PAQUETES INFORMATICOS EN LA MEJORA GENETICA DEL VACUNO DE CARNE

COMPUTER PACKAGES IN BEEF CATTLE BREEDING

Molina Alcalá, A.[1], M.C. Crespo Giráldez[2], J.V. Delgado Bermejo[1] y A. Rodero Franganillo[1]

[1]Departamento de Genética. Facultad de Veterinaria. Universidad de Córdoba. 14005 Córdoba. España.
[2]Inspección Veterinaria comarcal. Elvira nº 6. 41400 Ecija. Sevilla. España.

Palabras clave adicionales

Mejora genética animal. Vacuno Retinto. Programas de ordenador.

Additional keywords

Animal breeding. Retinta beef cattle. Software

RESUMEN

Asistimos en las últimas décadas a una creciente potencia de cálculo de los ordenadores y a un sensible abaratamiento de estos. Esto ha permitido el desarrollo de modelos matemáticos para la resolución de problemas en mejora genética animal cada vez más complejos, la creación de algoritmos más sofisticados para su solución y la aparición de gran número de paquetes informáticos capaces de satisfacer las crecientes necesidades en este campo. Por otra parte, la expansión de las redes informáticas, especialmente la red INTERNET, *la gran red de redes*, prácticamente a todas las partes del mundo y la creación de diferentes grupos de discusión en mejora, como es el caso de AGDG (Animal Geneticists Discussion Group) a nivel internacional o ACTEON a nivel nacional, han puesto a disposición de los genetistas un amplio abanico de soluciones informáticas, tendentes a la resolución de los dos principales problemas que se plantean en la mejora animal, la estimación de los componentes de la varianza-covarianza y la predicción del valor genético de los reproductores (resolución de las ecuaciones de los modelos mixtos).

En el presente trabajo realizamos un análisis crítico de los principales programas no comerciales capaces de resolver esta problemática en la mejora genética del vacuno de carne: ABTK, DMU, MTDFREML, CMMAT y CMIT, JAA, MTDFS, DFREML, JSPFS, PEST; así como de las plataformas y sus sistemas operativos donde se pueden ejecutar y del lenguaje en el que están programados.

SUMMARY

In the last decades, we have witnessed a growing calculation power of computers and a dramatic cheapening of these. This has permitted development of mathematical models for problem resolution in increasingly complex animal genetic improvement, the creation of more sophisticated algorithms for its solution and the appearance of a great number of computer packages capable of satisfying the growing needs in this field. On the other hand, the expansion of the data processing nets, especially INTERNET *the great net's net*, practical to all the parts of the world and the creation of different discussion groups in animal breeding, as is the case of AGDG (Animal

Geneticists Discussion Group) on an international level or ACTEON on thet national level. They have put at the disposal of the geneticists a wide fan of data processing solutions, tending to the resolution of the principal two problems that are outlined in animal improvement, the estimate of the components of the variance - covariance and the forecast of the animal genetic value (resolution of the mixed models equations).

In the present work we make a critical analysis of the principal non commercial programs capable of solving this problem in genetic breeding of beef cattle: ABTK, DMU, MTDFREML, CMMAT and CMIT, JAA, MTDFS, DFREML, JSPFS, PEST; as well as those of the platforms and their operative systems where they can be executed and of the language in which they are programmed.

INTRODUCCION

La evolución de las técnicas de computación utilizadas en la mejora animal ha sido posible gracias al progreso del *hardware* (ordenadores). Así, aunque los planteamientos teóricos sean más antiguos, hasta que la potencia de cálculo de los ordenadores no lo ha permitido, no se han podido simular estos modelos matemáticos, y posteriormente, una vez puestos a punto los programas informáticos, llevarlos a la práctica como modelos de rutina.

Los genetistas, aunque pueden utilizar *software* comercial como Matlab o SAS, generalmente necesitan la resolución de problemas tan específicos que no tienen cabida en estos programas estándares, por lo que se ven forzados, mucho más que en cualquier disciplina de la ciencia animal, a desarrollar sus propios programas. Por otra parte, en mejora animal los modelos matemáticos son cada vez más complejos y pocos grupos de investigación son capaces de dedicar grandes recursos para el desarrollo de programas cada vez más sofisticados.

Las redes de ordenadores pueden distribuir rápidamente cualquier programa informático disponible a través del mundo en cuestión de minutos. Esta misma red permite el correo electrónico entre los desarrolladores de estos programas y los usuarios. El compartir el *software* presenta una serie de ventajas, pero también inconvenientes para los desarrolladores. Entre las ventajas podríamos destacar que al ser evaluados los programas por muchos usuarios y bajo condiciones muy diversas, los posibles errores son rápidamente detectados, por otra parte los desarrolladores reciben un amplio reconocimiento por su trabajo, como desventaja, los programadores deben tener en cuenta la amplia variedad de situaciones (modelos, especies, plataformas ....) en los que podría ejecutarse su *software*, lo cual repercute en programas mas complicados y laboriosos.

