**CAPITULO 2**

**2. MARCO TEÓRICO**

Para poder obtener un análisis eficiente e importante, debemos antes agrupar un marco teórico el cual nos servirá para facilitar el proceso del reconocimiento de la afección de las enfermedades del camarón y su propagación, tanto en la zona costera de la Provincia del Guayas, como en el interior y zonas aledañas.

Empecemos puntualizando algunas definiciones que forman la base de la teoría a ser utilizada:

**2.1. Análisis Multivariado**

El análisis Multivariado es el conjunto de técnicas estadísticas que nos sirven cuando tenemos muchas variables de por medio y queremos explicar su variación o el grado en que un grupo de variables sigue un cierto comportamiento.

**2.1.1. Matriz de datos**

Sea **Χ** una matriz determinadas por n filas, las cuales corresponden al total de unidades investigadas y p columnas, donde se encuentran el número de variables que se investigan.

**Χ = **

La matriz de datos **Χ**, entonces, contiene los datos que consisten en todas las observaciones, en todas las variables.

**2.1.2. Vector de Medias**

Sea una matriz de 1 columna y p filas, denotaremos al vector que contiene los valores esperados o medias de las variables que se investigan, como **vector de medias** 



**2.1.3. Matriz de Varianza y Covarianzas**

Sea una matriz simétrica con p filas y p columnas, cuya diagonal principal se encuentren las varianzas de cada una de las variables que se investigan, y en la posición *i,j* obtenemos las covarianzas entre la i-ésima y la j-ésima variable.

****

**2.1.4. Matriz de Correlaciones**

Sea una matriz *p* donde se obtienen las correlaciones entre las p variables que se investigan, y esta a su vez tiene las mismas propiedades de la matriz de varianzas y covarianzas.

****

**2.1.5. Análisis de Componentes Principales**

**2.1.5.1. Características**

El **Análisis de Componentes Principales (ACP)** se utiliza para describir una matriz ***R*** de variables continuas del tipo individuos por variables. Es decir, una matriz que recoge el valor que toman cada una de las variables *j*, donde *j = 1,..., p* en cada uno de los individuos u observaciones *i*, *i= 1,...,n.*

**R=**

Las variables figuran en columnas y los individuos, en filas. Estos pueden ser individuos encuestados, observaciones, etc.

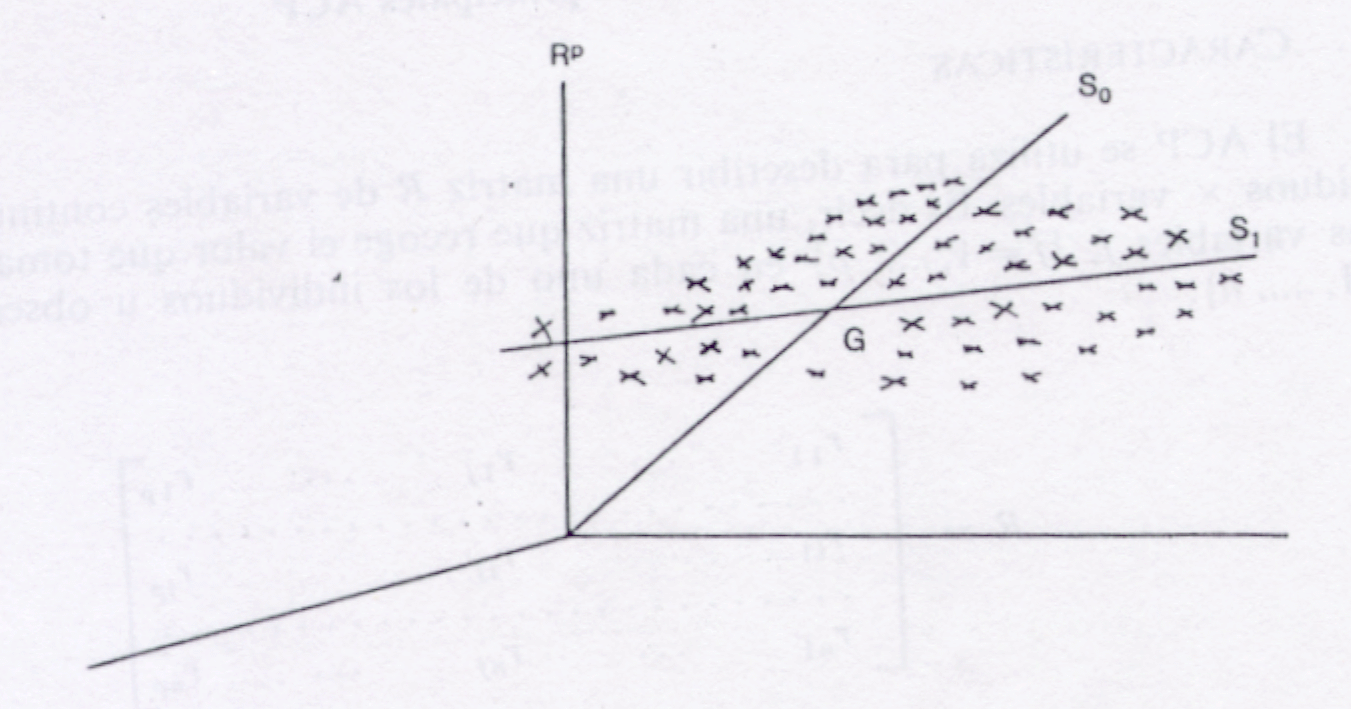
Esta matriz puede ser muy disimétrica, y las variables, muy heterogéneas, tanto en media como en desviación.

