# ALGORITMO E IMPLEMENTACIÓN DEL MÉTODO STATIS

Jorge González[1] y Oldemar Rodríguez[1]

**Abstract**

En este artículo se presenta un algoritmo en pseudocódigo para el método Statis, útil para el análisis de tablas múltiples de datos. Se presentan también algunos comentarios sobre la implementación desarrollada en C++. Finalmente se ilustra el método mediante un ejemplo.

**Palabras clave:** Statis, compromiso, programación orientada a objetos, tablas múltiples de datos.

## 1 Introducción

El método Statis (Estructura estadística de una tabla de tres índices) parte de un conjunto de tablas de datos (individuos×variables), obtenidas en situaciones diferentes, un caso común es el de disponer de las mismas observaciones en diferentes instantes. Disponemos entonces de $r$ tablas $X_1, X_2, \ldots, X_r$, donde $X_k$ es de tamaño $n \times p_k$ que tienen asociadas métricas $M_k$ para los espacios de los individuos $\mathbb{R}^{p_k}$ y de una métrica de pesos $D$ para el espacio de las variables $\mathbb{R}^n$.

Se podría efectuar un análisis separado para cada tabla pero esto no permitiría tener una idea clara de la evolución de los datos, ya que los diferentes planos principales no se pueden superponer. Por esta razón es necesario tener un sistema de referencia común, llamado en adelante ejes del compromiso, en el cual se representen adecuadamente las diferentes estructuras aportadas por cada tabla. Para lograr este sistema de referencia común existen varios métodos, entre ellos el método Statis,

[1]Escuela de Matemática, Universidad de Costa Rica.

estudiado por H. L'Hermier Des Plantes en 1976 [3] con base en los trabajos del profesor Y. Escoufier. Las propiedades del método Statis se pueden encontrar en [4] y una presentación del mismo en [2].

El aporte principal de este trabajo consiste en presentar un algoritmo numérico en pseudocódigo para este método, la correctitud de este algoritmo la hemos verificado mediante la programación del mismo en el lenguaje C++. Además se presenta un ejemplo ilustrativo del algoritmo.

# 2 El algoritmo

En esta sección se presenta un algoritmo para el método Statis en pseudocódigo el cual puede escribirse fácilmente en cualquier lenguaje de programación estructurado como Pascal o C. El algoritmo lo dividimos en: Entrada de datos, Intraestructura e Interestructura.

En la métrica $D$ consideramos todos los pesos iguales y $M_k$ la definimos como la identidad correspondiente.

**Entrada:** Se reciben $r$ tablas (matrices) de datos $X_k$ individuos$\times$variables de tamaño $n \times p_k$, centradas respecto a $D$, donde

$n$ es el número de individuos, el mismo en las $r$ mediciones.

$p_k$ es el número de variables en la $k-$ésima medición.

$D = diag(1/n, 1/n, \ldots, 1/n)$ es la matriz de pesos, invariable para las $r$ mediciones.

**Interestructura:** (pasos 1-4). Como los individuos en las diferentes mediciones han sido los mismos, comparamos su distribución espacial a través de las matrices $W_i = X_i X_i^t$. Para ello usamos la métrica definida por el producto interno

$$< W_i, W_j >_\phi = \text{traza}(W_i W_j) = < \vec{W}_i, \vec{W}_j >_I$$

donde $\vec{W}_i \in I\!R^{n^2}$ es un vector formado por todas las filas de la matriz $W_i$ para $i = 1, 2, \ldots, r$.

**Paso 1:** Calcular las matrices $W_i$ $n \times n$ dadas por $W_i = X_i X_i^t$ para $i = 1, 2, \ldots, r$

**Paso 2:** Genere la matriz

$$X = \left[ \begin{array}{ccccccc} \vec{W}_1 & | & \vec{W}_2 & | & \cdots & | & \vec{W}_r \end{array} \right]_{n^2 \times r}$$

**Paso 3:** Efectúe el Análisis en Componentes Principales del triplete $(X, D_{\frac{1}{\sigma^2}}, \frac{1}{n^2} I_{n^2})$ (para efectuar este Análisis en Componentes Principales utilizamos el algoritmo presentado en [5]).

