ingeniería civil
4-78

C.D.U.: 519.24:65.011.2

Un ensayo para el análisis y la clasificación de las fuentes de información mediante métodos de la estadística matemática

Por Jorge Luis Romeu
Lic. en Matemática

RESUMEN. Se desarrolla un proyecto de análisis estadístico que, aplicando ciertos métodos paramétricos y no paramétricos a una serie histórica de datos (correlaciones, regresión, análisis de varianza, covarianza y discriminante), clasifica al conjunto de las fuentes de información de acuerdo con la calidad de las mismas. Seleccionadas las mejores informaciones, pueden calcularse entonces parámetros de predicción, control y toma de decisiones. Se ilustra el método mediante un ejemplo numérico.

1. INTRODUCCION

El tratamiento de la información que fluye periódicamente a la dirección de un organismo es complicado e importante. Es complicado porque el volumen y la frecuencia de la información hacen difícil su procesamiento por métodos convencionales, y requieren el empleo de técnicas de computación para agilizarlo y disminuir posibles errores. Es importante por cuanto el flujo de información proporciona

SUMMARY OF ARTICLES

Anchored timberings

By José Menéndez Menéndez, Civil Engineer

The advantages of using anchored timberings are exposed and the necessary data for their project are specified, as well as the considerations regarding the different project elements. The construction process steps and the verifications in site are detailed.

Hydrogeophysics. Electrical control of the grout curtain in dams

By Roberto Puyada Teijeiro, Engineer

A series of works on some special applications of geophysical methods to hydrogeological studies are presented under the title of hydrogeophysics. These applications taken from seminaries, publications, experiences, etc., show the beginning of a new subject. A method and a technique for controlling the grouting work at the grout curtain area in dams are exposed.

An essay on the analysis and classification of information sources by statistical mathematics method

By Lic. Jorge Luis Romeu

We develop a statistical analysis project that, applying certain parametrical and non parametrical statistical methods to time series (correlation, regressions, analysis of variance, covariance and discriminant analysis) classifies the set of information sources according to the quality of the information. Hence, selecting the best of the information we can calculate the forecast, control and decision parameters. The method is illustrated through a numerical example.

a la dirección del organismo los parámetros y ecuaciones para controlar paso a paso su marcha (función de *feedback* o chequeo) y los parámetros y ecuaciones para proyectar sus actividades futuras (función de planificación o pronóstico).

Pero esos parámetros y ecuaciones dependen de la calidad o confiabilidad de la información procesada. Esta puede ser defectuosa por errores al recopilarla en la base, al consolidarla a nivel intermedio, etc., y en este caso su utilización para las dos funciones arriba mencionadas resulta perjudicial. Por lo tanto, la capacidad que pueda tener el organismo de dirección para clasificarla, filtrarla y desechar la parte deficiente de la misma resulta relevante.

El presente trabajo pretende, partiendo de una experiencia real, proporcionar algunos elementos para realizar estos análisis y la clasificación de la información, así como para la interpretación de la misma.

2. PLANTEAMIENTO ORIGINAL

A comienzos del año 1975 se encomendó al Grupo Vial Nacional del DAP hacer un análisis estadístico a un conjunto de informaciones que fluía mensualmente a dicho nivel de dirección, así como tratar de clasificarlas atendiendo a su grado de confiabilidad, ya que se sabía que ésta no era homogénea.

Entonces comenzó a estudiarse la posibilidad de aplicar ciertas técnicas estadísticas del control de la calidad y del análisis multidimensional al flujo continuo de datos.

La información mensual consistía de las tablas horarias por equipo de construcción que reportaban las brigadas en el antiguo modelo IP-16. También contábamos, como elemento auxiliar, con las cifras correspondientes al movimiento de tierra reportado por esas mismas brigadas para el mismo período. Y esta información era consolidada por provincias.

Se contaba pues con el número de horas productivas, de taller y mantenimiento, y las pérdidas de los equipos de tra-

bajo por lluvia, así como su fondo horario bruto, lo cual permitía trabajar esas variables en porcentajes. Se contaba con esta información para todos los equipos: camiones de volteo, cargadores, buldozers, traíllas, motoniveladoras, pipas y demás equipos integrantes de una unidad de trabajo usual.

2.1. *Parámetros y ecuaciones de control*

Antes de obtener los elementos para futuros chequeos y pronósticos había que asegurar la validez de la información, sin la posibilidad de realizar un muestreo con las fuentes de información.