**2.1.5.2. Análisis en Rp**

Para evitar que variables que toman valores muy altos tengan un peso muy importante en la determinación de los ejes, se realiza una transformación consistente en centrar los datos de la siguiente forma:



donde  es la media de la variable j. De esta manera se elimina la influencia del nivel general de las variables. Gráficamente, podemos comprobar la conveniencia de realizar esta operación. Supongamos que la representación es la Fig. 2.1



***Figura 2.1: Nube de puntos***

***Fuente: “Métodos multivariantes para el análisis comercial”. I. Grande - E. Abascal.***

Si buscamos el subespacio de dimensión reducida que, pasando por el origen, represente bien la nube de puntos, S0, no obtendremos una buena representación. Se produce entonces una deformación fuerte al proyectar los puntos individuo sobre S0. Es evidente que en un caso como la figura 2.1 esto se lograría mejor y obtendríamos una representación más fiel sobre *S1*, que no pasa por el origen sino por el centro de gravedad *G.*

Si además las dispersiones de las variables son muy diferentes, se hará necesario realizar otra transformación en los datos de partida.

Se tipifican dividiendo por su desviación:



Así, al calcular las distancias entre los individuos



al estar divididos los términos de sumatorio por la desviación correspondiente de la variable, se reduce el efecto sobre la distancia de las variables que tienen mucha dispersión. Todas las variables tendrán una contribución semejante a la determinación de las proximidades. No hay variables que por ser muy dispersas contribuyan más al cálculo de las distancias. El coeficiente  se introduce en la transformación con el objeto de que la matriz X’X coincida con la matriz de correlación *C.*



Una vez realizada la transformación de la tabla, se analiza mediante el AG la nueva tabla *X*. El análisis consistirá en obtener los vectores propios de la matriz de correlación *C=X´X*. Las proyecciones de los individuos sobre los ejes dirigidos por estos vectores propios son los componentes principales; se obtienen mediante



Este factor es una variable artificial combinación lineal de las variables iniciales y se denomina **componente principal**.

**2.1.5.3. Análisis en Rn**

La transformación realizada en la tabla de datos produce efectos diferentes en este espacio. En Rn la transformación produce una deformación de la nube de puntos.

El cambio de escala de cada variable sitúa todos los puntos variables a la distancia 1 del origen.



En efecto:

La distancia entre 2 puntos variables en el espacio Rn



donde  es el coeficiente de correlación entre las variables j y j´. Así las proximidades entre los puntos variables se pueden interpretar en términos de correlación.