Guardar el primer vector propio de este ACP, que denotamos por $u = (u_1, u_2, \ldots, u_r)$, y el primer valor propio $\lambda_1$ para usarlos en el compromiso.

**Paso 4:** Graficar el círculo de correlaciones (en este aparecerán representados los vectores $\vec{W}_i$. Puntos cercanos significará configuraciones de individuos parecidas y vectores paralelos representará configuraciones homotéticas (ver [2]).

**Intraestructura:** (pasos 5-13). El estudio evolutivo de individuos y variables lo realizamos con la factorización del compromiso $\sum_{k=1}^{r} \beta_k W_k$ obtenida del ACP: $(\tilde{X}, I_l, D)$, donde $l = \sum_{k=1}^{r} p_k$, $\beta_k$ está definido en el paso 5 y $\tilde{X}$ se define como sigue:

$$\tilde{X} = \left[ \begin{array}{c|c|c|c} \sqrt{\beta_1} X_1 & \sqrt{\beta_2} X_2 & \cdots & \sqrt{\beta_r} X_r \end{array} \right]_{n \times \sum_{k=1}^{r} p_k}$$

Dado que el operador $WD$ del ACP de $\tilde{X}$ es igual al compromiso, este nos da una representación simultánea de las $\sum_{k=1}^{r} p_k$ variables en la base $D$-ortonormal de sus vectores propios de $WD$ que denominamos ejes del compromiso.

Pero el ACP anterior no resuelve el problema de la representación simultánea de los individuos, para ello definimos las coordenadas de los individuos de la $k-$ésima tabla como la $D$-proyección de la matriz $W_k$ sobre los ejes del compromiso. Para una justificación de esta definición ver [4].

Una trayectoria es la representación de un mismo individuo a través de las diferentes tablas, en los ejes del compromiso.

**Evolución de las variables :** (pasos 5-10)

**Paso 5:** Calcular $\beta = (\beta_1, \beta_2, \ldots, \beta_r) = \frac{1}{\sqrt{\lambda_1}} u$.

**Paso 6:** Calcular la matriz por bloques $\tilde{X}$ como sigue:

$$\tilde{X} = \left[ \begin{array}{c|c|c|c} \sqrt{\beta_1} X_1 & \sqrt{\beta_2} X_2 & \cdots & \sqrt{\beta_r} X_r \end{array} \right]_{n \times \sum_{k=1}^{r} p_k}$$

**Paso 7:** Efectuar un $ACP(\tilde{X}, I_l, D)$ (no reducido).

**Paso 8:** Guardar la matriz de componentes principales, denotada por

$$C = [C_1, C_2, \ldots, C_l]$$

donde $l = \sum_{k=1}^{r} p_k$.

**Paso 9:** Graficar el círculo de correlación de este ACP, en el cual se podrá estudiar la evolución de las variables.

**Paso 10:** Graficar el plano principal de este ACP, en el que aparecen representados los $n$ individuos promedio (ver definición en [4]).

**Evolución de los individuos :** (pasos 11-13)

**Paso 11:** Calcular la matriz IND definida por bloques como sigue:

$$\mathsf{IND} = \begin{bmatrix} W_1 \\ W_2 \\ \vdots \\ W_r \end{bmatrix}_{(rn)\times r}$$

**Paso 12:** Calcular las coordenadas de los individuos a través de las diferentes tablas, denotada por $\Omega$, mediante el siguiente producto matricial:

$$\Omega = \mathsf{IND} \times C$$

**Paso 13:** Graficar el plano principal, en éste se podrá estudiar la evolución de los individuos.

# 3 Algunas observaciones sobre la implementación

El algoritmo para el método Statis presentado en la sección anterior fue implementado en lenguaje C++ como un módulo más del sistema `PIMAD` [5], esto permitió aprovechar el núcleo del sistema `PIMAD` para efectuar los Análisis en Componentes Principales que son necesarios en el método Statis y para generar los gráficos.