El número creciente de programas con similares propósitos hace necesaria su comparación, aunque es difícil seleccionar el mejor paquete informático para todo el mundo, ya que cada programa tiene sus puntos fuertes y sus puntos débiles y estos dependerán en parte de las necesidades del usuario final (modelo, tamaño de las bases de datos, plataformas etc). Además este programa informático evoluciona continuamente, por lo que la comparación de estos programas puede ser muy tediosa y complicada, ya que habría que probarlos en muchas situaciones, por lo que el análisis se realizará principalmente en base a su documentación.

RELACION ENTRE HARDWARE Y DESARROLLO DE APLICACIONES

EVOLUCION DE LAS PLATAFORMAS

Es difícil argumentar en contra de la importancia del ordenador en el desarrollo de la mejora. Su rápida evolución exige y va a exigir una continua transformación de los programas y de los modelos que se pueden computar de forma rutinaria. De forma simplificada podemos distinguir 3 grandes grupos de plataformas informáticas:

- SUPERCOMPUTADORES (ordenadores tipo CRAY) con una elevadísima potencia de cálculo, pero que debido a su gran coste sólo están disponibles para grandes empresas públicas y multinacionales.

- MINIORDENADORES, y ESTACIONES DE TRABAJO. Con una inferior potencia, pero a un costo que permite que lo posean los centros de cálculo científico o la mayoría de las empresas.

- MICRORDENADORES. Computadoras basadas en microprocesadores como el Motorola (APPLE principalmente) o Intel (PC). De una potencia mucho menor a los anteriores y muy difundidos tanto en el hogar como en cualquier puesto de trabajo.

Hasta hace relativamente pocos años los programas informáticos de aplicación en la mejora genética animal sólo se podían ejecutar en supercomputadoras o miniordenadores de la gama alta. En los últimos años asistimos a un espectacular aumento de potencia del *hardware*, de forma relativa más evidente en los

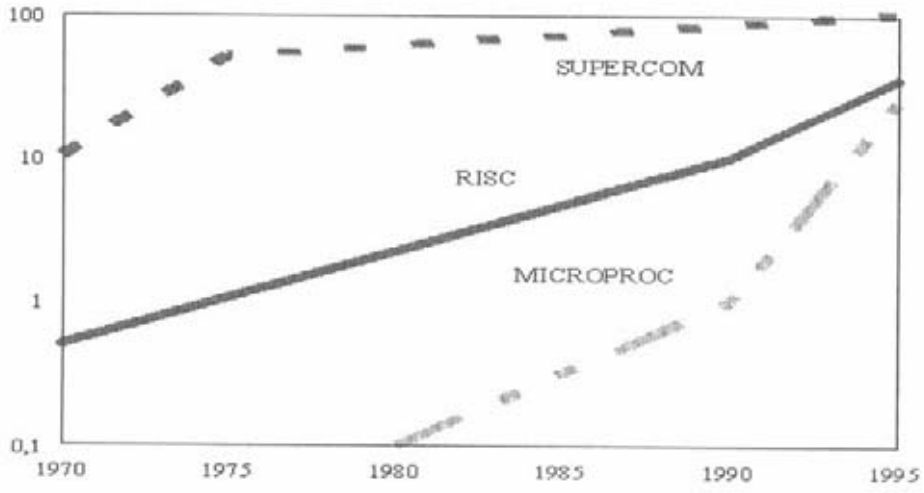

Figura 1. *Progreso relativo de las principales plataformas sobre las que se puede ejecutar software específico para la mejora genética animal.* (Relative progress of the principal platforms on those which specific software for animal breeding can be executed).

miniordenadores y estaciones de trabajo (*workstations*) en relación a los superordenadores **(figura 1)**. Esta circunstancia, unida a la drástica disminución de precios, RISC por debajo del millón (incluso menos en el caso del RISC POWERPC y POWERMAC) o de las 500000 pta en el caso del PENTIUM, han ocasionado que hoy día los grandes ordenadores se reserven a la investigación de nuevas metodologías (p.e. muestreo de GIBBS) o grandes procesos de simulación (p.e. Monte Carlo Markov), estando la evaluación de rutina en manos de los RISC para grandes volúmenes de datos (p.e. el Plan Nacional de Mejora del Vacuno en U.S.A.) o de microprocesadores en los pequeños-medianos.

Si evaluamos estos dos últimos, aunque la relación potencia/precio inicial está a favor de los RISC, el coste añadido de aprendizaje y mantenimiento es muy superior al de los PC. El inconveniente de estos últimos radica en que normalmente están ligados a un sistema operativo como el DOS *(Disk operating system)* que presenta graves inconvenientes para la ejecución de *software* específico para la mejora genética. No obstante, esta situación podría resolverse con el cambio al UNIX para PC (Motif, Linux etc...), OS/2 o al Windows-95 cuando madure lo suficiente.

