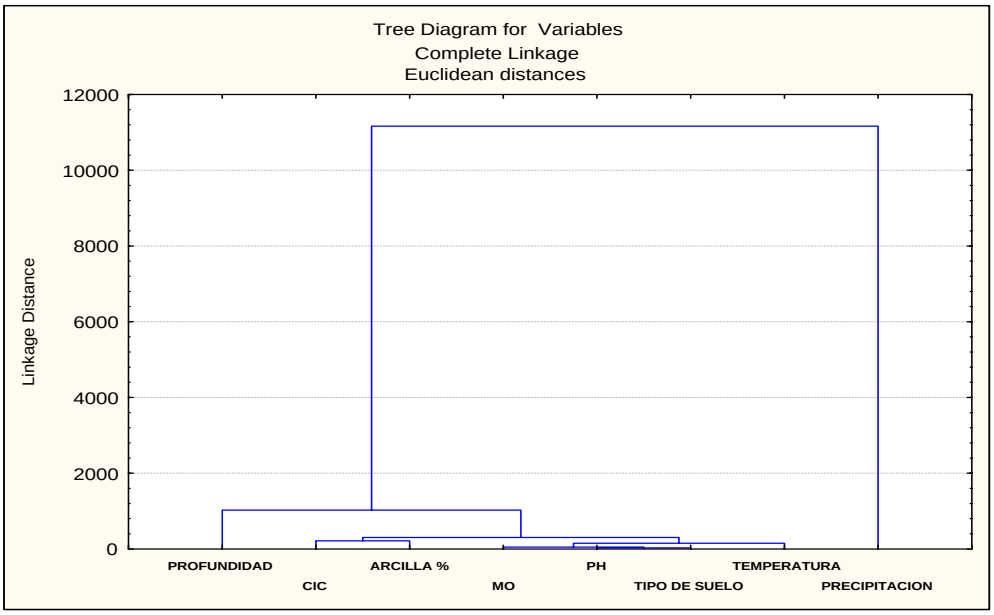


Figura 4.- Método de liga completa, aplicado a las variables (profundidad, arcilla %, cic %, pH, materia orgánica, tipo de suelo, temperatura y precipitación), donde se muestra que la variable precipitación está influenciando a las demás variables.

En el caso del dendograma de la temperatura se decidió tomar en cuenta únicamente seis conglomerados (ver figura 5), ya que se depuraron los grupos atípicos, es decir, interesan los elementos como miembros de grupos, no la excesiva *"individualidad"*. Se aprecia que en tres de los seis conglomerados hubo una gran heterogeneidad de suelos, siendo muy difícil identificar si existe un cambio evolutivo de unidades. En dos de los tres conglomerados restantes se agrupaban Phaeozem + Regosoles y Cambisol + Vertisol, en el último se agrupa a los Andosoles, siendo el grupo más homogeneo.

El dendograma correspondiente a la precipitación y el de precipitación y temperatura (ver figuras 6 y 7) mostraron resultados muy similares, lo que indica que en el agrupamiento tiene un mayor peso la precipitación que la temperatura.

Desde el punto de vista de la Ciencia del Suelo, este resultado es muy valioso, ya que demuestra claramente la influencia de la humedad en la formación y diferenciación de unidades.

Se puede observar que en ambos dendogramas se obtiene el mismo número de conglomerados, que en este caso es de siete (para el dendograma de precipitación y el dendograma donde se considera a la precipitación y temperatura), aquí se puede ver que existe una mejor agrupación de los suelos, es decir, que puede estar ocurriendo un cambio evolutivo, una involución de suelos o hay condiciones de sitio.

Respecto a la formación de conglomerados se puede apreciar con claridad que cuatro de los siete conglomerados agrupan a los suelos de una manera bastante interesante. Por ejemplo el conglomerado que agrupa Vertisoles y Luvisoles. Se sabe que ambas unidades se forman en condiciones climáticas diferentes. Los Vertisoles requieren un clima con una época seca larga (de más de 9 meses), mientras que los Luvisoles se forman en climas más húmedos. ¿Por qué entonces aparecen juntos?, esto puede estar relacionado a la dinámica paleoambiental, pues en los estudios realizados en paleosuelos del centro de México, la ocurrencia de Luvisoles antiguos es común (Sedov et al. 2001; Solleiro et al. 2003), es decir, esta unidad representa un paleosuelo, mientras que el Vertisol corresponde al suelo moderno que se está formando actualmente.

Existen dos grupos que asocian Regosol + Cambisol + Luvisol + Andosol + Phaeozem + Vertisol, la heterogeneidad puede explicarse a las condiciones locales, ya que se debe recordar que el clima no es el único factor formador. También influyen el relieve (quizá la pendiente), los organismos, el tiempo (edad de las rocas) y el material parental.

Existen otros dos conglomerados en los cuales hay una agrupación de Andosoles + Acrisoles + Cambisoles + Regosoles. Lo raro en este grupo es la formación de Acrisoles ya que se trata de suelos muy evolucionados a comparación de las otras unidades que son más jóvenes, por tanto sería apresurado decir si se está presentando una transición de un suelo a otro o simplemente son condiciones de sitio y para considerar esta posibilidad es necesario tomar otro tipo de variables en nuestro análisis. Probablemente, como en el caso de Vertisoles + Luvisoles, los Acrisoles representan paleosuelos.

El método de liga simple es un buen método, sin duda hay algunas agrupaciones bastante interesantes, que como resultado arroja, en ciertos casos, una transición evolutiva de algunas unidades de suelos. Al aplicar el método de liga completa se obtuvo una mejor agrupación de las unidades de suelos.

El método de liga completa se aplicó únicamente a la precipitación con todas las variables y a la precipitación, temperatura con todas las demás variables (ver figuras 8 y 9). No se consideró a la temperatura vs. variables, como en el caso de liga simple, ya que se concluyó que la variable que aporta mayor información es la precipitación (ver figuras 3 y 4), por lo tanto el dendograma de la temperatura no es relevante.

Ambos dendogramas resultantes son idénticos, es decir, no hubo cambios en la formación de los conglomerados. En estos dendogramas se analizaron ocho grupos, en uno de los grupos se formaron exclusivamente Andosoles, lo cual indica que es un grupo bastante homogéneo, y puede indicar sitios con características pedogenéticas similares.

Hay otros cinco grupos, en los cuales se presentan las siguientes formaciones: 1) Regosol + Cambisol + Andosol, 2) Leptosol + Regosol + Cambisol + Phaeozem, 3) Cambisol + Phaeozem + Vertisol, 4) Regosol + Andosol + Phaeozem + Luvisol y 5) Regosol + Cambisol + Andosol + Phaeozem + Vertisol + Luvisol + Acrisol.

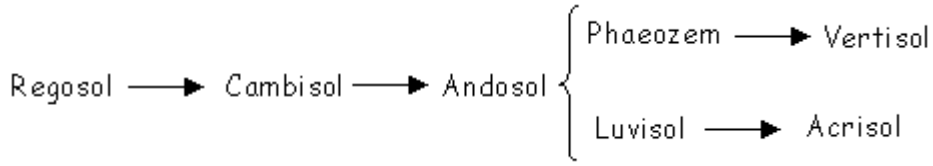
En el primer grupo (Regosol + Cambisol + Andosol) se aprecia la formación de una transición evolutiva de suelos, pues los Regosoles son suelos que se desarrollan sobre materiales no consolidados, alterados y que se encuentran en cualquier zona climática, pueden estar evolucionando hacia Cambisoles los cuales se desarrollan sobre materiales de alteración procedentes de un amplio abanico de rocas u horizontes y también se encuentran en cualquier tipo de clima y vegetación, y a su vez este tipo de suelos puede evolucionar hacia Andosoles.

En segundo grupo (Leptosol + Regosol + Cambisol + Phaeozem) se observa la formación de Leptosoles, Regosoles y Cambisoles, estos suelos son poco desarrollados. También se tiene el Phaeozem que es un suelo más desarrollado que los anteriores, aquí pueden estar ocurriendo dos cosas interesantes, la primera es una transición evolutiva de suelos o la segunda una degradación de los mismos; En el primer caso, la evolución se daría de la siguiente manera: Leptosoles ⟶ Regosoles ⟶ Cambisoles ⟶ Phaeozem, la cual ocurre en regiones de climas húmedos pero donde haya mayor evapotranspiración que precipitación. La explicación de la segunda idea es que los Phaeozem son suelos fértiles, extensamente cultivados. Esta cultivación causa degradación y como forma principal de este tipo, erosión, modificando la cubierta edáfica y dejando suelos que ya no clasifican como Phaeozem, sino como Regosoles o Leptosoles, e inclusive Cambisoles.

En el tercer grupo (Cambisoles + Phaeozem + Vertisol) se aprecia una transición evolutiva de suelos, ya que los Cambisoles son suelos que pueden evolucionar hacia Phaeozem el cual puede evolucionar o transformarse a un Vertisol, ya que ambos suelos se forman en condiciones ambientales similares.

En el cuarto grupo (Regosol + Andosol + Phaeozem + Luvisol), la evolución de Regosoles ⟶ Andosoles es común. Los Andosoles son típicos suelos derivados de materiales volcánicos formados en climas húmedos, no son suelos estables, sino transicionales (Fernández et al. 1987). Cualquier variación en las condiciones de sitio, modifica sus propiedades y cambia a otra unidad. Entonces, en climas húmedos, donde la precipitación sea menor que la evapotranspiración, se tendrán Phaeozem, mientras que en condiciones de mayor humedad (precipitación > evapotranspiración) se formarán Luvisoles. Esta transición de Andosoles a Luvisoles es común en el Eje Neovolcánico y está bien documentada (Sedov et al. 2003).

