X CONGRESO LATINOAMERICANO DE SOCIEDADES DE ESTADÍSTICA
CÓRDOBA, ARGENTINA. 16 A 19 DE OCTUBRE 2012

ESTIMACIÓN DE CURVAS DE LACTANCIA EN VACAS

MARÍA DUTTO

Instituto de Estadística, Facultad de Ciencias Económicas y de Administración, UDELAR

mdutto@iesta.edu.uy

RESUMEN

Este trabajo tiene como objetivo generar un modelo aproximado para las curvas de primera lactancia (fenotípicas) en vacas Holando. Se exploran distintos modelos, paramétricos y no paramétricos: Wood, Wilmink, Ali & Shaeffer, polinomios de Legendre, *smoothing splines* y *regression splines*. Los modelos se comparan usando algunos criterios de bondad de ajuste. En particular para Wood y Wilmink se analizan las distintas formas que pueden tomar las curvas y se evalúa la existencia de un efecto estacional. Finalmente, se realiza un análisis de *cluster* con las *smoothing splines* estimadas, buscando obtener una tipología de curvas y vincularla al mes en el que paren las vacas. Como conclusión se recomienda el uso de los polinomios de Legendre de tercer orden y de *regression splines* con dos nodos variables para cada vaca. El modelo de Wood también es adecuado, salvo para estimar las curvas de las vacas que paren entre febrero y abril.

PALABRAS CLAVE: curvas de lactancia / regresión / clustering

1. INTRODUCCIÓN

Este trabajo surge a partir de la pasantía de la Licenciatura en Estadística de la Facultad de Ciencias Económicas y de Administración de la Universidad de la República (UDELAR), tutoreada por Juan José Goyeneche y Gabriel Rovere. El objetivo general era generar un modelo aproximado para las curvas de primera lactancia (fenotípicas) en vacas Holando. Como objetivos específicos se plantearon:

1. Estimar y estudiar la bondad de ajuste de modelos paramétricos específicamente desarrollados para curvas de lactancia como la curva de Wood, de Wilmink y de Ali & Shaeffer, ajustados a vacas individuales, y analizar qué tipos de curvas se detectan.

2. Estimar curvas de lactancia con otros métodos más flexibles como *smoothing splines, regression splines* y polinomios de Legendre.

3. Realizar una tipología de curvas y analizar si existen diferencias de forma según el mes de parto.

La curva de lactancia muestra el comportamiento de la producción de leche de la vaca en función del tiempo, medido en días desde el parto. Conocer su forma es importante por varias razones. En primer lugar, porque permite gestionar de forma más eficiente el tambo, por ejemplo, para planificar la alimentación, decidir el momento apropiado para dejar de ordeñar a la vaca y monitorear la salud de los animales (Grossman y Koops, 1988; Silvestre et al., 2006; Macciotta et al., 2005).

Además conociendo la forma de la curva de lactancia y los primeros datos para una vaca individual se podría llegar a predecir la producción de leche para toda la lactancia. Finalmente, los modelos de curvas de lactancia fenotípicos son importantes también como insumo para los modelos que estudian el componente hereditario de la productividad de las vacas y permiten la selección genética (Grossman y Koops, 1988; Macciotta et al., 2005).

La curva de lactancia estándar es creciente hasta un pico que se da entre las 4 y las 8 semanas posteriores al parto y luego decreciente. Sin embargo, la bibliografía indica que la forma de la curva de lactancia puede variar según el mes de parto de la vaca (Macciotta et al., 2006; Urioste et al., 2002) y ser distinta a la estándar. Por ejemplo, Urioste et al. (2002) (refiriéndose a Uruguay) dicen que los partos de otoño “sistemáticamente muestran un doble pico de producción de

leche correspondiendo el primero al inicio de la lactancia y el segundo a la producción durante la primavera."

Como principales antecedentes se destacan los trabajos de Macciotta et al. (2005), Silvestre et al. (2006) y García-Muñiz et al. (2008). Macciotta et al. (2005) en Italia analizaron la relación entre la forma de las curvas de lactancia y las propiedades matemáticas de las funciones ajustadas: Wood, Wilmink, Ali & Schaeffer y polinomios de Legendre normalizados de cuarto orden. Usaron datos de 27.823 lactancias y 229.518 controles en vacas Simmental italianas. Ajustaron las distintas funciones para cada lactancia (de cada vaca) y estudiaron la forma de las que tenían un R^2 ajustado mayor a $0{,}75$. Concluyeron que las de Wood y Wilmink detectan principalmente dos grupos de curvas: las estándar y las atípicas. Los otros dos modelos (Ali & Schaeffer y polinomios de Legendre) detectan una variedad mayor de formas, pero son más sensibles a las variaciones locales, lo que según los autores, se evidencia en el sesgo en la estimación de la producción de leche al principio y al final de la lactancia (efecto borde).

Por otro lado, Silvestre et al. (2006) en Portugal modelaron las curvas de lactancia a nivel fenotípico con siete funciones matemáticas: Wood, Wilmink, Ali & Schaeffer, splines cúbicas y polinomios de Legendre normalizados de segundo, tercer y cuarto orden. Usaron los datos diarios de 144 lactancias completas, pero trabajaron sólo con los registros entre 5 y 305 DIM. Disponían de observaciones diarias y tomaron muestras de ocho maneras distintas, que representan las formas más comunes de registro. Los esquemas de muestreo se hicieron combinando diferentes tiempos para el primer control (8, 30, 60 y 90 días desde el parto) y dos intervalos distintos entre controles (4 y 8 semanas). En cada lactancia tenían entre 4 y 11 observaciones, dependiendo del esquema de muestreo.