Si dos variables están correlacionadas muy positivamente, la distancia es nula . Si están correlacionadas negativamente , la distancia es máxima . Si están incorrelacionadas , la distancia es intermedia.

Para obtener los factores no es necesario diagonalizar la matriz *XX´.* Los vectores propios de *XX´* asociados a los valores propios se obtienen a partir de las de *X´X* mediante



La proyección de los puntos variables sobre el eje α vendrá dada por el vector



La coordenada de un punto variable j sobre el eje factorial α es el coeficiente de correlación entre la variable j y el factor α,

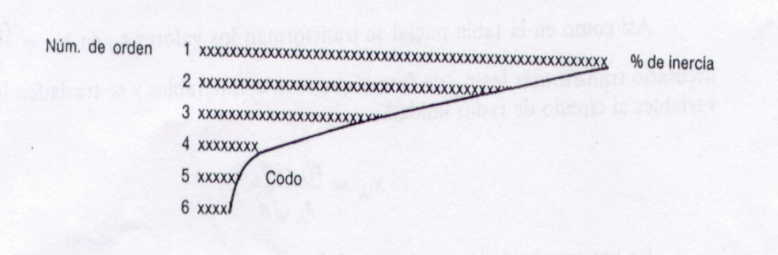


Esta expresión para las coordenadas se obtiene sustituyendo en la anterior  por su expresión en la fórmula que relaciona los dos espacios .

**2.1.5.4. Elección del número de ejes**

Existen distintos métodos para elegir el número de factores que se han de retener, aunque ninguno de ellos es determinante.

1. Si la nube inicial no tiene ninguna dirección privilegiada, los valores propios serán próximos: no diferirán mucho. Se conservará entonces un eje cuyo porcentaje de varianza sea netamente superior a , ya que éste es el valor que le correspondería si todos los ejes explicasen exactamente la misma cantidad. Se puede decidir conservar los ejes que restituyan una parte de la inercia superior a  o .
2. La inercia explicada por cada eje sucesivo debe ser decreciente. Si representamos el histograma de los valores propios con los números de los ejes en ordenadas, y los porcentajes de inercia explicada en abscisas, se pueden eliminar los ejes cuyo número de orden es posterior al “codo” que se produce en la curva, así en el ejemplo del gráfico nos quedaríamos con los cuatro primeros ejes.



***Figura 2.3: Curva de los***

***Componentes Principales***

***Fuente: “Métodos multivariantes para el análisis comercial”. I. Grande - E. Abascal.***

En la práctica, estas reglas dan el mismo resultado y la primera es mucho más fácil.

**2.2. Series de Tiempo**

Una **serie temporal**, llamada también *serie histórica o serie cronológica*, es una sucesión de valores observados de una variable referida a períodos de tiempo generalmente regulares. El análisis univariante de una serie temporal consiste en hacer uso de estos datos para elaborar un modelo que describa adecuadamente el comportamiento de esta variable en el pasado, y permita realizar **predicciones** satisfactorias.

Para poder encontrar el modelo adecuado para nuestro conjunto de datos, primero tenemos que conocer definiciones básicas de la teoría de series de tiempo.

**2.2.1. Procesos Estocásticos**

Un proceso estocástico *X* es una familia o colección de variables aleatorias, notado por:

## 

es decir, que para cada valor de *t* que pertenece a T se tiene que *X(t)* es una variable aleatoria. El subíndice *t* puede ser interpretado como el tiempo *X(t)* también se denomina *"estado de proceso al tiempo t".* A menudo a *X(t) se* le nota por *Xt*. Una realización del proceso estocástico X se denomina una *trayectoria o cambio muestral*.

Los elementos más importantes en un proceso estocástico son:

1.- **El espacio de estados :** Es el conjunto donde las v.a. *X(t)* toman sus valores.