SISTEMAS OPERATIVOS, LENGUAJES DE PROGRAMACION Y COMPILADORES

Históricamente, y aún hoy dia, la mayoría del *software* específico para mejora genética animal ha sido escrito en FORTRAN, en concreto el estándar del año 1977 (Fortran 77), un lenguaje estructurado muy eficiente en las operaciones numéricas y que ha ido acumulando un gran número de bibliotecas de subrutinas que facilitan mucho la programación, a pesar de que no es un lenguaje muy rico en órdenes y estructuras de programación y no posee gestores de memoria. Este último inconveniente hace que sea dificil de escribir programas largos (esto es especialmente grave en PC bajo DOS). Todos los programas analizados menos ABTK están escritos en Fortran 77.

Estas limitaciones hacen que muchos programadores utilicen extensiones al Fortran, haciendo que los programas no sean totalmente compatibles con otras plataformas, agravado por el hecho de que aunque el Fortran-77 es un estándar, presenta una serie de pequeñas diferencias dependiendo del compilador y de la plataforma donde se instale. Esto hace que un programa escrito para una plataforma necesite modificarse para ejecutarse en otras diferentes.

Una posible solución sería el desarrollar aplicaciones bajo FORTRAN-90 que soporta la sintaxis de Fortran-77, presenta una mejor gestión de la memoria y más estructuras de programación. Su gran inconveniente es que todavía está poco generalizado debido a la escasez de compiladores.

Actualmente la tendencia en todas las disciplinas es a programar en un lenguaje como el C (aunque fue desarrollado hace 20 años), que presenta un nivel de implementación más bajo, y por tanto mucho más potente. Este lenguaje resuelve muchas de las limitaciones del Fortran (p.e. los gestores de memoria,) aunque es más dificil de programar y aún tiene muy pocas librerías de subrutinas específicas disponibles. Hoy día el único paquete de los analizados desarrollado en C es ABTK. En el futuro asistiremos

a un auge del C++, lenguaje con todas las ventajas y potencia del C y que está orientado a objetos (el programador se centra en que hay que hacer, encargándose el lenguaje de definir como hacerlo).

En cuanto a los sistemas operativos, hoy día la mayoría de programadores y usuarios han emigrado a plataformas UNIX (SUN, APOLO ...) debido a su gran potencia, propiedades multitarea y multiusuario (lo que permite abaratar mucho los costes), numerosas utilidades y máximo aprovechamiento de la memoria, así como a su creciente evolución al no tener propietario. Pero el UNIX y sus compiladores no son 100 p.100 compatibles, por lo que la instalación de la mayoría de los programas no suele ser fácil y a menudo requiere elevados conocimientos de fortran, compiladores y del UNIX.

En general, los programas que corren bajo PC con el S.O. DOS, son mucho más fáciles de instalar que los de UNIX (p.e. la serie de Misztal), y cuanto más grande y complejo sea el paquete, más difícil será su instalación. PEST y VCE son los únicos que se venden preparados para cada plataforma específica (PC bajo DOS y OS/2, APPLE bajo System 7, SUN bajo UNIX etc.), por lo que las modificaciones que se necesitan son mínimas (especialmente debidas a las necesidades de cada usuario y a las diferencias entre compiladores).

PAQUETES INFORMATICOS

CONSIDERACIONES GENERALES

La tarea principal de los paquetes descritos aquí es obtener las soluciones de los modelos mixtos (solución de los factores fijos y del valor genético de los reproductores) y la estimación de los componentes de la varianza (para la estimación de los parámetros genéticos) mediante la metodología REML (máxima verosimilitud restringida). Como mínimo todos los programas soportan el BLUP (*Best linear unbiased predictor*) modelo animal (salvo CMIT y CMMAT) con efectos fijos, aleatorios, interacciones entre los factores y covariables. En el resto de propiedades, los programas difieren mucho; así unos pueden ser útiles para pequeñas bases de datos (hasta 10000 ecuaciones), medianas (10000-100000) o grandes (>100000); unos son unicarácter (en general todas las primeras versiones), mientras otros multicarácter (últimas versiones de todos los programas salvo CMIT, CMMAT, JAA y JSPFS); entre estos últimos, unos pueden permitir diferentes modelos para cada variable (PEST y VCE) y otros no (resto de programas). Otras características que pueden soportar son la consanguinidad (PEST, DFREML, MTDFREML, VCE, ABTK, y DMU), efectos maternos (DFREML, MTDFREML, VCE, ABTK, DMU, PEST, o MTDFS), grupos de padres desconocidos; hipótesis sobre los componentes de la varianza o calcular el error de las predicciones (PEV) para grandes conjuntos de datos (DMU, PEST, JAA). Algunos paquetes presentan un preprocesado de los datos, con recodificación de los factores alfanuméricos (PEST y VCE), validación de la consistencia del pedigrí en cuanto al orden padre-hijos y fechas de nacimiento, eliminación de animales no contribuyentes y asignación de grupos de padres desconocidos (PEST, VCE, DFREML, RENUM). Por último existen otras características técnicas que determinan la fiabilidad de los resultados

así como la velocidad de su obtención.

