UNIVERSIDAD DE QUINTANA ROO
DIVISIÓN DE CIENCIAS E INGENIERÍA

MÉTODOS ESTADÍSTICOS APLICADOS EN PROBLEMAS AMBIENTALES

Trabajo monográfico
PARA OBTENER EL GRADO DE
INGENIERO AMBIENTAL

CARRERA
INGENIERÍA AMBIENTAL

PRESENTA
BR. EVELIO SOSA BATÚN

supervisores
M.E.M. JOSÉ LUIS GONZÁLEZ BUCIO
DR. JAIME DIONISIO CUEVAS DOMÍNGUEZ
DR. JOSÉ MANUEL CARRIÓN JIMÉNEZ

CHETUMAL QUINTANA ROO, MÉXICO, JULIO DE 2018

UNIVERSIDAD DE QUINTANA ROO
DIVISIÓN DE CIENCIAS E INGENIERÍA

TRABAJO MONOGRÁFICO TITULADO
"MÉTODOS ESTADÍSTICOS APLICADOS EN PROBLEMAS AMBIENTALES"

ELABORADO POR
BR. EVELIO SOSA BATÚN

BAJO SUPERVISIÓN DEL COMITÉ DEL PROGRAMA DE LICENCIATURA Y APROBADO COMO REQUISITO PARCIAL PARA OBTENER EL GRADO DE:

INGENIERO AMBIENTAL

COMITÉ SUPERVISOR

SUPERVISOR: _____

M.E.M. JOSÉ LUIS GONZÁLEZ BUCIO

SUPERVISOR: _____

DR. JAIME DIONISIO CUEVAS DOMÍNGUEZ

SUPERVISOR: _____

DR. JOSÉ MANUEL CARRIÓN JIMÉNEZ

UNIVERSIDAD DE QUINTANA ROO
CONTROL ESCOLAR
TITULACIONES

Universidad de Quintana Roo
DCi DIVISIÓN DE CIENCIAS E INGENIERÍA

CHETUMAL QUINTANA ROO, MÉXICO, JULIO DE 2018

AGRADECIMIENTOS

GRACIAS DIOS POR FACILITARME LAS OPORTUNIDADES Y BRINDARME LA SABIDURÍA Y EL CORAJE PARA TOMARLAS.

A MIS PADRES MARTHA OLIVIA BATÚN SALAZAR Y EVELIO SOSA JIMÉNEZ, PORQUE NO PUDE HABER TENIDO UNOS MEJORES QUE A PESAR DE MIS DECIDÍAS SE MANTUVIERON SIEMPRE AL PENDIENTE E INSISTENTES EN COMPLETAR JUNTO CONMIGO ESTE LOGRO Y CON ESTE ÚLTIMO UN ÉXITO MÁS PARA ELLOS, GRACIAS PAPAS POR ESTAR SIEMPRE AHÍ PARA MÍ LO LOGRAMOS.

A MIS HERMANAS OLIVIA SOLEDAD SOSA BATÚN, mi hermana mayor, QUIEN CON SU FORMA ESTRICTA DE SER E INUSUAL FORMA DE EXPRESIÓN NO HA DEJADO DE ALENTARME A CONTINUAR, GRACIAS POR SER EL PRIMER EJEMPLO A SEGUIR. CON TU ÍMPETU HAS LOGRADO LLEGAR Y PERSEVERAR, MARTHA MARÍA SOSA BATÚN POR ESTAR EN TODO MOMENTO APOYANDO Y SIGUIENDO CADA UNA DE MIS ETAPAS, POR SER TAN SENSIBLE Y MÁS HUMANA, POR LAS PORRAS QUE ME HECHAS DÍA A DÍA Y DEMOSTRARA QUE, AUNQUE EL CAMINO NO ES FÁCIL DE RECORRER HOY ESTAMOS COSECHANDO UN ÉXITO MÁS QUE COMPARTO CONTIGO IGUAL, GRACIAS POR SER MI SEGUNDO EJEMPLO A SEGUIR MI HERMANA MENOR.

A MIS MAESTROS Y DOCTORES QUE CUANDO YA NO CREÍA PODER LOGRARLO ME IMPULSARON A NO DEJARLO Y CONTINUAR, DEJÁNDOME COMO ENSEÑANZA MÁS GRANDE QUE TODO SE PUEDE. GRACIAS DR. JAIME DIONISIO CUEVAS DOMÍNGUEZ, DR. JOSÉ MANUEL CARRIÓN JIMÉNEZ Y EN ESPECIAL AL M.E.M. JOSE LUIS GONZÁLEZ BUCIO POR EL VOTO DE CONFIANZA Y POR NO DEJAR QUE ESTE PROYECTO CAIGA EN VERDAD MIL GRACIAS.

A ADELAIDA GUADALUPE LLANOS PUC, COMPAÑERA, AMIGA Y CONFIDENTE POR LAS PORRAS QUE ME HECHAS, POR SOPORTAR CADA UNO DE LOS ALTIBAJOS QUE TODO ESTE PROCESO LLEVO, PERO MÁS AÚN GRACIAS POR ESTAR AHÍ ASESORÁNDOME Y ANIMÁNDOME.

A EMMANUEL ÁNGEL GUTIÉRREZ GARCÍA POR ALENTARME A CONTINUAR AUN CUANDO TODO SE VEÍA PERDIDO NUNCA DUDASTE DE UN SERVIDOR, GRACIAS POR LAS ASESORÍAS, LAS CORRECCIONES Y LOS REGAÑOS QUE HOY RINDEN SUS FRUTOS. GRACIAS POR LLEGAR Y QUEDARTE

CONTENIDO

CAPITULO I.- INTRODUCCIÓN

Debido a los graves problemas de contaminación de ambientes marinos en todo el mundo, se hace imprescindible la protección de los recursos naturales asociados a estos ecosistemas. Muchos de estos ecosistemas acuáticos se encuentran contaminados, producto de la actividad humana, algunos incluso, con sustancias tóxicas como los metales pesados e hidrocarburos.

En la actualidad ha aumentado incontrolablemente la contaminación ambiental, producto de las actividades antropogénicas, por lo que los cuerpos de agua se han visto afectados por la contaminación, producto de los vertimientos de desechos industriases y de aguas servidas.

En este sentido, vemos la importancia de estudiar a través de la aplicación de técnicas estadísticas modernas como el análisis multivariado; análisis de Clúster y Análisis de componentes principales (ACP), diferentes trabajos medioambientales con estas técnicas.

Dentro de este marco, se dividió en los siguientes capítulos:

Capítulo I, que lleva el nombre de “aplicación de la estadística en problemas ambientales”, en el cual se presentan los elementos conceptuales de la estadística con sus distintos componentes. Considerando que las estadísticas ambientales son un instrumento importante para alcanzar el desarrollo sustentable, resulta conveniente hacer alguna referencia a este tema. Las estadísticas del medio ambiente deben facilitar la formulación y evaluación de programas y proyectos como asimismo ofrecer datos básicos para investigaciones especializadas.

En capítulo II, denominado “correlación lineal y regresión” el objetivo de este capítulo es analizar el grado de la relación existente entre variables utilizando modelos matemáticos y representaciones gráficas. Así pues, para representar la relación entre dos o más variables desarrollaremos una ecuación que permitirá estimar una variable en función de la otra. También, contemplaremos dicho grado de relación entre dos variables en lo que llamaremos análisis de correlación. Para representar esta relación utilizaremos una representación gráfica llamada diagrama de dispersión y, finalmente, veremos un modelo matemático para estimar el valor de una variable basándonos en el valor de otra, en lo que llamaremos análisis de regresión.

Los objetivos que se deben plasmar en este capítulo son:

• Aprender a calcular la correlación entre dos variables.

• Saber dibujar un diagrama de dispersión.

• Representar la recta que define la relación lineal entre dos variables.

• Saber estimar la recta de regresión por el método de mínimos cuadrados e interpretar su ajuste.

• Realizar inferencia sobre los parámetros de la recta de regresión.

• Construir e interpretar intervalos de confianza e intervalos de predicción para la variable dependiente.

En el capítulo III, titulado “Varianza y Covarianza” el análisis de la covarianza es una técnica estadística que, utilizando un modelo de regresión lineal múltiple, busca comparar los resultados obtenidos en diferentes grupos de una variable cuantitativa, pero "corrigiendo" las posibles diferencias existentes entre los grupos en otras variables que pudieran afectar también al resultado (covariantes). Sin lugar a duda, la medida más usada para estimar la dispersión de los datos es la desviación típica; ésta es especialmente aconsejable cuando se usa la media aritmética como medida de tendencia central. Al igual que la desviación media, está basada en un valor promedio de las desviaciones respecto a la media. En este caso, en vez de tomar valores absolutos de las desviaciones, para evitar así que se compensen desviaciones positivas y negativas, se usan los cuadrados de las desviaciones. Esto hace además que los datos con desviaciones grandes influyan mucho en el resultado final.

Finalmente, en el capítulo IV, que se titula “Análisis Multivariado” es importante el trabajar con volúmenes grandes de datos generados, en análisis ambiental ha traído como consecuencia la necesidad de utilización de métodos especiales del procesamiento de los resultados, entre las principales herramientas quimiométricas se encuentran los métodos multivariados, como son; el análisis de Discriminantes (AD), las correlaciones Canónicas y el Análisis por Componentes Principales (ACP). Estas han sido empleadas en los últimos 20 años, en la interpretación de datos numéricos resultantes de los diseños de experimentos y de la medición de numerosas variables simultáneamente.

1.1 Objetivo general

Revisión documental donde se apliquen métodos estadísticos (análisis multivariado) en problemas ambientales.

1.2 Objetivos particulares

Analizar diferentes trabajos donde apliquen métodos estadísticos (Análisis de Componentes Principales, Clúster) en situaciones ambientales.

Entendimiento e interpretación del estudiante a la Regresión múltiple, Análisis de Componentes Principales y Clúster.

CAPITULO II.- REGRESIÓN Y CORRELACIÓN

2.1. INTRODUCCIÓN

Es común que las personas tomen decisiones personales y profesionales basadas en predicciones de sucesos futuros. Para hacer estos pronósticos, se basan en la relación intuitiva y calculada entre lo que ya se sabe y lo que se debe estimar. Si los responsables de la toma de decisiones pueden determinar cómo lo conocido se relaciona con un evento futuro, pueden ayudar considerablemente al proceso de toma de decisiones. (http://asesorias.cuautitlan2.unam.mx)

Cualquier método estadístico que busque establecer una ecuación que permita estimar el valor desconocido de una variable a partir del valor conocido de una o más variables, se denomina análisis de regresión. Los análisis de regresión y correlación mostrarán como determinar la naturaleza y la fuerza de una relación entre dos variables. (http://asesorias.cuautitlan2.unam.mx)

El término regresión fue utilizado por primera vez por el genetista y estadístico inglés Francis Galton (1822-1911) en 1877 Galton efectúo un estudio que demostró que la altura d los hijos de padres altos tendía a retroceder, o "regresar", hacia la talla media de la población. *Regresión* fue el nombre que le dio al proceso general de predecir una variable, (la talla de los niños) a partir de otra (la talla de los padres). (http://asesorias.cuautitlan2.unam.mx)

Hoy en día, esta tendencia de miembros de cualquier población que están en una posición extrema (arriba o debajo de la media poblacional) en un momento, y luego en una posición menos extrema en otro momento, (ya sea por sí o por medio de sus descendientes), se llama efecto de regresión. (http://asesorias.cuautitlan2.unam.mx)

El análisis de regresión se desarrolla una ecuación de estimación, es decir, una fórmula matemática que relaciona las variables conocidas con las desconocidas. Luego de obtener el patrón de dicha relación, se aplica el análisis de correlación para determinar el grado de relación que hay entre las variables. (http://asesorias.cuautitlan2.unam.mx)

2.1.1 REGRESIÓN LINEAL SIMPLE

"Una técnica estadística que establece una ecuación para estimar el valor desconocido de una variable, a partir del valor conocido de otra variable, (en vez de valores de muchas otras variables) se denomina análisis de regresión simple." (http://asesorias.cuautitlan2.unam.mx)

Entonces, el análisis de regresión lineal simple, trata de obtener una variable (Y) a partir de una (X)

Las relaciones entre las variables pueden ser directas o también inversas.

 Relación directa: la pendiente de esta línea es positiva, porque la variable Y crece a medida que la variable X también lo hace.

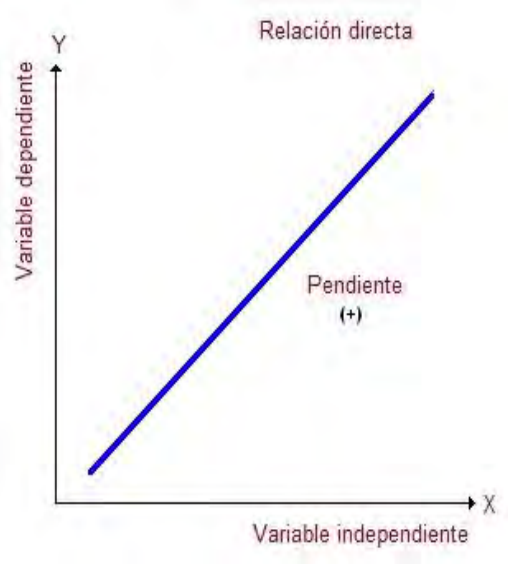

Ilustración 1 Relación Directa

 Relación inversa: La pendiente de esta línea es negativa, porque a medida que aumenta el valor de la variable Y, el valor de la variable X disminuye.

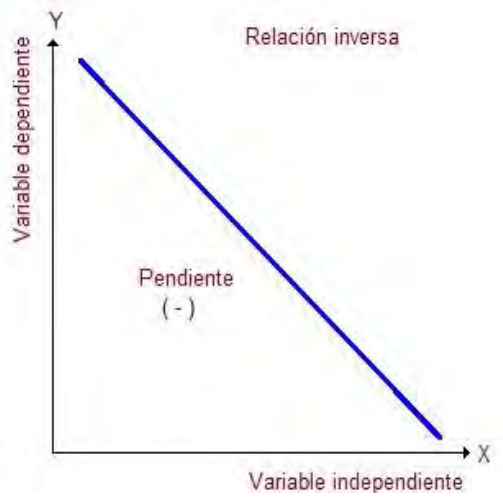

Ilustración 2 Relación Inversa

2.1.2 VARIABLE INDEPENDIENTE (X)

En el análisis de regresión una variable cuyo valor se suponga conocido y que se utilice para explicar o predecir el valor de otra variable de interés se llama variable independiente; se simboliza con la letra X. Podremos nombrar la variable independiente de otras formas como, variables explicatoria, variable predictora o variable regresora.

2.1.3 VARIABLE DEPENDIENTE (Y)

En el análisis de regresión una variable cuyo valor se suponga desconocido y que se explique o prediga con ayuda de otra se llama variable dependiente y se simboliza con la letra Y.

Al igual que la variable independiente a la variable dependiente se le puede denominar: variable explicada o variable pronosticada.

