SISTEMAS DE PROGRMACION PARA TRATAMIENTO ESTADISTICO DE DATOS

Por J.A. Martínez Carrillo

La elaboración estadística de datos fue la primera aplicación de las máquinas de proceso de información (Hollerith y las tarjetas perforadas) que más interés ha suscitado a estudiosos y profesionales del proceso de información o beneficiarios del mismo.

Sin embargo, y a pesar del desarrollo de los lenguajes que emplean traductores para la puesta en máquina de los programas, la programación en estadística es difícil y poco fiable. El panorama del usuario se agrava con la proliferación de métodos, en particular en análisis multivariacional y en clasificación de datos. Estos métodos frecuentemente no han sido estudiados suficientemente o son equivalentes o caso particular de otros ya conocidos.

Estas circunstancias han movido a prestigiosas instituciones tales como la B.S.S. y la N.B.S. a promover estudios de revision y comparativos para marcar líneas claras y seguras.

Es dentro de esta idea en la que planteamos el presente trabajo. Nuestro propósito es comparar las características de los recursos de programación más ampliamente difundidos y fomentar la actitud crítica en este área del proceso de datos.

Consideramos los siguientes puntos:

1º) Sistemas de programas

2º) Sistemas de programación

3º) Lenguajes y estadística

4º) Precisión en los cálculos estadísticos

5º) Métodos no usuales

1- SISTEMAS DE PROGRAMAS

Entenderemos por sistemas de programas, colecciones construidas con un criterio unificado en el diseño. Los programas así escritos tienen la unidad que les da ese criterio pero son piezas independientes, en el sentido que no se pueden evocar unas a otras y solo a través del sistema operativo pueden encadenarse. Este encadenamiento depende también de la forma que tenga la base de datos comunes a todos ellos. Las deficiencias de intercomunicación son cubiertas con programas intermediarios para la reestructuración del fichero intermedio y utilizando la lectura de formatos en ejecución.

En este sentido se pueden considerar como sistemas de programas, entre los de uso difundido y accesibles, BMD (Dixon, 1968) y todos los que de el se han derivado por transformaciones menores.

De todas estas bibliotecas unificadas de programas o sistemas de programas sobresale BMD por la variedad de algoritmos ampliamente usados. Es a la vez uno de los esfuerzos mas antiguos y sostenidos. Precisa en las implantaciones usuales 128K (octetos). Comprende dos

series de programas: la serie principal, muy estudiada y la serie experimental (BMD - X) en estudio. La primera comprende 44 programas, cada uno de los cuales cubre la totalidad de esa memoria. Están organizados en seis clases:

clase D - descripción y tabulación

clase M - análisis multivariacional

clase R - análisis de regresión

clase S - programas especiales

clase T - análisis de series temporales

clase V - análisis de la varianza

En la serie BMD - X hay unos veinte programas. Su contenido varía de acuerdo con la experiencia ganada sobre ellos, según la cual pasan a la serie principal.

Esta colección ha servido por su difusión para estudiar la eficacia relativa de otros métodos y técnicas. También ha sido la base para numerosas transformaciones o reprogramaciones que han hecho resaltar cuestiones particulares.

Otra colección importante es SSUPAC. Fueron desarrollados hasta 1964 por la Statistical Service Unit de la Universidad de Illinois en 7094 con 1301 Disk File y dos 1401 y posteriormente transportados a otras máquinas. Es una colección de 47 programas de muy variadas características. Los cuatro programas de análisis factorial han sido muy estudiados y utilizados. Tiene buenas facilidades para el proceso de ficheros extensos, extracción de informes estadísticos y representaciones gráficas.

Hay otras muchas colecciones extensas o parciales; en gran medida derivadas de BMD, algunas de las cuales han sido industrializadas. La actitud en general mas que a usar programas se dirige a la redacción de nuevas codificaciones que introducen modificaciones de detalle. Son mas o menos utilizadas las siguientes: la librería de rutinas del proyecto MAC 7094, la de C-E-I-R Multi-Access Computer Services, la del Dartmouth College, la del National Bureau of Standard, el paquete de rutinas S.S.Pde IBM, los 1108 Univac Math-Pack programas, los del National Bureau of Standard, y los de algunas entidades industriales tales como Sandia Co. y Esso.