El quinto caso (Regosol + Cambisol + Andosol + Phaeozem + Vertisol + Luvisol + Acrisol) agrupa una mayor cantidad de unidades. Como ya se menciono anteriormente, los caminos evolutivos pueden ser:



La asociación de dos caminos climáticamente diferentes puede deberse a la presencia común de paleosuelos del tipo Luvisol y Acrisol asociados con los Phaeozem, Vertisoles y Andosoles.

Por último hubo dos conglomerados en los que se presentó la agrupación: de Regosoles + Cambisoles + Andosoles + Acrisoles, que también pueden representar caminos evolutivos de Regosol ⟶ Cambisol ⟶ Andosol ⟶ Acrisol o bien esta última unidad representa a un paleosuelo.

Al analizar el dendograma de la liga completa considerando el tipo de suelo de 100 en 100, se pudo apreciar una mejor agrupación de los suelos a comparación de cuando se habían categorizado, con valores de 1 al 8, esto significa que si el valor numérico de esta variable es cada vez mayor se presenta una mejor agrupación de las unidades de suelo, lo mismo sucedió con el dendograma de liga completa tomando en cuenta la variable tipo de suelo de 100 en 100 (ver figura 11).

Los grupos son más homogéneos (ver dendogramas 8 y 9) y el efecto del clima es más contundente, pues se tienen tres conglomerados con 1.- Andosoles, 2.- Phaeozem y 3.- Vertisoles, además de las siguientes asociaciones:

4.- Leptosol + Regosol + Cambisol, los cuales son suelos poco evolucionados.
5.- Phaeozem + Vertisol, son suelos formados bajo condiciones climáticas similares (precipitación < evapotranspiración).
6.- Luvisol + Acrisol, son suelos formados en climas húmedos y conectados evolutivamente (precipitación > evapotranspiración).
7.- Regosol + Cambisol + Andosol, se observa claramente la transición de suelos de poco a moderado desarrollo.

En la figura 11, también hay una muy buena agrupación de los suelos, de hecho se aprecia con una mejor claridad que existen secuencias evolutivas mucho más definidas considerando a las unidades de suelos de 100 en 100, por ejemplo, tenemos las siguientes formaciones: 1) leptosol + regosol + cambisol, 2) phaeozem + vertisol, 3) regosol + cambisol + andosol, 4) phaeozem + vertisol + luvisol, 5) phaeozem + vertisol + luvisol + acrisol y 6) acrisol, del 1) al 5) se trata de una secuencia evolutiva, mientras que 6) no proporciona información.

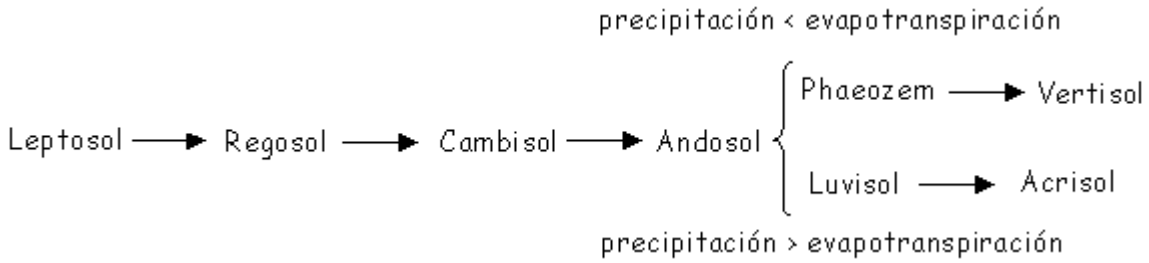
CONCLUSIONES.

El modelo de regresión logística se usa con frecuencia para modelar la probabilidad de que una unidad experimental caiga en un grupo particular, en función a la información referente en la propia unidad, sin embargo con este método los resultados no fueron los esperados por lo tanto hubo la necesidad de buscar otros caminos.

Al aplicar la regresión lineal a los datos proporcionados por INEGI, tampoco se obtuvieron resultados significativos ya que el análisis que arrojó el paquete estadístico STATA 7.0, daba valores muy pequeños para r-cuadrada debido a que las variables muestran tendencias tipo polinomial.

Cuando se aplicó el análisis de conglomerados se obtuvieron mejores resultados que con los métodos anteriormente mencionados. Se compararon dos métodos en esta etapa; 1) el método de liga simple y 2) el método de liga completa, con respecto al primer método se utilizó a la temperatura y todas las demás variables, sin tomar en cuenta a la precipitación y no hubo resultados satisfactorios. Los dendogramas de precipitación vs. variables y precipitación más temperatura vs. variables, fueron muy similares, concluyéndose que es la precipitación la que juega un papel más determinante en la formación de los conglomerados.

Al emplear el método de liga completa, se apreció una mejor agrupación de las unidades de suelo, permitiendo ver que existía una transición evolutiva de las unidades de suelo, Los resultados del agrupamiento coinciden con los aspectos teóricos de la génesis de suelos, en donde se tiene una secuencia evolutiva para suelos derivados de materiales volcánicos:



Al encontrar grupos que asocian a los extremos de la serie, se concluye que se está detectando la presencia de paleosuelos.

Ambos métodos resultan interesantes, el que tiene la escala de 1 al 8, permite ver la heterogeneidad, la cual debe ser explicada agregando otros factores formadores. El método con escalas mayores (con escala de 100 en 100) da una visión más particular.

El uso de esta primera aproximación puede ser útil para interpretaciones paleoclimáticas o bien hacer predicciones sobre el efecto del cambio climático en el suelo.

GLOSARIO

Adsorción. Proceso por el cual una capa de átomos o moléculas de una sustancia se incorpora a la superficie de otra (sólida o líquida) ejemplo: materia orgánica, arcilla. La capa adsorbida puede sostenerse por enlaces químicos que involucran cargas.

Eluviación e iluviación. Ambos procesos están muy relacionados. Eluviación es el proceso de remoción de constituyentes de un horizonte de suelo, capa o zona por solución o lavado, casi siempre con agua. Los horizontes donde la eluviación es dominante pueden ser referidos como horizontes E eluviales. La iluviación puede ser definida como el proceso que permite recibir o acumular materiales movidos por eluviación.

Horizontes. Materiales transformados por los procesos edafogénicos, dispuestos de manera horizontal o paralelamente a la superficie del suelo; los cuales se extienden continua o discontinuamente en la unidad edáfica. El conjunto de horizontes constituye el "perfil".

Horizonte eluvial. Horizonte del que ha sido removido material ya sea en solución o en suspensión.

Horizonte iluvial. Horizonte que recibe y acumula material en solución o suspención de alguna otra parte del suelo.

Perfil del suelo. Es un corte plano del suelo en dos dimensiones (largo y ancho) que se extiende verticalmente desde la superficie del suelo, de tal manera que se expongan todos los horizontes (o capas superpuestas) presentes en él, y parte del material relativamente inalterado.

Suelo poligenético. El que se ha formado por dos o más procesos contrastantes, de tal manera que todos sus horizontes no están relacionados, genéticamente, entre sí.

ANEXO A

TIPO DE SUELO	TEMPERATURA	PRECIPITACION		TIPO DE SUELO	TEMPERATURA	PRECIPITACION
PLANOSOL	15	693		ACRISOL	17	1660
PLANOSOL	15	730		ACRISOL	17	1662
PLANOSOL	15	732		ACRISOL	17	1661
PLANOSOL	15	930		ACRISOL	19	1400
PLANOSOL	15	930		ACRISOL	19	1401
PLANOSOL	12	950		ACRISOL	19	1401
PLANOSOL	14	980		ACRISOL	21	1150
PLANOSOL	14	981		ACRISOL	21	1151
VERTISOL	14	740		ACRISOL	18	1170
VERTISOL	14	741		ACRISOL	18	1170
VERTISOL	14	743		ACRISOL	18	1171
VERTISOL	14	745		ACRISOL	18	1188
VERTISOL	15	735		ACRISOL	18	1189
VERTISOL	14	790		ACRISOL	18	1188
VERTISOL	14	792		ACRISOL	12	960
VERTISOL	14	793		ACRISOL	13	960
VERTISOL	14	795		ACRISOL	13	961
VERTISOL	14	811		ACRISOL	13	960
VERTISOL	19	1164		LITOSOL	19	1100
VERTISOL	19	1165		LITOSOL	25	955
VERTISOL	19	1165		LITOSOL	25	956
VERTISOL	20	1120		LITOSOL	14	550
VERTISOL	21.81	650		LITOSOL	9	1480
VERTISOL	24	448		LITOSOL	14	695
VERTISOL	22	1650		LITOSOL	14	980
VERTISOL	24	199		LITOSOL	14	980
VERTISOL	14	900		LITOSOL	14	980
VERTISOL	14	650		LITOSOL	14	830
VERTISOL	14	751		LITOSOL	13	570