MODELOS Y ESPECIES

Los modelos utilizados para analizar los datos dependen generalmente de la especie animal en cuestión. Así en vacuno de leche los registros productivos suelen analizarse mediante un modelo unicarácter con repetibilidad, mientras que los datos de conformación por medio de un modelo multicarácter con repetibilidad. El efecto del grupo de padres desconocidos es considerado importante, mientras que la consanguinidad no. Para cerdos y aves el modelo adecuado suele ser multicarácter. En vacuno de carne es considerado esencial un modelos multicarácter con efectos maternales, no considerándose importante la consanguinidad ni los grupos de padres desconocidos.

FACILIDAD DE USO

Se podría dividir a su vez en facilidad de aprendizaje y de uso una vez que se maneja con soltura. PEST (y VCE) es el único que presenta una interfaz *amigable*, aunque al ser necesario su programación no es demasiado fácil de aprender (sin llegar a niveles de complejidad como ABTK o DMU), salvo si se es programador de SAS, ya que presenta una interfaz muy parecida a éste. DMU utiliza una combinación de programas que a su vez llaman a determinadas rutinas, por lo que es un programa fácil de usar pero no de aprender ya que no se comprende lo que se hace (sólo se ejecutan una serie de programas numerados en cadena). El resto de programas, DFREML, MTDFREML, la serie de Misztal (JAA etc.) se pueden usar interactivamente o mediante ficheros de parámetros. Dentro de estos los más fáciles de usar son los de Misztal, aunque a costa de ser menos potentes en la mayoría de los aspectos.

CONSIDERACIONES TECNICAS

RECODIFICACION DE LOS DATOS

Todos los programas aceptan al menos el formato libre para los ficheros de datos. PEST (y VCE) es el único capaz de recodificar los datos de tipo alfanumérico, aunque hay algunas utilidades como RENUM (Misztal, 1993), capaces de recodificar, aunque sólo sea los códigos numéricos de los animales. El resto de los paquetes aceptan sólo datos numéricos y además los factores deben ser recodificados previamente (proceso muy tedioso, especialmente para factores combinados). PEST, VCE, DFREML y utilidades como RENUM permiten eliminar los animales del pedigrí que no contribuyen a las evaluaciones y comprobar que los padres aparecen antes que los hijos; además son capaces de asignar grupos de padres desconocidos en base a la fecha de nacimiento y al periodo medio entre generaciones.

COSTE COMPUTACIONAL

Los programas están compuestos por muchos bloques como son la preparación de datos, creación de las EMM (ecuaciones de los modelos mixtos), cálculo de soluciones, etc. Cada una de estas tareas puede resolverse siguiendo diferentes estrategias, las cuales van a definir las características del programa.

ESTRATEGIAS PARA LA RESOLUCION DEL SISTEMA DE ECUACIONES, DEL REML

Generalmente el procedimiento más costoso es el cálculo de determinantes y

trazas. Aunque existen diversas estrategias para este cálculo, todos los paquetes analizados utilizan la factorización de la matriz (una excepción serían las primeras revisiones del DFREML que utilizaba la absorción). PEST utiliza un módulo de factorización comercial incluido en el coste del paquete, el resto de paquetes realizaban la inversión de la matriz hasta la aparición del módulo de libre distribución FSPAK (Pérez-Enciso, Misztal y Elzo, 1994).

En cuanto a la velocidad y exactitud de las estimaciones REML, dependen básicamente de la estrategia de maximización utilizada. La popular maximización *libre de derivadas* (DF) utilizada por DFREML, MTDFREML y DMU, es muy lenta en modelos multicarácter, necesitando muchas iteraciones hasta llegar a la convergencia, mientras que los basados en la derivada (D) no dependen del número de caracteres. Dentro de estas dos estrategias también dependerá del algoritmo usado, asi p.e. en los que usan la derivada (D) de la función de verosimilitud, pueden reducir mucho el coste en tiempo utilizando el algoritmo EM (Maximización de la Expectación) acelerado.

Programas que usen la Derivada no son frecuentes ya que los algoritmos DF son mucho más fáciles de programar y la inversión (antes de la llegada de FSPAK) era muy complicado; además necesita mucha memoria y tiempo. El único programa que utiliza D (algoritmo de Newton-Raphson) es el DMU. Por otra parte las soluciones DF son menos precisas que el D ya que debe encontrar un máximo iterativamente (en vez de derivarlo la función de verosimilitud) y depende de los puntos de partida y superficie de búsqueda (con la posibilidad de llegar a máximos locales).

En modelos con más de 4 caracteres el DF puede ser demasiado lento e impreciso, siendo hoy día el único procedimiento adecuado la transformación canónica (CT) donde el coste computacional crece de forma lineal en relación al número de caracteres, en vez de forma cuadrática. No obstante sólo se soportan determinados modelos (los factores deben ser idénticos, todos los caracteres recogidos y un sólo efecto aleatorio). Este método CT es soportado por DFREML, VCE y MTDFS, aunque MTDFS mediante una modificación del CT es capaz de utilizar varios efectos aleatorios extendiéndose a modelos con repetibilidad.