El siguiente fragmento de código permite ver la definición de la "class" o clase `Statis` como una clase derivada de la clase `TablaDatos` del sistema `PIMAD`, lo que permite reutilizar todas las características de las tablas de datos ya implementadas, tales como, calcular la matriz de varianzas, calcular vectores y valores propios, generar gráficos etc.

```
class Statis : public TablaDatos {
  float  beta[MAX];
  int NTecho, MTecho;
  int NumeroIndividuos;
  int NumeroVariables;
  int NumeroTablas;
```

```
    TNombreArchivo NomArchs[MAX];
  public:
    Statis(int InitN);
    boolean GeneraTablaBin();
    void CalculaBeta(float  *z[MAX],float  d[MAX]);
    boolean GeneraTablaXTechoTBin();
    boolean GeneraIND();
};
```

Como se nota en esta definición de la clase `Statis` fue necesario utilizar un archivo binario para almacenar la tabla de datos, es decir, la matriz $X$ del algoritmo anterior. Esto fue necesario debido a que, como puede verse en el paso 2 del algoritmo de la sección anterior, la matriz $X$ tiende a ser sumamente grande lo cual hace que algunas veces sea imposible almacenarla en la memoria RAM del computador.

El hecho de que la tabla de datos esté en un archivo hace que el sistema sea bastante más lento en algunos procesos de cálculo, por ejemplo al calcular la matriz de varianzas, pero garantiza que el sistema funcione adecuadamente aún cuando las dimensiones de las tablas de datos sean muy grandes o bien se tengan muchas tablas de datos iniciales.

El sistema `PIMAD` proporciona en el menú `Archivo` una opción para cargar los datos del método Statis, como puede verse en la Figura 1. Una vez que se selecciona esta opción el sistema le pedirá una secuencia de archivos para las tablas de datos, un archivo de etiquetas para las variables y otro para las etiquetas de los individuos. Todos estos archivos deben estar en código `ASCII` en formato `PIMAD` (ver [5]).

Figure 1: Datos para el método Statis en `PIMAD`

Una vez que el sistema `PIMAD` ha cargado los datos, pueden ejecutarse las opciones del menú `Statis` que se muestran en la Figura 2. Estas opciones permiten correr paso a paso el algoritmo de la sección anterior, permitiendo incluso visualizar los resultados por pantalla, imprimirlos o generar código LaTeX.

Figure 2: Opción `Statis` en `PIMAD`

# 4 Ejemplo

Con el objeto de mostrar los diferentes pasos del algoritmo, así como las diferentes salidas del programa, aplicaremos Statis a los datos recolectados en cuatro muestreos por el proyecto "*Desarrollo y aplicación de métodos efectivos a bajo costo para el monitoreo biológico de los ríos de Costa Rica*" de la Universidad Nacional. Uno de los objetivos de este proyecto es la definición de un índice para medir la calidad del agua, que se base en la presencia-ausencia de macroinvertebrados bentónicos y que sea adecuado para las condiciones tropicales. El índice se ha denominado **CRBI** y se encuentra en una etapa exploratoria. La idea es utilizar las trayectorias y la evolución de las variables para compararlo con otros índices como: **BBI, GBI, BMWP, ASPT**, usados en otras latitudes (consultar [6], [7]). Además se usan, como control, variables Fisico-Químicas de comportamiento conocido respecto de la calidad del agua.