Observaron que las splines cúbicas, Ali & Schaeffer y los polinomios de Legendre de cuatro orden fueron los que mostraron mejor ajuste a los datos diarios. Concluyeron que el desempeño de los modelos de Wood, Wilmink y Ali & Schaeffer está muy afectado por la reducción del tamaño de la muestra, especialmente cuando aumenta el intervalo entre el parto y el primer control, aunque igual encuentran una variación considerable de ajustes dentro de cada esquema de muestreo.

En México, García-Muñiz et al. (2008) evaluaron la bondad de ajuste de 16 ecuaciones para modelar curvas de lactancia para seis genotipos bovinos: Pardo Suizo Americano, Bos Indicus, Bos Taurus y combinaciones de los dos últimos. Utilizaron 2076 lactancias y estimaron las

ecuaciones para cada vaca individual por regresión lineal y no lineal. Dentro de cada genotipo, jerarquizaron los modelos en base a los cuadrados medios de los residuos. También usaron como criterios la proporción de casos con producción estimada de leche diaria anormal (negativa o extrema) y los casos con autocorrelación positiva (estadístico de Durbin-Watson). Concluyeron que la curva que tiene mejor ajuste (sopesando distintos indicadores) es una reparametrización de la ecuación de Wood, que considera la primera fecha de control como el tiempo cero, aunque si tomaban solo los cuadrados medios de los residuos los mejores eran Ali & Shaeffer y los polinomios de Legendre de cuarto orden.

2. METODOLOGÍA

Se partió de una base de datos del Instituto Nacional de Mejoramiento Lechero que contiene 17.948 primeras lactancias (158.926 controles) de vacas que parieron entre el 2000 y el 2008.

En primer lugar, se procedió a la edición de los datos, eliminando las inconsistencias y dejando sólo aquellos datos que permitieran realizar los análisis posteriores (se eliminaron los controles menores a 5 días desde el parto y mayores a 305 y las lactancias con menos de 6 controles). Luego de la descripción de los datos se ajustaron los modelos para cada una de las vacas (Wood, Wilmink, Ali & Shaeffer, polinomios de Legendre, *smoothing splines* y *regression splines*). Siguiendo a Macciotta et al. (2005), para los modelos de Wood y Wilmink, dentro de las curvas que tenían un R^2 ajustado $> 0{,}75$, se analizó la proporción de curvas consideradas atípicas. Se intentó ver también en forma descriptiva si algunos criterios de bondad de ajuste (R^2, R^2 ajustado, proporción de predicciones diarias negativas y proporción de predicciones diarias "atípicamente grandes") están vinculados con la estación del año en la que paren las vacas o con la disponibilidad de información.

Finalmente, se realizó un análisis de *cluster* con la derivada de las *smoothing splines* evaluada en diez momentos $(30, 60, \ldots, 270, 300)$, con el fin de realizar una tipología de curvas. Nuevamente se intentó ver si los grupos detectados seguían un patrón estacional. Todos los programas se hicieron en R (2009).

Modelos paramétricos Los modelos paramétricos que se usaron son los de Wood (1967), Wilmink (1987) y Ali & Shaeffer (1987), cuyas expresiones se presentan en el Cuadro 1.

Cuadro 1: Modelos paramétricos utilizados

Nombre	Fórmula
Wood	$Y(t) = a.t^b.e^{ct}$
Wilmink	$Y(t) = a + be^{-kt} + ct$, en general se usa k fijo.
Ali & Shaeffer	$Y(t) = a + b(t/305) + c(t/305)^2 + d\log(305/t) + e[\log(305/t)]^2$

Los modelos de Wood y Wilmink permiten ajustar únicamente cuatro tipos de curvas, todas unimodales. Las formas posibles según el signo de los coeficientes se muestran en el cuadro 2.

Cuadro 2: Tipos de curva que pueden ajustar las funciones de Wood y Wilmink

Forma de la curva	Wood		Wilmink	
	b	c	b	c
Curva estándar	$+$	$-$	$-$	$-$
Curva continuamente decreciente (atípica)	$-$	$-$	$+$	$-$
Curva estándar invertida (U)	$-$	$+$	$+$	$+$
Curva continuamente creciente	$+$	$+$	$-$	$+$

Tomado de Macciotta et al.(2005)

Se ajustaron también tres modelos de regresión no paramétrica: polinomios de Legendre, *regression splines* y *smoothing splines*. Los dos primeros parten de una base de funciones y estiman los coeficientes asociados.

Polinomos de Legendre Cualquier polinomio de grado n puede ser escrito como una combinación lineal de los polinomios de Legendre de grado 0 hasta n. El polinomio de Legendre de grado n puede escribirse como:

$$P_n(x) = \frac{1}{2^n n!}\frac{d^n[(x^2-1)^n]}{dx^n}$$

con n entero (Bronshtein y Semendiaev, 1993).

Los polinomios de Legendre son ortogonales con respecto al producto escalar definido en L_2 en el intervalo $[-1, 1]$, es decir

$$\int_{-1}^{1} P_m(x)P_n(x)\,dx = 0 \quad \text{si } m \neq n$$

Como la ortogonalidad de los polinomios de Legendre se cumple en $[-1, 1]$, para las curvas de lactancia se estandarizan los tiempos respecto del parto (DIM) de la siguiente forma:

$$w(t) = 2\frac{(t - t_{min})}{(t_{max} - t_{min})} - 1$$

donde t_{min} es el tiempo mínimo y t_{max} el máximo (Silvestre et al., 2006); en este trabajo 5 y 305 días respectivamente.