En cuanto a los algoritmos para resolver los sistemas de ecuaciones (EMM) incluyen resolución directa en memoria (DM), iteración en memoria (IM), en disco (ID) o en disco sobre los datos (IDa). Los primeros dan soluciones muy precisas, peros sólo son adecuados para sistemas de pocas ecuaciones (10000-30000 ecuaciones) debido al elevado coste en términos de memoria (DFREML, MTDFREML y ABTK, VCE, JSPFS, MTDFS y CMMAT utilizan este sistema). Los métodos iterativos permiten una menor exactitud y velocidad pero reducen drásticamente las necesidades de memoria. Los métodos de iteración en memoria (IM) son apropiados para resolver de 30000 a 500000 ecuaciones (PEST, VCE y DMU). Los métodos de iteración en disco (ID) son más lentos pero están menos condicionados por los límites de memoria, aunque consumen gran cantidad de espacio en disco (PEST y VCE). Como los sistemas de ecuaciones ocupan mucho más espacio que los propios datos, la iteración directa sobre es-

tos (IDa), con creación de una ecuación en cada iteración, ahorra mucho espacio en disco. Es por lo tanto el método más adecuado para grandes cantidades de datos, permitiendo la resolución de más de 500000 ecuaciones (JAA, DMU, PEST, VCE y CMI).

OPERACIONES EN DISCO

Los programas iterativos donde se leen y escriben en el disco los datos o los coeficientes de las matrices, conllevan un elevado conste computacional (tiempo), llegando incluso a suponer el 95 p.100 del tiempo total. Por tanto la eficacia de estos programas dependerá básicamente de la plataforma y las características del ordenador donde se ejecute (p.e. PC-AT<<PC PENTIUM; HD 28ms<<8ms, IDE<<SCSI).

Una de las características a tener en cuenta es que las entradas sin formato (libres) son mucho más lentas que las entradas con formato (si bien hay que perder el tiempo preparando datos, aunque tendrían la ventaja adicional de que no tiene repercusión los datos vacíos). Esto es más evidente cuanto mayor sea la base de datos. Todos los paquetes capaces de resolver grandes cantidades de ecuaciones son capaces de leer los datos en formato libre y todos salvo el PEST utilizan algoritmos de lectura acelerada.

DESCRIPCION DE LOS PRINCIPALES PROGRAMAS

En las **tablas I, II** y **III** se pueden observar las principales características de los programas analizados. En la **tabla IV** presentamos a modo de anexo la

Tabla I. *Resumen de las principales características de los paquetes para la valoración de reproductores analizados.* (Summary of the main characteristics of the software packages for breeding value prediction analyzed).

	CMIT		CMMAT		JAA		PEST	
	v1	v2	v1	v2	v1	v20	v1	v3
Blup animal	-	-	-	-	+	+	+	+
Multicarácter	-	-	-	-	-	-	-	+
Efectos maternos	-	+	-	+	-	-	+	+
Consanguinidad	-	-	-	-	-	-	+	+
P.E.V.	-	-	-	-	-	+	+	+
Facilidad	+	+	+	+	+	+	+	+
Recodifica factores	*	*	*	*	*	*	+	+
Recodifica pedigrí	*	*	*	*	*	*	+	+
REML	DF	DF	DF	DF	DF	DF	DF	DF
Método multicarácter								TC,dCh
E.M.M.	IDa	IDa	DM	DM	IDa	IDa	DM,IM,ID,IDa	DM,IM,ID,IDa

*Con RENUM (With RENUM); DF= Algoritmo libre de derivada; CT= Transformación canónica; DM= Resolución directa en memoria; dCh= Descripción de Chelosky; IDa= Iteración en disco sobre los datos; IM= Iteración en memoria; ID= Iteración sobre el disco

Tabla II. *Resumen de las principales características de los paquetes para estimación de parámetros genéticos analizados.* (Summary of the main characteristics of the software packages for the genetic parameter estimation analyzed).

	DFREML		MTDFREML	VCE	JSPFS	MTDF	
	v1	v2				v1	v3
Blup animal	+	+	+	+	+	+	+
Multicarácter	-	+	+	+	-	+	+
Método multicarácter	CT	CT	dCh,CT		DF	CT	
REML	DF	DF	DF	DF	DF	DF	DF
E.M.M.	DM	DM	DM	DM,IM,ID,IDa	DM	DM	DM
Efectos maternos	+	+	+	+	-	-	+
Consanguinidad	-	+	+	+	-	-	-
Recodifica factores	-	-	-	+	*	*	*
Recodifica pedigrí	+	+	-	+	*	*	*
Facilidad	--	-	-	-	+	+	+

*Con RENUM (With RENUM); DF= Algoritmo libre de derivada; CT= Transformación canónica; DM= Resolución directa en memoria; dCh= Descripción de Chelosky; IDa= Iteración en disco sobre los datos; IM= Iteración en memoria; ID= Iteración sobre el disco

dirección del correo electrónico (*e-mail*) de sus autores y los servidores de FTP anonymous donde se pueden obtener (salvo PEST y VCE que al tener un pequeño costo se tienen que solicitar directamente a los autores).