2.1.4 DIAGRAMAS DE DISPERSIÓN

Un diagrama de dispersión es una ilustración gráfica que se usa en el análisis de regresión. Consta de una dispersión de puntos tal que cada punto representa un valor de la variable independiente (medido a lo largo del eje horizontal), y un valor asociado de la variable dependiente (medido a lo largo del eje vertical). (http://asesorias.cuautitlan2.unam.mx)

El diagrama de dispersión, también llamado nube de puntos, brinda dos tipos de información, visualmente se pueden determinar los patrones que indican como las variables están relacionadas (lineal o mediante una curva) y por otro lado si existe una relación entre ellas visualizando la clase de línea o ecuación de estimación que describe a dicha relación. (http://asesorias.cuautitlan2.unam.mx)

A continuación, se ilustran algunas relaciones en los diagramas de dispersión:

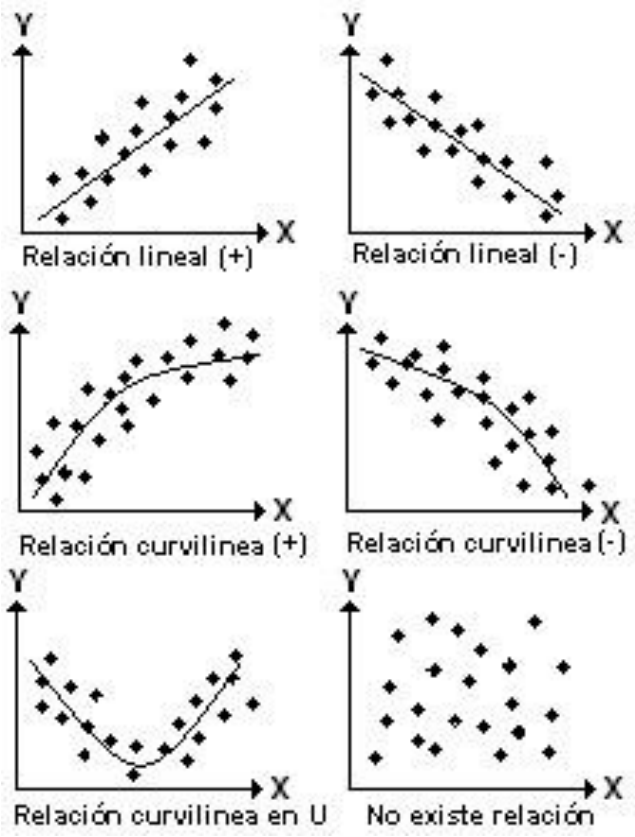

Ilustración 3 Diagramas de Dispersión

2.1.5 MÉTODO DE MÍNIMOS CUADRADOS

El método que por lo común se utiliza para ajustar una línea a los datos muéstrales indicados en el diagrama de dispersión, se llama método de mínimos cuadrados. La línea se deriva en forma tal que la suma de los cuadrados de las desviaciones verticales entre la línea y los puntos individuales de datos se reduce al mínimo. (http://asesorias.cuautitlan2.unam.mx)

El método de mínimos cuadrados sirve para determinar la recta que mejor se ajuste a los datos muéstrales, y los supuestos de este método son:

- El error es cero.
- Los datos obtenidos de las muestras son estadísticamente independientes.
- La varianza del error es igual para todos los valores de X.

Una línea de regresión calculada a partir de los datos muéstrales, por el método de mínimos cuadrados se llama línea de regresión estimada o línea de regresión muestral. Dicha línea recta es la que mejor se ajusta al conjunto de datos (X, Y) y es aquella en que la distancia que hay entre los datos y la supuesta recta es la menor posible, y se calcula mediante la siguiente formula: $\hat{y} = a + bx$.

- Para calcular el valor de b (pendiente), que representa el grado de inclinación que tiene la recta, se emplea la siguiente formula:

$$b = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n(\bar{x})^2}$$

- Para calcular el valor de a (ordenada al origen), que representa el punto en que la recta corta al eje de las Y, se emplea la siguiente formula:

$$a = \bar{y} + \bar{b}x$$

Las variables a y b son constantes numéricas que son las que se calculan mediante el método de mínimos cuadrados. (http://asesorias.cuautitlan2.unam.mx)

2.1.6 ERROR ESTANDAR DE ESTIMACIÓN

El siguiente paso en el análisis de la regresión lineal simple necesario, el de medir la confiabilidad de la ecuación de estimación que se ha desarrollado.

El error estándar de estimación mide la variabilidad o dispersión de los valores observados alrededor de la línea de regresión y se representa como S_e. Su fórmula es la siguiente:

$$S_e = \sqrt{\frac{\sum y^2 - \left(a \bullet \sum y\right) - \left(b \bullet \sum xy\right)}{n-2}}$$

Cuanto mayor sea el error estándar de la estimación, más grande será la dispersión (o esparcimiento) de puntos alrededor de la línea de regresión. Por el contrario, si $\mathbf{S_e} = 0$, se espera que la ecuación de estimación sea un estimador "perfecto" de la variable dependiente, en este caso todos los puntos caerían directamente sobre la línea de regresión y no habría puntos dispersos, como se muestra en la siguiente figura:

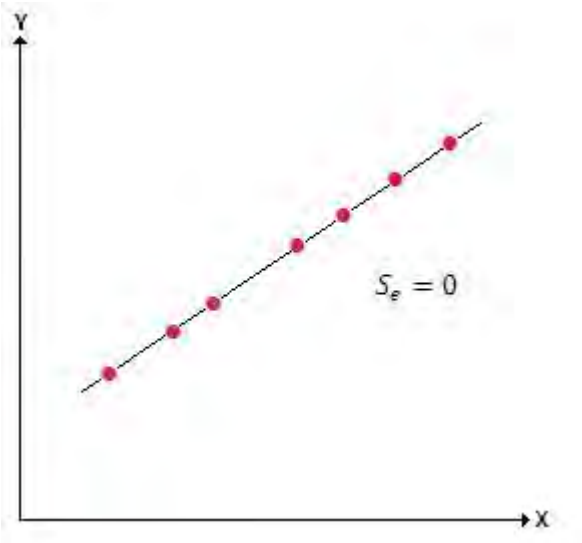

Ilustración 4 Estimación Perfecta

El error estándar de estimación tiene la misma aplicación que de la desviación estándar que se vio en los temas anteriores. Esto es, suponiendo que los puntos observados tienen una distribución normal alrededor de la recta de regresión, podemos esperar que:

- 68% de los puntos están dentro de ± $1s_e$
- 95.5% de los puntos están dentro de ± $2s_e$
- 99.7% de los puntos están dentro de ± $3s_e$

El error estándar de la estimación se mide a lo largo del eje "Y", y no perpendicularmente desde la recta de regresión.

Las suposiciones son:

1. Los valores observados para Y tienen distribución normal alrededor de cada valor estimado de $\hat{y}$
2. La varianza de las distribuciones alrededor de cada valor posible de $\hat{y}$ es la misma.

Si esta segunda suposición no fuera cierta, entonces el error estándar en un punto de la recta de regresión podría diferir del error estándar en otro punto. (http://asesorias.cuautitlan2.unam.mx)

2.1.7 CORRELACIÓN SIMPLE

El análisis de correlación es la herramienta estadística que se utiliza para describir el grado o fuerza en el que una variable esta linealmente relacionada con otra. Mientras que el análisis de regresión simple se establece una ecuación precisa que enlaza las dos variables. De acuerdo a la medida cuantitativa podemos afirmar, que tan cerca se pueden observar dos variables, lo que nos permite saber cuál es la confiabilidad que tiene una variable con ayuda de otra.

Una técnica estadística que establece un índice que proporciona, en un solo número, una medida de la fuerza de asociación entre dos variables de interés, se llama análisis de correlación simple. El análisis de correlación es la herramienta estadística de que nos valemos para describir el grado de relación que hay entre dos variables. (http://asesorias.cuautitlan2.unam.mx)

En la mayoría de casos el análisis de correlación simple se puede utilizar con el análisis de regresión lineal simple esto con la finalidad de medir la eficacia con que la línea de regresión explica la variación de la variable dependiente (Y).

Diagramas de dispersión con correlación débil y fuerte.

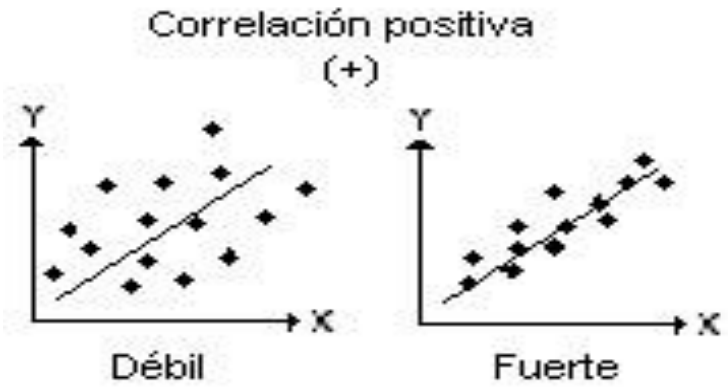

Ilustración 5 Correlación Positiva

Existen dos medidas para describir la correlación entre dos variables: el *coeficiente de determinación* y el *coeficiente de correlación.* (http://asesorias.cuautitlan2.unam.mx)

2.1.8 COEFICIENTE MUESTRAL DE DETERMINACIÓN

Otra medida importante que también ajusta la línea de regresión estimada en los datos muéstrales de los cuales se basa, es el coeficiente de determinación muestral, el cual es igual a la proporción de la variación total de los valores de la variante dependiente, (Y), la cual se explica por medio de la asociación de Y con X medida por la línea de regresión estimada.

El coeficiente de determinación es la manera primaria de medir el grado, o fuerza, de la relación que existe entre dos variables, X y Y. El coeficiente de determinación muestral se representa como r^2, y mide exclusivamente la fuerza de una relación lineal entre dos variables. (http://asesorias.cuautitlan2.unam.mx)

El Cálculo del coeficiente de determinación se lleva a cabo con la siguiente formula:

$$r^2 = \frac{(a \bullet \sum y) + (b \bullet \sum xy) - (n \bullet \bar{y}^2)}{\sum y^2 - (n \bullet \bar{y}^2)}$$

2.1.9 COEFICIENTE MUESTRAL DE CORRELACIÓN

La raíz cuadrada del coeficiente de determinación muestral, $r = \sqrt[2]{2}$, es un índice alternativo común del grado de asociación entre dos variables cuantitativas. Esta mediad se llama coeficiente de correlación muestral (r) y es un estimador puntual del coeficiente de correlación poblacional (ρ). El coeficiente de correlación muestral es la segunda medida con que puede describirse la eficacia con que una variable es explicada por otra, así pues, el signo de ***r*** indica la dirección de la relación entre las dos variables X y Y. (http://asesorias.cuautitlan2.unam.mx)

El siguiente esquema representa adecuadamente la intensidad y la dirección del coeficiente de correlación muestral.

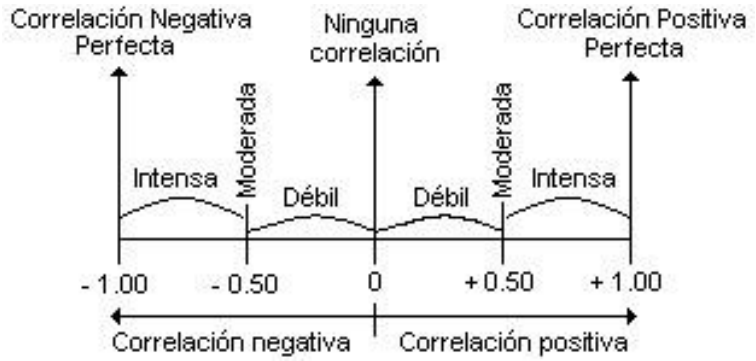

Ilustración 6 Intensidad y Dirección del Coeficiente de Correlación Muestral

El cálculo del coeficiente de correlación muestral se lleva a cabo con la siguiente formula:

$$r = \sqrt{r2}$$

2.1.10 INTERVALO DE CONFIANZA

Para dar seguridad a nuestros cálculos es necesario elaborar un intervalo de confianza ya que la recta estimada de regresión, no es del todo real.

Como se ha visto, cuando se utilice el método de mínimos cuadrados, los coeficientes de regresión, *a* y *b* son estimadores insesgados, eficientes y consistentes de α y β, también aquí es muchas ocasiones es deseable establecer intervalos de confianza. (http://asesorias.cuautitlan2.unam.mx)

Los intervalos de confianza se calculan con la siguiente fórmula:

$$y_c = \hat{y} \pm t_{\alpha/2}, gl_{n-2}\left(\frac{S_e}{\sqrt{n}}\right)$$

2.1.11 INTERVALO DE PREDICCIÓN

El intervalo de predicción, como su nombre lo indica, se utiliza para predecir un intervalo de valores de Y, dado un valor de X. (http://asesorias.cuautitlan2.unam.mx)

 El intervalo de predicción se calcula con la siguiente fórmula:

$$y_p = \hat{y} \pm \left(t_{\alpha/2}(n-2)\right) \bullet S_e \bullet \sqrt{1 + \frac{1}{n} + \frac{(X-\bar{x})^2}{\sum x^2 - n(\bar{x})^2}}$$

2.2 APLICACIÓN EN LA RESOLUCIÓN DE PROBLEMAS AMBIENTALES

Un ejemplo de correlación es el que publica (Ramos-Herrera S. *et al*., 2010), estudio estadístico de la correlación entre contaminantes atmosféricos, que serán nuestras variables independientes (Y) y variables meteorológicas, que serán nuestra variables (X) en la zona norte de Chiapas, México; donde aplica el análisis de regresión múltiple, un método estadístico empleado en muchas áreas del conocimiento. En este estudio, dicho análisis se aplicó a los datos de concentraciones de cuatro contaminantes atmosféricos (SO_2, NO_2, H_2S y PM_{10}), monitoreados en tres estaciones que se ubican en la Zona Norte del estado de Chiapas. El periodo que abarcó el estudio fue de enero 2001 a febrero 2005. El objetivo fue proponer funciones de regresión para describir la concentración en función del tiempo y/o las variables meteorológicas. Se empleó un análisis de regresión lineal múltiple, paso a paso en la selección de variables regresoras. Las más importantes fueron la temperatura, la humedad relativa y la dirección del viento. Se obtuvieron funciones de regresión de la concentración anual, mensual y diaria de estos contaminantes. Se obtuvo una regresión lineal simple para explicar la concentración anual de SO_2, alcanzando un coeficiente de determinación de 0.927. Los modelos de la concentración mensual alcanzaron un coeficiente de determinación entre 0.417 y 0.846; mientras que para los de la concentración diaria, este coeficiente varió de 0.285 a 0.581. De este estudio paramétrico, se concluyó que las variables meteorológicas describieron adecuadamente la concentración anual y mensual, pero no la concentración diaria.

En esta primera tabla observamos cómo se fueron dando las concentraciones de los contaminantes atmosféricos (Y) de acuerdo a la dirección del viento (X), esto para cada una de las dos estaciones expuestas

Tabla 1. Correlación de la concentración diaria con la dirección del viento.