Además

$$\int_{-1}^{1} P_n(x)^2 \, dx = \frac{2}{2n+1}$$

Entonces, los polinomios de Legendre normalizados $\Phi_n(x)$ se obtienen de la siguiente manera:

$$\Phi_n(x) = \sqrt{\frac{2n+1}{2}} \, P_n(x)$$

Con estos elementos se define el modelo de polinomios de Legendre de orden k como:

$$Y(t) = \alpha_0\Phi_0(w(t)) + \alpha_1\Phi_1(w(t)) + \ldots + \alpha_k\Phi_k(w(t))$$

Regression splines Los polinomios muchas veces presentan demasiadas oscilaciones no deseadas porque cada dato afecta el ajuste globalmente. Una solución a este problema es partir el rango de la función en intervalos y en cada uno de ellos ajustar un polinomio para aproximar la función, agregando además que la función en su conjunto cumpla ciertas condiciones globales de "suavidad". En eso consisten las splines (Faraway, 2002; Ma et al., 2005; Györfi et al., 2002).

De manera más formal, siguiendo a Györfi et al. (2002), se define el espacio spline $S_{u,M}([u_0, u_K))$ como el conjunto de funciones $f : [u_0, u_K) \to \mathbb{R}$ que son un polinomio de grado M o menos en cada intervalo $[u_i, u_{i+1}),\ i = 0, \ldots, K-1$ y continuamente diferenciables $M-1$ veces en $[u_0, u_K)$ (si $M > 1$), siendo M el grado del espacio de splines y $u = \{u_j\},\ j = 0, 1, \ldots, K$ el vector de nodos ($u_i \in \mathbb{R}, u_0 < u_1 < \ldots < u_K$). En *regression splines* los nodos $\{u_j\}$ son elegidos arbitrariamente (y no necesariamente están en las observaciones como en *smoothing splines*, como se verá más adelante).

Regression splines son métodos de regresión no paramétricos que usan bases del espacio spline para estimar funciones. Sea $\{B_1(x), B_2(x), \ldots\}$ una base de funciones de $S_{u,M}([u_0, u_K))$, entonces

$$f(x) = \sum_i B_i(x)\alpha_i$$

Los coeficientes $\alpha_i \in \mathbb{R}$ se estiman por mínimos cuadrados ordinarios. Para este trabajo se usó la base B-splines.

Smoothing splines Las *smoothing splines* aparecen como solución al siguiente problema de regresión: dados los pares de observaciones $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ se desea encontrar la función $f(x) : [a, b] \to \mathbb{R}$ ($a \in \mathbb{R}, b \in \mathbb{R}, a < b$) que tenga las dos primeras derivadas continuas y que minimice la suma de cuadrados penalizada (Fox, 2002):

$$SCP(f) = \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int_a^b f''(x)^2 \, dx$$

El primer término es la suma de cuadrados de los errores y el segundo es una penalización, que es grande cuando la función $f(x)$ es *rough* (tiene cambios rápidos de pendiente). Este método se conoce como mínimos cuadrados penalizados.

Modificando λ controlamos el balance entre la bondad de ajuste a los datos y la "suavidad" de la curva. Por eso a λ se le llama parámetro de "suavizado". Cuando $\lambda = 0$ si todos los valores x_i son distintos, obtenemos una función $\hat{f}(x)$ que interpola los datos (Faraway, 2006; Déjean et al., 2007). Esta es una estimación demasiado *rough*, ya que las observaciones son medidas con error y por lo tanto la función objetivo no debería pasar exactamente por todos ellos. En el otro extremo, cuando λ es muy grande, $\hat{f}(x)$ va a ser elegida de tal manera que $\hat{f}''(x)$ sea cero siempre, y por lo tanto, $\hat{f}(x)$ va a ser lineal (o casi), porque obliga a que la derivada segunda tienda a cero (Fox, 2002; Déjean et al., 2007).

Se demuestra que el resultado de la minimización de la suma de cuadrados penalizada en la clase de funciones diferenciables con derivada primera absolutamente continua es la spline cúbica natural (SCN) con nodos en los valores observados x_i (Green y Silverman, 1993; Fox, 2002). Según Lin y Carroll (2008), una de las ventajas de las *smoothing splines* es que no hay que elegir nodos (están en los valores observados). Sin embargo, el mismo autor afirma que cuando el tamaño de la muestra es grande se vuelve difícil de manejar computacionalmente.

***Cluster* de curvas** A la hora de hacer un análisis de *cluster* con datos longitudinales la clave está en definir una medida de disimilaridad entre curvas adecuada al objetivo del trabajo. En este caso necesitamos una medida que agrupe según la forma de la curva y no según el nivel de producción. Queremos distinguir las curvas con forma de U invertida, de las bimodales o siempre decrecientes, por ejemplo.

Para calcular ciertas medidas de disimilaridad se necesita una grilla de puntos que correspondan al mismo momento del tiempo con respecto al parto para cada vaca. Como los datos proporcionados no cumplen con esta característica es necesario estimar primero la curva completa para cada vaca con alguno de los modelos propuestos y luego discretizarla (o en su defecto hacer interpolación lineal o cúbica). Según D'Urso (2000) la disimilaridad longitudinal confronta las trayectorias comparando la intensidad del cambio entre dos instantes de tiempo consecutivos. Para las vacas l y m es:

$$d_{lm}^2 = \sum_{t=2}^{10} [(x_{l,t} - x_{l,t-1}) - (x_{m,t} - x_{m,t-1})]^2$$

es decir, la distancia euclídea entre los incrementos en las curvas discretizadas. Es cero si los incrementos de las dos curvas son los mismos para todos los tiempos. Si la discretización es fina o las funciones son regulares, esto implica que la distancia es cero cuando las curvas tienen la misma forma, es decir, una es una traslación de la otra en sentido vertical.