ESTIMACION DE LOS COMPONENTES DE LA VARIANZA

Conjunto de programas informáticos cuya finalidad es la estimación de los parámetros genéticos (heredabilidad, correlaciones fenotípicas, ambientales y genéticas ...) mediante la estimación de los componentes de la varianza-covarianza.

LSML ("HARVEY", LSMLMW)

Programa desarrollado en la década de los 60 por Walter Harvey. Fue el primer programa comercial de estimación de C.V. y solución de modelos mix-

Tabla III. *Resumen de las principales características de los paquetes de doble propósito analizados.* (Summary of the principal characteristics of the dual purpose software packages analyzed).

	ABTK	DMU
Blup animal	+	+
Multicarácter	+	+
Método multicarácter	TC	
REML	DF	D
E.M.M.	DM	IM,IDa
Efectos Maternos	+	+
Consanguinidad	+	+
P.E.V.	+	+
Recodifica factores	-	-
Recodifica pedigrí	-	-
Facilidad	---	--

D= Algoritmo basado en la derivada de la función de la máxima verosimilitud; DF= Algoritmo libre de derivada; CT= Transformación canónica; DM= Resolución directa en memoria; IDa= Iteración en disco sobre los datos; IM= Iteración en memoria; ID= Iteración sobre el disco

Tabla IV. *Direcciones del correo electrónico de los autores de los programas analizados y de los servidores de FTP Anonymous donde se pueden obtener.* (Electronic mail address of the authors of analyzed programs and FTP Anonymous servers where they can be obtained).

CORREO ELECTRONICO

K. Meyer	kmeyer@didgeridoo.une
v. Vleck	ansc418@unlvm.bitnet
E. Groeneveld	eg@jupiter.tzv.fal.d400.de
I. Misztal	ignacy@uiuc.edu
B. Golden	bgolden@cgel.agsci.colostate.edu
J. Jensen	lofjust@vm.uni-c.dk

SERVIDORES DE FTP ANONYMOUS

AGDG	cgel.agsci.colostate.edu
	metz.une.edu.au
	misz.animal.uiuc.edu
DMU	rs580.foulum.min.dk

tos, de amplísima difusión por todo el mundo, con diferentes versiones casi hasta la década de los 90. Su cometido era el cálculo de los estadísticos simples, la solución de los factores fijos, diversos contrastes de estos, la estimación de los componentes de la varianza (Henderson III) mediante siete tipos de modelos, la estima directa de parámetros genéticos (h^2, correlaciones genéticas, ambientales y fenotípicas) etc. Hoy día prácticamente no se usa ya que es poco flexible, no admite estimaciones REML ni el modelo Animal.

DFREML

Programa desarrollado por Karin Meyer en 1988 (en 1991 apareció la segunda versión). Fue el primer paquete público que implementaba el algoritmo DF del REML (Máxima verosimilitud restringida usando el algoritmo libre de derivadas). Tal vez sea el programa de estimación de los componentes de la varianza más citado internacionalmente y sigue siendo el estándar de comparación para el resto de programas con igual cometido. Destaca el test para las estimas de los componentes de la varianza.

Su amplísimo uso hace que se le considere prácticamente libre de errores y su único punto débil podría ser que soporta sólo diez modelos predefinidos, aunque todos los principales modelos están incluidos. No es adecuado para grandes bases de datos al resolver las EMM directamente en memoria. Su documentación es muy extensa y detallada. Su instalación, aprendizaje y uso no son demasiado fáciles.

MTDFREML

Programa más moderno (1993) desarrollado por Bolman, Kriese, Furgón Vleck y Kachman a partir de una versión de DFREML. Comparado con este es más fácil de instalar y usar, presenta modelos más generales, aunque tiene menos opciones adicionales (la última versión ya soporta los grupos de padres desconocidos y la consanguinidad). Presenta también la misma limitación en cuanto al tamaño de las bases de datos derivada del método de resolución de EMM.

VCE

Es un programa para estimar los C.V. desarrollado por los mismos investigadores que el PEST. Aunque se puede utilizar independientemente, es más eficaz con el PEST (este aporta todas las utilidades para la preparación de los

datos). Soporta varios algoritmos de cálculo y debe ser modificado y recompilado cada vez que utilizamos un modelo o unos datos diferentes mediante *script* UNIX.

JSPFS

Programa desarrollado de forma altruista por Ignacy Misztal. Es tal vez el programa más fácil de todos los analizados, aunque a costa de una menor potencia (unicarácter, sin efectos maternos, polinomios, tampoco es adecuado para grandes bases de datos etc.).

MTDFS

Programa desarrollado por I. Misztal, algo más completo (multicarácter), y con un manejo similar al de todos los programas desarrollados por este autor (JSPFS, JAA...), aunque no admite nada más que un efecto aleatorio (por lo tanto no calcula los efectos maternos), tampoco admite polinomios, no tiene en cuenta la consanguinidad, ni es válido para grandes bases de datos.

Existe una modificación de Nicolás Gengler (**MTC20**) capaz de soportar varios efectos aleatorios aunque con un solo modelo para todos las variables.