Dirección del viento	Estación Reforma				Estación Giraldas			
	SO_2	NO_2	H_2S	PM_{10}	SO_2	NO_2	H_2S	PM_{10}
N	0.05	0.04	0.04	0.04	0.01	-0.03	-0.04	0
NNE	0.05	0.05	0.075*	0.04	0.073*	0.06	-0.01	0.04
NE	0.076*	0.01	0.01	0.070*	0.152**	0.04	0.08	-0.01
ENE	-0.01	0.142**	0.080*	0.083*	0.086*	0.01	0.142**	-0.098**
E	-0.075*	-0.131**	-0.074*	0.02	0	-0.02	0.087*	-0.04
ESE	-0.05	-0.05	-0.092**	0	0.01	0.089**	0.06	-0.02
SE	-0.01	-0.05	-0.086*	0	0.06	0.092**	0.06	0.121**
SSE	0.02	-0.01	-0.03	-0.01	-0.03	0.070*	-0.05	0.077*
S	-0.02	0	0	0.03	0.05	0.06	-0.04	0.142**
SSO	0.02	-0.01	0.02	-0.06	-0.01	0.03	-0.01	0.05
SO	0.01	0.01	-0.01	-0.02	-0.083*	-0.02	-0.093*	-0.01
OSO	0.03	0.094**	0.05	-0.170**	-0.04	-0.06	-0.142**	-0.02
O	0	-0.05	0.05	-0.112**	-0.06	-0.05	-0.06	-0.068*
ONO	0.01	0	0.122**	-0.01	0.03	-0.075*	-0.07	-0.094**
NO	0.06	0.04	0	0.084*	0.06	0.01	-0.04	0.076*

En los siguientes diagramas de dispersión se dan a conocer la relación que existe entre ambas estaciones con respecto a la variable meteorológica: temperatura - temperatura, presión - presión, radiación - radiación y humedad relativa – humedad relativa.

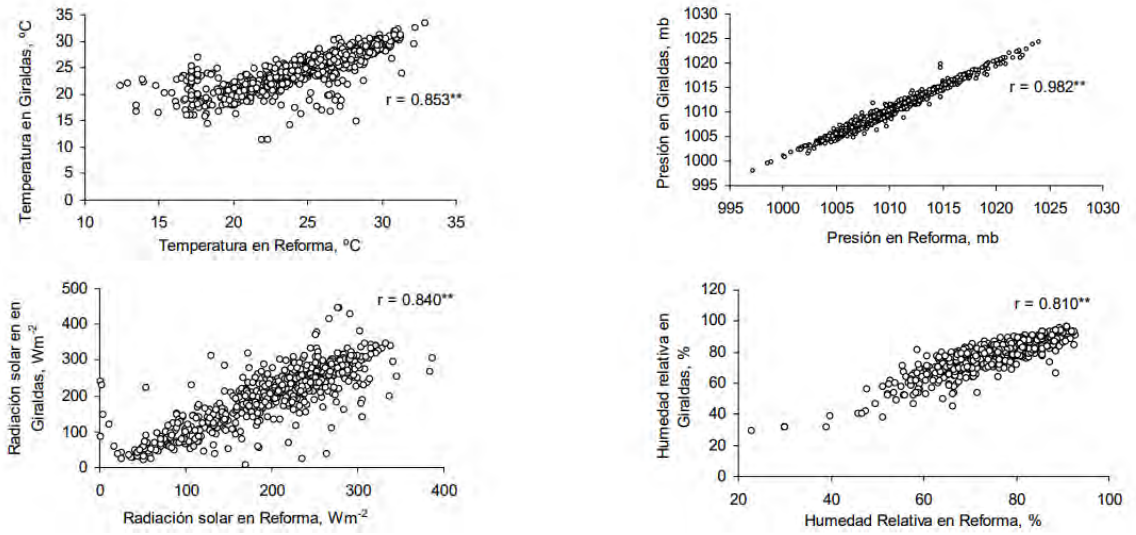

Ilustración 7. Gráficas de dispersión de las variables meteorológicas en Giraldas vs Reforma.

CAPITULO III.- ANALISIS DE LA VARIANZA

3.1 INTRODUCCIÓN

En múltiples ocasiones el analista o investigador se enfrenta al problema de determinar si dos o más grupos son iguales, si dos o más cursos de acción arrojan resultados similares o si dos o más conjuntos de observaciones son parecidos. Pensemos por ejemplo en el caso de determinar si dos niveles de renta producen consumos iguales o diferentes de un determinado producto, si las notas de dos grupos en una asignatura son similares, si tres muestras de análisis químico de una sustancia son iguales, o si los municipios de cuatro provincias colindantes tienen el mismo nivel de paro. (https://uam.es/departamentos/economicas/econapli/anova.pdf)

Una aproximación simple sería comparar las medias de estos grupos y ver si las medias aritméticas de la variable estudiada son parecidas o diferentes. Pero tal aproximación no es válida ya que la dispersión de las observaciones influirá en la posibilidad de comparar los promedios o medias de cada grupo. Así, supongamos que tenemos una variable X (consumo) y dos grupos (nivel de renta alto y medio) y que tenemos dos resultados distintos correspondientes a dos provincias

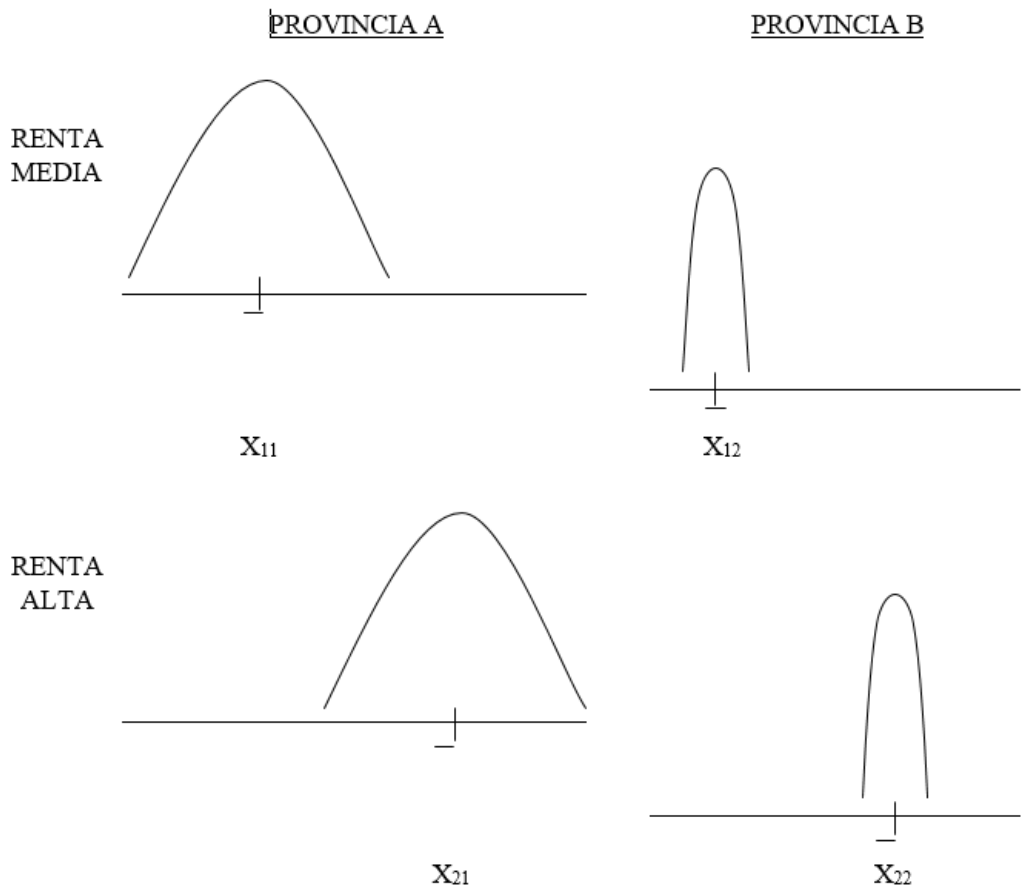

Ilustración 8 Ejemplo de Dispersión ANOVA

Es evidente que la conclusión de que con renta alta el consumo es mayor que con renta media es más rotundo en la provincia B que en la A. En la provincia A existen familias de renta media con un consumo superior a otras familias de renta alta, aunque en promedio $\overline{X}_{21} > \overline{X}_{11}$. Esta situación no se produce en la provincia B donde todas las observaciones de renta alta son superiores a las de renta media. En consecuencia, la dispersión deberá tenerse en cuenta para realizar una comparación de medias o de grupos y esto es lo que se pretende con el Análisis de la Varianza. https://uam.es/departamentos/economicas/econapli/anova.pdf)

El Análisis de la Varianza puede contemplarse como un caso especial de la modelización econométrica, donde el conjunto de variables explicativas son variables ficticias y la variable dependiente es de tipo continuo. En tales situaciones la estimación del modelo significa la realización de un análisis de la varianza clásica (ANOVA), de amplia tradición en los estudios y diseños experimentales. Una ampliación a este planteamiento es cuando se dispone de una variable de control que nos permite corregir el resultado del experimento mediante el análisis de la variación con la variable a estudiar. En tal situación nos encontramos frente a un análisis de la covarianza (ANCOVA). A continuación, se expondrán ambos procedimientos, ANOVA, ANCOVA, precedidos de un ejemplo que facilita su compresión. (https://uam.es/departamentos/economicas/econapli/anova.pdf)

3.1.1 ANÁLISIS DE LA VARIANZA: ANOVA

Ejemplo: Pretendemos medir la influencia que tiene en la venta de un producto de alimentación, la posición en que se exhibe al público dentro del establecimiento. https://uam.es/departamentos/economicas/econapli/anova.pdf

Las posiciones establecidas son:

- ALTA: por encima de los ojos.
- MEDIA: nivel de los ojos.
- BAJA: por debajo del nivel de los ojos.

Para la realización del experimento se han seleccionado 12 autoservicios de dimensiones similares. Los autoservicios se agrupan en tres conjuntos de cuatro elementos cada uno, procediendo de forma aleatoria en su asignación. Con ello suponemos que los tres conjuntos son de características similares, colocándose el producto en cada uno de ellos, de una de las formas anteriormente descritas y registrando sus ventas durante veinte días. Las ventas resultantes, en unidades, quedan recogidas en el cuadro I. Se pretende responder a las siguientes preguntas:

1º. ¿Tiene alguna influencia el posicionamiento del producto en la venta del mismo?

2º. ¿Qué posicionamiento es más eficaz?

3º. ¿Son estadísticamente significativas las diferencias obtenidas?

Cuadro I. Ventas en autoservicios por tipo de tratamiento

POSICIÓN PRODUCTO	**ESTABLECIMIENTO**	**VENTAS (unidades)**
ALTA	A	663
	B	795
	C	922
	D	1056
MEDIA	E	798
	F	926
	G	1060
	H	1188
BAJA	I	528
	J	660
	K	792
	L	924

Ilustración 9. Ventas en Autoservicios por Tipo de Tratamiento

Este sencillo ejemplo nos presenta el caso de tener un único factor a analizar (posición del producto) y tres niveles del factor (ALTO, MEDIO, BAJO). Sin embargo, podemos

encontrarnos con múltiples factores a estudiar simultáneamente. Al mismo tiempo, podemos distinguir tres tipos de modelos según sean de:

Efectos fijos: donde sólo estudiamos determinados niveles del factor (es el caso de las tres alturas) y únicamente perseguimos sacar conclusiones para éstos. (Situación más común en las Ciencias Sociales).

Efectos aleatorios: en este caso los niveles son infinitos y estudiamos una muestra de los mismos. Sus resultados también serán aleatorios.

Efectos mixtos: cuando nos encontramos con uno o más factores de las clases anteriores.

Como vemos, ANOVA será especialmente útil en aquellos supuestos en los que queramos analizar distintas situaciones o alternativas de actuación y donde de alguna forma podemos intervenir en la realización del experimento. A diferencia del análisis econométrico habitual, donde las series históricas son dadas y no podemos repetir la situación, ni modificar alguna de las condiciones o variables (pensemos en el P.I.B., inflación, etc.) para estudiar sus efectos, en el contexto ANOVA y ANCOVA nos encontraremos la mayoría de las veces ante datos experimentales (controlables y/o repetibles en mayor o menor grado). (https://uam.es/departamentos/economicas/econapli/anova.pdf)

Si bien los desarrollos clásicos de ANOVA y ANCOVA se han efectuado desde el análisis de variación de las variables y su descomposición (variaciones entre - intragrupos), podemos efectuar una sencilla aproximación desde el análisis de regresión múltiple, con idénticos resultados. Dado que suponemos al alumno familiarizado con la aproximación tradicional de ANOVA, en base a explicaciones de otras asignaturas, aquí nos limitaremos a un breve recuerdo de la misma. (https://uam.es/departamentos/economicas/econapli/anova.pdf)

El modelo ANOVA tradicional tiene la expresión:

$$Yij = m + t\,j + eij$$

Y_{ij} = es la variable objeto de estudio y que en nuestro caso es la venta para el establecimiento i del nivel j.

m= es una constante e indica la respuesta media de todos los niveles.

t_j = es el efecto diferencial del nivel j. Recoge la importancia de cada tratamiento y es el objetivo del análisis. Dado que los t_j son efectos diferenciales sobre $m$tenemos que

$$\sum t_j = 0$$

e_{ij} = es un término de error, considerado como variable aleatoria $N \sim (0, s^2)$

Por tanto, las ventas de un autoservicio, Y_{ij} se descomponen en una parte que es común a todos los tratamientos, *(m)*, o en otras palabras el efecto medio de todos ellos y otra parte, (t_j) que es el efecto diferencial de poner el producto a una determinada altura y que es propio de ese nivel. Un tercer componente es lo no recogido por los dos anteriores y que denominamos error.

No olvidemos que el subíndice i nos indica el elemento o autoservicio:

$i = 1,2,.....n_j$

Para cada nivel j.

j = 1,2,.....g

En nuestro ejemplo, g es igual a tres niveles (ALTO, MEDIO Y BAJO) y n_j es igual a cuatro para cualquier nivel j (cuatro establecimientos para cada nivel). (https://uam.es/departamentos/economicas/econapli/anova.pdf)

El ANOVA tradicional parte de descomponer la variación total de la muestra, en dos componentes:

VARIACIÓN TOTAL	=	VARIACIÓN ENTRE	+	VARIACIÓN INTRA

Ilustración 10 ANOVA Tradicional

Esta igualdad básica nos indica que la variación total es igual a la suma de la variación o dispersión entre los grupos, más la variación o dispersión dentro de cada grupo. Los grupos están definidos por los niveles de factor.

La anterior igualdad puede expresarse por:

$$\underbrace{\sum_{j=1}^{g}\sum_{i=1}^{n_j}(Y_{ij}-\bar{Y}_{..})^2}_{\text{V. TOTAL}} = \underbrace{\sum_{j=1}^{g} n_j(\bar{Y}_{.j}-\bar{Y}_{..})^2}_{\text{V. ENTRE}} + \underbrace{\sum_{j=1}^{g}\sum_{i=1}^{n_j}(Y_{ij}-\bar{Y}_{.j})^2}_{\text{V. INTRA}}$$

Correspondiendo cada término de la suma a las anteriores variaciones y siendo $\bar{Y}_{..}$ la media total e $\bar{Y}_{.j}$ la media de grupo o nivel j.