2- SISTEMAS DE PROGRAMACION

Llamamos sistemas de programación en estadística a programas capaces de realizar las acciones sobre las tablas de datos que puedan ser formuladas en términos de los comandos u órdenes que constituyen en lenguaje de programación del sistema.

Están formados por un programa central de control que interpreta las secuencias de comandos presentadas al sistema para ser ejecutadas y de una serie de rutinas que entran en ejecución oportunamente evocadas por el programa central.

Estos sistemas tienen actualmente algunas características básicas

cas:

a) Necesitan memorias extensas

b) Tienen facilidades para manipular los ficheros

c) Permiten emitir descripciones estadísticas usuales

d) Tienen serias limitaciones en la reformulación de los algoritmos y reprogramación de los mismos.

e) La introducción de un nuevo programa, posibilidad que siempre se da por supuesta, no siempre es posible practicamente por la necesidad de establecer compatibilidad con el resto del sistema.

Hay muchos sistemas (Biomed, Ascop, Geustat, P-Stat, Statpack) de esta naturaleza, buena parte de ellos desarrollados por empresas de informática industrial. Los más conocidos son:

S.P.S.S. Statistical Package for the Social Sciences
Universidad de Chicago.

DATA-TEXT Dept. of Social Relations
Universidad de Harvard.

OSIRIS del que hay dos versiones desarrolladas por Institute of Social Research (ISR) y por The Inter-University Consortium for Polit. Research.
Ambas de: Ann Arbor - Michigan.

S.P.S.S. tiene un único programa de control y necesita de 160 a 230 octetos y está escrito en ensamblador y FORTRAN y hay versiones para CD6 6600, Univac 110B, RCA Spectra, PDP 10, B 6500. Hay realizados actualmente unas 80 implantaciones. Su gramática ha sido formulada. El lenguaje está formado por series de órdenes.

DATA-TEXT (distinto de DATATEXT, IBM) fue inicialmente puesto en 7090/94. Está especialmente orientado hacia la reestructuración de las tablas de datos. Tiene un solo programa monitor y ha sido reprogramado para CD6 6000 y IBM 360. Escrito en ensamblador y FORTRAN. Necesita de 200 a 250 K byts.

OSIRIS II: La última versión de OSIRIS necesita de 100 a 200 k (byts). Esta escrito en ensamblador FORTRAN y PLS. Se usa en IBM 360, CDC 6000, SIEMENS y RCA. muy recientemente ha sido anunciada una nueva versión: OSIRIS III.

Estos sistemas tienen previstas tres tipos de procesos básicos: elaboración de los ficheros, estadística descriptiva y análisis estadístico.

Tratamiento de ficheros: Suponiendo que es necesario manejar ficheros del orden de 10^8 a 10^{10} caracteres tales como se requiere en estadísticas nacionales o en grandes encuestas colectivas las po-

sibilidades de SPSS se ven ampliamente superadas por DATA-TEXT y - OSIRIS en particular en transformación de datos, extracción de subficheros y adición de subficheros.

Estadística descriptiva: Esta función es posiblemente la mejor estudiada y cubierta con igual eficacia por los programas existentes. Las diferencias son pequeñas y los procesos muy seguros y - depurados de errores.

Análisis estadístico: Los tres sistemas de programación son muy deficitarios de este tipo de programas. Sería difícil hacer un catálogo de todos los métodos disponibles en la literatura estadística y cuales de ellos están presentes. De todos ellos OSIRIS II es el más extenso. De todas formas no sería acertado elegir sin una previa revisión cuidadosa de la precisión de los resultados que a veces difiere en un 100% de los calculados a mano con gran precisión.

Entre los sistemas de programación ocupa un lugar especial el llamado OMNITAB de H. Himtelratz de N.B.S. Trabaja de forma inusual. Dispone de una tabla única de trabajo en la cual era un sencillo lenguaje de comandos, pueden hacerse una amplia gama de transformaciones predefinidas en las matrices de datos con alta precisión. Este - es uno de los pocos sistemas donde se ha estudiado exigentemente la precisión numérica de los resultados.