RESOLUCION DE MODELOS MIXTOS

Programas cuya finalidad es la valoración genética de los reproductores y obtención de las soluciones de los factores fijos mediante la resolución de las ecuaciones de los modelos mixtos.

CMIT y CMMAT

Programas desarrollados por I. Misztal, que aunque no soportan el BLUP modelo animal (sólo el modelo abuelo materno), son los únicos capaces de resolver modelos umbrales (*threshold models*) para datos de tipo categórico. La única diferencia entre ambos es la forma de obtención de las soluciones, mientras que CMIT itera directamente en los datos (es más lento pero es capaz de manejar grandes volúmenes de datos), CMMAT utiliza la absorción de la matriz en memoria (mucho más rápido pero sólo útil para bases de datos pequeñas).

En el año 1991 apareció la segunda versión de ambos (**CMIT y CMMAT V.2.0**) que permite estimar la (co) varianza para efectos directos y maternales.

JAA (MIXIT)

Programa desarrollado altruistamente por Ignacy Misztal para la obtención de la solución de modelos mixtos en modelos unicarácter mediante iteración directa sobre los datos, lo cual permite el procesado de grandes cantidades de información. También destaca que es capaz de dar las varianzas del error de la predicción (PEV) con el modelo animal en grandes bases de datos (**JAA20**, modificación de Nicolás Gengler). No soporta efectos maternos y el manual es muy pobre, asi mismo presenta pocas opciones adicionales.

PEST

Desarrollado por Eildert Groeneveld, Milena Kovack y Tialin Wang, es un programa para la resolución de modelos mixtos de gran potencia. Se le considera muy depurado, prácticamente libre de errores, ya que ha evolucionado mucho desde la primera versión (1990) hasta la v. 3.1 actual (más de 10 revisiones). Han sido bien implementados en muy diversas plataformas (SUN workstation, VAX, CRAY, Macintosh, estaciones IBM), presenta una buena documentación y

diversos test para verificar la instalación. Como contrapartida hay que pagar un pequeño precio (alrededor de 250$).

Está indicado tanto para pequeñas como para grandes bases de datos. Acepta los datos en diversos formatos y es fácil de usar debido a su interfaz semejante a la del S.A.S. Cubre modelos fijos, aleatorios, y mixtos. Permite cualquier número de efectos fijos, aleatorios, covariables, y polinomios de hasta orden 20. Es capaz de recodificar los códigos de los niveles de los factores, así como los animales y presenta diversos tratamientos de las casillas vacías.

Soporta el BLUP modelo macho, padre-madre, abuelo materno, animal, etc. con tratamientos de grupos genéticos y de grupos de padres desconocidos. En los modelos multicarácter permite diferentes matrices de incidencia (diferentes factores para cada variable). Permite, si los modelos tienen la misma matriz de incidencia, la descomposición de Cholesky (dCh), que es más rápida en la convergencia que DF REML o D REML. Además, si tenemos un sólo efecto aleatorio, sin datos vacíos, y con la misma matriz de incidencia, permite la transformación canónica, que es el método actualmente más rápido para modelos multicarácter.

Realiza test de hipótesis uni y multicarácter, contrastes entre los niveles de los factores, da el PEV para las predicciones BLUP, y los errores típicos para las estimas BLUE, y por último es el único con tratamiento de las varianzas heterogéneas.

PAQUETES MIXTOS

Conjunto de paquetes y utilidades informáticas, que permiten tanto la valoración de reproductores como la estimación de parámetros genéticos.

ABTK

Paquete de herramientas independientes desarrollado en lenguaje C por Golden, Snelling y Mallinckrodt para el S.O. Unix. Posee un gran número de herramientas para manipulación de datos, operaciones matriciales, etc., cada una de ellas con muchas opciones, por lo que prácticamente se puede realizar cualquier tarea, pero se necesitan unos conocimientos de álgebra matricial, resolución de modelos mixtos, BLUP etc., así como de UNIX, muy elevados. Es por lo tanto muy difícil de instalar y aprender, pero muy potente. Sus puntos débiles, en comparación con el PEST, son las escasas utilidades para la preparación de los datos y que no es adecuado para grandes bases de datos.

DMU

Colección de programas desarrollados en Dinamarca por Jensen y Madsen, para la investigación y evaluación de rutina. Son programas difíciles de aprender por la gran cantidad de detalles críticos a tener en cuenta y por los ficheros de parámetros que hay que crear para cada análisis. Se compensa por la cantidad de diagnósticos exhaustivos que presenta y lo completo del paquete. Es un programa que está en continua evolución y cuando esté maduro será de los mejores (p.e. la versión 4 soporta PEV para aproximadamente unas 200000 ecuaciones). Incluye un módulo de estimación de CV (**DMUAI**) que usa un D-REML con el algoritmo de Newton-Raphson (tal vez el más rápido en el análisis unicarácter).