Los grados de libertad (número de observaciones – parámetros a estimar) correspondientes a cada uno de los componentes de la variación total son:

- Variación ENTRE: g – 1
- Variación INTRA: n – g
- Variación TOTAL: n – 1

Dado que a través del Análisis de la Varianza se persigue saber si los distintos niveles de un factor influye en los valores de una variable continua (en nuestros ejemplo queremos saber si la posición de un producto en una estantería influye en las ventas de ese producto), para que efectivamente sí haya diferencias en los valores de la variable continua según el nivel del factor, se tiene que dar simultáneamente que el comportamiento de la variable continua sea lo más distinto posible para los distintos niveles del factor, y a su vez, que dentro de cada grupo (determinado por los niveles del factor) los valores sean lo más homogéneos posibles. En otras palabras, se tiene que dar que la variación intragrupos sea mínima, y que la variación entre-grupos sea máxima. (https://uam.es/departamentos/economicas/econapli/anova.pdf)

Por tanto, el análisis de la varianza se va a basar no sólo en la descomposición de la variación total, sino además en la comparación de la variación ENTRE-grupos y la variación INTRA-grupos, teniendo en cuenta sus correspondientes grados de libertad.

Se demuestra que:

$$E\left[\frac{VARIACIÓN\quad ENTRE\,/\,g-1}{VARIACIÓN\quad INTRA\,/\,n-g}\right] \approx F_{g-1,n-g}$$

Por tanto, un valor elevado de este cociente significará que mayores son las diferencias entre los distintos grupos (niveles del factor), cumpliéndose asimismo que la variación dentro de cada grupo sea mínima, y por tanto la probabilidad de que los niveles del factor influyan en los valores de la variable continua será mayor. (https://uam.es/departamentos/economicas/econapli/anova.pdf)

Dado que dicho cociente se distribuye como una F de S nedecor con g-1,n-g grados de libertad, el valor para el cual podremos asumir que sí existen efectos diferenciales entre los niveles dependerá del valor de tablas de la función F para un nivel de significación de al menos el 5%. Si el valor calculado es mayor que el valor de tablas significará que sí hay efectos diferenciales entre los grupos y por tanto aceptaremos la hipótesis de que existe dependencia entre las variables. (https://uam.es/departamentos/economicas/econapli/anova.pdf)

Por el contrario, si el valor calculado es inferior al valor de tablas de una $F_{g-1,n-g}$ aceptaremos que no existen efectos diferenciales entre los grupos, o en otras palabras:

$$t1 = t2 = \quad = tn = 0$$

Así, la hipótesis nula a contrastar a través del Análisis de la Varianza puede ser establecida como igualdad de efectos:

$$H0 = t1 = t2 = \quad = tg = 0$$

Siendo la hipótesis alternativa (H_1) que alguno de los efectos diferenciales sea distinto de cero.

Resumiendo, diremos:

Si $F > F_{g-1,n-g} \rightarrow H_1$ (Existen diferencias entre los tratamientos)

Si $F = F_{g-1,n-g} \rightarrow H_0$ (No existen diferencias entre los tratamientos)

En nuestro ejemplo los resultados de la aproximación tradicional se presentan en el cuadro adjunto. Recordemos que la fuente de variación "explicada" corresponde a la *entre* grupos y la "residual" a la *intra* grupos. Los grados de libertad correspondientes son:

$g - 1 = 2$ (g = 3 niveles)

$n - g = 9$ (n = 12 establecimientos)

Corregido por los grados de libertad podemos obtener por cociente el ratio F que en este caso es 2,492. Si comparamos este valor con el obtenido en las tablas, encontramos que para un 95% de probabilidad $F_\varepsilon = 4{,}26$ luego aceptaríamos la hipótesis nula de que todos los efectos del factor altura son iguales. (https://uam.es/departamentos/economicas/econapli/anova.pdf)

Tabla 2 Ejemplo de ANOVA

VARIACIÓN	SUMA DE CUADRADOS	GRADOS DE LIBERTAD	MEDIA CUADRÁTICA	F
ENTRE (Explicada)	142578.667	2	71289.333	2.492
INTRA (Residual)	257438.000	9	28604.222	
TOTAL	400016.667	11	36365.152	

3.2 UTILIZACIÓN DEL PROGRAMA SPSS

A continuación, se describirán cuáles son los pasos necesarios para realizar el Análisis de la Varianza utilizando la aplicación del SPSS para Windows. Para nuestra aplicación utilizaremos el ejemplo en el que se intenta determinar si el posicionamiento del producto influye o no en sus ventas, por lo que generamos una nueva variable que denominaremos posición y que diferencia los niveles del factor. (https://uam.es/departamentos/economicas/econapli/anova.pdf)

Tabla 3 Ejemplo de ANOVA en SPSS

Establecimiento	**Ventas**	**Posicionamiento del Producto**
A	663	1
B	795	1
C	922	1
D	1056	1
E	798	2
F	926	2
G	1060	2
H	1188	2
I	528	3
J	660	3
K	792	3
L	924	3

3.2.1 ANÁLISIS DE LA VARIANZA CON UN SOLO FACTOR

Opción recomendable cuando deseamos aplicar un Análisis de la Varianza en el que utilizamos un sólo factor como variable explicativa. Para ello, una vez abierto nuestro archivo de datos e introducidas las variables “posición” y “ventas”, nos introducimos en la opción de "Analizar" y pinchamos en “Comparar Medias”, seleccionando la opción "ANOVA de un factor" que nos permitirá realizar el Análisis de la Varianza. (https://uam.es/departamentos/economicas/econapli/anova.pdf)

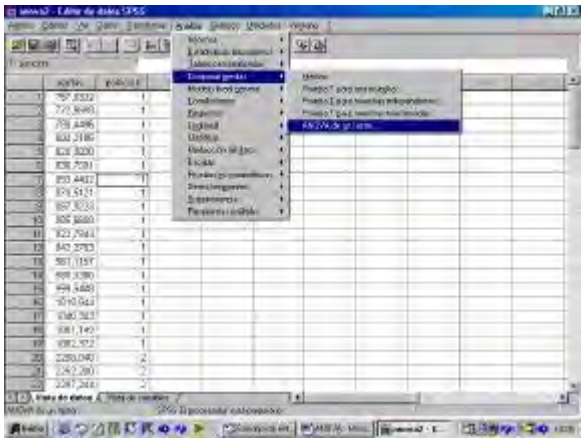

Ilustración 11 ANOVA de un Factor en SPSS

Una vez seleccionada esta opción aparece el cuadro de diálogo del Anova de un Factor, donde debemos especificar cuál es la variable dependiente (Ventas) y el Factor o variable independiente (Posición). Inicialmente las variables aparecerán en el cuadro blanco de la parte izquierda de la imagen; nosotros deberemos desplazar dichas variables a su casilla correspondiente utilizando los iconos de las flechas. En nuestro ejemplo deberemos introducir la variable "Ventas" en la casilla correspondiente a "Variables dependientes", y la variable "Posición" en la casilla que dice "Factor", tal y como se muestra en la imagen.

Ilustración 12 ANOVA de un Factor en SPSS

A continuación, podemos seleccionar una serie de opciones, pulsando en cada uno de los tres botones del cuadro de dialogo inicial (Contrastes, Post hoc y Opciones). Pulsando el botón Contrastes permite dividir la suma de cuadrados entre-grupos en componentes de tendencia o especificar contrastes a priori para que se contrasten mediante el estadístico t.

Cuando el ANOVA rechace la hipótesis nula (es decir cuando aceptemos la hipótesis de que los niveles del factor influyen sobre la variable endógena) será interesante realizar un análisis ex-post. Este tipo de análisis se basa en comparaciones múltiples por parejas entre las medias de los distintos grupos, para así conocer a qué grupos exactamente se deben las diferencias observadas entre ellos. El botón Post Hoc nos permite seleccionar distintas pruebas para realizar este tipo de análisis, entre las que se encuentran el test de la Diferencia Mínima Significativa (DMS), Bonferroni, Sidak, etc... (https://uam.es/departamentos/economicas/econapli/anova.pdf)

Ilustración 13 ANOVA de un Factor en SPSS

Pulsando el botón Opciones aparece la siguiente pantalla, cuyas distintas alternativas se explican a continuación:

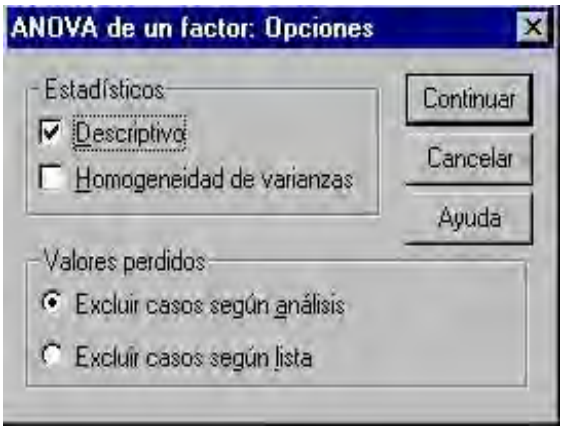

Ilustración 14 ANOVA de un Factor en SPSS

♦ **Descriptivos**: Muestra el número de casos, la media, la desviación típica, el error típico, los valores mínimo y máximo y el intervalo de confianza al 95% para cada variable dependiente en cada grupo.

♦ **Homogeneidad de varianzas:** Contrastan las violaciones del supuesto de igualdad de varianzas utilizando la prueba de Levene.

♦ **Excluir casos según análisis:** Excluye los casos que tienen valores perdidos en la variable implicada en esa prueba.

♦ **Excluir casos según lista:** Excluye los casos que tienen valores perdidos en cualquiera de las variables utilizadas en cualquiera de los análisis.

Una vez seleccionadas todas las opciones que consideremos necesarias para nuestro análisis ya estaremos en condiciones para realizar al análisis de la varianza (ANOVA), pulsando la tecla Aceptar. Los resultados del ANOVA aparecerán en el Navegador de resultados de SPSS. (https://uam.es/departamentos/economicas/econapli/anova.pdf)

A continuación, se muestran la salida de SPSS correspondiente al Análisis de la Varianza con los datos propuestos en el ejemplo habiendo seleccionado únicamente las opciones de Estadísticos descriptivos en el botón de Opciones:

3.2.2 ANOVA DE UN FACTOR

Descriptivos

Tabla 4 ANOVA de un Factor en SPSS

		N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
						Límite inferior	Límite superior		
VENTAS	POSICIONALTA	4	859,0000	168,6120	84,3060	590,7046	1127,2954	663,00	1056,00
	MEDIA	4	993,0000	168,3528	84,1764	725,1170	1260,8830	798,00	1188,00
	BAJA	4	726,0000	170,4113	85,2056	454,8416	997,1584	528,00	924,00
	Total	12	859,3333	190,6965	55,0493	738,1706	980,4961	528,00	1188,00

ANOVA

Tabla 5 ANOVA de un Factor en SPSS

		Suma de cuadrados	gl	Media cuadrática	F	Sig.
VENTAS	Inter-grupos	142578,67	2	71289,333	2,492	,138
	Intra-grupos	257438,00	9	28604,222		
	Total	400016,67	11			

La primera tabla muestra la media, la desviación típica, el error típico, y los valores máximo y mínimo para cada uno de los grupos. Los valores de esta tabla nos permiten conocer en qué posición sobre la estantería, las ventas del producto son mayores. Dados estos resultados se puede observar a primera vista que las ventas en la posición media son mayores que las ventas en las posiciones baja y alta, y que cuando el producto se coloca en la posición baja

las ventas del producto son las menores. (https://uam.es/departamentos/economicas/econapli/anova.pdf)

La siguiente tabla es la salida básica de un Análisis de la Varianza: a través de los datos que muestra podremos saber si realmente existe una relación de dependencia entre las variables objeto de estudio o no, podremos saber si los distintos niveles de la variable cualitativa o factor (posición del producto sobre la estantería) determinan el valor de la variable cuantitativa (ventas del producto). (https://uam.es/departamentos/economicas/econapli/anova.pdf)

Lo que en la tabla aparece como "Inter-grupos" es el valor de la VARIACIÓN ENTRE, y el valor de "Intra-grupos", es la VARIACIÓN INTRA. También aparece el valor de la VARIACIÓN TOTAL. A continuación, la salida muestra los grados de libertad, que para el caso de la "Variación Entre" son g – 1 = 2 y en el caso de la "Variación Intra" son n – g = 9. La columna "Media cuadrática" muestra los valores del cociente de la Variación Entre y la Variación Intra por sus correspondientes grados de libertad. Recordemos que cuanto más se aproximen la media cuadrática factorial (Variación Entre/g-1) y la media cuadrática residual (Variación Intra/n-g) mayor será la probabilidad de aceptar la hipótesis nula (H_0) o no influencia del factor. (https://uam.es/departamentos/economicas/econapli/anova.pdf)

Por último, la salida del SPSS nos muestra el valor calculado del estadístico F y su nivel de significación. El nivel de significación nos va a permitir aceptar o rechazar la hipótesis nula (independencia entre las variables) sin necesidad de tener que comparar el valor de la F con su valor real de las tablas estadísticas de una F de Snedecor. (https://uam.es/departamentos/economicas/econapli/anova.pdf)

El valor que nos sirve de referencia a la hora de aceptar o rechazar la hipótesis nula es el nivel de significación. Si el nivel de significación es mayor que 0,05, aceptaremos la hipótesis nula de independencia entre las variables (no existen efectos diferenciales entre los tratamientos). Si el nivel de significación es menor que 0,05 rechazaremos la hipótesis nula y aceptaremos la hipótesis alternativa, es decir, concluiremos que existe una relación de dependencia entre las variables, y en este caso podremos decir que los distintos niveles del factor sí influyen sobre los valores de la variable cuantitativa. El nivel de significación como

se expuso en el capítulo anterior es la probabilidad de rechazar la hipótesis nula siendo cierta. (https://uam.es/departamentos/economicas/econapli/anova.pdf)

En nuestro caso, dado que el valor del nivel de significación es 0,138 y este valor es mayor que 0,05 aceptaremos la hipótesis nula de que no existen efectos diferenciales entre los tratamientos. Esto querrá decir que la posición del producto sobre la estantería no hace que las ventas del mismo sean estadísticamente diferentes. (https://uam.es/departamentos/economicas/econapli/anova.pdf)

3.3 APLICACIÓN EN LA RESOLUCIÓN DE PROBLEMAS AMBIENTALES

En el trabajo de (Rodríguez-Rosales V. *et al*., 2005), modelación atmosférica de la calidad del aire en la ciudad de chihuahua, se emplean el uso de la varianza y covarianza y, describe; el estudio del comportamiento de los parámetros contaminantes CO, O_3, NO_2 y PM_{10}, utilizando una herramienta del análisis multivariado, conocida como análisis de componentes principales (PCA), en el que se han tomado en cuenta además factores meteorológicos tales como: velocidad y dirección del viento, temperatura atmosférica, humedad relativa, radiación solar, presión barométrica y precipitación pluvial. Se generó un modelo empírico que permite predecir los niveles de estos parámetros contaminantes a partir de datos históricos, y se identificaron las variables que contribuyen principalmente a la contaminación atmosférica. Se requiere usar mayor variabilidad en los datos para tener un mejor modelo capaz de predecir eficientemente el comportamiento de las variables medidas.

Tabla 6 Porciento de varianza capturada por el modelo ACP.