VERTISOL	14	751		LITOSOL	13	571
VERTISOL	16	711		CAMBISOL	21	999
VERTISOL	16	711		CAMBISOL	21	1510
VERTISOL	13	880		CAMBISOL	21	1511
VERTISOL	13	980		CAMBISOL	24	1193
VERTISOL	15	675		CAMBISOL	24	1190
VERTISOL	15	675		CAMBISOL	16	1300
VERTISOL	15	625		CAMBISOL	16	1300
VERTISOL	15	625		CAMBISOL	21	1310
LUVISOL	12	829		CAMBISOL	24	1000
LUVISOL	12	830		CAMBISOL	24	1000
LUVISOL	12	831		CAMBISOL	17	1420
LUVISOL	12	832		CAMBISOL	15	590
LUVISOL	16	1502		CAMBISOL	13	700
LUVISOL	15	1204		CAMBISOL	13	810
LUVISOL	15	1204		REGOSOL	22	1600
LUVISOL	15	1205		REGOSOL	22	1490
LUVISOL	15	1205		REGOSOL	23	1275
LUVISOL	14	1100		REGOSOL	26	1185
ANDOSOL	14	825		REGOSOL	22	1265
ANDOSOL	13	935		REGOSOL	18	1209
ANDOSOL	13	960		REGOSOL	24	445
ANDOSOL	13	962		REGOSOL	25	445
ANDOSOL	13	964		REGOSOL	12	831
ANDOSOL	13	980		REGOSOL	12	832
ANDOSOL	10	1015		REGOSOL	14	1105
ANDOSOL	13	1195		REGOSOL	14	1105
ANDOSOL	17	1195		REGOSOL	14	1106
ANDOSOL	17	1196		REGOSOL	14	880
ANDOSOL	17	1195		REGOSOL	12	1000
ANDOSOL	17	1196		REGOSOL	12	1001
ANDOSOL	16	1198		REGOSOL	12	1000
ANDOSOL	16	1198		REGOSOL	14	888

ANDOSOL	16	1199		REGOSOL	13	550
ANDOSOL	17	1300		REGOSOL	14	565
ANDOSOL	17	1301		FEOZEM	25	1196
ANDOSOL	17	1301		FEOZEM	23	1183
ANDOSOL	18	880		FEOZEM	25	1170
ANDOSOL	18	1000		FEOZEM	18	1300
ANDOSOL	18	1780		FEOZEM	22	990
ANDOSOL	18	1781		FEOZEM	23	545
ANDOSOL	18	1781		FEOZEM	23	546
ANDOSOL	18	1782		FEOZEM	23	544
ANDOSOL	12	860		FEOZEM	25	550
ANDOSOL	12	920		FEOZEM	25.07	550
ANDOSOL	13	1300		FEOZEM	24	999
ANDOSOL	10	1020		FEOZEM	24	501
ANDOSOL	6	1345		FEOZEM	13	850
ANDOSOL	6	1346		FEOZEM	14	990
ANDOSOL	6	1347		FEOZEM	14	750
ANDOSOL	6	1346		FEOZEM	12	1009
ANDOSOL	14	1020		FEOZEM	12	1009
ANDOSOL	14	1021		FEOZEM	12	1010
ANDOSOL	12	1950		FEOZEM	13	940
ANDOSOL	10	1205		FEOZEM	18.93	730
ANDOSOL	9	1230		FEOZEM	15	555
ANDOSOL	9	1220		FEOZEM	14	880
ANDOSOL	14	1260		FEOZEM	15	300
ANDOSOL	10	1400		FEOZEM	13	1103
ANDOSOL	10	1400		FEOZEM	13	1103
ANDOSOL	10	1401		FEOZEM	13	1102
ANDOSOL	10	1390		FEOZEM	15	640
ANDOSOL	10	1391		FEOZEM	13	605
ANDOSOL	10	1390		FEOZEM	13	605
ANDOSOL	10	1480		FEOZEM	14	635
ANDOSOL	10	1300		FEOZEM	13	640

ANDOSOL	11	1010		FEOZEM	14	675
ANDOSOL	15	1500		FEOZEM	14	805
ANDOSOL	15	1500		FEOZEM	15	780
ANDOSOL	14	1300		FEOZEM	14	830
ANDOSOL	14	900		FEOZEM	14	830
ANDOSOL	14	901		FEOZEM	13	685
ANDOSOL	14	902		FEOZEM	13	685
ACRISOL	15	1200		FEOZEM	12	700
ACRISOL	15	1500		FEOZEM	13	680
ACRISOL	15	1501		FEOZEM	13	670
ACRISOL	15	1501		FEOZEM	12	680
ACRISOL	19	1125		FEOZEM	12	680
ACRISOL	19	1126		FEOZEM	12	680
ACRISOL	19	1126		FEOZEM	15	890
				FEOZEM	14	665

Tabla 2.- Datos de los estados: Pachuca, Morelia, Morelos, Ciudad de México, Ciudad Altamirano y Veracruz.