Otro enfoque es el que tomaron Déjean et al. (2007), para hacer un *cluster* de las curvas de intensidad de la expresión genética a lo largo del tiempo. Estudiaron la expresión de 200 genes medida en 11 momentos del tiempo entre las 0 y las 72 horas. No les interesaba el nivel absoluto de la expresión del gen sino la forma de la curva. Como primer paso estimaron las curvas con *smoothing splines* para obtener funciones regulares y diferenciables para la expresión de cada gen en función del tiempo. En este paso realizaron dos supuestos: que las mediciones tienen ruido y que la expresión de los genes es una función regular.

Optaron por usar el mismo parámetro de "suavizado" (λ) para todas las curvas, ya que observaron que si elegían por validación cruzada uno distinto para cada caso obtenían pobres resultados en el *clustering* posterior. La elección del λ común la hicieron de forma heurística combinando análisis de componentes principales obtenido para distintos valores de λ y la interpretación biológica de las curvas obtenidas. De esta manera seleccionaron $\lambda = 0{,}6$.

Para terminar realizaron un análisis de conglomerados con las derivadas de las *smoothing splines* con $\lambda = 0{,}6$ en los 20 puntos, usando un algoritmo jerárquico con la distancia euclídea y el método de Ward. Eligieron el número de grupos tomando en cuenta el dendrograma y la interpretación biológica de los *clusters* obtenidos con cada configuración. Luego hicieron *k-means* tomando como valores iniciales los centros de los grupos anteriores. El resultado para los autores fue satisfactorio desde el punto de vista biológico.

Finalmente, Abraham et al. (2003) aplicaron un análisis de *cluster* a las curvas de acidificación en la producción de quesos (Ph en función del tiempo). Para cada unidad observacional disponían de distinta cantidad de datos y en diferentes momentos del tiempo. Propusieron un análisis en dos etapas: estimar las funciones con *B-splines* y particionar los coeficientes estimados de la base B-splines con el algoritmo *k-means*. Usaron el mismo grado de las splines y el mismo vector de nodos para todas las curvas, o sea, la misma base de funciones (por lo tanto, cada coeficiente tenía el mismo significado).

3. RESULTADOS

Descripción De la estadística descriptiva de los datos cabe resaltar que el primer control de cada lactancia se da en promedio a los $31{,}5$ días (desvío $21{,}4$) aunque el máximo es 154 días. Además, el intervalo entre controles en promedio es de $32{,}9$ días (desvío $10{,}5$), con un máximo de 150 (Q3 $= 34$ días).

Como puede verse en la Figura 1, los partos son bastante menos frecuentes en los meses de noviembre, diciembre y enero y también en junio y julio. Los modos de la variable "mes de parto" se ubican en marzo y en setiembre. Las frecuencias son similares a las presentadas por Urioste et al. (2002) para las primeras, segundas y terceras lactancias en su conjunto.