Componente principal	Eigenvalor de Cov(x)	% Varianza del PC	% Varianza total
1	3.42e+000	31.13	31.13
2	1.94e+000	17.66	48.79
3	1.50e+000	13.61	62.40
4	1.03e+000	9.41	71.81
5	7.66e-001	6.97	78.77
6	7.00e-001	6.37	85.14
7	5.76e-001	5.23	90.37
8	4.48e-001	4.07	94.44
9	2.80e-001	2.54	96.99
10	1.96e-001	1.78	98.76
11	1.36e-001	1.24	100.00

CAPÍTULO IV ANÁLISIS DE COMPONENTES PRINCIPALES

4.1 INTRODUCCIÓN

El análisis de componentes principales (ACP) es un método estadístico multivariado de simplificación o reducción de la dimensión de una tabla de casos o variables con datos cuantitativos, para obtener una tabla de menor número de variables, combinación lineal de las iniciales que se denominan componentes (CPs). La ACP reduce el número de variables en un conjunto de datos, conteniendo toda la información original, agrupando aquellas que fornecen informaciones semejantes y que están altamente correlacionadas. Su aplicación es directa sobre cualquier conjunto de variables, permite describir la estructura y las interrelaciones de las variables originales en los fenómenos en estudio a partir de los componentes obtenidos. La reducción de muchas variables a pocos componentes puede simplificar la ampliación sobre otras técnicas multivariadas (Pérez, C. 2001).

El principal uso del ACP está relacionado con la explicación de la estructura de varianzas y covarianzas de una serie de variables originales, mediante unas pocas combinaciones lineales de ellas, buscando generar nuevas variables que puedan expresar la información contenida en el conjunto original de datos y reducir la dimensionalidad del problema estudiado. Las nuevas variables generadas se denominan componentes principales. El análisis es esencialmente descriptivo y tiene una interpretación geométrica. (Lema, 1996, Smith, 2002)

4.1.1 OBTENCIÓN DE LAS COMPONENTES PRINCIPALES

Sea $\mathbf{X} = [X_1;:::;X_p]$ una matriz de datos multivariantes. Lo que sigue también vale si $\mathbf{X}$ es un vector formado por p variables observables. (Cuadras, 2017)

Las componentes principales son unas variables compuestas incorrelacionadas tales que unas pocas explican la mayor parte de la variabilidad de $\mathbf{X}$*:* (Cuadras, 2017)

De inicio

Las componentes principales son las variables compuestas

$$Y_1 = \mathbf{Xt}_1; Y_2 = \mathbf{Xt}_2;:::; Y_p = \mathbf{Xt}_p$$

(Cuadras, 2017)

Tales que:

1. *var*(Y_1) *es máxima condicionado a* $\mathbf{t'}_1\mathbf{t}_1 = 1$*:*
2. *Entre todas las variables compuestas Y tales que cov* ($Y_1;Y$) = 0*; la variable* Y_2 *es tal que var*(Y_2) *es máxima condicionado a* $\mathbf{t'}_2\mathbf{t}_2 = 1$*:*
3. *Si p* 3*; la componente* Y_3 *es una variable incorrelacionada con* $Y_1;Y_2$ *con varianza máxima.*
4. *Análogamente se definen las demás componentes principales si p*> 3.

Si $\mathbf{T} = [\mathbf{t}_1;\mathbf{t}_2;:::;\mathbf{t}_p]$ es la matriz *p x p* cuyas columnas son los vectores que definen las componentes principales, entonces la transformación lineal

X Y

Y = **XT**

Se llama transformación por componentes principales.

Teorema 5.1.1 *Sean* $\mathbf{t}_1;\mathbf{t}_2;:::;\mathbf{t}_p$ *los p vectores propios normalizados de la matriz de covarianzas* **S***,*

$\mathbf{St}_i = {}_i\mathbf{t}_i;\ \mathbf{t'}_i\mathbf{t}_i = 1;\ i = 1;:::;p:$

Entonces:

1. *Las variables compuestas* $Y_i = \mathbf{Xt}_i;i = 1;:::;p;$ *son las componentes principales.*
2. *Las varianzas son los valores propios de* **S**

   $\text{var}(Y_i) = x_i;\ i = 1;:::;p:$

3. *Las componentes principales son variables incorrelacionadas:*

   $\text{cov}(Y_i;Y_j) = 0;\ i \neq j = 1;:::;p:$

Demostración.: Supongamos $x_1 > > x_p > 0$. Probemos que las variables $Y_i = \mathbf{X}\mathbf{t}_i$, $i = 1;:::;p;$ están incorrelacionadas:

$$\begin{aligned} \operatorname{cov}(Y_i, Y_j) &= \mathbf{t}_i' \mathbf{S} \mathbf{t}_j = \mathbf{t}_i' \lambda_j \mathbf{t}_j = \lambda_j \mathbf{t}_i' \mathbf{t}_j, \\ \operatorname{cov}(Y_j, Y_i) &= \mathbf{t}_j' \mathbf{S} \mathbf{t}_i = \mathbf{t}_j' \lambda_j \mathbf{t}_i = \lambda_i \mathbf{t}_j' \mathbf{t}_i, \end{aligned}$$

$$\Rightarrow (\lambda_j - \lambda_i)\mathbf{t}_i' \mathbf{t}_j = 0, \Rightarrow \mathbf{t}_i' \mathbf{t}_j = 0, \Rightarrow \operatorname{cov}(Y_i, Y_j) = \lambda_j \mathbf{t}_i' \mathbf{t}_j = 0, \text{ si } i \neq j.$$

Además0-r, para $i = j;$ la varianza de Y_i es:

$$\operatorname{var}(Y_i) = \lambda_i \mathbf{t}_i' \mathbf{t}_i = \lambda_i.$$

Sea ahora $Y = \mathrm{P}^p_{i=1}\, a_i X_i = \mathrm{P}^p_{i=1}\, i Y_i$ una variable compuesta tal que ${}^p i{=}1\ {}^2 i = 1:$ Entonces:

$$\operatorname{var}(Y) = \operatorname{var}\left(\sum_{i=1}^{p} \alpha_i Y_i\right) = \sum_{i=1}^{p} \alpha_i^2 \operatorname{var}(Y_i) = \sum_{i=1}^{p} \alpha_i^2 \lambda_i \leq (\sum_{i=1}^{p} \alpha_i^2)\lambda_1 = \operatorname{var}(Y_1),$$

4.2 VARIABILIDAD EXPLICADA POR LAS COMPONENTES

Que prueba que Y_1 tiene varianza máxima.

Consideremos ahora las variables Y incorrelacionadas con Y_1: Las podemos expresar como:

$$Y = \sum_{i=1}^{p} b_i X_i = \sum_{i=2}^{p} \beta_i Y_i \text{ condicionado a } \sum_{i=2}^{p} \beta_i^2 = 1 \quad :$$

Entonces:

$$\operatorname{var}(Y) = \operatorname{var}\left(\sum_{i=2}^{p} \beta_i Y_i\right) = \sum_{i=2}^{p} \beta_i^2 \operatorname{var}(Y_i) = \sum_{i=2}^{p} \beta_i^2 \lambda_i \leq \left(\sum_{i=2}^{p} \beta_i^2\right) \lambda_2 = \operatorname{var}(Y_2),$$

Y por lo tanto Y_2 está incorrelacionada con Y_1 y tiene varianza máxima.

Si p 3; la demostración de que $Y_3;:::;Y_p$ son también componentes principales es análoga (Cuadras, 2017).

4.2.1 VARIABILIDAD EXPLICADA POR LAS COMPONENTES

La varianza de la componente principal Y_i es var(Y_i) = $_i$ y la variación total es tr$(\mathbf{S}) = \sum_{i=1}^{p} \lambda_i$ *:* Por lo tanto:

1. Y_i contribuye con la cantidad $_i$ a la variación total tr(**S**)*:*
2. Si $m<p;Y_1;:::;Y_m$ contribuyen con la cantidad$\sum_{i=1}^{m} \lambda_i$ a la variación total tr(**S**)*:*
3. El porcentaje de variabilidad explicada por las *m* primeras componentes principales es:

$$P_m = 100\frac{\lambda_1 + \cdots + \lambda_m}{\lambda_1 + \cdots + \lambda_p}.$$

En las aplicaciones cabe esperar que las primeras componentes expliquen un elevado porcentaje de la variabilidad total. Por ejemplo, si $m = 2 <p;$ y $P_2 = 90\%;$ las dos primeras componentes explican una gran parte de la variabilidad de las variables. Entonces podremos sustituir $X_1;X_2;:::;X_p$ por las componentes principales $Y_1;Y_2:$ En muchas aplicaciones, tales componentes tienen interpretación experimental.

Representación de una matriz de datos:

Sea **X** =[$X_1;:::;X_p$] una matriz *n p* de datos multivariantes. Queremos representar, en un espacio de dimensión reducida *m* (por ejemplo, *m* = 2), las las $\mathbf{x}_1', \mathbf{x}_2', \ldots, \mathbf{x}_n'$ de **X***:* Necesitamos introducir una distancia. (Cuadras, 2017)

5.3.1 *La distancia euclidea (al cuadrado) entre dos las de* **X**

$$\mathbf{x}_i' = (x_{i1}, \ldots, x_{ip}), \quad \mathbf{x}_j' = (x_{j1}, \ldots, x_{jp}),$$

Es

$$\delta_{ij}^2 = (\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j) = \sum_{h=1}^{p}(x_{ih} - x_{jh})^2.$$

La matriz Δ = (δ_{ij}) *es la matriz n x n de distancias entre las filas.*

Podemos representar las *n* las de **X** como *n* puntos en el espacio R^p distanciados de acuerdo con la métrica $_{ij}$*:* Pero si *p* es grande, esta representación no se puede visualizar. Necesitamos reducir la dimensión.

La variabilidad geométrica de la matriz de distancias es el promedio de sus elementos al cuadrado

$$V_\delta(\mathbf{X}) = \frac{1}{2n^2}\sum_{i,j=1}^{n}\delta_{ij}^2 \quad :$$

Si **Y** = **XT** *es una transformación lineal de* **X***, donde* **T** *es una matriz p m de constantes,*

$$\delta_{ij}^2(q) = (\mathbf{y}_i - \mathbf{y}_j)'(\mathbf{y}_i - \mathbf{y}_j) = \sum_{h=1}^{q}(y_{ih} - y_{jh})^2$$

Es la distancia euclidea entre dos las de **Y***: La variabilidad geométrica en dimensión m p es*

$$V_\delta(\mathbf{Y})_m = \frac{1}{2n^2}\sum_{i,j=1}^{n}\delta_{ij}^2(m) \quad :$$

4.3 REPRESENTACION DE UNA MATRIZ DE DATOS

Teorema 5.3.1 *La variabilidad geométrica de la distancia euclidea es la traza de la matriz de covarianzas*

$$V_\delta(\mathbf{X}) \;=\; tr(\mathbf{S}) = \sum_{h=1}^{p}\lambda_h \quad :$$

Demostración.: Si $x_1;:::;x_n$ es una muestra univariante con varianza s^2, entonces

$$\frac{1}{2n^2}\sum_{i,j=1}^{n}(x_i - x_j)^2 = s^2 \quad :$$

En efecto, si $\bar{x}$ es la media

$$\begin{aligned}\tfrac{1}{n^2}\sum_{i,j=1}^{n}(x_i-x_j)^2 &= \tfrac{1}{n^2}\sum_{i,j=1}^{n}(x_i-\overline{x}-(x_j-\overline{x}))^2\\ &= \tfrac{1}{n^2}\sum_{i,j=1}^{n}(x_i-\overline{x})^2+\tfrac{1}{n^2}\sum_{i,j=1}^{n}(x_j-\overline{x})^2\\ &\quad +\tfrac{2}{n^2}\sum_{i,j=1}^{n}(x_i-\overline{x})(x_j-\overline{x})\\ &= \tfrac{1}{n}ns^2+\tfrac{1}{n}ns^2+0=2s^2.\end{aligned}$$

(Cuadras, 2017)

Aplicando (5.3) a cada columna de **X** y sumando obtenemos

$$V_\delta(\mathbf{X})=\sum_{j=1}^{p}s_{jj}=\mathrm{tr}(\mathbf{S})$$

Una buena representación en dimensión reducida *m* (por ejemplo, *m* = 2) sera aquella que tenga máxima variabilidad geométrica, a n de que los puntos están lo más separados posible. Teorema 5.3.2 *La transformación lineal* **T** *que maximiza la variabilidad geométrica en dimensión m es la transformación por componentes principales* **Y** = **XT***; es decir,* **T** = [$\mathbf{t}_1$*;:::;*$\mathbf{t}_m$] *contiene los m primeros vectores propios normalizados de* **S***:* (Cuadras, 2017).

Demostración.: Utilizando (5.3), la variabilidad geométrica de **Z** = **XV***;* donde **V** = [$\mathbf{v}_1$*;:::;*$\mathbf{v}_m$] es *p m* cualquiera, es

$$V_\delta(\mathbf{Z})_m=\sum_{j=1}^{m}s^2(Z_j)=\sum_{j=1}^{m}\mathbf{v}_j'\mathbf{S}\mathbf{v}_j,$$

Siendo$s^2(Z_j)=\mathbf{v}_j'\mathbf{S}\mathbf{v}_j$ la varianza de la variable compuesta Z_j*:* Alcanzamos la máxima varianza cuando Z_j es una componente principal: $s^2(Z_j)$ *j:* As :

$$\text{máx}\, V_\delta(\mathbf{Y})_m=\sum_{j=1}^{m}\lambda_j$$
:

El porcentaje de variabilidad geométrica explicada por **Y** es

$$P_m = 100\frac{V_\delta(\mathbf{Y})_m}{V_\delta(\mathbf{X})_p} = 100\frac{\lambda_1 + \cdots + \lambda_m}{\lambda_1 + \cdots + \lambda_p}$$

Supongamos ahora $m = 2$*:* Si aplicamos la transformación, la matriz de datos **X** se reduce a

Tabla 7 Matriz de Datos Cuando m = 2

$$\mathbf{Y} = \begin{pmatrix} y_{11} & y_{12} \\ \vdots & \vdots \\ y_{i1} & y_{i2} \\ \vdots & \vdots \\ y_{n1} & y_{n2} \end{pmatrix}$$

Entonces, representando los puntos de coordenadas $(y_{i1};y_{i2});i = 1;:::;n;$ obtenemos una representación optima en dimensión 2 de las de **X***:* (Cuadras, 2017)

Número de componentes principales

En esta sección presentamos algunos criterios para determinar el número $m<p$ de componentes principales (Cuadras, 2017).

4.4 CRITERIO DEL PORCENTAJE

El número m de componentes principales se toma de modo que P_m sea próximo a un valor especificado por el usuario, por ejemplo, el 80%. Por otra parte, si la representación de $P_1;P_2;:::;P_k;:::$ con respecto de k prácticamente se estabiliza a partir de un cierto m, entonces aumentar la dimensión apenas aporta más variabilidad explicada. (Cuadras, 2017)

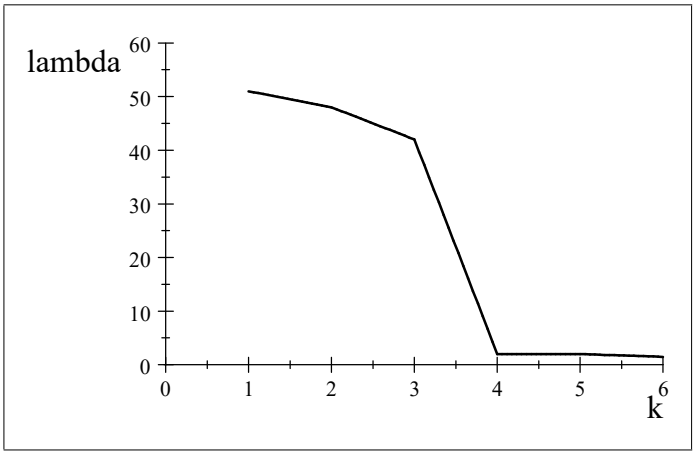

Ilustración 15 Representación de Valores Propios

En esta figura se ejemplifica la representación de los valores propios, que indicar a tomar las 3 primeras componentes principales.