2.- **El conjunto índice T:** Si T es un conjunto contable (como los números naturales, etc.) se dice que el proceso estocástico *X* es **discreto***.* Si T es igual a los números reales o T es igual a algún subconjunto del mismo, se dice que *X* es un proceso estocástico **continuo***.*

En nuestro análisis este tipo de procesos nos va a ser de mucha ayuda, para así obtener proyecciones a futuro de las enfermedades a través de una **serie temporal** o datos en el tiempo.

### 2.2.2. Procesos Estacionarios con incrementos independientes

### Sí para todo v.a.



son independientes, se dice que *X* es un *proceso* **estacionario con incrementos independientes.**

Si además las v.a.



tienen la misma distribución de probabilidad, se dice que *X* es un proceso estocástico estacionario con incrementos independientes, es decir, en un proceso es un proceso estacionario la distribución de probabilidad de sólo depende de s y no del valor de t.

**2.2.3. Procesos Estrictamente Estacionarios**

Un proceso estocástico  donde T puede ser los reales, , los naturales o Z, se dice *estacionario en sentido estricto (o estrictamente estacionario)* si las funciones de distribución conjuntas de las familias de v.a.

### 

### son las mismas para todo h>0 y para valores arbitrarios de t1, t2,..., tn los cuales pertenecen a T.

Esta condición establece la existencia de un cierto equilibrio probabilístico y que los *"tiempos"* (particulares) en que examinamos el proceso no tienen ninguna relevancia. Es más, nos asegura que la distribución de *Xt* es la misma para todo .

**2.2.4. Procesos Estacionarios en Covarianza**

Un proceso estocástico  se dice **estacionario en sentido amplio o estacionario en Covarianza** si el proceso posee:

1.- Una *media* constante: 

2.- Momentos de segundo orden finitos: 

3.- Una Covarianza



que depende solamente de la diferencia de tiempo *.*

Un proceso estocástico en sentido estricto que tienen momentos de segundo orden finitos es un **proceso estacionario en sentido amplio**.

* + 1. **Autocovarianza y Autocorrelación**

Para un proceso estacionario (con la condición débil), se define la **Autocovarianza de** **orden h** como



donde  es un proceso retrocedido h pasos y 

Se define la **Autocorrelación de orden h** como



teniendo en cuenta que .

El valor de este coeficiente está acotado por



* + 1. **Ecuaciones de diferencias de primer orden**

##### Consideremos la **ecuación de diferencias de primer orden**



Esta se resuelve mediante un método recursivo



Los comportamientos posibles de la solución son:

* Con el valor  la solución es explosiva.
* Con el valor  la solución es explosiva y presenta oscilaciones.
* Con el  la solución decae exponencialmente.
* Con el valor  la solución decae exponencialmente con oscilaciones.

**2.2.6.1. Diferenciación**

Muchas situaciones de no estacionariedad en media pueden resolverse satisfactoriamente diferenciando los datos. Se dice que las series que tienen esta propiedad exhiben una no estacionariedad homogénea, de forma que una serie es homogénea de grado d si la serie transformada: es estacionaria en media.

**2.2.7. Ruido Blanco**

Dada la secuencia de variables aleatorias  a esta se la denomina **procesos de ruido blanco.** El proceso de ruido blanco  es supuesto a una transformación del proceso , el cual es estacionario en sentido débil y además.

El interés de este proceso radica en que, si sólo se cuenta con información muestral acerca de su propio pasado, no puede realizarse ninguna previsión mejor que su esperanza incondicional. Por tanto, los procesos de ruido blanco de esperanza nula resultan útiles para caracterizar las propiedades ideales del término de error de un modelo estocástico dinámico.

**2.2.8. Procesos Autoregresivos de orden p o AR(p)**

Un modelo estocástico que puede ser sumamente útil en la representación de cierta serie es el **Modelo Autoregresivo**.

En este modelo, el valor actual del proceso se expresa como un agregado finito, lineal de valores anteriores del proceso y el error .

Dado que  no son independientes, entonces



es un **Proceso Autoregresivo de orden p o AR(p),** en donde es el **ruido blanco** del proceso**.** Se describe en forma funcional como



donde  y es inversible si sus raíces son en modulo mayor que 1 ( ).