3- LENGUAJES DE PROGRAMACION EN ANALISIS ESTADISTICO

Los estadísticos profesionales encuentran a los sistemas de - programación extremadamente rígidos. Las razones que encuentran son:

a) Solo los diseñadores pueden de hecho hacer correcciones o adiciones y a veces ellos tampoco. Por otra parte no se encuentran habitualmente inclinados a hacer estas modificaciones que no suelen ser interesantes para ello aunque son vitales para otros. De los conocidos solo P-Stat da información de como hacer modificaciones y no es cuestión sencilla.

b) Operaciones sencillas en un lenguaje son prácticamente difíciles en los sistemas de programación. Se puede manipular vectores pero es difícil a veces manejar escalares.

c) No se pueden hacer proceso cíclico o condicionales. Tampoco puede hacerse designación simbólica de identificadores.

d) No hay definiciones amplias de las posibles estructuras de datos que van a ser presentados al ordenador. Las descripción depende demasiado estrechamente de los casos establecidos sobre la ficha perforada.

Otra forma de trabajar en el proceso de datos en estadística consiste en enriquecer los compiladores con colecciones de rutinas - o macrogeneradores especiales o incluso diseñar lenguajes especiales. Esta es una dirección prometedora y poco explorada.

Como lenguajes de programación han sido usados los subconjuntos del lenguaje de K. Iverson. Las facilidades del APL 1130 y 360 han sido utilizadas extensamente en la Universidad de Alberta donde se ha construído una colección de programas. La utilización de APL da lugar a programas de redacción fáicl pero de difícil lectura. Si bien los lenguajes de programación en general son escritura fácil y lectura difícil, en APL esta diferencia se hace más marcada. Esta será quizás la más importante restricción en el uso de este lenguaje. Sin embargo, la ausencia de ciclos y las compiladores incrementales le hacen muy atrayente para los estadísticos.

En cuanto a PL I no añade posibilidades importantes respecto a ALGOL 60 o a FORTRAN IV. Podríamos decir que es FORTRAN el lenguaje más conveniente en cuanto al uso de ficheros externos tan frecuentemente usados.

Estos ficheros tienen estructuras sencillas que no precisa las facilidades adicionales de COBOL o PL I, por lo cual no dejan de ser interesantes en las fases iniciales de depuración de datos en los procesos extensos.

De hecho FORTRAN IV es el lenguaje más ampliamente aceptado, a veces incrementado con bibliotecas de rutinas.

Son interesantes varios estudios realizados para construir lenguajes derivados más o menos de ALGOL 60 para hacer estadística computacional. En particular los de Dautsing y OL/II de Michigan para partir grandes matrices se han desarrollado recientemente.

Para subsanar todas las dificultades de una forma unificada. Gower ha construído un lenguaje de características avanzadas (derivado del Algol 68) y el compilador necesario.

4- PRECISION EN LOS CALCULOS ESTADISTICOS

Solo en pocos programas ha sido estudiado este problema. Muchos de los programas se han escrito de acuerdo con procedimientos de cálculo establecidos en máquinas de despacho con 13 o 15 cifras significativas.

Al ejecutarse estos programas en máquinas de 16, 24 o 32 bits la pérdida de precisión es a veces muy grande.

Wampler ha estudiado la precisión en aproximaciones de funciones lineales por mínimos cuadrados en varios tipos de máquinas y algoritmos. A visto que, precisamente los métodos más usados y recomendados como son los de eliminación, son peligrosos e imprecisos, mientras que los menos extendidos de Householder y de Gram-Schmidt se muestran como más precisos.

Estas consideraciones no han sido repetidas en otras áreas tales como el análisis factorial y análisis de la varianza donde serían muy útiles. Parece ser como si los programadores estadísticos no tuvieran frecuentemente en cuenta los avances del cálculo numérico. Por descontado los usuarios son todavía bastante menos críticos en estos asuntos.

5- PROCEDIMIENTOS NO USUALES

Otras formas no usuales de hacer cálculo estadístico son:

a) El empleo de procedimientos de generación automática de codificaciones usando programas de definición. Se ha estudiado en análisis de varianza.

b) Métodos conversacionales de definición. Igualmente empleados en análisis de varianza.