4.5 CRITERIO DE KÁISER

Obtener las componentes principales a partir de la matriz de correlaciones **R** equivale a suponer que las variables observables tengan varianza 1. Por lo tanto, una componente principal con varianza inferior a 1 explica menos variabilidad que una variable observable. El criterio, llamado de Káiser, es entonces: (Cuadras, 2017)

Retenemos las *m* primeras componentes tales que $_m 1$*;* donde $_1 p$ son los valores propios de **R***;* que también son las varianzas de las componentes. Estudios de Montecarlo prueban que es más correcto el punto de corte x*= 0*:*7*;* que es más pequeño que 1. Este criterio se puede extender a la matriz de covarianzas. Por ejemplo, *m* podría ser tal que $x_{m\geq} v$*;* donde *v* =tra(**S**)/*p* es la media de las varianzas. También es aconsejable considerar el punto de corte 0.7 x *v*. (Cuadras, 2017)

4.5.1 TEST DE ESFERICIDAD

Supongamos que la matriz de datos proviene de una población normal multivariante $N_p(\mu, \Sigma)$. Si la hipótesis

$$H_0^{(m)} : \lambda_1 > \cdots > \lambda_m > \lambda_{m+1} = \cdots = \lambda_p$$

Es cierta, no tiene sentido considerar más de *m* componentes principales. En efecto, no hay direcciones de máxima variabilidad a partir de *m;* es decir, la distribución de los datos es esférica. El test para decidir sobre $H_0^{(m)}$ está basado en el estadístico ji-cuadrado y se aplica secuencialmente: Si aceptamos$H_0^{(0)}$ es decir, *m* = 1*;* todos los valores propios son iguales y no hay direcciones principales. Si rechazamos $H_0^{(0)}$, entonces repetimos el test con

$H_0^{(1)}$*:* Si aceptamos$H_0^{(1)}$ entonces *m* = 1*;* pero si rechazamos$H_0^{(1)}$ repetimos el test con$H_0^{(2)}$, y así sucesivamente. Por ejemplo, si *p* = 4*;* tendríamos que *m* = 2 si rechazamo$H_0^{(0)}$, $H_0^{(1)}$s y aceptamos$H_0^{(2)} : \lambda_1 > \lambda_2 > \lambda_3 = \lambda_4$*:* ((Cuadras, 2017))

4.5.2 CRITERIO DEL BASTÓN ROTO

La suma de los valores propios es V_t =tr(**S**)*;* que es la variabilidad total. Imaginemos un bastón de longitud V_t*;* que rompemos en *p* trozos al azar (asignando *p* 1 puntos uniformemente sobre el intervalo (0*;*V_t)) y que los trozos ordenados son los valores propios $l_1 > l_2 > \ldots > l_p$*:* Si normalizamos a V_t = 100*;* entonces el valor esperado de l_j es

$$E(L_j) = 100 \times \frac{1}{p} \sum_{i=1}^{p-j} \frac{1}{j+i}.$$

Las *m* primeras componentes son significativas si el porcentaje de varianza explicada supera claramente el valor de $E(L_1) + + E(L_m)$*:* Por ejemplo,

si *p* = 4*;* los valores son:

Porcentaje	$E(L_1)$	$E(L_2)$	$E(L_3)$	$E(L_4)$
Esperado	52*:*08	27*:*08	14*:*58	6*:*25
Acumulado	52*:*08	79*:*16	93*:*74	100

Si V_2 = 93*:*92 pero V_3 = 97*:*15*;* entonces tomaremos los dos componentes. (Cuadras, 2017)

4.6 BIPLOT

Un *biplot* es una representación, en un mismo gráfico, de las (individuos) y las columnas (variables) de una matriz de datos **X**(*n p*)*:*

Suponiendo **X** matriz centrada, el biplot clásico (debido a K. R. Gabriel), se lleva a cabo mediante la descomposición singular

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{V}0;$$

Donde **U** es una matriz *n p* con columnas ortonormales, **V** es una matriz *p p* ortogonal, y es una matriz diagonal con los valores singulares de **X** ordenados de mayor a menor. Es decir, $\boldsymbol{U}0\boldsymbol{U} = \boldsymbol{I}n; \boldsymbol{V}0\boldsymbol{V} = \boldsymbol{V}\boldsymbol{V}0 = \boldsymbol{I}p;$

=diag($_1$*;:::;* $_p$)*:* Como **X**0**X** = **U**0 ^{2}**U** vemos que **XV** = **U** es la transformación en componentes principales (5.1), luego las coordenadas para representar las *n* las están contenidas en **U** *:* Las coordenadas de las *p* columnas son las de la matriz **V***:* Filas y columnas se pueden representar (tomando las dos primeras coordenadas) sobre el mismo gráfico (Cuadras, 2017).

González, *et al*, 2013, realizó un trabajo en el Río Hondo de Quintana Roo, el cual transporta desechos agroquímicos y metales pesados que se vierten de las zonas cañeras y poblaciones de la zona fronteriza México-Belice. En este trabajo se determinó el contenido de mercurio (Hg), plomo (Pb), cadmio (Cd) y cinc (Zn) en sedimentos y plantas de lirio acuático (Nymphaea ampla) en cinco sitios de esta zona sobre la ribera de este Río para investigar si el contenido de dichos metales pudiera influenciar negativamente a las comunidades que consumen los alimentos faunísticos que dependen del Río, entre ellas, las poblaciones humanas ubicadas en la ribera del Río y que utilizan esta agua para preparar sus alimentos, actividades domésticas y para riego. Los resultados demuestran que Nymphaea ampla puede acumular altas cantidades de Hg y en menor grado Pb y Zn, por lo que puede ser una especie con potencial para la fitorremediación de aguas con alto contenido de estos metales. Además, debido a la cantidad de metales pesados adsorbidos en los sedimentos y ligeros cambios en la temperatura o pH, podría promover la desorción de estos compuestos de los sedimentos a la columna de agua, afectando la microbiota, fauna y demás comunidades que dependen del agua del Río.

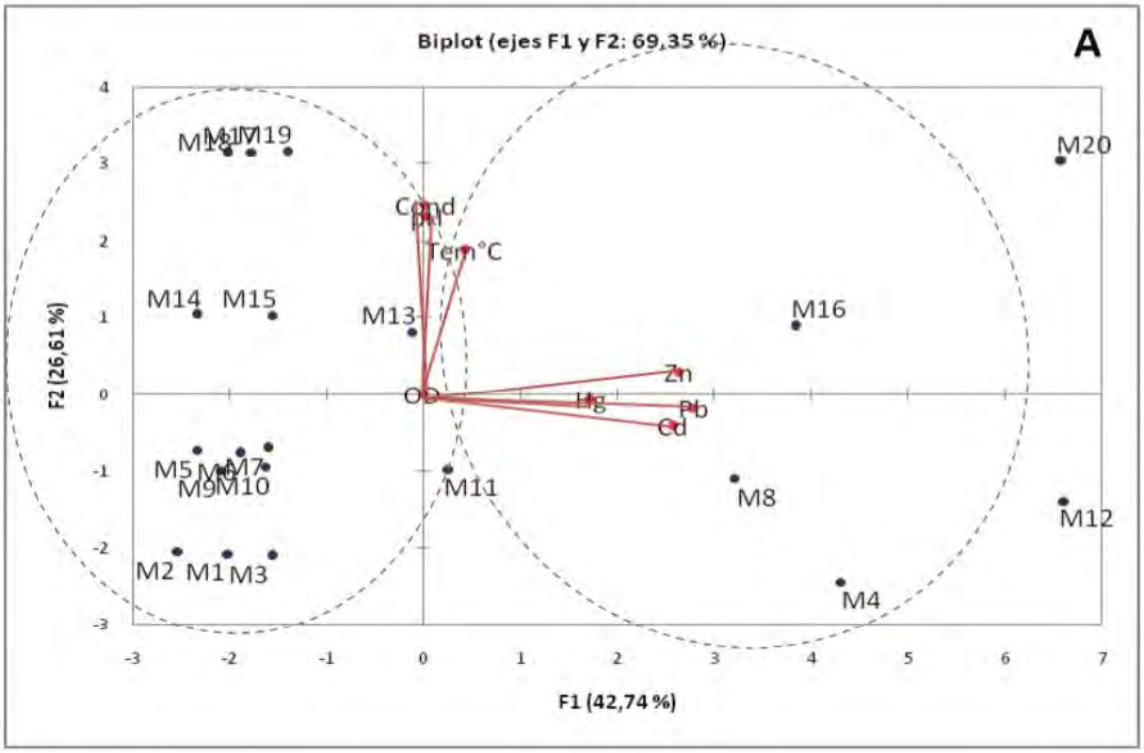

Ilustración 168 Representación de un ACP.

En esta figura observamos la aplicación del análisis de componentes principales ACP.

4.7 APLICACIÓN EN LA RESOLUCIÓN DE PROBLEMAS AMBIENTALES

Ávila Pérez H., *et al*., 2015, trabajó con; Análisis de Componentes Principales, como herramienta para interrelaciones entre variables fisicoquímicas y biológicas en un ecosistema léntico de Guerrero, México, en ello se aplicó el Análisis de Componentes Principales (ACP), con el objeto de conocer las interrelaciones entre las variables analizadas que determinan el grado de alteración del agua de la Laguna de Coyuca de Benítez, en Guerrero, México. Se consideraron variables fisicoquímicas y biológicas como: Temperatura, pH, Oxígeno disuelto, conductividad, sólidos totales disueltos, salinidad, diversidad, abundancia, clase 1, 2, 3 (requerimientos de oxígeno por insectos) y ordenes de insectos colectados. Los resultados permiten ilustrar las aplicaciones del ACP, que permitieron encontrar que las variables Temperatura, pH y Oxígeno disuelto presentaron mayores interrelaciones en este sistema lentico.

Para la obtención de los grupos de los sitios a partir de los promedios de valores de los factores fisicoquímicos, así como de la diversidad y abundancia, se aplicaron las técnicas de Análisis de Varianza de una Sola Vía (ANOVA), donde previamente se obtuvo la prueba de Homogeneidad de varianzas de Levene´s. Cuando las varianzas resultaron ser no homogéneas, se realizaron transformaciones utilizando logaritmos vulgares (log) y neperianos (Ln) para lograr la estabilización de las varianzas y definición de los grupos.

Tabla 8 Primer Análisis de Componentes Principales.

Variables	Componentes rotados		
	1	2	3
Temperatura	0.268	0.051	0.810
pH	0.464	-0.027	- 0.652
Oxígeno disuelto	0.165	0.540	-0.070
Conductividad	0.929	0.112	0.088
S.T.D.	0.936	0.036	-0.008
Salinidad	0.890	0.070	-0.009
Diversidad	-0.318	0.761	0.002
Abundancia	0.223	0.740	0.230

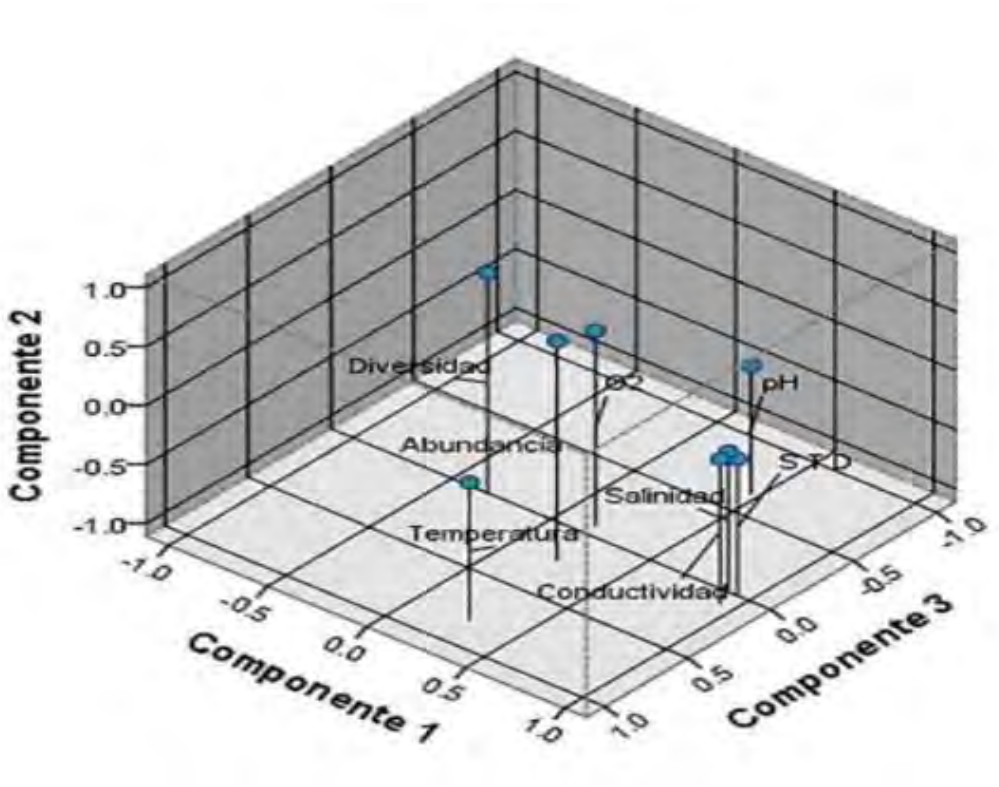

Ilustración 17. Gráfica de componentes rotados en tercera dimensión del primer ACP.

4.8 ANÁLISIS DE CLUSTER

El análisis de Clúster, tiene como objetivo proporcionar métodos cuya finalidad sea el estudio conjunto de datos multivariantes y se utiliza para clasificar los objetos o casos en grupos homogéneos llamados conglomerados con respecto a algún criterio de selección predeterminado.

Los conglomerados de objetos resultantes deberían mostrar un alto grado de homogeneidad interna (dentro del conglomerado) y un alto grado de heterogeneidad externa (entre conglomerados). Por tanto, si la clasificación es acertada, los objetos dentro de los conglomerados estarán muy próximos cuando se presenten gráficamente, y los diferentes grupos estarán muy alejados (Hair J.F. *et al*., 2000).

El valor teórico del análisis de clúster es el conjunto de variables que representan las características utilizadas para comparar objetos en análisis de clúster. Dado que el valor teórico del análisis de clúster incluye solo las variables utilizadas para comparar objetos,

determina el carácter de los objetos. El análisis de clúster es la única técnica multivariante que no estima el valor teórico empíricamente sino que utiliza el valor teórico especificado por el investigador. El análisis de clúster es comparable al análisis factorial en su objetivo de evaluar la estructura. Pero el análisis de clúster difiere del análisis factorial en que el análisis de clúster agrupa objetos, mientras que el análisis factorial se centra principalmente en la agrupación de variables. Un análisis de clúster es muy útil cuando el investigador desea desarrollar la hipótesis concerniente a la naturaleza de los datos o para examinar las hipótesis previamente establecidas (Hair J.F. *et al*., 2000).