ANEXO B

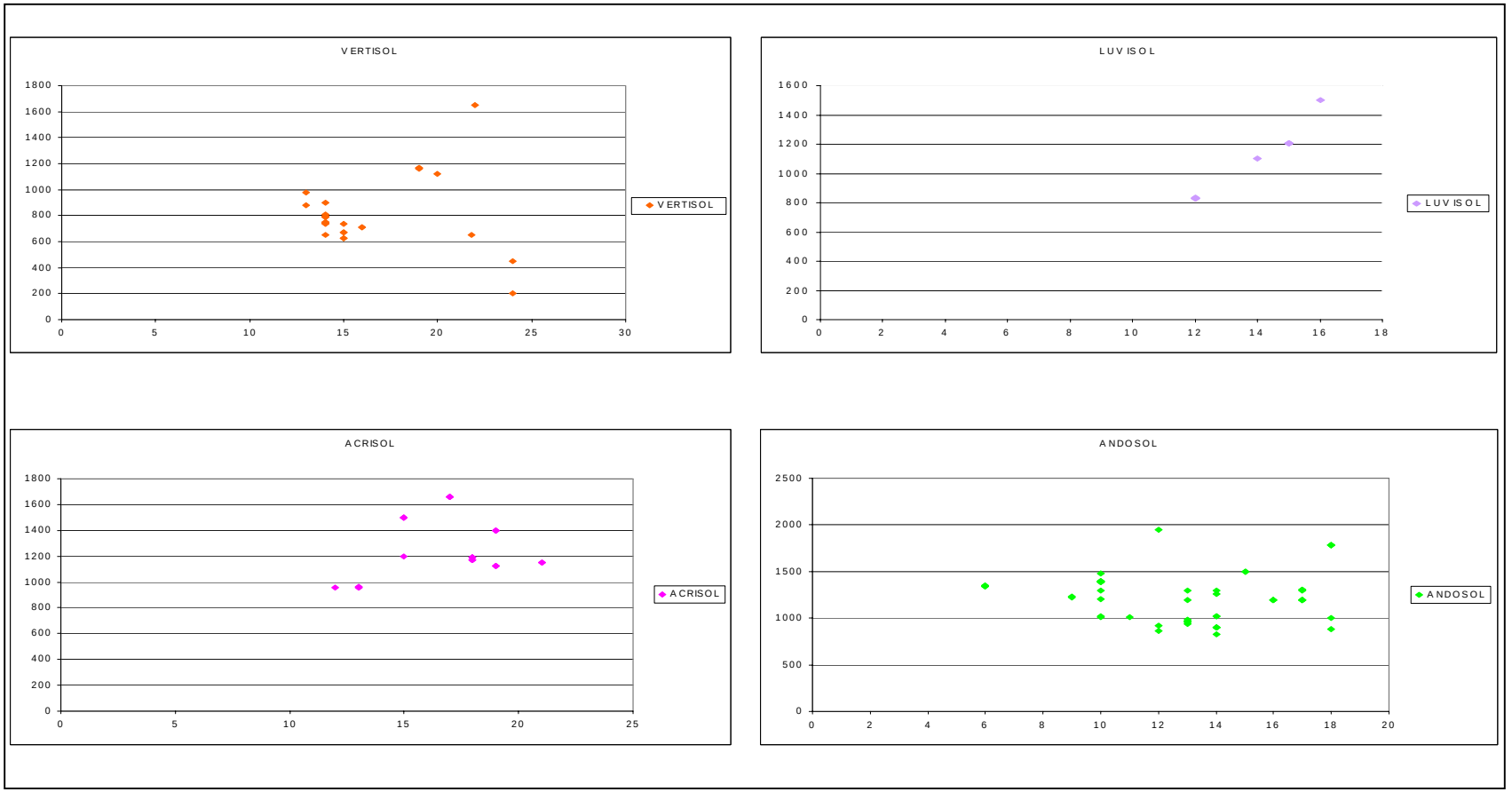
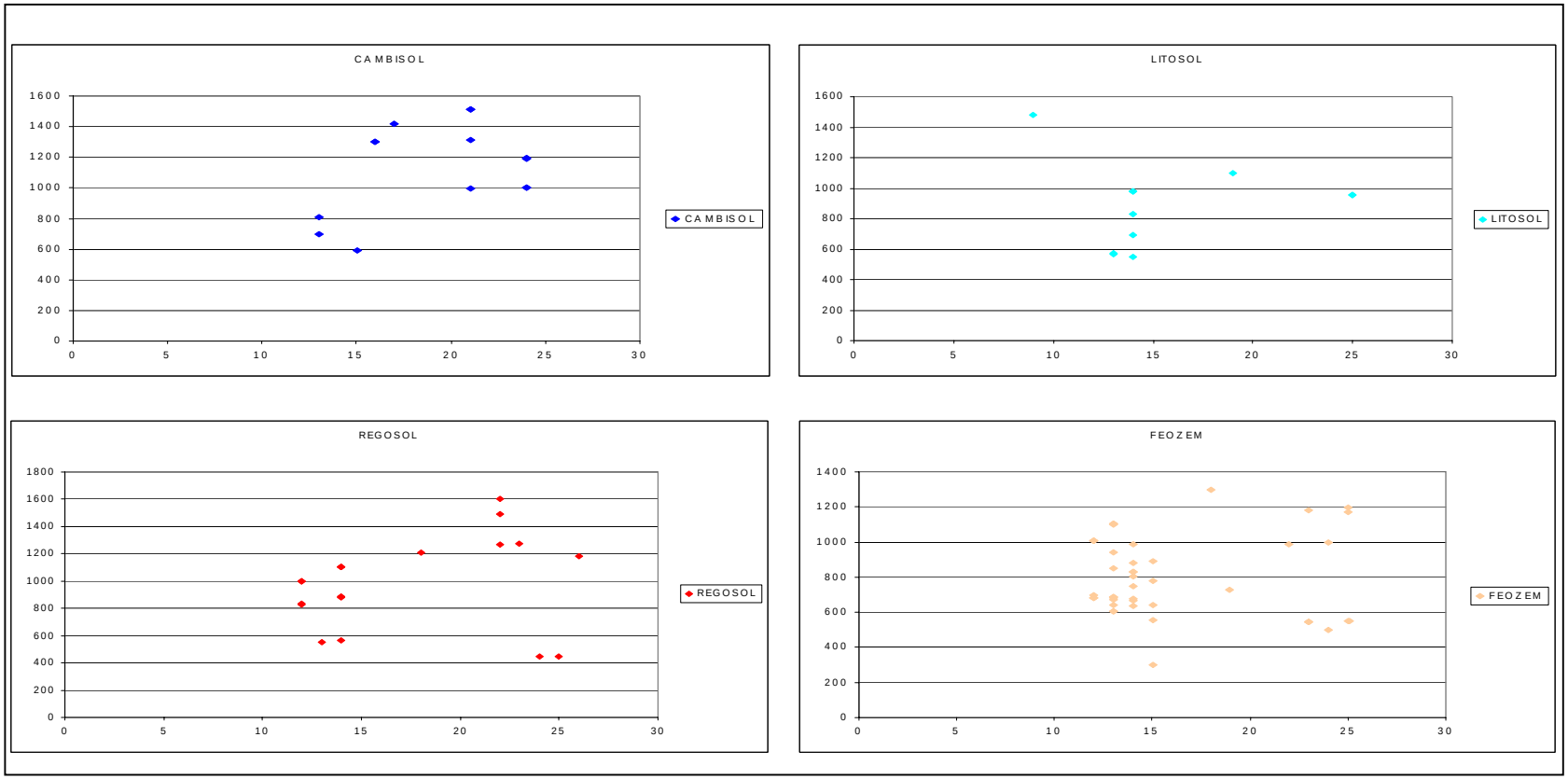
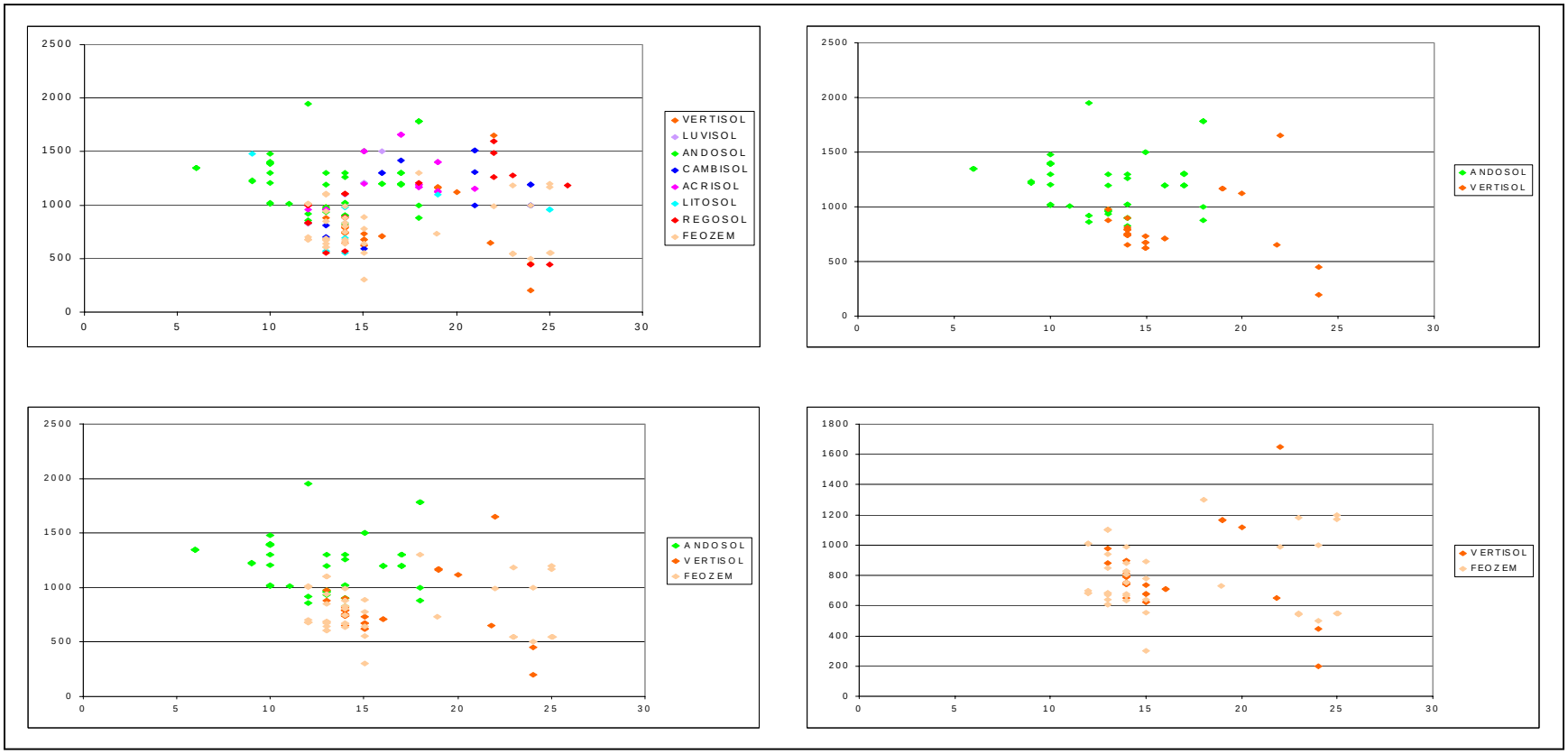
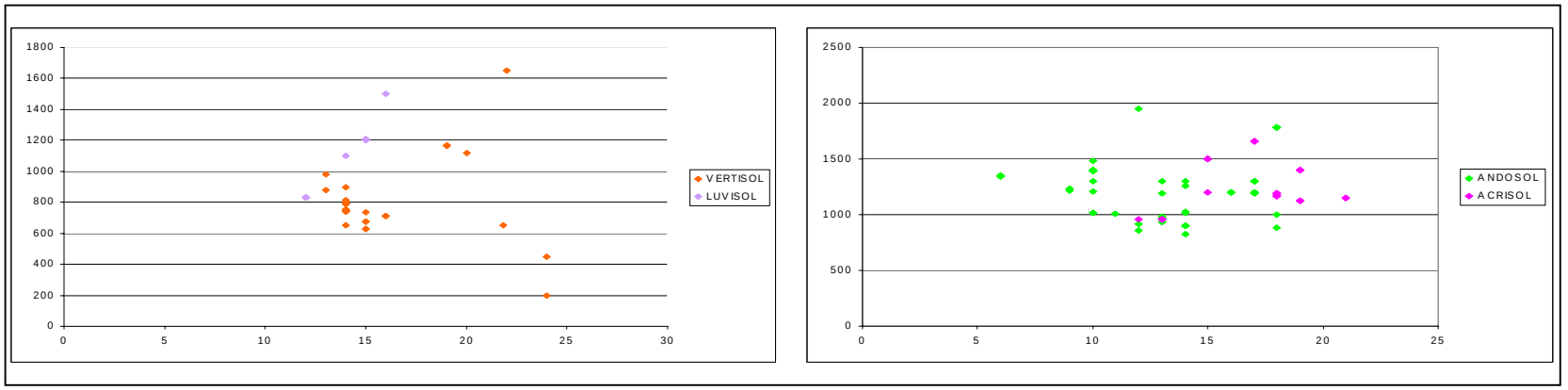


Figura 1. Gráficas de temperatura Vs precipitación correspondientes a Vertisol, Luvisol, Andosol, Acrisol, Litosol, Cambisol, Regosol, Phaeozem, Andosol + Vertisol, Andosol + Vertisol + Phaeozem, Vertisol + Phaeozem, Vertisol + Luvisol, Andosol + Acrisol y la gráfica con todo das las unidades de suelo.