1. **Entrada de los datos**

   Se tienen cuatro tablas de tamaño $18 \times 20$, que corresponden a 18 puntos de muestreo en la *Cuenca del río Tárcoles* y a 20 variables, de las cuales 9 son Fisicoquímicas y 4 son índices de calidad de agua.

   Las tablas corresponden a los meses de: febrero, setiembre, octubre y diciembre.

2. **Interestructura**

   El programa `PIMAD` permite imprimir la matriz del paso 2 del algoritmo, así como imprimir el círculo de correlación del paso 4, el cual se presenta en la Figura 3.

   Los cuatro puntos corresponden a cada una de las tablas. Se observan comportamientos similares en los meses de febrero y diciembre, difiriendo estos de

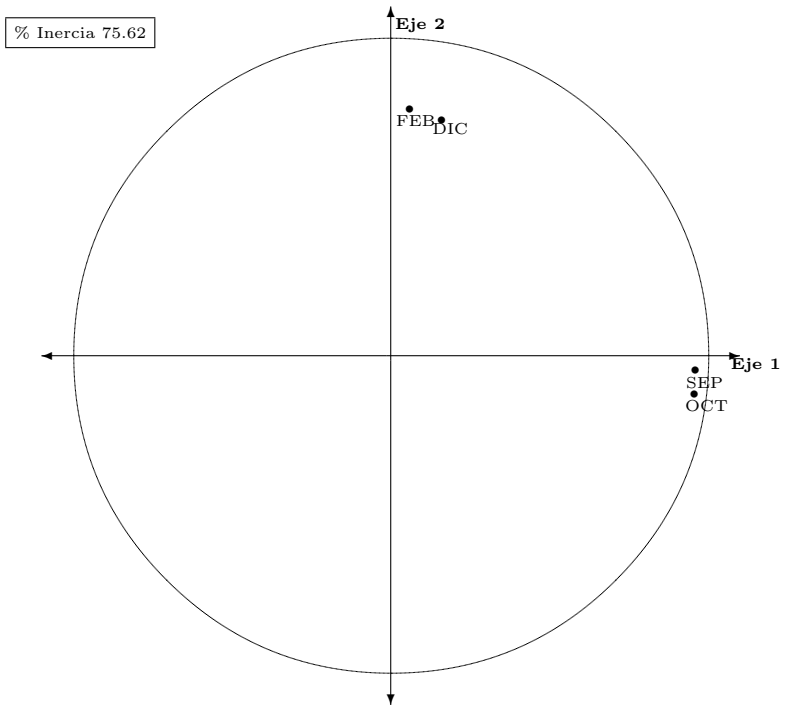


Figure 3: Círculo del correlación, corresponde al paso 2 del algoritmo

los comportamientos similares de septiembre y octubre.

3. **Intraestructura**

   El programa `PIMAD` permite imprimir la matriz $\tilde{X}$, paso 8 del algoritmo, y todos los pasos intermedios del ACP del paso 9, tales como: la matriz de varianzas, los vectores propios, los valores propios, y la matriz de componentes principales. Se presenta únicamente, en la Figura 4, la salida más importante de esta etapa del algoritmo, es decir, el gráfico con la evolución de las variables.

   En la intraestructura también es posible imprimir un plano principal en el que se representan los individuos promedio, en este caso el baricentro de los puntos de muestreo, así como las coordenadas de este gráfico, es decir las componentes principales. En la Figura 5 se presenta el plano principal de los individuos promedio.

   Finalmente el sistema `PIMAD` permite imprimir un plano en que se presenta la trayectoria de los individuos, en este caso la trayectoria de los puntos de muestreo, este gráfico se presenta en la Figura 6. Además `PIMAD` también permite imprimir o visualizar por pantalla todos los cálculos necesarios para producir este plano.