Se utiliza la información de una serie de variables para cada sujeto u objeto y, conforme a estas variables se mide la similitud entre ellos. Una vez medida la similitud se agrupan en: grupos homogéneos internamente y diferentes entre sí. La "nueva dimensión" lograda con el cluster se aprovecha después para facilitar la aproximación "segmentada" de un determinado análisis. https://www.uam.es/personal_pdi/economicas/rmc/documentos/cluster.PDF

CONVIENE TENER CLARO DESDE EL PRINCIPIO:

♦ Que la técnica no tiene vocación / propiedades inferenciales

♦ Que, por tanto, los resultados logrados para una muestra sirven sólo para ese diseño (su valor atañe sólo a los objetivos del analista): elección de individuos, variables relevantes utilizadas, criterio similitud utilizado, nivel de agrupación final elegido.... definen diferentes soluciones.

♦ Que Cluster y discriminante no tiene demasiado en común: el discriminante intenta explicar una estructura y el Cluster intenta determinarla. https://www.uam.es/personal_pdi/economicas/rmc/documentos/cluster.PDF

OBJETIVOS BÁSICOS:

♦ Análisis "taxonómico" con fines exploratorios o confirmatorios.

♦ Cambio (simplificación) de la dimensión de los datos (*lo descrito al inicio de este documento: agrupación de objetos individuales en nuevas estructuras de estudio (grupales)(https://www.uam.es/personal_pdi/economicas/rmc/documentos/cluster.PDF).*

CLUSTER

(Objetivo) Una empresa desea clasificar a sus consumidores en "tipos" según sus distintas percepciones de determinados atributos de la marca: CALIDAD GLOBAL, NIVEL SERVICIO, PRECIO, SERVICIO POSTVENTA Y VARIEDAD. https://www.uam.es/personal_pdi/economicas/rmc/documentos/cluster.PDF

(Diseño) Para ello, se diseña una muestra con 100 compradores a los que cuestiona sobre su percepción, en una escala de intervalo, de las anteriores 5 características de los productos de la empresa. https://www.uam.es/personal_pdi/economicas/rmc/documentos/cluster.PDF

(Resultado) La idea final consiste en diseñar distintas estrategias de promoción en función de sus diversos perfiles, si es que estos existen. https://www.uam.es/personal_pdi/economicas/rmc/documentos/cluster.PDF

ETAPAS DE UN ANÁLISIS CLUSTER

1.-SELECCIÓN DE LA MUESTRA DE DATOS

2.-SELECCIÓN y TRANSFORMACIÓN DE VARIABLES A UTILIZAR

3.-SELECCIÓN DE CONCEPTO DE DISTANCIA O SIMILITUD Y MEDICIÓN DE LAS MISMAS

4.-SELECCIÓN y APLICACIÓN DEL CRITERIO DE AGRUPACIÓN

5.-DETERMINACIÓN DE LA ESTRUCTURA CORRECTA

Elección del número de grupos (https://www.uam.es/personal_pdi/economicas/rmc/documentos/cluster.PDF).

1.- SELECCIÓN DE LA MUESTRA

♦ Adecuar al máximo la muestra al objeto de análisis.

♦ Depuración de atípicos (interesan elementos como miembros de grupos, no interesa la excesiva "individualidad")

2.- SELECCIÓN DE VARIABLES

CANTIDAD

♦ No elegir variables indiscriminadamente: RECORDAMOS: cada estructura se manifiesta en una serie de variables y cada grupo de variables revela, sólo, una determinada estructura.

♦ Resultado muy sensible a la inclusión de alguna variable irrelevante.

♦ La inclusión indiscriminada de variables aumenta la probabilidad de atípicos.

¿TRANSFORMACIÓN?

♦ Depende / Afecta a muchas decisiones posteriores (medida de distancia / similitud empleada, por ejemplo).

♦ Estandarización por variable: aunque resulta útil para mediciones posteriores de distancia puede afectar al resultado del análisis y no se recomienda si las diferencias de medidas reflejan alguna cualidad natural de interés conceptual.

♦ Estandarización por encuestado: singular, pero en baterías de indicadores elimina patrones de respuesta en los sujetos, ofreciendo la importancia relativa de cada indicador.

♦ Factorización: puede resultar interesante factorizar previamente las variables y realizar el Cluster con factores en lugar de con variables.

♦ El tipo de escala de medida afectará a fases posteriores del procedimiento. (https://www.uam.es/personal_pdi/economicas/rmc/documentos/cluster.PDF).

3.- MEDIDAS DE SIMILITUD O DISTANCIA

TIPOS

A.- CORRELACIÓN: Se traslada el concepto tradicional de covariación, de conexión entre variables, de "pautas" de transición (por ejemplo, el cálculo de un coeficiente de correlación) aplicándolo a las observaciones de los sujetos como si fuesen observaciones de variables.

B.- Medidas de SIMILITUD / DISTANCIA: Definen proximidad, no Covariación, y su elección (tipos) viene determinada por la escala de medida de las variables: binaria u ordinal o de intervalo/razón.

♦ Medidas de distancia para escalas ordinales, de intervalo o razón; amplia variedad,

♦ Medidas de similitud para variables nominales binarias: reciben el nombre de medidas de asociación (https://www.uam.es/personal_pdi/economicas/rmc/documentos/cluster.PDF).

UNA ADVERTENCIA BÁSICA:

¡¡¡¡¡ El resultado final del Cluster depende radicalmente de la medida de ASOCIACIÓN / SIMILITUD / DISTANCIA utilizada. Se recomienda, en cada contexto, observar

empíricamente esas diferencias. !!!!!

https://www.uam.es/personal_pdi/economicas/rmc/documentos/cluster.PDF

ALGUNAS MEDIDAS DE DISTANCIA

EUCLÍDEA (para "t" variables)

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

♦ **Problemas con las unidades de medida:** normalización previa de variables recomendable. *Ojo: en SPSS obtenemos por defecto su cuadrado.*

MANHATTAN (o función de la distancia absoluta, o City-Block)

$$d_{ij} = \sum_{k=1}^{t} \left| X_{ik} - X_{jk} \right|$$

♦ **Problemas con la colinealidad.** *En SPSS esta medida aparece con el nombre de*

FORMULACIÓN GENERAL DE POWER (s,r)

$$d_{ij} = \left(\sum_{k=1}^{t} \left(X_{ik} - X_{jk} \right)^s \right)^{1/r}$$

♦ *En SPSS aparece como Power. Su variante más clásica es la de Minkowski (s=r).*

D ^{2} DE MAHALANOBIS

$$d_{ij} = \left(X_i - X_j \right)' \sum^{-1} \left(X_i - X_j \right)$$

Donde Xi y Xj son matrices fila (1 x p) de observaciones para cada sujeto y S es la matriz de varianzas - covarianzas de las variables consideradas.

♦ Dos ventajas de la D^2:

1.- **Se consigue mitigar el problema de las unidades en la medida** en que cada variable entra en el cálculo de distancia corregida por su variabilidad (*función del tamaño*).

2.- Se elimina la información redundante. La más correcta en caso de elevada multi - colinealidad.

ALGUNAS MEDIDAS DE ASOCIACIÓN

Tabla 9 Medidas de Asociación

	VARIABLE				
INDIVIDUO	**A**	**B**	**C**	**D**	**E**
I	1	0	0	1	1
J	1	1	0	1	1
K	0	1	1	0	1

Convenimos: (a) si los individuos I y J tienen la variable, (b) si el individuo I tiene la variable y J no, (c) el individuo J tiene la variable e I no, (d) los individuos I y J no tienen la variable y $p = a + b + c + d$

(* *) SEMEJANZA SIMPLE: (a+d) /p

JACARD: a / (a+b+c)

DICE: 2a/ (2a+b+c+)

(* *) RUSSELL Y KAO: a/p

ALGORITMO DE AGRUPACIÓN – DIVISIÓN PARA LA OBTENCIÓN DE CONGLOMERADOS

I. JERÁRQUICOS (ESTRUCTURA PROGRESIVA ARBOL)

I.A. JERÁRQUICOS AGLOMERATIVOS

I.A.1.- Distancia mínima (single linkage)

I.A.2.- Distancia máxima (complete linkage)

I.A.3.- Distancia entre centros (centroid)

I.A.4.- Distancia mediana (median)

I.A.5.- Distancia promedio

- simple (average linkage)
- entre grupos (between groups)
- intragrupos (within groups)

I.A.6.- Método de Ward

I.B. JERÁRQUICOS DIVISIVOS

I.B.1.- Por cálculo iterativo de centros

I.B.2.- Monothetic

I.B.3.- Polythetic

II. NO JERÁRQUICOS (K-MEDIAS):

II.B. UMBRAL SECUENCIAL II.C.- UMBRAL PARALELO III.D.- OPTIMIZACIÓN

MÉTODOS JERÁRQUICOS

♦ Definición: la agrupación se realiza mediante proceso un con fases de agrupación o desagrupación sucesivas. El resultado final es una jerarquía de unión completa en la que cada grupo se une o separa en una determinada fase. (https://www.uam.es/personal_pdi/economicas/rmc/documentos/cluster.PDF).

Método jerárquico aglomerativo:

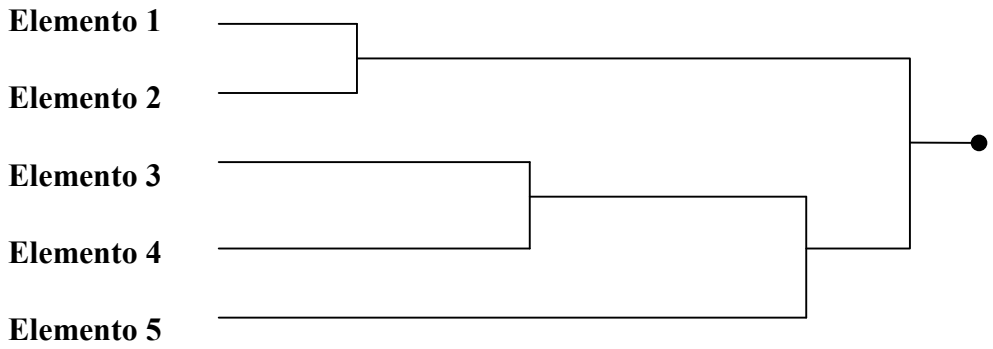

Ilustración 18 Método Jerárquico Aglomerativo

Método jerárquico divisivo:

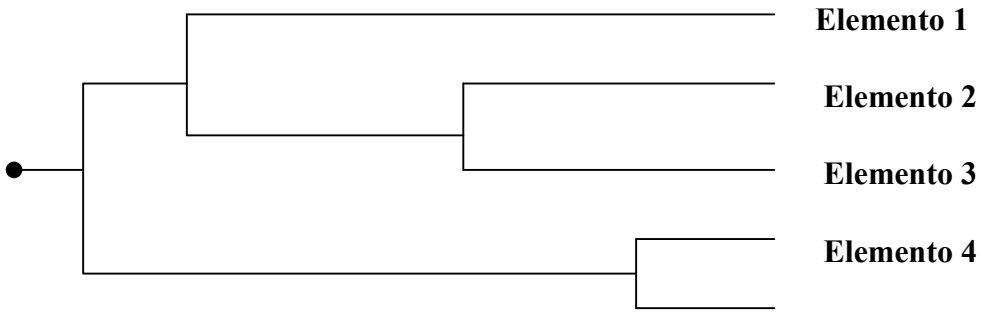

Ilustración 19 Método Jerárquico Divisivo

Elemento 5

DISTINTOS MÉTODOS AGLOMERATIVOS (Ejemplos)

♦ La selección de uno u otro método se basa en la forma en que la distancia se considera en el algoritmo de agrupación:

I.A.1.- Distancia mínima (single linkage)

Los grupos se unen considerando la menor de las distancias existentes entre los miembros más cercanos de distintos grupos. (https://www.uam.es/personal_pdi/economicas/rmc/documentos/cluster.PDF).

(Crea grupos más homogéneos pero permite cadenas de alineamientos entre sujetos muy lejanos)

I.A.2.- Distancia máxima (complete linkage)

Los grupos se unen considerando la menor de las distancias existentes entre los miembros más lejanos de distintos grupos. https://www.uam.es/personal_pdi/economicas/rmc/documentos/cluster.PDF

(Resuelve el anterior problema aunque los grupos son más heterogéneos)

I.A.6.- Método de Ward

IDEA BÁSICA: Se trata de ir agrupando de forma jerárquica elementos de modo que se minimice una determinada función objetivo.

FUNCIÓN A MINIMIZAR: Se perseguirá la minimización de la Variación Intra Grupal de la estructura formada. https://www.uam.es/personal_pdi/economicas/rmc/documentos/cluster.PDF

(Tiende a generar conglomerados demasiado pequeños y demasiado equilibrados en tamaño)

$$SCI = \sum_{k=1}^{h} SCI_K$$

Partiendo de "h" grupos y "m" variables:

Para cada grupo

$$SCI_K = \sum_{i=1}^{m} \sum_{j=1}^{n_k} \left(X_{ijk} - \overline{X}_{ik} \right)^2$$

Suma cuadrática intra del grupo **"k"**

Suma de desviaciones en todas las variables (m) para todos los sujetos (n_j) dentro del grupo "k".

Media de la variable "i" en el grupo "k"

Valor de la variable "i" para cada sujeto "j" perteneciente al grupo "k" .

Ilustración 20 Suma Cuadrática Intra

DISTINTOS MÉTODOS NO JERÁRQUICOS (Ejemplos)

II.B.- UMBRAL SECUENCIAL

Se seleccionan una tras otra, "semillas" de conglomerado agrupando en torno a ellas todos los objetos que caen dentro de una determinada distancia. Cada objeto ya asignado no se considera para posteriores asignaciones. (https://www.uam.es/personal_pdi/economicas/rmc/documentos/cluster.PD).

II.B.- UMBRAL PARELELO

Similar al anterior, pero se generan todas las semillas al mismo tiempo y los umbrales mínimas de aceptación en cada grupo. (https://www.uam.es/personal_pdi/economicas/rmc/documentos/cluster.PD).

III.D.- OPTIMIZACIÓN

Similares a los jerárquicos, pero no se clasifican como tales porque en las etapas sucesivas se permite la reasignación de sujetos. (https://www.uam.es/personal_pdi/economicas/rmc/documentos/cluster.PD).

NÚMERO ÓPTIMO DE GRUPOS

- **No existen criterios objetivos y ampliamente válidos.**
- **Hay una IDEA importante:** A medida que vamos formando grupos estos son menos homogéneos (las distancias para las que se forman los grupos iniciales son menores que las de los grupos finales). Pero la estructura es más clara.
- **Por tanto, podemos fijar un OBJETIVO:** Identificar el punto de equilibrio entre la estructura incompleta y la estructura mezclada o confusa.
- **No obstante, tenemos un problema.....:** Es difícil definir conceptualmente y más aun estadísticamente la situación de estructura correcta, no confusa, o la contraria de falta de estructura. (Estructura por asociación o diferenciación)
- **NOS APOYAREMOS, PARA DEFINIR LA ESTRUCTURA,** en la observación, tanto de las variables iniciales, como de la definición inicial de los sujetos y el significado de cada una de las etapas del proceso de agrupación.
- **Podemos, además, utilizar alguna herramienta técnica:** discriminante, caída brusca en la similitud o en la homogeneidad, dendograma. (https://www.uam.es/personal_pdi/economicas/rmc/documentos/cluster.PDF).