En tales condiciones, a fin de emitir criterios sobre la calidad de la información de cada fuente, se decidió estudiar las correlaciones entre las variables horarias (*input*) y el movimiento de tierra realizado durante el mismo período (*output*) como si fuera un proceso de producción continuo en una fábrica; se estableció un paralelo entre el control de la calidad en aquélla y de la información en ésta.

En la medida en que fueran consistentes los resultados de los análisis con ciertas hipótesis de trabajo preestablecidas se daría un diagnóstico referente a la confiabilidad de la fuente de información.

Se pretendía, con la parte admisible de la información, construir después, para la predicción y el control por equipo y por tipo de vía (socioeconómica, autopista, agropecuaria, etc.), ecuaciones de Regresión Lineal del tipo:

$$Y_i = \alpha_0 + \alpha_1 X_i + e_i \ , \ 1 \leqslant i \leqslant n$$

Modelo I: Y_i — movimiento de tierra reportado en el i-ésimo mes;

X_i — horas productivas totales de un equipo reportado en el i-ésimo mes;

e_i — error aleatorio con las características requeridas en el modelo de Regresión Lineal.

Semejantes ecuaciones para cada equipo de construcción y para cada tipo de vía permitirían, utilizando el intervalo de predicción construido con la Regresión, chequear si un reporte mensual futuro de una brigada se encontraría dentro de los límites de lo posible, o en caso contrario chequear si el incremento o decremento fuesen producto de alteraciones en los sistemas de trabajo o de errores en la información.

Por ejemplo, si para el equipo buldozer en carreteras del tipo agropecuario se hubiese obtenido la ecuación:

$$Y_i = 25 + 10\,X_i \;\; ; \text{ para los valores}$$

$$\sigma^2 = 25 \;\; , \;\; \overline{X} = 200 \;\; , \;\; n = 200 \;\; , \;\; \Sigma (X_i - \overline{X})^2 = 50$$

y una de las brigadas de construcciones agropecuarias hubiese reportado en el mes i-ésimo: $Y_i = 2\,000\ m^3$ y $X_i = 190$ horas productivas, se hubiera podido detectar un desajuste (por un cambio en el sistema de trabajo o un error en la información suministrada), ya que:

$$\sigma^2(\hat{Y}_i) = \sigma^2 \left\{ 1 + \frac{1}{n} + \frac{(X_i - \overline{X})^2}{\Sigma (X_i - \overline{X})^2} \right\} =$$

$$= 25 \left\{ 1 + \frac{1}{20} + \frac{(190 - 200)^2}{50} \right\} = \frac{81}{5} + 50 = 66{,}8$$

$$\hat{Y}_i = 25 + 10\,X_i = 25 + 10 \times 190 = 25 + 1900 = 1925$$

Y el Y_i informado contra X_i de horas productivas se encuentra fuera del intervalo de predicción para el 95 % construido mediante el modelo de regresión, según podemos ver:

$$\hat{Y}_i + t_{0,025}\,\sigma(\hat{Y}_i) < Y_i < \hat{Y}_i + t_{0,975}\,\sigma(\hat{Y}_i)$$

$$1925 - 17{,}14 < Y_i < 1925 + 17{,}14$$

$$1907{,}86 < Y_i < 1942{,}14$$

con:

$$\sigma(\hat{Y}_i) = \sqrt{66{,}8} = 8{,}16$$

$$t_{0,975}\ (18) = -\ t_{0,025}\ (18) = 2{,}101$$

Así, estos elementos junto con las medias y varianzas calculadas para cada variable horaria X_i servirían para estimar las capacidades reales de trabajo con determinado parque de equipo, en determinado tipo de vía, planificar futuros trabajos y/o analizar los intercambios de equipos en las brigadas, según fuesen surgiendo nuevas necesidades constructivas.

2.2. *Consolidación*

A fin de obtener el mayor número de informaciones para cada análisis (sólo se contaba con 24 meses en la serie histórica), era necesario reunir equipos iguales en tipos de vía con características similares, pero desigual parque de equipos. Por ejemplo, reunir toda la información relativa a camión de volteo para las carreteras agropecuarias del país.

Para ello era necesario primero considerar el efecto α_i introducido por una provincia i-ésima y demostrar que era nulo. Había pues que considerar el modelo de Análisis de Covarianza, ya que el número de equipos en cada provincia, y por consiguiente el número de horas productivas, sería desigual:

$$Y_{ij} = \mu + \alpha_i + \beta_i X_{ij} + e_{ij}\ ,\ 1 \leqslant i \leqslant n\ ,\ 1 \leqslant j \leqslant k$$

Modelo II: Y_{ij} — movimiento de tierra, mes j-ésimo, provincia i-ésima;

μ — media general;

α_i — efecto de la provincia i-ésima;

β_i — coeficiente productivo de la provincia i-ésima;

X_{ij} — horas productivas totales de un equipo, mes j-ésimo y provincia i-ésima;

e_{ij} — error aleatorio con las características requeridas en el modelo de Análisis de Covarianza.