Figura 1: Distribución de los partos por mes

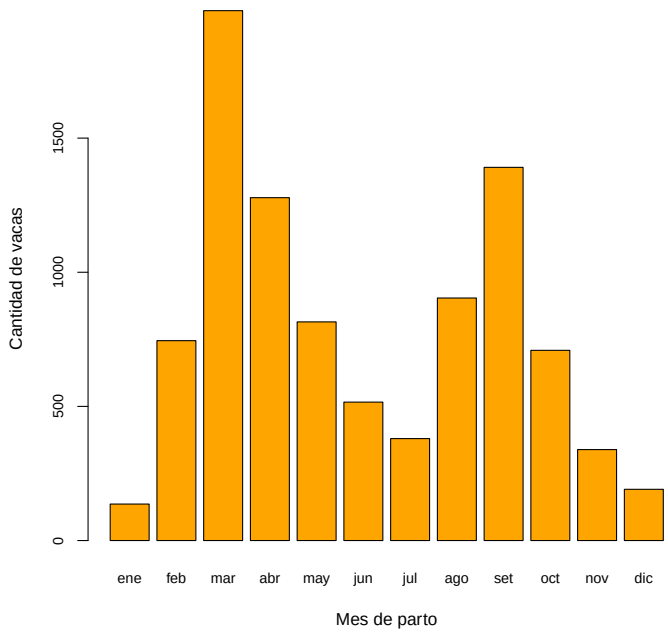
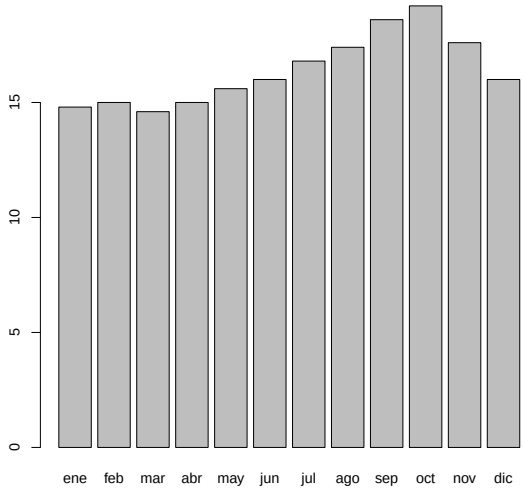
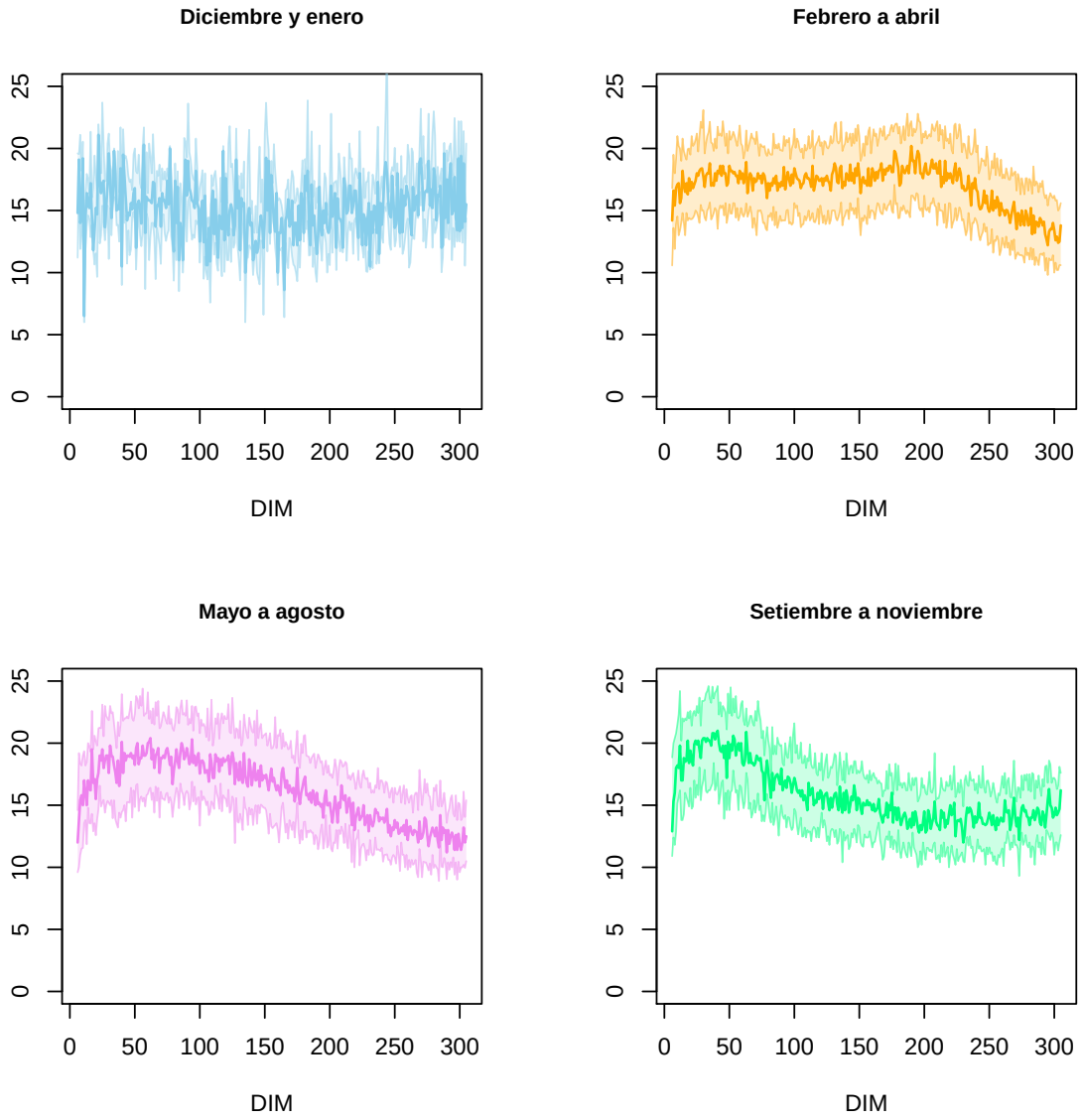


Figura 2: Mediana de la producción de leche (en kg) por mes de control



La Figura 2 muestra que la producción de leche es mayor en los meses de la primavera (con un pico en octubre) y menor de enero a marzo, aunque obviamente este dato se ve afectado por la distribución de los partos.

Figura 3: Cuartiles de la producción de leche por DIM según estación de parto



Para explorar si es razonable en estos datos suponer que la forma de la curva se ve modificada con el mes en el que paren las vacas se realizó el gráfico de la Figura 3. En los cuatro gráficos se pueden apreciar curvas de forma marcadamente distinta (aunque hay que tener en cuenta que en diciembre y enero solo parieron 327 vacas a lo largo de los nueve años considerados, por lo que el gráfico correspondiente no da mucha información sobre el patrón de producción). Las que parieron entre febrero y abril parecen tener curvas bifásicas. Las que parieron entre mayo y agosto tienen curvas con la forma típica (un solo pico alrededor de los 50 días y con concavidad negativa), mientras que en las vacas que parieron entre setiembre y noviembre el pico parece darse antes y luego de éste tienen concavidad positiva.

Ajuste de los modelos A continuación se muestra el resumen del ajuste de cada uno de los modelos propuestos a las curvas de lactancia de cada vaca.

Cuadro 3: Distribución de las curvas estimadas según su R^2 ajustado por modelo

	R^2 ajustado					
Modelo	$\leq 0{,}2$	$(0{,}2; 0{,}4]$	$(0{,}4; 0{,}6]$	$(0{,}6; 0{,}8]$	$(0{,}8; 1]$	Total
Wood	33,7 %	14,9 %	17,8 %	20,2 %	13,4 %	100 %
Wilmink	38,4 %	14,6 %	16,9 %	18,3 %	11,9 %	100 %
Ali & Shaeffer	21,1 %	10,6 %	16,4 %	22,5 %	29,4 %	100 %
Legendre 3	24,5 %	13,1 %	18,5 %	23,9 %	20,0 %	100 %
Legendre 4	21,6 %	11,4 %	16,2 %	22,6 %	28,2 %	100 %
Regression splines	24,3 %	13,2 %	18,4 %	24,0 %	20,2 %	100 %
Smoothing splines[a]	13,1 %	12,5 %	19,6 %	29,0 %	25,8 %	100 %

[a] En este caso el R^2 ajustado se define como como $1 - \frac{SCRes/(n-edf)}{SCT/(n-1)}$, donde n es el número de observaciones y edf son los grados de libertad equivalentes que se calculan como la traza de la matriz S: $\hat{y} = S * y$ (Hastie, 1993).