<u>TÉCNICAS DE AYUDA PARA DETERMINAR LA AGRUPACIÓN ÓPTIMA</u>

- **Observación de la variación intragrupal**

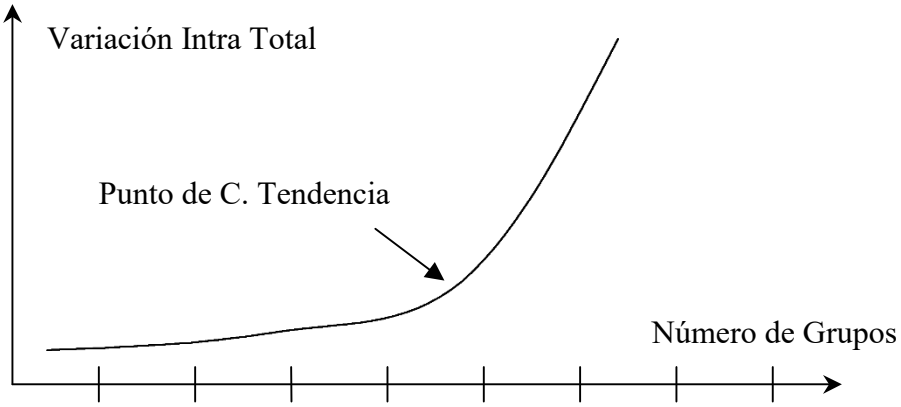

Ilustración 21 Observación de la Variación intergrupal

…….. 7 6 5 4 3 2 1

- **Dendograma**

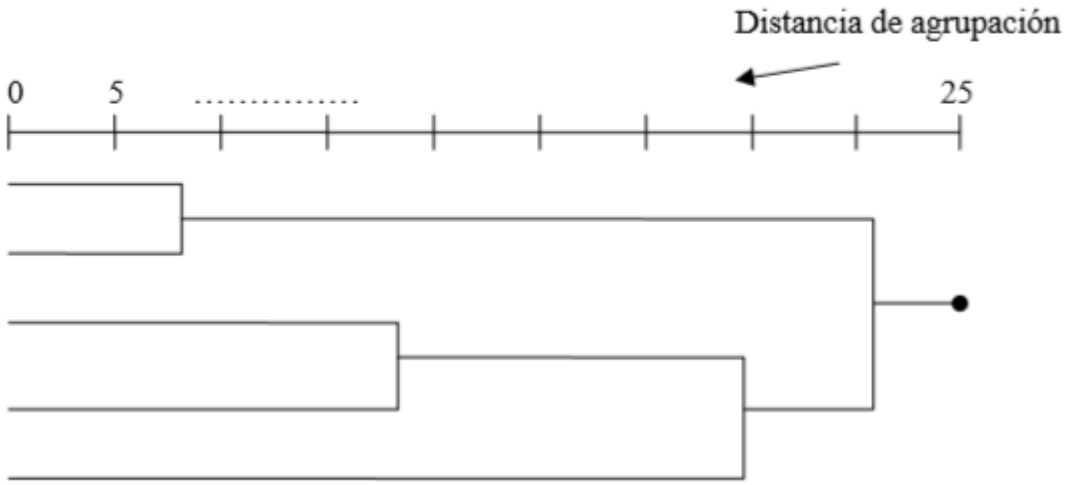

Ilustración 22 Dendograma

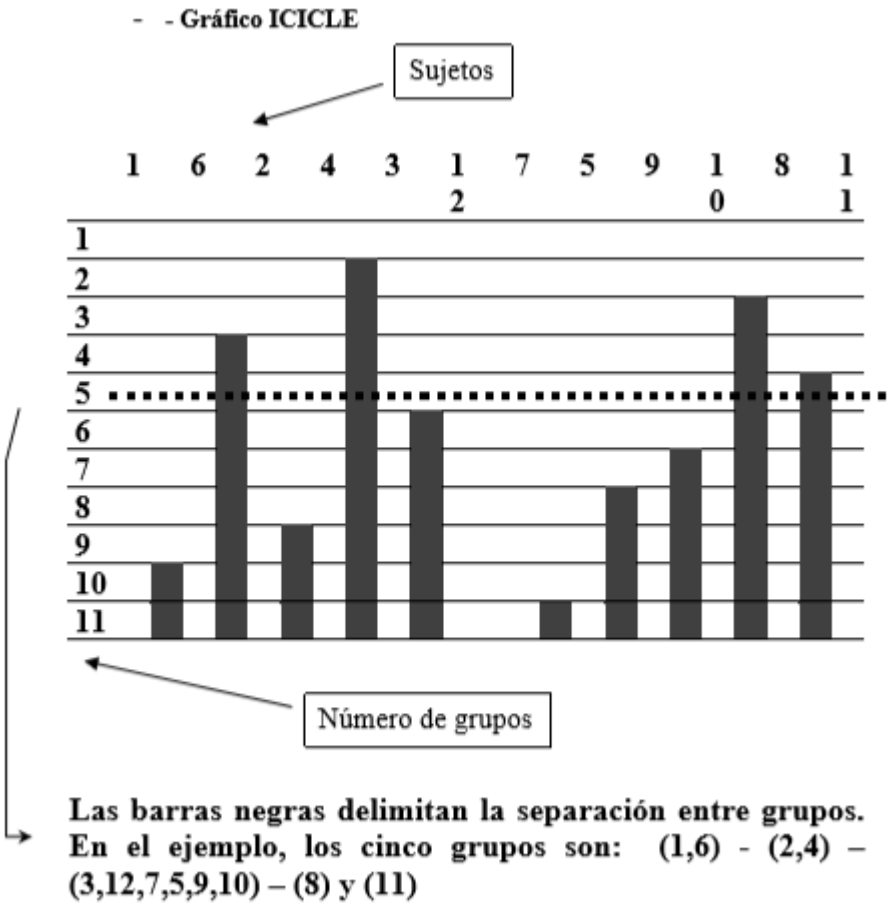

Ilustración 23. Grafico ICICLE

4.9 APLICACIÓN EN LA RESOLUCIÓN DE PROBLEMAS AMBIENTALES

El autor (Herrera Murillo J. *et al*., 2009), en "Aplicación de técnicas quimiométricas para clasificar la calidad de agua superficial de la microcuenca del río Bermúdez en Heredia, Costa Rica" aplicó técnicas quimiométricas: análisis de cluster, análisis de componentes principales y análisis de factores, para clasificar la calidad del agua de los ríos y evaluar datos de contaminación. En este trabajo se monitorearon 14 parámetros fisicoquímicos en 10 estaciones localizadas en la microcuenca del río Bermúdez de agosto de 2005 a febrero de 2007. Los resultados permitieron determinar la existencia de dos clusters naturales de sitios de monitoreo con características similares de contaminación e identificar la DQO, DBO,

NO3 -, SO4 -2 y SST, como las principales variables que discriminan entre los sitios de muestreo.

Análisis de cluster, es una herramienta que se utiliza para agrupar objetos (sitios de monitoreo) en clases (clusters), sobre la base de las similitudes entre los miembros de una misma clase y las disimilitudes entre los diferentes grupos. Los resultados de este tipo de análisis ayudan en la interpretación de los datos e indican patrones de comportamiento. En el agrupamiento hierarchical, los clusters se forman secuencialmente iniciando con los pares de objetos más similares para luego formar clusters más grandes paso a paso (Einax, 1992).

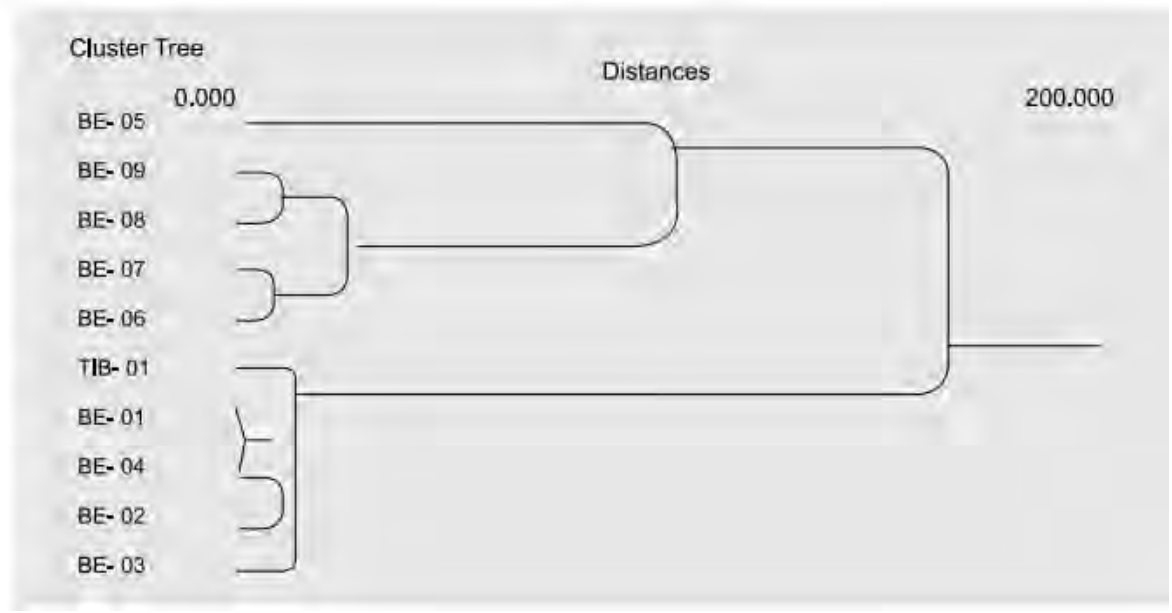

Ilustración 24. Dendrograma obtenido a partir del análisis de clúster aplicado a los resultados del monitoreo de parámetros químicos en la microcuenca del río Bermúdez.

CONCLUSIONES

Estadística es el estudio de los métodos para coleccionar, resumir, organizar, presentar y analizar información de datos. El término estadística también se refiere a la derivación de conclusiones válidas y a la formación de decisiones razonables. En la colección de datos de un grupo de observaciones, a menudo es imposible o impráctico observar toda la población.

De manera qué, en lugar de examinar el grupo en su totalidad, llamado la población o universo, es conveniente examinar solamente una parte de la población llamada muestra. Por ello, importante la reducción de los datos en una muestra infinitesimal o muy grande.

El uso de la estadística se ha extendido, no tan solo a las áreas tradicionales universitarias o escolásticas, sino también a todos los campos de la ingeniería, la agricultura, la biología, la química, en las encuestas políticas, entre otros. A través del tiempo, la estadística ha logrado un gran desarrollo en toda su metodología lo cual la hace una herramienta muy útil y esencial en cualquier área de estudio, además brinda un método práctico de organizar y analizar investigaciones, logrando de esta manera una mayor eficacia en un desempeño profesional.

Los fenómenos con los cuales trabajará el futuro Ingeniero, existe algún grado de incertidumbre; por ejemplo, demandas de tráfico máximo, la contaminación ambiental de un cuerpo de agua, precipitación anual máximo, precipitación pluvial anual, resistencia del acero, son fenómenos que no siempre tendrán exactamente los mismos valores observados, aún bajo condiciones aparentemente idénticas.

Es de suma importancia reconocer el alcance que ha tenido la estadística como herramienta principal en la resolución de problemas de tipo ambiental, ya que esta es la base fundamental para poder definir e inclusive, en algunos caso, resolver o ser parte de la solución de estos mismos.

A lo largo de la presente investigación fuimos dando cuenta como cada una de las características que engloban esta herramienta son indispensables para dar solución, de algunas interrogantes que se van dando en torno a cada problemática desde la importancia de la recolección de una muestra, la valoración de los datos obtenidos y la interpretación de los mismos.

Día a día desde la implementación en la resolución de problemas de todo tipo hasta nuestros tiempos hemos observado la evolución que esta ha tenido dentro de cada uno de sus ámbitos, no debemos perder de vista el crecimiento que se ha tenido también en la implementación de

las nuevas tecnologías, estas últimas han facilitado mediante programas como SPSS, XLSTAT y los sistemas de información geográfica (SIG), el procesamiento de grandes volúmenes de datos en menor tiempo.

BIBLIOGRAFÍA

Crespo, L.F. (2017). Componentes Principales. 2017, de Fuente Rebollo obtenido de: http://www.fuenterrebollo.com/Master-Econometria/Componentes_Principales.pdf

Cuadras, C.M. (2017). *Universitat de Barcelona.* Obtenido de http://www.ub.edu.

Einax J.E. Multivariate data analysis in environmental analytical chemestry. GIT-Fachz-Lab, 36(8):815, 1992.

Fondo Social Europeo. (2017). *Universidad Autonoma de Madrid.* Obtenido de https://www.uam.es.

González, J.L.B. Et. Al. (2013). Evaluación de la Acumulación de Hg, Pb, Cd y Zn en Sedimentos y Lirio Acuático (Nymphaea ampla) en el Río Hondo de Quintana Roo. TECNOCULTURA, 30, 24 - 32.

Grisa, A.M.C., Et Al. (2010). Análisis Multivariado de Parámetros Fisicoquímicos del Relleno Sanitario de São Giacomo de Caxias do Sul, RS en la de Degradación de Polipropilene. Polimeros: Ciencia y Tecnologia, 20, 1 - 6.

Hair, J.T., Et. Al. (2014). Multivariate Data Analysis. Inglaterra: Pearson.

Herrera, S.R., Et. Al. (2010). Estudio estadístico de la correlación entre contaminantes atmosféricos y variables meteorológicas en la zona norte de Chiapas, México. Junio 2018, de división académica de ciencias biológicas. UJAT obtenido de: www.ujat.mx/publicaciones/uciencia.

Lema, T. A. (1996). Borrador para algunos elementos de estadistica multivariada, Diseño de Muestreo, Universidad Nacional de Colombia, Medellín, Colômbia.

Murillo, J.H., Et. Al. (2009). Aplicación de técnicas quimiométricas para clasificar la calidad de agua superficial de la microcuenca del río Bermúdez en Heredia, Costa Rica. Tecnología en Marcha, 22, 1 – 11.

Otero, J.V., Et. Al. (2017). *Universidad Autonoma de Madrid.* Obtenido de http://www.uam.es.

Pérez, C. (2001). Técnicas Estadísticas con SPSS", Prentice Hall, Madrid.

Pérez, H.A., Et. Al. (2015). Análisis de Componentes Principales, como herramienta para interrelaciones entre variables fisicoquímicas y biológicas en un ecosistema lentico de Guerrero, México. Iberoamericana de Ciencias, 2, 1 – 11.

Rosales, V.R., Et. Al. (2006). Modelación atmosférica de la calidad del aire en la ciudad de Chihuahua. Revista Mexicana de Ingeniería Química, 5, 1 - 8.

Smith, L. I. (2002). “A tutorial on principal components analysis”, University of Otago, Dunedin.

Universidad Nacional Autonoma de Mexico. (2017). *Facultad de estudios superiores cuautitlan.* Obtenido de http://www.cuautitlan.unam.mx.

ÍNDICE DE ILUSTRACIONES

ÍNDICE DE TABLAS