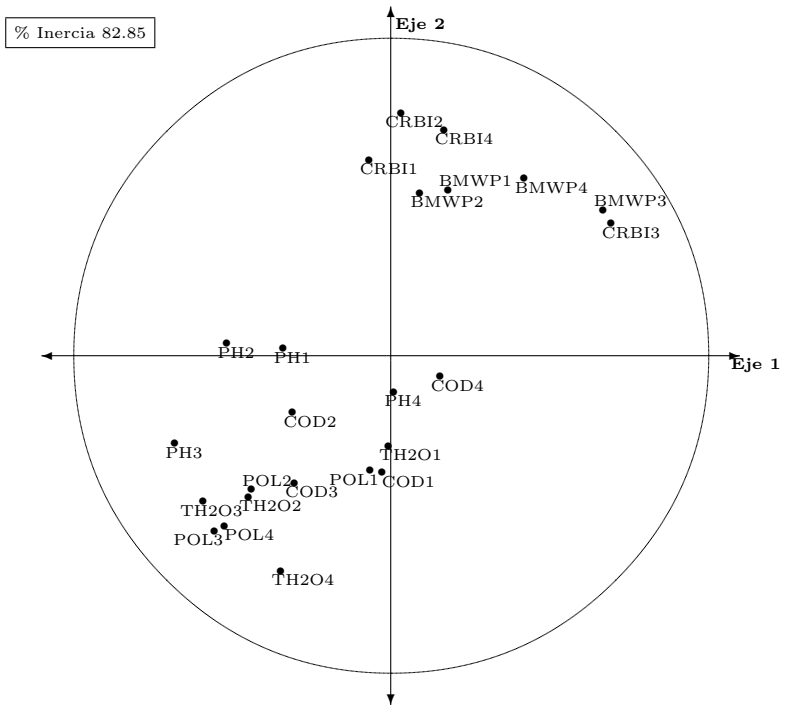


Figure 4: Evolución de las variables, corresponde al paso 11 del algoritmo

# 5 Perspectivas

Otro ejemplo importante en cual este algoritmo será usado es en el estudio de la evolución de la opinión de pública en Costa Rica, para este propósito se usarán las bases de datos del proyecto de investigación de la Escuela de la Matemática de la Universidad de Costa Rica *Estructuras de la opinión pública en Costa Rica*.

Actualmente se trabaja en el diseño de un algoritmo similar para el Análisis Factorial Múltiple (AFM), el cual será implementado en C++ e incorporado como un módulo más del sistema `PIMAD`.

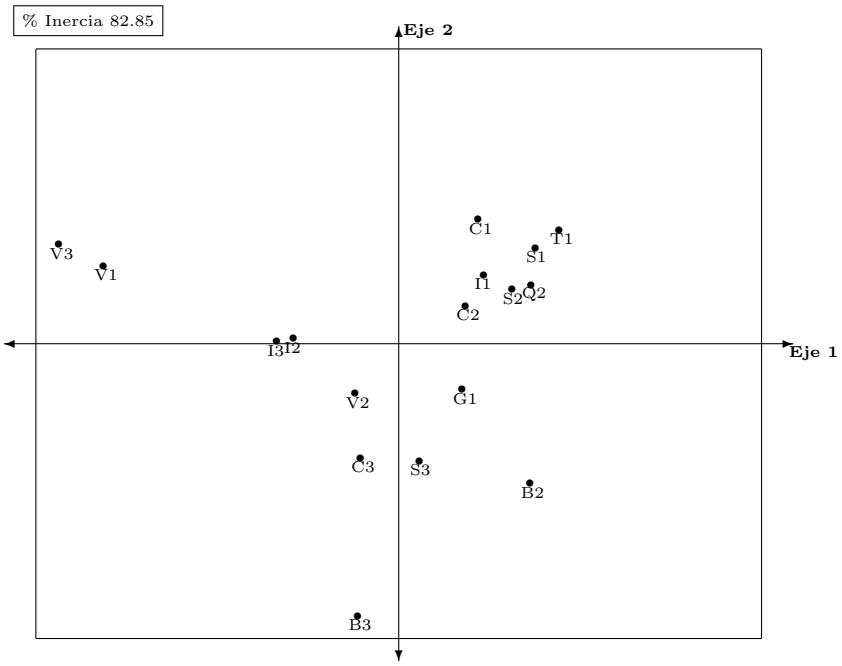


Figure 5: Baricentro de los puntos de muestreo, corresponde al paso 12 del algoritmo