Se debían constatar en el modelo dos hipótesis sucesivas:

Hipótesis I: $H_0 : \alpha_i = 0 \, , \, 1 \leqslant i \leqslant n \, ; \, H_1 : \alpha_i \neq 0$ para algún i

Hipótesis II: $H_0 : \beta_i = \beta \, , \, 1 \leqslant i \leqslant n \, ; \, H_1 : \beta_i \neq \beta$ „ „ „

Mediante la Hipótesis I se analizaba la influencia debida a la distribución provincial (la diferencia en movimientos de tierra sería debida al efecto de condiciones naturales particulares de cada provincia). Se podían reunir las informaciones, en caso de aceptar H_0, en un solo Análisis de Regresión para un determinado equipo y tipo de vía (Modelo I). En caso contrario se podría sustraer de cada observación mensual el efecto $\hat{\alpha}_i$ estimado, introducido por la provincia i-ésima. Así ya se podrían analizar como un solo grupo las distintas provincias en un plano de igualdad.

Mediante la Hipótesis II se analizaba si la efectividad β_i en la provincia i-ésima era constante e igual a β. Así se probaba igual eficiencia para un tipo de vía en todo el país. Y se podía utilizar un mismo parámetro $\hat{\beta}$ para recalcular los valores.

2.3. *Hipótesis de trabajo*

Para emitir criterios sobre la calidad de la información, sobre la base de las correlaciones entre el movimiento de tierra y las horas informadas por los distintos equipos, así como la correlación entre las horas reportadas de los equipos, tomados dos a dos, se consideraron las siguientes hipótesis:

a) Que de los elementos de que se disponía, las horas informadas constituían la mejor medida de la cantidad de trabajo invertido (*input*).

b) Que por la misma razón el movimiento de tierra constituía la mejor medida del trabajo ejecutado (*output*).

c) Que la correlación entre los dos elementos anteriores debía ser, por lo tanto, positiva.

d) Que el encadenamiento de las labores de los distintos equipos que constituyen una unidad de trabajo (sobre todo buldozers, cargadores y camiones que fueron incluidos en el análisis por estar siempre presentes en todas ellas) es muy grande.

e) Que, por lo tanto, la correlación entre las horas reportadas por los equipos, tomadas dos a dos, debía ser significativamente positiva.

f) Finalmente y, como consecuencia de lo anterior, que una correlación baja o negativa de c) o e) sería tomada como elemento de juicio para emitir un criterio desfavorable con respecto a su fuente de información.

3. EJECUCION

Al empezar la elaboración de los análisis se comenzó a obtener resultados que impedían llevar a cabo los modelos I y II del epígrafe 2.

Las variables Y (movimiento de tierra) y X (variables horarias) no seguían una distribución normal, sino sufrían perturbaciones de simetría y de curtosis. Esto violaba una de las hipótesis fundamentales de los modelos I y II e impedía igualmente el cálculo de la correlación de Pearson. Recordemos que para los métodos paramétricos se exige, generalmente, la distribución normal de las variables. Por lo tanto, se acudió a los métodos no paramétricos que obvian este supuesto.

Para calcular la correlación entre las variables Y y X se utilizó el método de correlación no paramétrica de Spearman, que emplea los rangos. Se siguió idéntico método para calcular las correlaciones entre las variables horarias de diferentes equipos. Y para sustituir el modelo II se utilizó el Análisis de Varianza no paramétrico de Kruskall-Wallis, también por rangos, según el modelo:

$$X_{ij} = \mu + \alpha_i + e_{ij}$$

Modelo III: X_{ij} — variable horaria, dada en porcentaje, del mes j-ésimo y provincia i-ésima;

μ — media general;

α_i — efecto de la provincia i-ésima;

e_{ij} — error aleatorio del análisis de Kruskall-Wallis.

Nótese que se sustituyeron las horas informadas como productivas, de taller y mantenimiento, y pérdidas por lluvia, por su porcentaje con relación al fondo horario bruto, para balancear la diferencia en el número distinto de equipos por provincia.

Se observó finalmente, al superponer algunos gráficos de correlación inter-equipos vs mes, por provincia y tipo de vía, que se producían algunas incongruencias significativas que permitían calificar la fuente de información. Por ejemplo, para el mes de mayo, de máximas precipitaciones, se observó en algunos casos que todos los equipos, salvo uno, sufrían sensibles pérdidas por causa del agua (figura 1).