En cuanto a la distribución del R^2 ajustado, que se muestra en el Cuadro 3, se ve que el modelo de Ali & Shaeffer, los polinomios Legendre de cuarto orden y las *smoothing splines* son los que tienen mayor proporción de valores superiores a $0{,}6$ (en más del $50\,\%$ de las curvas ajustadas), aún cuando son los que tienen mayor número de parámetros para estimar. En cambio, los de Wood y Wilmink son los que tienen mayor frecuencia de curvas con R^2 ajustado menor o igual a $0{,}2$ y casi la mitad de las curvas con valores menores a $0{,}4$.

Los datos presentados por Macciotta et al. (2005) sobre la distribución del R^2 ajustado por modelo (para Wood, Wilmink, Ali & Shaeffer y polinomios de Legendre de cuarto orden) muestran una bondad de ajuste muy superior a la que se ve en el Cuadro 3; es más, en su trabajo en todos los modelos más de la mitad de las curvas estimadas tienen un R^2 ajustado superior a $0{,}8$. De todas maneras, a grandes rasgos se mantiene la tendencia de que Wood y Wilmink tienen peor *performance* en este indicador que Ali & Shaeffer y los polinomios de Legendre de cuarto orden.

Por otro lado, el modelo de Ali & Shaeffer y los polinomios de Legendre de tercer y cuarto orden

tienen mayor R^2 ajustado entre junio y setiembre (y peor en enero y febrero). Algo similar ocurre con las *splines*, que tienen su mejor desempeño entre junio y octubre (y peor enero y febrero). Para el modelo de Wilmink los meses que tienen mejor ajuste son julio y agosto, lo que se mantiene en el caso de Wood agregando junio. En definitiva, todos los modelos muestran mejor distribución del R^2 ajustado en invierno y peor en verano.

Si el primer control disponible es muy distante del parto, o sea, si es posterior al pico de producción, la forma de la curva de lactancia estimada puede ser muy diferente de la real (Silvestre et al., 2009). Esto no se puede comprobar para este caso, pero en los modelos de Wood y Wilmink se observó que si el primer control es posterior a los 60 días desde el parto, el R^2 ajustado empeora levemente.

De acuerdo a la bibliografía, también se utilizaron otros indicadores de bondad de ajuste: la proporción de estimaciones diarias negativas (dado que la producción de leche no puede ser menor que cero) y la proporción de estimaciones diarias "atípicamente grandes" (en este caso se tomó 50 kg como punto de corte). Los resultados se muestran en el Cuadro 4.

Cuadro 4: Otros indicadores de bondad de ajuste por modelo

Modelo	Estimaciones negativas			Estimaciones >50 kg		
		% días			% días	
	%[1]	Media	Desvío	%[2]	Media	Desvío
Wood	0,0	0,00	0,00	14,5	0,45	1,93
Wilmink	38,7	2,58	5,48	20,1	1,36	4,10
Ali & Shaeffer	59,8	4,41	7,25	35,1	1,95	4,63
Legendre 3	11,2	1,30	4,88	3,6	0,37	2,46
Legendre 4	23,1	2,93	7,42	8,1	1,00	4,33
Regression splines	4,2	0,46	2,90	0,0	0,04	0,80
Smoothing splines	2,4	0,26	2,18	0,0	0,00	0,13

[1] Se refiere al porcentaje de curvas que tienen al menos una estimación negativa (para algún DIM).

[2] Se refiere al porcentaje de curvas que tienen al menos una estimación mayor a 50 kg (para algún DIM).

Como se ve en el Cuadro 4, el modelo de Wood no genera ninguna estimación diaria negativa y las *splines* en sus dos formas generan muy pocas (tienen alguna en menos del 5 % de las curvas y el promedio de días con estimaciones negativas es inferior a uno). En cambio, los modelos de Wilmink y de Ali & Shaeffer tienen alguna estimación diaria negativa en el $38{,}7\,\%$ y $59{,}8\,\%$ de los casos, respectivamente. Con respecto a las estimaciones de leche "atípicamente grandes", los que tienen un mejor desempeño son los polinomios de Legendre de tercer orden, las *regression splines* y las *smoothing splines*. El modelo de Ali & Shaeffer falla también en este indicador, con más de un tercio de las curvas con alguna estimación diaria por encima de los 50 kg, aunque el promedio del porcentaje de estimaciones "atípicamente grandes" no es alarmante.

Los datos del Cuadro 4 se pueden comparar con los presentados por Silvestre et al. (2006) para los esquemas de muestreo SG2 y SG3[1], que son los más parecidos al esquema de los datos usados para este trabajo. A Silvestre et al. (2006) el porcentaje de estimaciones negativas les dio menor a $0{,}06$ para todos los modelos y en los dos esquemas de muestreo, lo que contrasta marcadamente con lo obtenido en este trabajo (con excepción del modelo de Wood). En cuanto al porcentaje de estimaciones diarias "atípicamente grandes", los datos del Cuadro 4 son similares a los de Silvestre et al. (2006), y en los casos de Wood y las *splines* incluso menores.