CAMBISOL
1600
1400
1200
1000
800
600
400
200
0
0
5
10
15
20
25
30
CAMBISOL
LITOSOL
1600
1400
1200
1000
800
600
400
200
0
0
5
10
15
20
25
30
LITOSOL
REGOSOL
1800
1600
1400
1200
1000
800
600
400
200
0
0
5
10
15
20
25
30
REGOSOL
FEOZEM
1400
1200
1000
800
600
400
200
0
0
5
10
15
20
25
30
FEOZEM


2500
2000
1500
1000
500
0
0 5 10 15 20 25 30
VERTISOL
LUVISOL
ANDOSOL
CAMBISOL
ACRISOL
LITOSOL
REGOSOL
FEOZEM
ANDOSOL
VERTISOL
1800
1600
1400
1200
1000
800
600
400
200


1800
1600
1400
1200
1000
800
600
400
200
0
0
5
10
15
20
25
30
VERTISOL
LUVISOL
2500
2000
1500
1000
500
0
0
5
10
15
20
25
ANDOSOL
ACRISOL

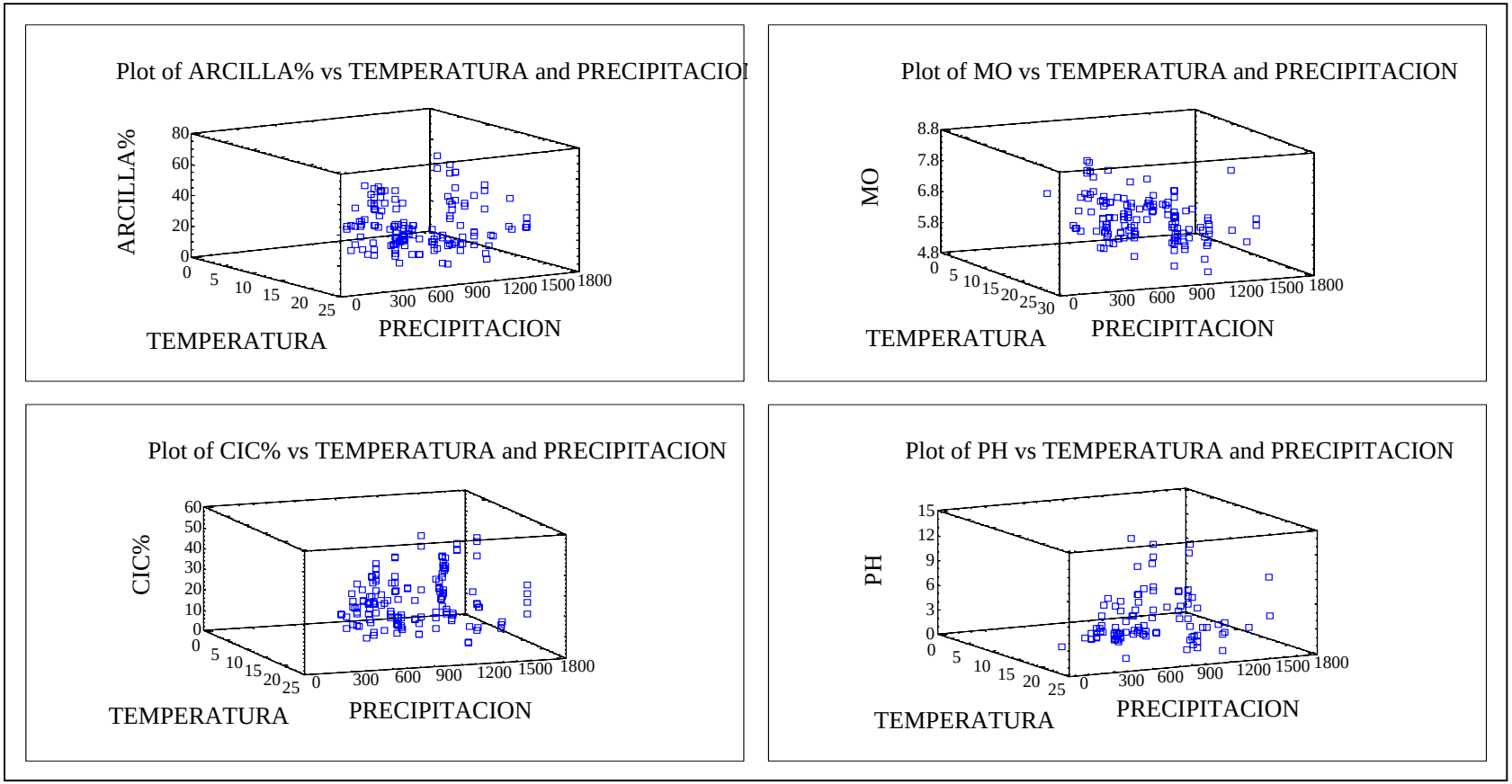


Figura 2.- Gráficas en tres dimensiones considerando a las variables temperatura y precipitación como variables independientes y como variables dependientes a las propiedades de los suelos, con el fin de evaluar sus tendencias.

ANEXO C

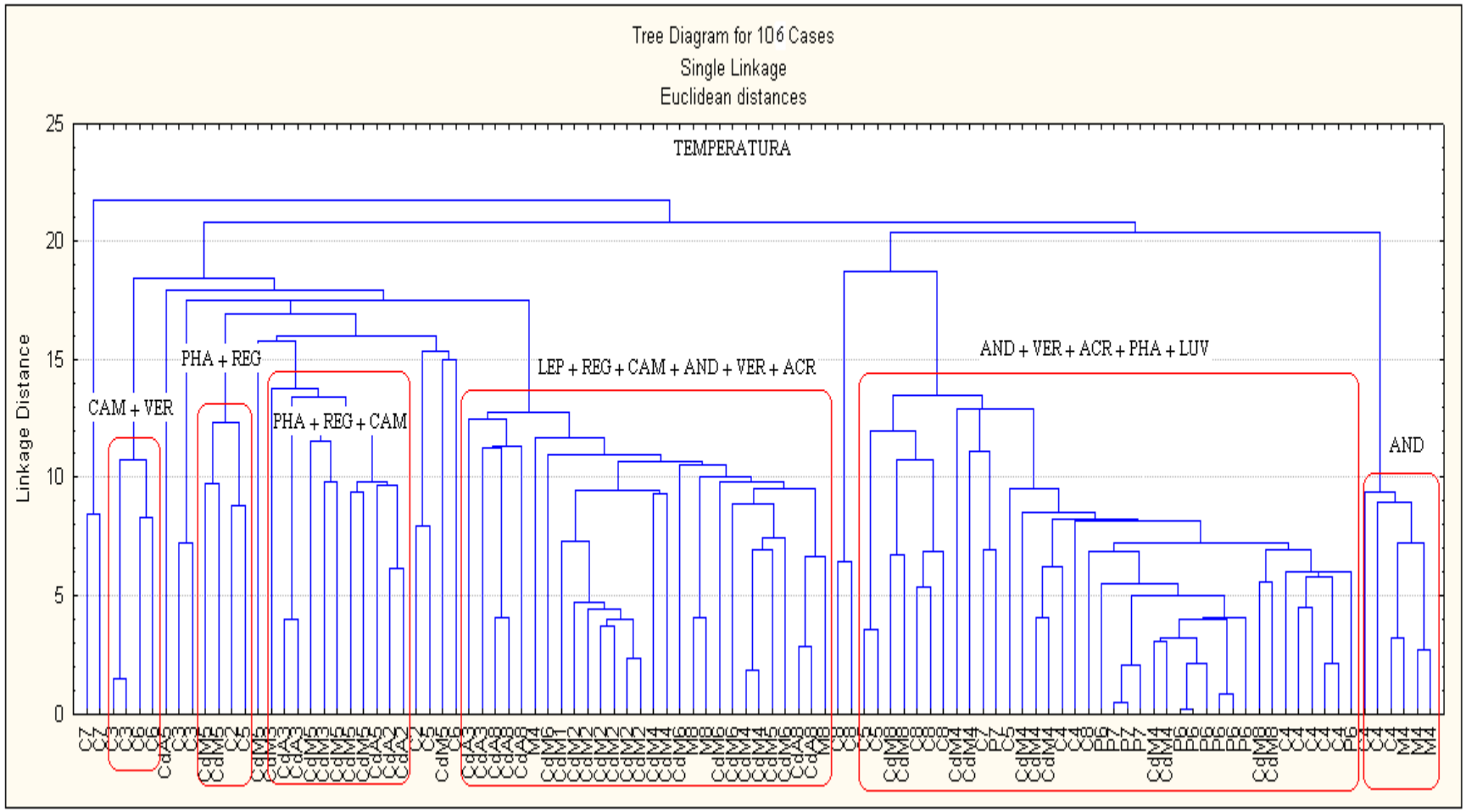


Figura 5.- Representación del dendograma del método de la liga simple, correspondiente a la temperatura y las propiedades del suelo, sin considerar a la precipitación y donde se aprecia la formación de 6 conglomerados.