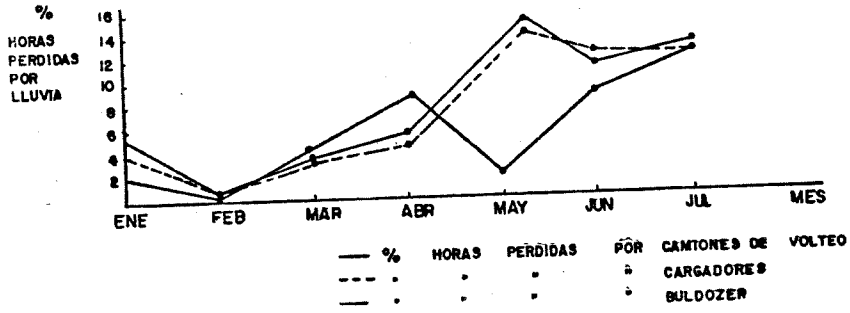

Figura 1

Con esos elementos de análisis se elevó en aquella fecha a la Dirección del Grupo Vial un informe estadístico de los resultados obtenidos, así como un diagnóstico evaluativo de las distintas fuentes de información, de acuerdo con la calidad y la consistencia de las mismas, con lo cual el trabajo se dio por terminado.

4. PROPOSICION

Pero la experiencia estaba incompleta. Proponemos entonces tratar de encontrar un método que permita clasificar objetivamente las fuentes de información para, tomando solamente aquellas calificadas como buenas, construir los parámetros y ecuaciones desarrollados en el epígrafe 2.1, ya que entonces sí sería meritorio el estudio de las transformaciones necesarias a los datos para cumplir las hipótesis de los modelos que se incumplían.

Y con las fuentes calificadas como deficientes realizar un trabajo para su control y/o mejora.

4.1. *Supuestos*

Para ilustrar mejor este trabajo se partirá de un ejemplo hipotético cuyos valores numéricos no tienen significado alguno, aunque sí han sido basados, desde el punto de vista de su concepción teórica, en la experiencia previamente esbozada.

Se define la división provincial, que representará a los individuos en el análisis discriminante que sigue, como la que existe en el país. Constituirá todas las fuentes de información posible.

Se definen las variables del análisis, asociadas a los individuos, como:

Z_1 : correlación de Spearman entre las variables Y y X_1.
Z_2 : „ „ „ „ „ „ Y „ X_2.
Z_3 : „ „ „ „ „ „ Y „ X_3.
Z_4 : „ „ „ „ „ „ X_1 „ X_3.

Donde:

Y — es el movimiento de tierra reportado en el mes i-ésimo.

X_1 — horas productivas reportadas en el mes i-ésimo en camión de volteo.

X_2 — horas productivas reportadas en el mes i-ésimo en cargador.

X_3 — horas productivas reportadas en el mes i-ésimo en buldozer.

Las variables Z_1, Z_2, Z_3 y Z_4 resultaron las más significativas de los análisis efectuados en el trabajo esbozado en el epígrafe 3. Las demás correlaciones de aquel análisis no fueron consideradas como variables para esta parte del trabajo por su bajo valor, y sólo fueron tomadas en consideración, como criterio general, para evaluar las provincias y formar los dos grupos que veremos más adelante.

Supongamos que los valores numéricos de las variables Z_1, Z_2, Z_3 y Z_4 asociados a cada provincia (o individuo) sean los que aparecen en la Tabla I.

TABLA I

Variables por provincias

Provincia	Z_1	Z_2	Z_3	Z_4	Orden
Pinar del Río	0,29	0,50	0,09	0,65	1
La Habana	0,78	0,58	0,70	0,43	2
Ciudad de La Habana	0,13	0,32	0,29	0,31	3
Matanzas	0,63	0,64	0,66	0,72	4
Cienfuegos	0,18	—0,21	0,43	0,05	5
Villa Clara	0,85	0,84	0,87	0,76	6
Sancti Spiritus	0,71	0,49	0,52	0,21	7
Ciego de Avila	0,07	0,43	0,27	—0,05	8
Camagüey	0,06	0,06	0,39	—0,18	9
Las Tunas	—0,32	0,00	0,28	0,51	10
Holguín	—0,17	—0,26	—0,28	0,11	11
Granma	0,57	0,62	0,77	0,37	12
Santiago de Cuba	0,52	0,53	0,61	0,84	13
Guantánamo	0,13	0,40	0,38	0,36	14

Supongamos que del análisis conjunto de todos los elementos manejados en la primera parte del trabajo, esbozada en el epígrafe 2 (correlaciones inter-equipos y con el movimiento de tierra, gráficos, Análisis de Varianza, etc.) se han

podido dar como las mejores provincias, por la calidad de su información:

Grupo I:
4. Matanzas.
6. Villa Clara.
12. Granma.
13. Santiago de Cuba.