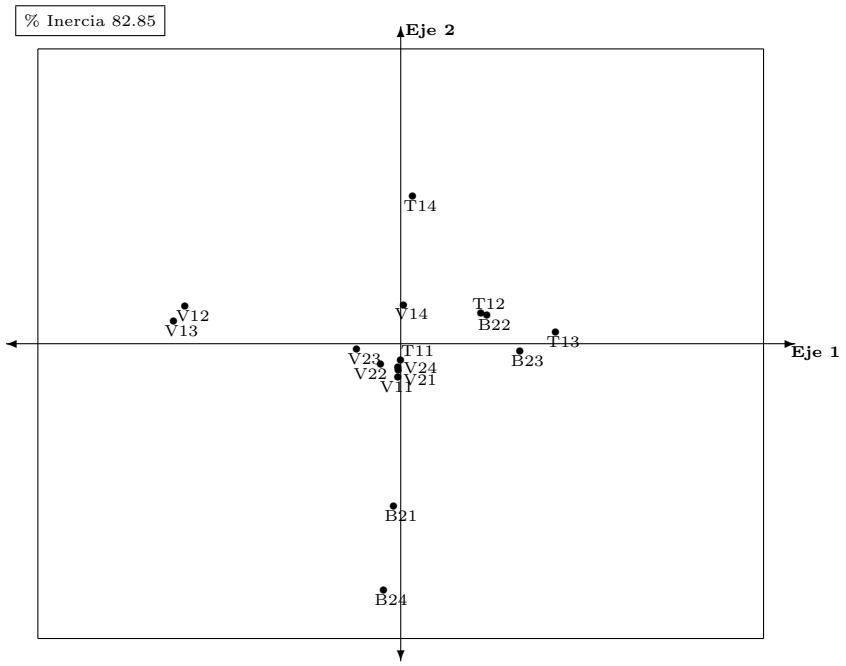


Figure 6: Trayectoria de los puntos de muestreo, corresponde al paso 15 del algoritmo

# References

[1] González J. (1994) *Estudio evolutivo de la calidad del agua*, Memoria del II Encuentro Centroamericano de Investigadores en Matemáticas, G. Mora (ed.), San Ramón, pp. 425-434.

[2] Castillo W y González J. (1994) *Análisis de tablas múltiples de datos*, Revista de Matemáticas: Teoría y Aplicaciones, vol. 1 $N^o1$, pp. 47-55.

[3] L'Hermier des Plantes H. (1976) *Structuration des tableux trois indices de la statistique: théorie et application d'une méthode d'analyse conjointe*, Thése presentée l'Université des Sciences et Techniques du Languedoc, Montpelier.

[4] Lavit, Ch. (1988) *Analyse Conjointe de Tableaux Quantitatifs.* Ed. Masson, Paris.

[5] Rodríguez O. (1993) *Desarrollo orientado a objetos: una aplicación al análisis de datos*, Tesis de maestría presentada en el Instituto Tecnológico de Costa Rica, Cartago.

[6] Simpson K.W. and Cloquhoun J.R. (1985) *The macroinvertebrate fauna of an acid-stressed headwater stream in the Adirondack Mountains*, New York, Freshwater Biology, $N^o$15, pp. 671-681.

[7] Wade K.R., Ormerod S.J. and Gee A.S. (1989) *Classification and ordination of macroinvertebrade assemblages to predic stream acidity in upland Wales*, Hydrobilogia, $N^o$171, pp. 59-78.