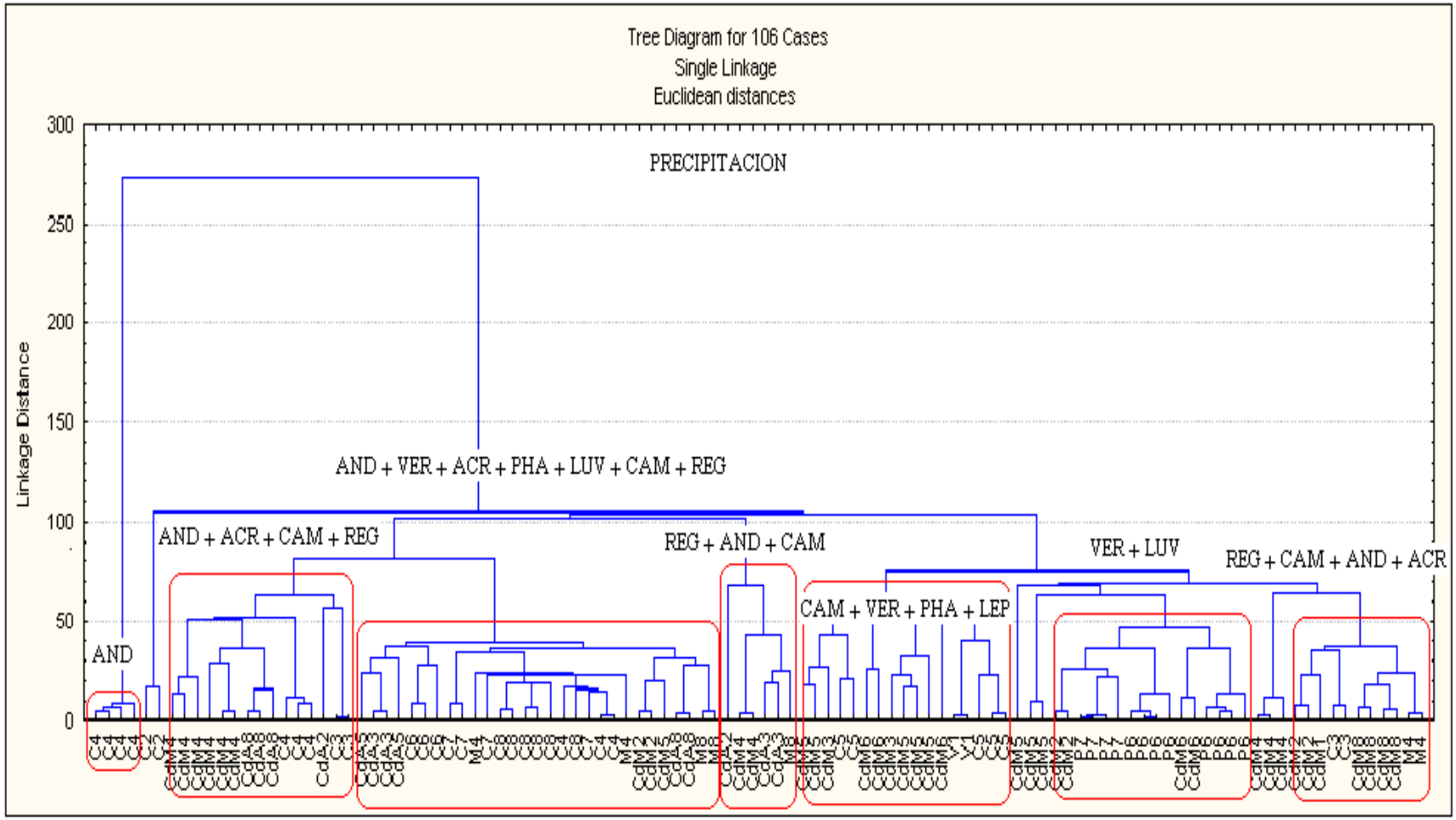


Figura 6.- Representación del dendograma del método de la liga simple, correspondiente a la precipitación y las propiedades del suelo, sin considerar a la temperatura y donde se aprecia la formación de 7 conglomerados.

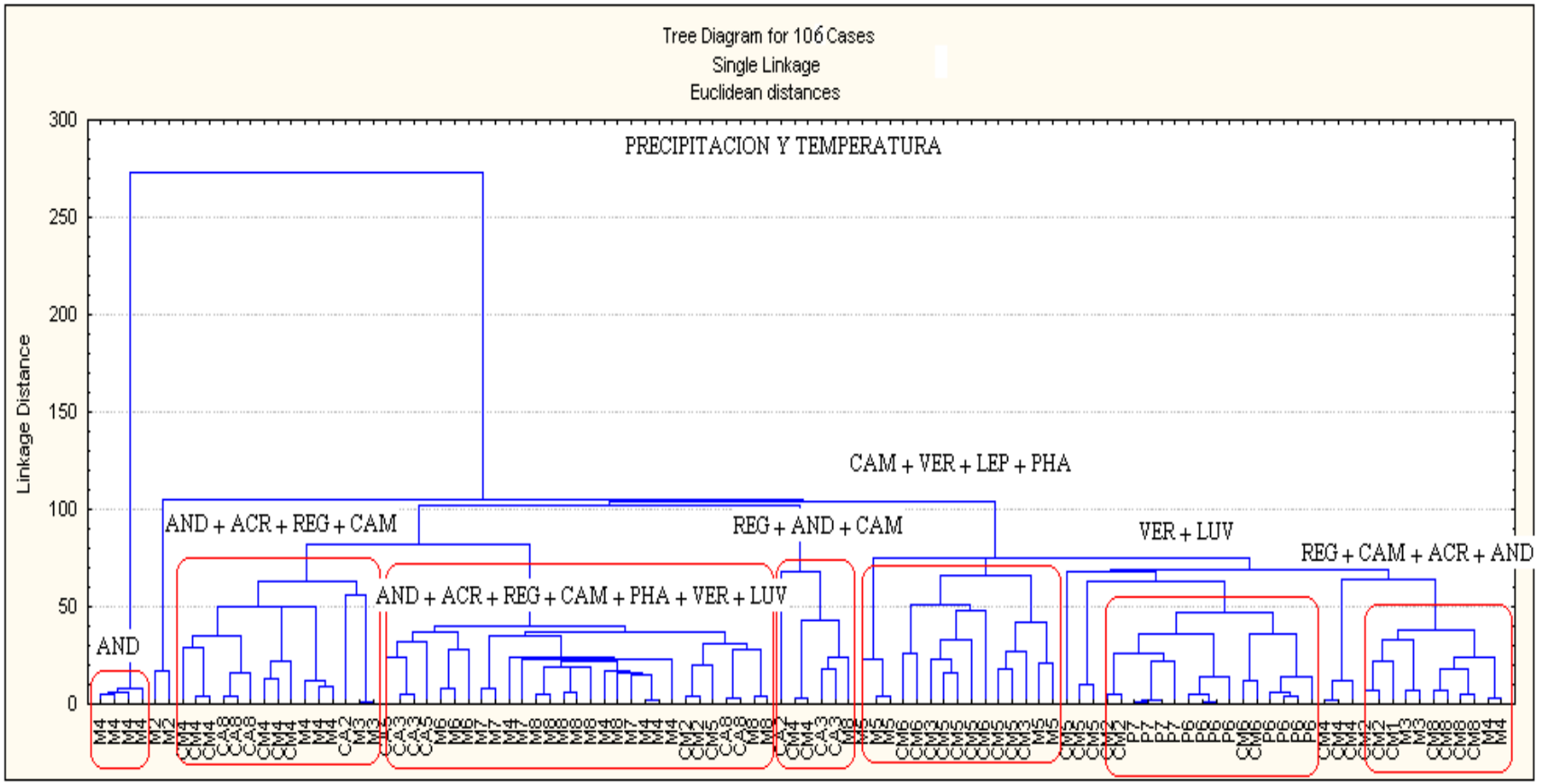


Figura 7.- Representación del dendograma del método de la liga simple, correspondiente a la precipitación, temperatura y las propiedades del suelo, y donde se aprecia la formación de 7 conglomerados, al igual que la figura 6.

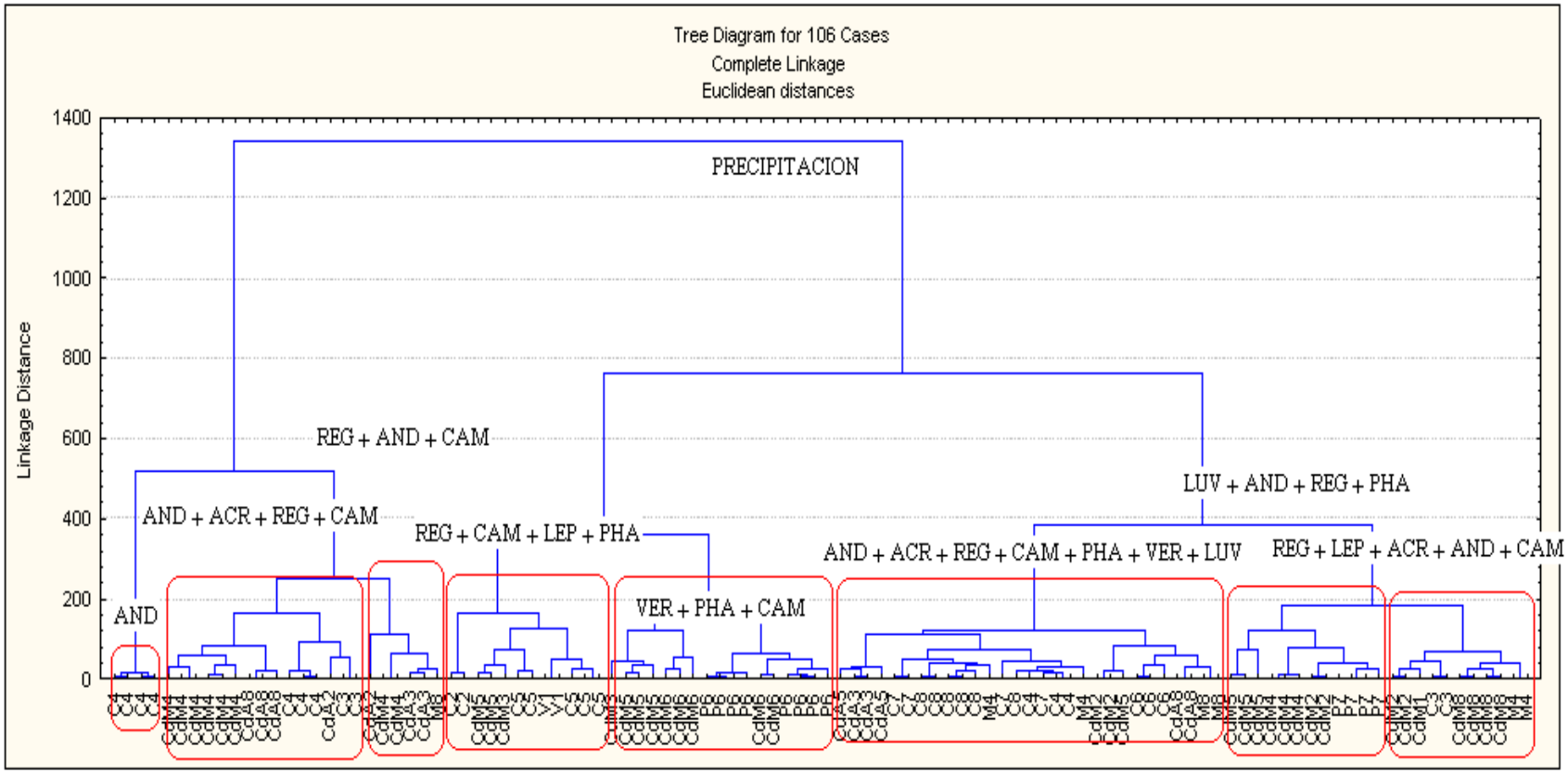


Figura 8.- Representación del dendograma del método de la liga completa, correspondiente a la precipitación y las propiedades del suelo, sin considerar a la temperatura y donde se aprecia la formación de 8 conglomerados.