Por un criterio similar y opuesto se pudieron clasificar como las provincias de más deficiente información:

Grupo II:
1. Pinar del Río.
8. Ciego de Avila.
10. Las Tunas.
11. Holguín.

Las demás provincias se encontraban, según el supuesto, en una posición intermedia entre los dos grupos sin que se pudiera clasificarlas, con los expresados elementos, por la calidad de su información, categóricamente en uno de los dos grupos: I ó II.

Vale añadir que en esa clasificación deben entrar también, y supondremos que en el caso que nos ocupa así se hizo, los análisis y criterios de los distintos especialistas que conocen tanto el trabajo de campo como las fuentes de información.

Lo importante es lograr dos grupos bien definidos de individuos o provincias, uno *bueno* y otro *malo*, para aplicarle el modelo de Análisis Discriminante a estas dos poblaciones.

En lo que sigue, y para evitar implicaciones que este trabajo no conlleva, se llamará a las provincias por su número de orden y a los grupos *bueno* y *malo* por grupos I y II.

4.2. *El Análisis Discriminante*

Se ha visto que las variables Y, X_1, X_2, X_3 no siguen necesariamente una distribución normal, pero las variables Z_1,

Z_2, Z_3 y Z_4 definidas como antes, se pueden suponer como distribuidas normalmente sobre la base de las hipótesis del epígrafe 2.3.

Entonces: $Z_i \sim N(\mu_i, \sigma_i^2)$, $1 \leqslant i \leqslant 4$

Tendremos entonces dos poblaciones:

π_1 el Grupo I o la población de las fuentes de información aceptables.

π_2 el Grupo II o la población de las fuentes de información inaceptables.

Hemos observado $p = 4$ variables como máximo:

$$(Z_p)\ 1 \leqslant p \leqslant 4.$$

Deseamos construir una función $U\ (Z_1, Z_2, Z_3, Z_4)$, tal que se pueda con ella clasificar a un individuo $Z^t \equiv (Z_1, Z_2, Z_3, Z_4)$ en uno de los dos π_i dados, $i = 1, 2$.

El individuo z es, en nuestro caso, una fuente de información, vale decir, una de las 14 provincias, a la que se le han medido las 4 variables (Z_1, Z_2, Z_3, Z_4) definidas en el epígrafe anterior.

Buscaremos, con ayuda de la Teoría de la Decisión Estadística, una regla de decisión $U\ (Z)$, basada en el siguiente criterio:

Particionando el espacio $\mathbb{R}^4$ (espacio euclidiano de 4 dimensiones) en dos regiones disjuntas R_1, R_2 tales que $R_1 \cup R_2 = \mathbb{R}^4$, asignamos al individuo z a π_1 si $z \in R_1$ y asignamos el individuo z a π_2 si $z \in R_2$.

Está claro que podemos cometer dos tipos de errores utilizando esta regla de decisión. Denominemos entonces:

α_1 el riesgo de clasificar un elemento de π_1 como perteneciente a π_2.

α_2 el riesgo de clasificar un elemento de π_2 como perteneciente a π_1.

El criterio para optimizar consistirá en minimizar $\alpha_1 + \alpha_2$ los riesgos totales.

Se puede demostrar que una regla de decisión que minimiza estos riesgos es la Regla de Bayes:

$$U(Z) = a^t Z - \tfrac{1}{2} (\mu^{(1)} - \mu^{(2)})^t a$$

donde:

$$\text{Si } Z \in \pi_1 \Rightarrow Z \sim N (\mu^{(1)}, \Sigma)$$

$$\text{Si } Z \in \pi_2 \Rightarrow Z \sim N (\mu^{(2)}, \Sigma)$$

$$a = \Sigma^{-1} (\mu^{(1)} - \mu^{(2)}) = \Sigma^{-1} \delta$$

Entonces las regiones R_1, R_2 quedarán como sigue: $c \in \mathbb{R}$

$$R_1 = \{x \in \mathbb{R}^4 : c \leqslant U(x) < + \infty\}$$
$$R_2 = \{x \in \mathbb{R}^4 : - \infty < U(x) < c\}$$