Las proporciones de estimaciones diarias negativas y "atípicamente grandes" se cruzaron gráficamente con el momento del primer control (en tramos), el mes de parto y el número de controles de la lactancia. Se vio que para los modelos de Wilmink y Ali & Shaeffer la proporción de estimaciones "atípicamente grandes" aumenta a medida que aumenta el mínimo DIM de la lactancia (aunque la mediana se mantiene en 0 en los tres grupos). Esto se ve de manera similar en Wood y polinomios de Legendre de cuarto orden pero solo para el grupo de vacas que tienen el primer control posterior a los 60 días. En cuanto a la proporción de estimaciones diarias negativas y el momento del primer control sucede algo parecido, con la diferencia de que en Wilmink y Ali & Shaeffer las medianas de cada grupo de primer DIM son todas distintas de cero. Además, en el caso de Wood no se observa diferencia según el momento del primer control ya que el modelo (por su formulación) no habilita estimaciones negativas.

Con respecto al mes de parto lo único que se observa es que en el caso de los modelos de Wood y Wilmink en algunos meses la proporción de estimaciones diarias "atípicamente grandes" es mayor al resto (aunque la mediana está en cero en todos los meses). Para Wood esto se da entre

[1]Los dos esquemas muestrean la producción de leche cada cuatro semanas, la diferencia es que en el SG2 el primer control se da a los 30 días, mientras que en el SG3 se da a los 60.

octubre y diciembre y para Wilmink entre setiembre y enero. En el caso de la proporción de estimaciones diarias negativas solo se ve un aumento (aunque no muy grande) para el modelo de Wilmink en los meses de mayo a julio.

Según Silvestre et al. (2006) a nivel individual es posible que 9 o 10 controles sean inadecuados para representar con precisión algunas curvas de lactancia, lo que se confirma en los modelos estimados. Se observa que al aumentar la cantidad de controles por lactancia hay cada vez menos curvas con una gran proporción de estimaciones diarias mayores a 50 kg o negativas, con la excepción de las *smoothing splines* y de Wood en el caso de las estimaciones negativas y de las *regression splines* en el caso de las estimaciones "atípicamente grandes" (porque prácticamente no hay).

Para cada vaca se realizó un *ranking* entre los modelos según cada uno de los criterios de bondad de ajuste ya comentados (R^2 ajustado, porcentaje de estimaciones diarias negativas y porcentaje de estimaciones diarias superiores a 50 kg). En el caso en que varios modelos empataran en algún indicador, se le dio el mérito a todos ellos (por eso no todas las filas del Cuadro 5 suman 100). Las *smoothing splines*, por ejemplo, son el mejor modelo para el $41\,\%$ de las vacas según el R^2 y para el $40\,\%$ según el R^2 ajustado, pero son el peor modelo para el $24\,\%$ de las vacas por el criterio de la proporción de estimaciones diarias negativas y para la mitad de las vacas por el criterio de la proporción de estimaciones diarias "atípicamente grandes". En cambio el de Wood es el mejor modelo según la proporción de estimaciones diarias negativas para todas las vacas (recordar que no daba estimaciones negativas) y para el $85\,\%$ de los animales según el porcentaje de estimaciones mayores a 50 kg, mientras que es el peor para el $30\,\%$ de las vacas según el R^2 (aunque esto mejora si se usa el R^2 ajustado). Llama la atención la mala posición de los modelos de Wilmink y de Ali & Shaeffer en la proporción de estimaciones negativas y la proporción de estimaciones "atípicamente grandes".

Cuadro 5: Resumen del desempeño de los modelos según cuatro indicadores

	R^2		R^2 ajustado		ENNeg[a]		ENSup[b]	
Modelo	Mejor	Peor	Mejor	Peor	Mejor	Peor	Mejor	Peor
Wood	1,4 %	30,1 %	7,5 %	15,2 %	100 %	23,0 %	85,5 %	53,0 %
Wilmink	0,2 %	48,0 %	5,0 %	29,9 %	61,3 %	44,1 %	79,9 %	64,5 %
Ali & Shaeffer	28,3 %	0,0 %	18,4 %	13,1 %	40,2 %	66,5 %	64,9 %	79,1 %
Legendre 3	0,0 %	1,9 %	5,7 %	4,3 %	88,8 %	26,0 %	96,4 %	52,1 %
Legendre 4	25,1 %	0,0 %	15,4 %	12,9 %	76,9 %	33,0 %	91,9 %	53,9 %
Regression splines	3,9 %	8,2 %	7,9 %	9,0 %	95,8 %	23,1 %	99,6 %	50,4 %
Smoothing splines[c]	41,1 %	11,8 %	40,0 %	15,7 %	97,6 %	24,0 %	100,0 %	50,4 %

[a] Es la proporción de estimaciones no negativas

[b] Es la proporción de estimaciones que está por debajo de los 50 kg

[c] En este caso el R^2 y el R^2 ajustado se definen de forma particular, como fue explicado en el Cuadro 3.

A la hora de interpretar estos resultados es importante tener en cuenta que ajustes pobres pueden deberse tanto a una mala elección de la función como a las perturbaciones aleatorias sobre el componente regular (Macciotta et al., 2005). Debido a enfermedad de la vaca o errores en el registro, un control reportado puede no ser representativo de la producción real de la vaca (Silvestre et al., 2009). Por otro lado, la gran variación biológica entre animales es una limitación para el enfoque funcional para ajustar curvas de lactancia individuales, además de la variación local debido a efectos ambientales (Silvestre et al., 2009).

Finalmente se vio que para los modelos de Wood y Wilmink, entre las curvas con un R^2 ajustado $> 0{,}75$, más del $90\,\%$ de las curvas obtenidas tenían forma de U invertida o continuamente decrecientes. Los promedios de los parámetros en los dos modelos coinciden a grandes rasgos con los de Macciotta et al. (2005), salvo en el b de Wilmink. La variabilidad es mucho más alta en los datos estudiados.