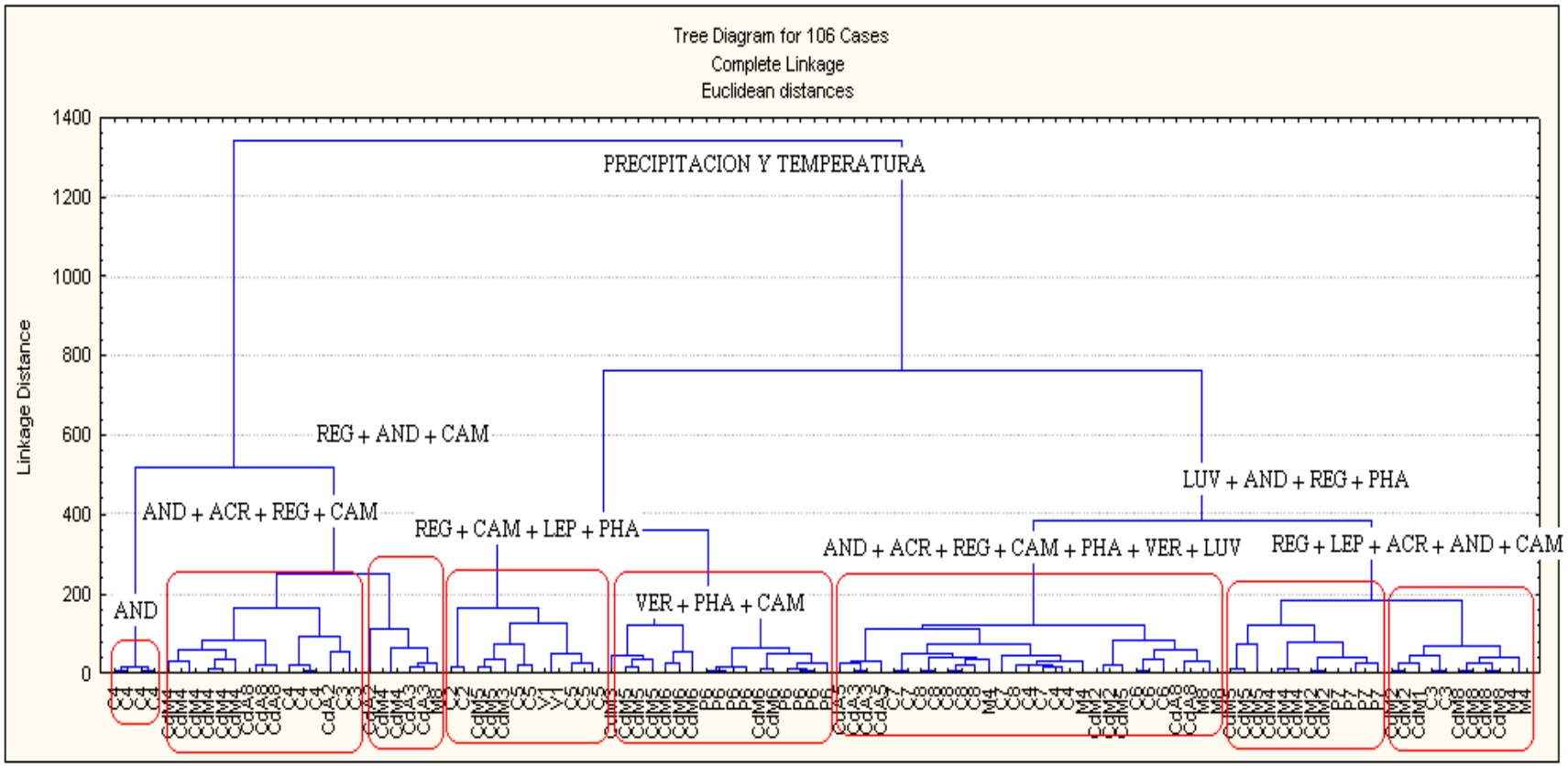


Figura 9.- Representación del dendograma del método de la liga completa, correspondiente a la precipitación, temperatura y las propiedades del suelo, y donde se aprecia la formación de 8 conglomerados.

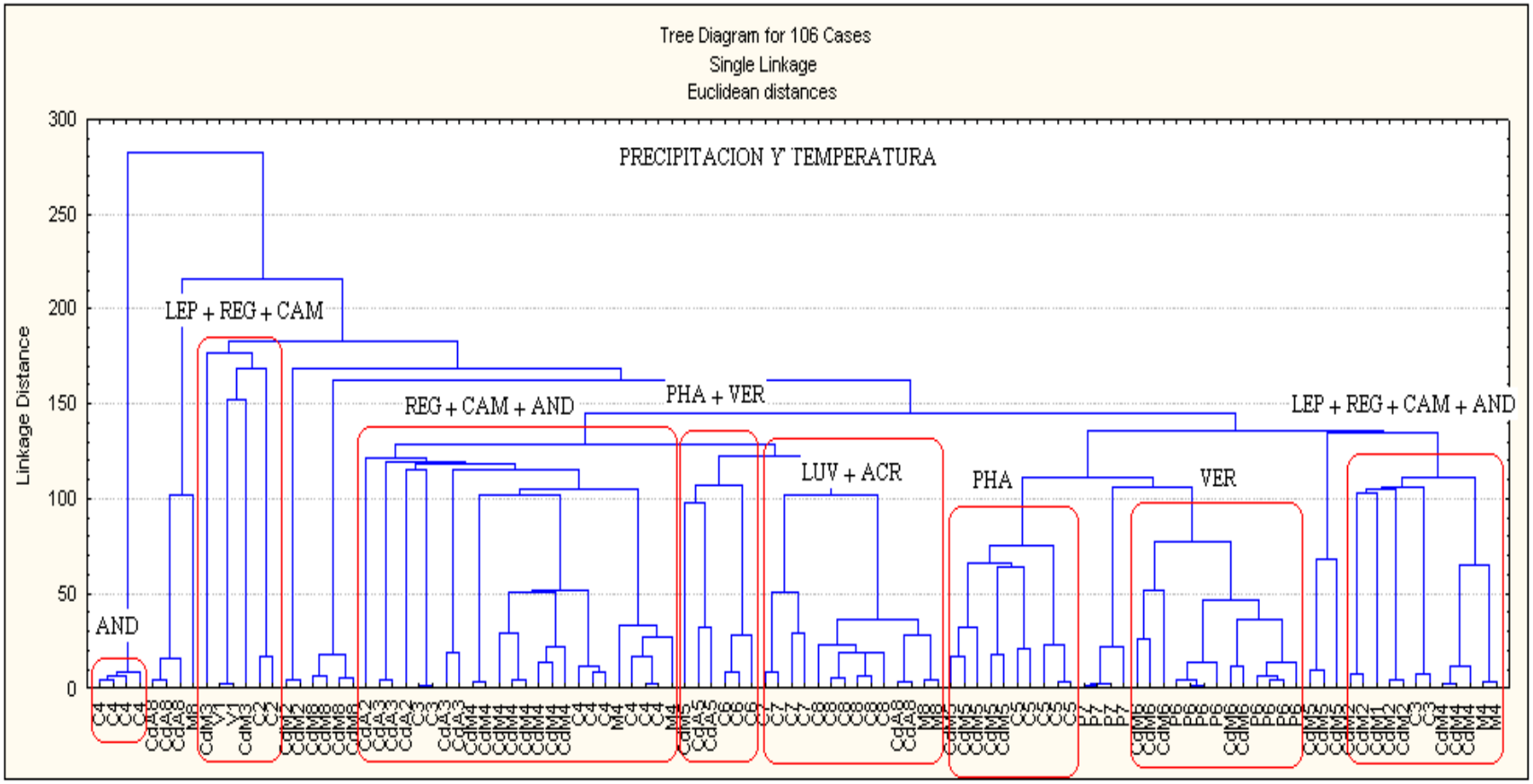


Figura 10.- Representación del dendograma del método de la liga simple, correspondiente a la precipitación, temperatura y las propiedades del suelo, considerando a la variable tipo de suelo de 100 en 100.