Y vemos que $U(Z)$ también sigue la distribución normal, con varianza expresada por:

$$\mathbb{V}\{U(Z)\} = \mathbb{E}\{a^t (Z - \mu^{(i)}) (Z - \mu^{(i)})^t a\}$$

$$= a^t \mathbb{E}\{(Z - \mu^{(i)}) (Z - \mu^{(i)})^t\} a$$

$$= a^t \Sigma a = (\delta^t \Sigma^{-1}) \Sigma (\Sigma^{-1} \delta) = \delta^t \Sigma \delta = \Delta_p^2$$

para: $\qquad Z \in \pi_i \ , \ i = 1, 2$

Esta Δ_p^2 es la distancia que separa, utilizando el criterio $U(Z)$, a las dos poblaciones π_1, π_2 y se conoce como *Distancia de Mahalanobis.*

En la medida en que el criterio discriminante $U(Z)$ sea mejor, la distancia que separará a las dos poblaciones será mayor y, por lo tanto, la probabilidad de clasificar erróneamente a un individuo utilizando $U(Z)$ será menor.

Entonces, una estimación sesgada de la probabilidad de mala clasificación será:

$\alpha_1 = \alpha_2 = \Phi\left(\frac{-1}{2}\Delta_p\right)$, donde Φ es la distribución Normal Estándar.

4.3. *Estimaciones*

Generalmente se conoce poco o nada de los parámetros del modelo anteriormente esbozado.

Partiremos de que Z_i, $1 \leqslant i \leqslant 4$ sigue la distribución Normal, pero desconocemos los parámetros Σ y $\mu_i^{(j)}$, $j = 1, 2$.

Trabajaremos entonces con sus estimaciones:

$$\bar{Z}^{(j)} = (\bar{Z}_1^{(j)}, \bar{Z}_2^{(j)}, \bar{Z}_3^{(j)}, \bar{Z}_4^{(j)})^t$$

Son las medias muestrales de: $Z \in \pi_j$, $j = 1, 2$.

S es la matriz de varianzas y covarianzas muestral.

Se puede demostrar que, si se toma como variable dependiente:

$$Y_k^{(1)} = \frac{n_2}{n_1 + n_2} \quad ; \quad 1 \leqslant k \leqslant n_1$$

cuando un individuo z es dado como perteneciente al Grupo I y n_1 es el total de individuos del Grupo I.

$$Y_k^{(2)} = \frac{-n_1}{n_1 + n_2} \quad ; \quad 1 \leqslant k \leqslant n_2$$

cuando un individuo z es dado como perteneciente al Grupo II y n_2 es el total de individuos en el Grupo II.

La Regresión Múltiple usual de las variables Z_i , $1 \leqslant i \leqslant 4$ sobre la $Y_k^{(i)}$ así definida es equivalente a la función discriminante de Fischer $U(Z)$ arriba definida, y las regiones R_1, R_2 para esta nueva función quedarán:

$$R_1 = \{x \in \mathbb{R}^4 : 0 \leqslant U(x) < +\infty\}$$

$$R_2 = \{x \in \mathbb{R}^4 : -\infty < U(x) < 0\}$$

Dado lo reducido de su número, y por la información adicional que pueda brindarnos en nuestro caso, escogemos el método de *todas las regresiones posibles* para seleccionar la función que conformará nuestro criterio.

De la Tabla II podemos ver cómo, para una sola variable independiente, las Regresiones Lineales Simples en Z_1, Z_2, Z_3, respectivamente, son muy significativas y con un ajuste R^2 satisfactorio. No ocurre así con la variable independiente Z_4, cuya información para separar la población en dos grupos es pobre. Observamos también que las dos variables más significativas en esta parte del análisis han resultado Z_1 y Z_3 (ver las funciones 1 al 4 en la Tabla II).

TABLA II

Funciones discriminantes

Orden	Variables independientes	R^2	Prueba F	Pruebas T				Distancia de Mahalanobis
				Z_1	Z_2	Z_3	Z_4	
1	Z_1	0,765	18,60	4,42	—	—	—	9,77
2	Z_2	0,521	6,52	—	2,55	—	—	3,26
3	Z_3	0,767	19,84	—	—	4,45	—	9,91
4	Z_4	0,373	3,56	—	—	—	1,88	1,78
5	Z_1, Z_3	0,843	13,50	1,56	—	1,58	—	16,20
6	Z_2, Z_3	0,769	8,436	—	—0,19	2,32	—	10,02
7	Z_1, Z_2	0,804	10,236	2,68	—0,99	—	—	12,28
8	Z_3, Z_4	0,788	9,29	—	—	3,13	0,69	11,15
9	Z_1, Z_4	0,776	8,69	3,00	—	—	0,50	10,39
10	Z_2, Z_4	0,592	3,63	—	1,63	—	0,93	4,35
11	Z_1, Z_2, Z_3	0,969	42,30	5,11	—4,05	4,65	—	93,77
12	Z_1, Z_2, Z_3, Z_4	0,969	24,01	4,19	—3,47	3,93	0,16	93,77