CONCLUSIONES

Se comparan un grupo de programas

disponibles a través de INTERNET, capaces de resolver modelos mixtos y estimar componentes de la varianza. Cada uno tiene su punto fuerte: DFREML es el único que permite comprobar hipótesis sobre las estimas de componentes de la varianza; MTDFREML es un potente programa para el análisis de datos experimentales relativamente fácil de usar; PEST, el único programa semicomercial analizado, tiene la mejor interfaz de usuario y es el que presenta mejor soporte; JAA es un programa muy simple capaz de estimar los valores genéticos y el PEV de gran número de reproductores; MTDFS es el más simple y puede computar estimas multicarácter de los componentes de la varianza para varios caracteres en modelos con repetibilidad; CMIT y CMMAT son los únicos capaces de resolver modelos umbrales (*threshold model*); ABTK introduce utilidades que ofrecen gran flexibilidad para los usuarios experimentados; finalmente DMU es un programa de ámbito general, eficiente tanto para bases de datos pequeñas como grandes y el único con soporte para el algoritmo REML de Newton-Rapson.

La elección del paquete viene regido principalmente por la aplicación que se le vaya a dar (modelo matemático y genético, tamaño de las bases de datos etc.), y por la disponibilidad en cuanto a personal cualificado, tipo de ordenador, características de este ..., si bien muchas veces no se tienen en cuenta una serie de características muy importantes como son la facilidad de aprendizaje y uso, la flexibilidad (el programa debe ser capaz de resolver modelos simples pero cuando los necesitemos también complejos), y las utilidades de preparación y verificación de los datos (especialmente del pedigrí).

Por último, la creación de grupos de discusión en mejora genética animal, como AGDG o ACTEON, la expansión del correo electrónico, la creación de servidores de *FTP Anonymous* donde se puede compartir cualquier programa, y la rápida evolución tecnológica, han permitido el paso de la *Edad Media* a un *Renacimiento* de la mejora genética animal en nuestro país, no justificándose actualmente la selección por parte de las Asociaciones de Criadores de un determinado Grupo de Investigación para la evaluación genética de sus animales en base exclusivamente a que posean un programa BLUP capaz de resolver un determinado modelo genético, como ha ocurrido en el pasado.

AGRADECIMIENTOS

Al grupo de discusión en mejora genética animal AGDG, y muy especialmente a Ignacio Misztal de la Universidad de Illinois (U.S.A.) y Bruce Golden de la Universidad de Colorado (U.S.A.) por la puesta a disposición a todos los interesados de los programas analizados y algunos de los artículos mencionados en la bibliografía.

BIBLIOGRAFIA

Bolman, K., L. Krieso, L. Van Vleck y S. Kachman. **1993.** MTDFREML: Multitraits programs to

estimate variance components by restricted maximum likelihood using a derivate-free algorithm. Manual de usuario. USDA-ARS, Clay Center, Nebraska.

Golden, B., 1994. Future needs in computing strategies. 5th World Congress on Genetics Applied to Livestock Production, Guelph, Canada, August 7-12.

Golden, B., W. Snelling and C. Mallinckrodt. 1994. ABTK: Animal Breeder's Tool Kit. User's Guide and Reference Manual. Department of Animal Sciences, Colorado State University, Ft Collins.

Groeneveld, E., M. Kovack y T. Wang. 1993. PEST: A general purpose BLUP package for multivariate prediction and estimation. Institute of Animal Husbandry and Animal Behaviour. Federal Agricultural Research Centre. Germany.

Groeneveld, E., M. Kovack y T. Wang. 1993. VCE: Variance components estimation. Institute of Animal Husbandry and Animal Behaviour. Federal Agricultural Research Centre. Germany.

Harvey, W. 1987. LSMLMW: Mixed Model Least-Squares and Maximun Likelihood Computer Program. User's guide. Polycopy.

Jensen, J. y P. Madsen. 1993. DMU: Multivariate mixed model package. *National Institute of Animal Science.* Institute of Animal Science. Research Centre Foulum. Denmark.

Meyer, K. 1991. DFREML: Programs to estimate variance components by restricted maximum likelihood using a derivate-free algorithm. User Notes. Institute of Animal Genetics, Edinburgh University. Scotland.

Misztal , I. 1991. CMIT y CMMAT: Programs for analysis of mixed linear and threshold models with support for REML-type variance component estimation and maternal grandsire model. Manual de usuario. Comunicación personal.

Misztal , I. 1992. JSPFS: Single-trait REML program for animal models using sparse matrix solver. Manual de usuario. Comunicación personal.

Misztal, I. and N. Gengler. 1992. JAA (MIXIT): Mixed model program using iteration on data with support for animal model. Manual de usuario. Comunicación personal.

Misztal , I. and N. Gengler. 1992. MTDFS: Multitraits REML estimation of variance components program. Manual de usuario. Comunicación personal.

Misztal, I. 1993. RENUM: Data preparation program for sire and animal models. Manual de usuario. Comunicación personal.

Misztal, I. 1994. Software packages in animal breeding. 5th World Congress on Genetics Applied to Livestock Production, Guelph, Canada, August 7-12.

Perez-Enciso, M., I. Misztal and M. Elzo. 1994. FSPAK: An interface for public domain sparse matrix subroutines. 5th World Congress on Genetics Applied to Livestock Production, Guelph, Canada, August 7-12.