El vector de retrasos es



y el vector de parámetros es



**2.2.8.1. Ecuaciones de Yule-Walker**

###### Desarrollado la expresión anterior para *k =1,2,...,p*



y teniendo en cuenta que y que , se obtiene:

****

que se conoce como el **sistema de ecuaciones Yule-Walker.** Este resultado permite obtener con facilidad:

1.- Las autocovarianzas o coeficientes de autocorrelación correspondientes a unos valores determinados de los parámetros .

2.- Los valores de los parámetros  correspondientes a unas autocovarianzas o coeficientes de autocorrelación determinados.

**2.2.9. Procesos de Medias Móviles de orden q o MA (q)**

El modelo de **Medias Móviles de orden q** o **MA (q)** cumple para el proceso



donde es independiente de y en general los no son independientes. Se puede escribir en forma funcional como



donde 

El **vector de retrasos de el ruido blanco** es



El **vector de parámetros** es



con



La **media** del proceso es



La **función autocovarianza** del proceso es



**2.2.10. Modelo Autoregresivo y de Medias Móviles o ARMA(p,q)**

Un modelo Autoregresivo con Media Móvil de orden p,q conocido también por **ARMA(p,q)** para , es un modelo que sigue la ecuación



es decir, la unión de un proceso AR(p) y MA(q), donde



Si  es invertible en un con se tiene



 sigue un proceso .

Si  es invertible en un 



 sigue un proceso .

Un modelo ARMA(2,1) tendría la siguiente forma:



**2.2.11. Modelos Autoregresivo de Medias Móviles Integrado o ARIMA (p,d,q).**

Un proceso de **medias móviles autoregresivo integrado** **de orden (p,d,q)** es un proceso  que sigue el modelo funcional



donde:

 es un operador autoregresivo de orden p.

 es un operador de medias móviles.

∇ es el operador de diferencias 

* + 1. **Modelos SARIMA(p,d,q)(P,D,Q)s**

En algunas ocasiones nos topamos con el caso de que existen dos tipos de relaciones, una entre sus observaciones cada mes sucesivo de un año particular y otra entre las observaciones para el mismo mes en años siguientes. Esto nos lleva a la interrogante que si un modelo ARIMA puede tener retardos múltiplos de 12. Los **modelos multiplicativos de tipo** son:



donde s es periodo de estacionalidad (s = 12,4,6,...)

y las diferencias son 

 son polinomios de grado p, q, P y Q.

El ruido blanco es .

Uno de los procesos que satisface, la anterior ecuación es conocido como **SARIMA (p,d,q)(P,D,Q)s**

* + 1. **Predicción**

La teoría de la predicción de modelos de series temporales se basa en:

1.- Calcular previsiones de la serie estacionaria expresando de la forma



donde  y son pesos obtenidos de la ecuación de coeficientes



y el modelo lineal puede ser escrito en términos de Zt para el filtro lineal



y la expresión de los pesos se la pude generalizar con el operador autorregresivo , es decir



el correspondiente modelo ARMA o a su vez ARIMA si la serie no es estacionaria y sus diferencias y

2.- A partir de ellas, construir las previsiones correspondientes para la serie original.

El uso de la forma antes mencionada, se debe a que resulta más sencillo construir para modelos de media móvil que para modelos Autoregresivos. Por otra parte, el que primero se obtengan previsiones para la serie estacionaria se debe a que la obtención de las formas





donde los modelos pueden ser escritos en forma invertida, así



de un proceso estocástico requiere que éste sea estacionario e invertible.

Supongamos que se desea prever el valor de la variable en *t* dada la información disponible hasta *t-k*. A partir de la forma



se obtiene



consecuentemente, el error de previsión es



y su varianza será 

A partir de los resultados anteriores resulta evidente que, bajo la hipótesis de normalidad, se cumple:



de manera que un intervalo de confianza de 95% para el verdadero valor de puede expresarse como:



Los resultados anteriores permiten obtener, de forma inmediata, aproximaciones de la previsión óptima, tanto por punto como por intervalo.