De las combinaciones para regresiones en dos variables vemos que la número 5 de la Tabla II, en Z_1 y Z_3, es la más significativa. Las regresiones que incluyen las variables Z_2 y Z_4 son menos significativas y la prueba T para el coeficiente de Z_2 y Z_4 es muy baja (números 6 a 10 de la Tabla II). Se puede considerar que este coeficiente es nulo, o sea, que las

En el ejemplo que desarrollaremos a manera de ilustración tenemos:

$$n_1 = n_2 = 4$$

$$Y_k^{(1)} = \frac{4}{8} = 0{,}5 \qquad 1 \leqslant k \leqslant n_1$$

$$Y_k^{(2)} = \frac{-4}{8} = -0{,}5 \qquad 1 \leqslant k \leqslant n_2$$

O sea, una regresión tomando como variable dependiente el valor $0{,}5 = Y_k^{(1)}$ para todos los individuos del Grupo I y el valor $-0{,}5 = Y_k^{(2)}$ para los individuos que pertenecen al Grupo II.

Hemos utilizado este camino para calcular todas las funciones discriminantes que se relacionan más abajo, utilizando como estimador de la Distancia de Mahalanobis el siguiente:

$$D_p^2 = \frac{n_1 + n_2 - 2}{\lambda^2} \cdot \frac{R^2}{1 - R^2}$$

Donde: $\lambda^2 = \dfrac{n_1 \cdot n_2}{n_1 + n_2}$ y R^2 es el ajuste de la Regresión.

4.4. *Los resultados*

Nuestro objetivo, desde el inicio, ha sido obtener una clasificación de las fuentes de información (individuos), de acuerdo con su confiabilidad.

Poseemos, basados en un criterio a priori subjetivo pero sólido, dos grupos bien delimitados de cuatro individuos: los dos grupos de provincias y un grupo de seis individuos o provincias que no nos sentimos capaces de clasificar con los criterios y elementos expresados. Tenemos, además, un conjunto de cuatro variables (Z_i) , $1 \leqslant i \leqslant 4$ asociadas a cada uno de los individuos. Deseamos encontrar un criterio, basado en las más significativas de esas cuatro variables, que los clasifique.

variables Z_2, Z_4, respectivamente, carecen de influencia en esas regresiones.

Hacemos una regresión en tres variables para Z_1, Z_2 y Z_3 (número 11 de la Tabla II) y se obtienen altos ajustes, $R^2 = 0{,}9694$ y prueba $F = 43{,}30$, así como pruebas T, para los coeficientes de Z_1, Z_2 y Z_3 significativos al 95 %.

Pero el signo de Z_2 es negativo, lo cual quiere decir que mientras mayor correlación haya entre el movimiento de tierra y las horas productivas del equipo cargador menor será el valor de esta función y, por lo tanto, mayor la probabilidad de clasificar esa fuente de información como deficiente. Esto va en contra de nuestras hipótesis de trabajo expresadas en el epígrafe 2.3. Por lo tanto, decidimos no considerar esta ecuación como el criterio discriminante buscado.

Finalmente, hacemos una regresión para las cuatro variables (número 12 de la Tabla II) y se obtienen valores R^2 similares al de la anterior y de la prueba F más bajos. También en esta función la variable Z_2 entra con signo negativo. Por otra parte, vemos cómo la contribución de Z_4 es insignificante, según se desprende de su prueba T correspondiente. Por lo tanto, tampoco consideraremos esta ecuación como criterio de clasificación.

La solución acertada debe encontrarse entre las regresiones en dos variables. Salta a la vista la de valores más significativos: la regresión para las variables Z_1 y Z_3 (número 5 de la Tabla II). Su prueba $F = 13{,}50$, significativa a todos los niveles, y su ajuste $R^2 = 0{,}8438$ que es satisfactorio. Las pruebas T para los coeficientes de Z_1 y Z_3 son significativas al 95 %, y estas variables entran con signo positivo en la función, lo cual está de acuerdo con nuestras hipótesis de trabajo. O sea, mientras más alta es la correlación entre movimiento de tierra y horas productivas reportadas, mayor es el valor de la función discriminante y, por lo tanto, mayor la probabilidad de clasificar la fuente como del Grupo I, o de información fidedigna. Por último, la Distancia de Mahalanobis es la mayor de las funciones que no han sido eliminadas.