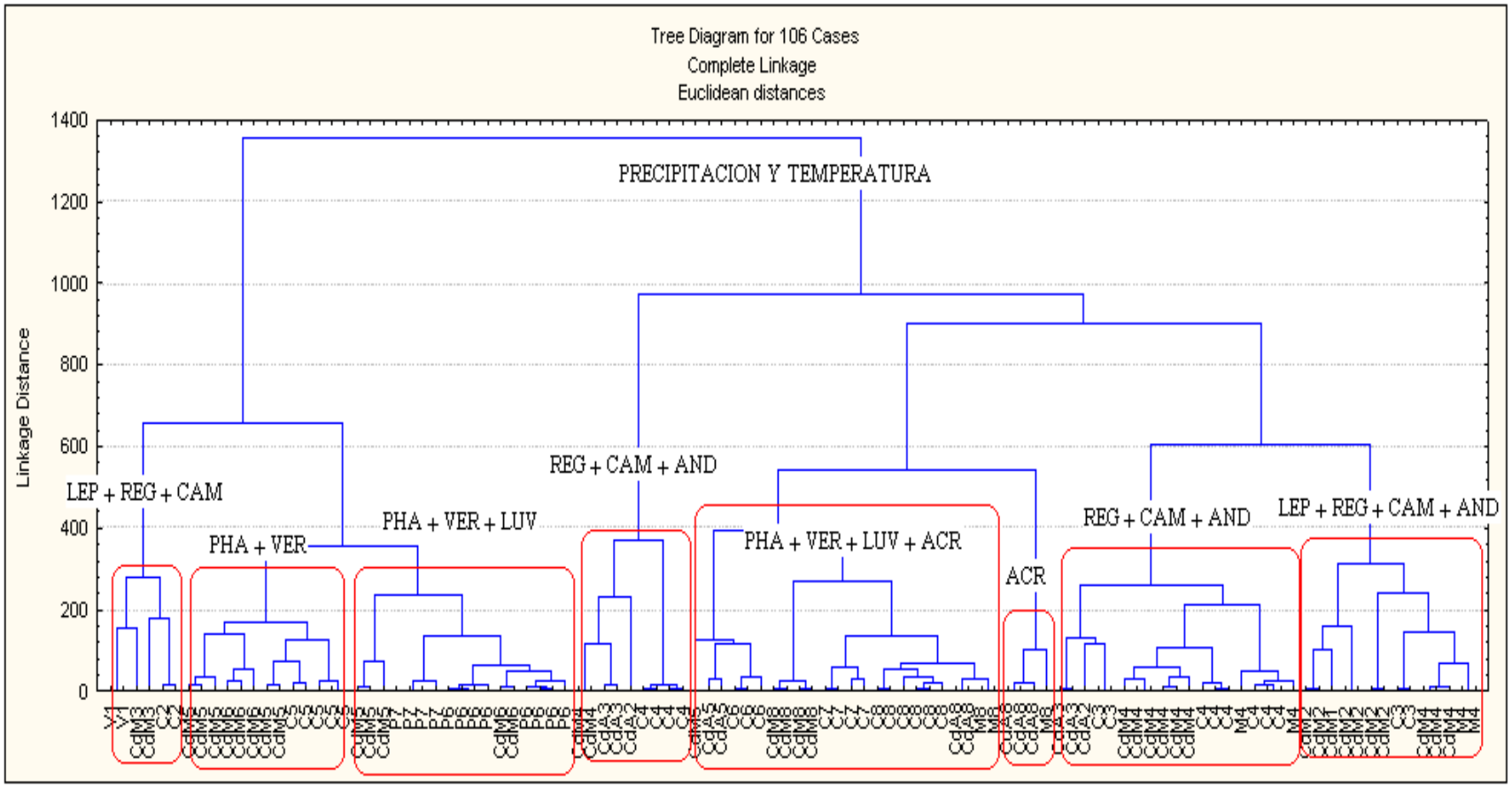


Figura 11.- Representación del dendograma del método de la liga completa, correspondiente a la precipitación, temperatura y las propiedades del suelo, considerando a la variable tipo de suelo de 100 en 100, además se puede ver que existe una mejor agrupación con respecto a los primero dendogramas.

BIBLIOGRAFIA.

Benzecri, J.P., 1976. L´Analyse des Données. I. La Taxinomie. L´Analyse des Données. II. L´Analyse des correspondances. Dunod, París.

Chang, D.H., Islam, S., 2000. Estimation of Soil Physical Properties Using Remote Sensing and Artificial Neuronal Network. Remote Sensing of Enviorement Vol. 74: 534-544 p.

Childs, E.C., Collis, N.G., 1950. The Permeability of Porous Materials. Proc. Roy. Soc., Ser. A(201): 392-405 p.

Cole, A.J., (ed.) 1969. Numerical Taxonomy. Proc. Coll. Num. Taxon., University of St. Andrews, 1968, Ac. Press, London.

Cox, G.M., Martin, W.M., 1937. Use of a Discriminant Function for Differentiating Soils With Different Azotobacter Populations. Iowa Experimental J451, 323-332 p.

Cuadras, C.M., 1996. Métodos de Análisis Multivariante. Universidad de Barcelona. 644 p.

Duchaufour, Ph., 1977. Atlas ecológico de los suelos del mundo. Barcelona; Masson, S.A., 178 p.

Fanning, D.S., and M.C.B. Fanning., 1989. Soil. Morphology, genesis and classification. John Wiley and Sons, New York; 395 p.

Fernandez, E. Hernández, J. 1987. Volcanic Soils. Weatherig and Landscape Relations Ships of Soils on Tephra and Basalt. Catena, Vol. 7, 7-23 p.

Fuentes, R.C., 1993. La Geometría Fractal en la Unificación de los Modelos de la Conductividad Hidráulica de los Suelos no Saturados. Contribución de la Coordinación de Tecnología de Riego y Drenaje del Instituto Mexicano de Tecnología del Agua (IMTA). 1-26 p.

Gama, C.J.E., Palacios, M.S., Villegas, S.M., 1990. Evaluación de la Hidroerosión en la Provincia de la Sierra Madre del Sur-Sistema Terrestre Tepetzingo, Estado de Morelos. Contribuciones a la Edafología Mexicana, Instituto de Geología, UNAM. 65-91 p.

Gama, C.J.E., Carreón F.D., Palacios, M.S., Solleiro, R.E., 1998. Genesis, Identificación y uso de los Suelos de México, Instituto Mexicano del Transporte y el Instituto de Geología, UNAM; 188-197 p.

Hastie, T.J., Tibshirani, R., Friedman, J., 2001. The elements of Statistical Learning: Data Mining, Inference and Prediction. Springer Series in Statistics. Springer-Verlag, New york.

Hodson, F.R., Kendall, D.G., Tautu, P., 1971. Mathematics in the Archaelogical and Historical Sciences. Proc. Anglo-Romaine Conf., Mamaia, 1970, Univ. Press, Edinburgh.

Jenny, H., 1941. Factors of Soil Formation, A System of Quantitative Pedology. McGraw-Hill, New York.

Jenny, H., 1980. The Soil Resource: Origin and Behavior. Ecological Studies 37, Springer Verlag, New York.

Johnson, D.L., Rockwell, T.K., 1982. Soil Geomorphology: Theory, Concepts and Principles with Examples and Applications on Alluvial and Marine Terraces in Coastal California. Geol. Soc. of Amer. Programs with Abstracrs 14, 176.

Johnson, D.L., Watson-Stegner, D., 1987. Evolution Model for Interpreting Quaternary Soils. Quaternary Research, Vol. 33: 306-9 p.

McBratney, A.B., Mendonça Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma, Vol. 117: 3-52 p.

Millington, R.J., Quirk, J.P., 1961. Permeability of Porous Solids. Trans. Faraday Soc. Vol. 57: 1200-1206 p.

Minasny, B., McBratney, A.B., 2002. The Neuro-m Method for Fitting Neuronal Network Parametric Pedotransfer Functions. Soil Science Society of America Journal Vol. 66: 352-361 p.

Oleschko, K., Figueroa, B., Miranda, M.E., Vuelvas, M.A., Solleiro, R.E., 2000. Mass Fractal Dimensions and Some Selected Physical Properties of Contrasting Soil and sediments of México. Soil & Tillage Research. Vol. 55: 43-61 p.

Phillips, J.D., 1993. Stability Implications of the State Factor Model of Soil as a Nonlinear Dynamical System. Geoderma, Vol. 58: 1-15 p.

Retallack, G.J., 1984. Completeness of the Rock and Fossil Record: Some Estimates Using Paleosols. Paleobiology, 10: 59-78 p.

Richter, J., 1987. The Soil as a Reactor. 1-193 p.

Romney, A.K., Shepard, R.N., Nerlove, S.B., 1972. Multidimensional Scaling. Theory and Applications in the Behavioral Sciences. Vol. II. Applications. Seminar Press, New York.

Ruhe, R.V., 1965. Quaternary Paleopedology. In: H.E. Wright, Jr. and D.G. Frey (eds.). The Quaternary of the United States. Pricenton Univ. Press, Pricenton, 922, 755-764 p.

Runge, E.C.A., 1973. Soil Development Sequences and Energy Models. Soil Sci, Vol. 115: 183-193 p.

Sedov, S, Solleiro, E, Gama, E., Andosol to Luvisol Evolution in Central México: Timing, mechanisms and environmental Setting. Catena, Vol. 54 (2003) 495-513p.

Shoji, S., Dahlgren, R., Nanzyo, M., 1993. Genesis of Volcanic ash Soils. In Shoji, S., Dahlgren, R., Nanzyo, M., (eds.): Volcanic ash Soils. Genesis, Properties and Utilization. Developments in Soil Science Vol. 21: 31-71 p.

Sneath, P.H.A., Sokal, R.S., 1973. Numerical Taxonomy. W.H. Freeman and Co., San Francisco.

Solleiro, R.E., 1997. Modelo Pedogenético para Establecer la Edad Evolutiva de Andisoles, Tesis de Doctorado, UNAM; 31-33 y 57-86 p.

Targulian V.O., Goriachkin S.V., 2004. Soil Memory: Typos of Record, Carriers, Hierarchy and Diversity. Revista Mexicana de Ciencias Geológicas. Vol. 21(1): 1-8.

Wilde, S.A., 1946. Forest Soils and Forest Growth, Walthur, Chronica Botánica.

Zhu, A.X., 2000. Mapping Soil Landscape as Spatial Continua: The Neuronal Network Approach. Water Resources Research, Vol. 36: 663-677 p.