Por lo tanto, estamos en presencia de la ecuación discriminante:

$$U(Z) = -0{,}46187 + 0{,}61966\, Z_1 + 0{,}66773\, Z_3$$

Su Distancia de Mahalanobis es:

$$\Delta_p^2 = \frac{n_1 + n_2 - 2}{\lambda^2} \cdot \frac{R^2}{1 - R^2} = \frac{6}{2} \cdot \frac{0{,}8438}{0{,}1562} = 16{,}2061$$

La probabilidad de clasificar mal utilizando este criterio es:

$$\alpha_1 + \alpha_2 = 2\,\Phi\left(\frac{-1}{2}\Delta_p\right) = 2\,\Phi\left(\frac{-\sqrt{16.20}}{2}\right)$$

$$= 2 \times \Phi\,(-2{,}01) \approx 0{,}0404$$

O sea, aproximadamente un 5 %.

La clasificación de todas las provincias, siguiendo este criterio, será la que aparece en la Tabla III.

TABLA III

Clasificación de provincias por la calidad de su información

Lugar	Provincia	$U\,(Z_1, Z_3)$
1	Villa Clara	0,645674
2	La Habana	0,488800
3	Granma	0,405444
4	Matanzas	0,369167
5	Sancti Spíritus	0,325249
6	Santiago de Cuba	0,267637
7	Cienfuegos	—0,063179
8	Guantánamo	—0,127539
9	Camagüey	—0,164227
10	Ciudad de La Habana	—0,187632
11	Pinar del Río	—0,222052
12	Ciego de Avila	—0,238156
13	Las Tunas	—0,473084
14	Holguín	—0,754071

Y la interpretación geométrica de esta clasificación puede verse en la figura 2.

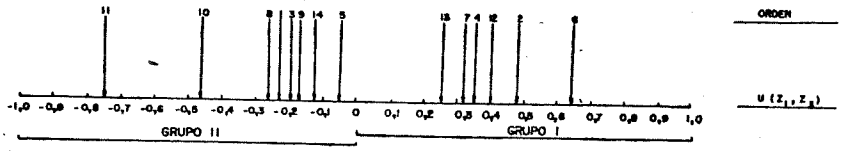

Figura 2

5. CONCLUSIONES

Hemos obtenido los siguientes resultados:

a) Una clasificación y un ordenamiento objetivos de todas las provincias o fuentes de información.

b) Una función discriminante que permite clasificar cualquier otra fuente de información similar que llegue posteriormente.

c) Un criterio sobre las variables Z_1 y Z_3 que lleva implícito el tomar sus componentes, las variables Y, X_1 y X_3, definidas en el epígrafe 4.1., como las que mayor confiabilidad tienen, ya que ha sido sobre ellas tres que se ha realizado la clasificación.

Podemos retomar ahora los individuos o provincias pertenecientes al Grupo I y trabajar con ellos las tres variables Y, X_1 y X_3, derivando las ecuaciones y parámetros según los métodos descritos en el epígrafe 2.1.

Podemos retomar los individuos o provincias clasificados en el Grupo II, hacer un trabajo acerca de ellos para analizar las causas que inciden en su pobre información y ponerles remedio, si es que ha habido deficiencias.

Podemos observar finalmente que, en la medida que se aumenta el número de individuos en la muestra, más confiabilidad podemos obtener en los resultados. Por lo tanto, subdividiendo las provincias o individuos en los elementos que la integran y que son independientes o autónomos, podremos tener criterios más sólidos con respecto a la calidad de su

información, ya que la provincia es la reunión de las brigadas constructoras que la componen.

Entonces, realizando este mismo trabajo por brigada constructora, lograríamos estimaciones más finas, probabilidades de mala clasificación más pequeñas y, por lo tanto, resultados más completos.

6. RECONOCIMIENTO

Agradecemos al profesor Josef Machek, del departamento de Matemática Aplicada de la Universidad de Carolina, de Praga, la lectura del manuscrito original y sus valiosas acotaciones y sugerencias.

7. BIBLIOGRAFIA

1. *Díaz, Elba. Prof.:* Conferencias de Análisis Multidimensional, Universidad de La Habana. 1973.
2. *Romeu, Jorge Luis. Lic.:* Informe Estadístico del IME. Grupo Vial Nacional, DAP, 1975.
3. *Tomassone, R. Prof.:* Conferencias de Análisis Discriminante, Universidad de La Habana. 1973.