Análisis de *cluster* Para el análisis de *cluster* se usaron las *smoothing splines* estimadas con para cada vaca $\lambda = 0{,}0018$, elegido gráficamente para que las curvas no fueran demasiado suaves (que permitieran dos modos) ni demasiado rugosas. Se evaluó la derivada primera de las *smoothing splines* estimadas en los DIM $30, 60, \ldots, 270, 300$ y con esas diez variables se realizó un análisis de *cluster* jerárquico con el método de Ward y la distancia euclídea.

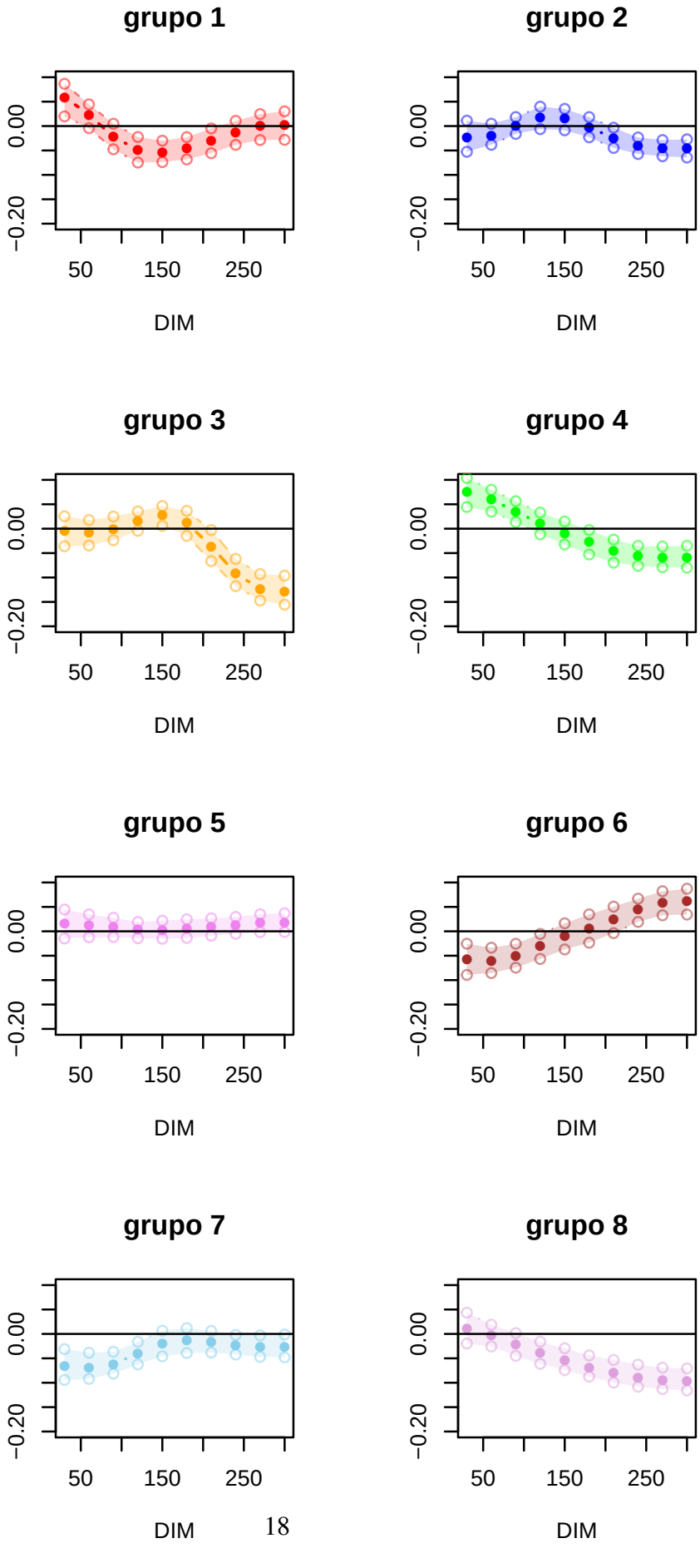
En el Cuadro 6 se presenta la distribución de vacas por grupo. Como se ve, los grupos son bastante homogéneos en cantidad de vacas, cosa que en general sucede con el método de Ward, salvo el grupo 7 que tiene bastante más observaciones que el resto.

Cuadro 6: Cantidad de vacas en cada grupo

Grupo	1	2	3	4	5	6	7	8
Vacas	1276	1399	1139	1294	760	939	1700	872
%	*13,6*	*14,9*	*12,1*	*13,8*	*8,1*	*10,0*	*18,1*	*9,3*

Para caracterizar los grupos primero se realizó un gráfico (Figura 4) con el primer cuartil, el promedio y el tercer cuartil de las derivadas en los DIM $30, 60, \ldots, 270, 300$ (que eran las variables usadas para el análisis de conglomerados) para cada *cluster*. Los puntos se unieron para mejorar la visualización. Allí se observa que el grupo 1 tiene un comportamiento marcadamente diferente del resto; es creciente al comienzo y luego decreciente, pero con un cambio de concavidad (a diferencia del grupo 4 que prácticamente siempre tiene concavidad negativa). A su vez, el grupo 8 también tiene una concavidad negativa casi siempre (forma de U invertida), pero es decreciente a partir del día 50, mientras que el grupo 4 es decreciente a partir del 150 aproximadamente. Otro grupo que es en promedio casi siempre decreciente es el 7, con un cambio de concavidad en la mitad de la lactancia.

Figura 4: Promedio de las derivadas de las splines en 10 puntos por grupo

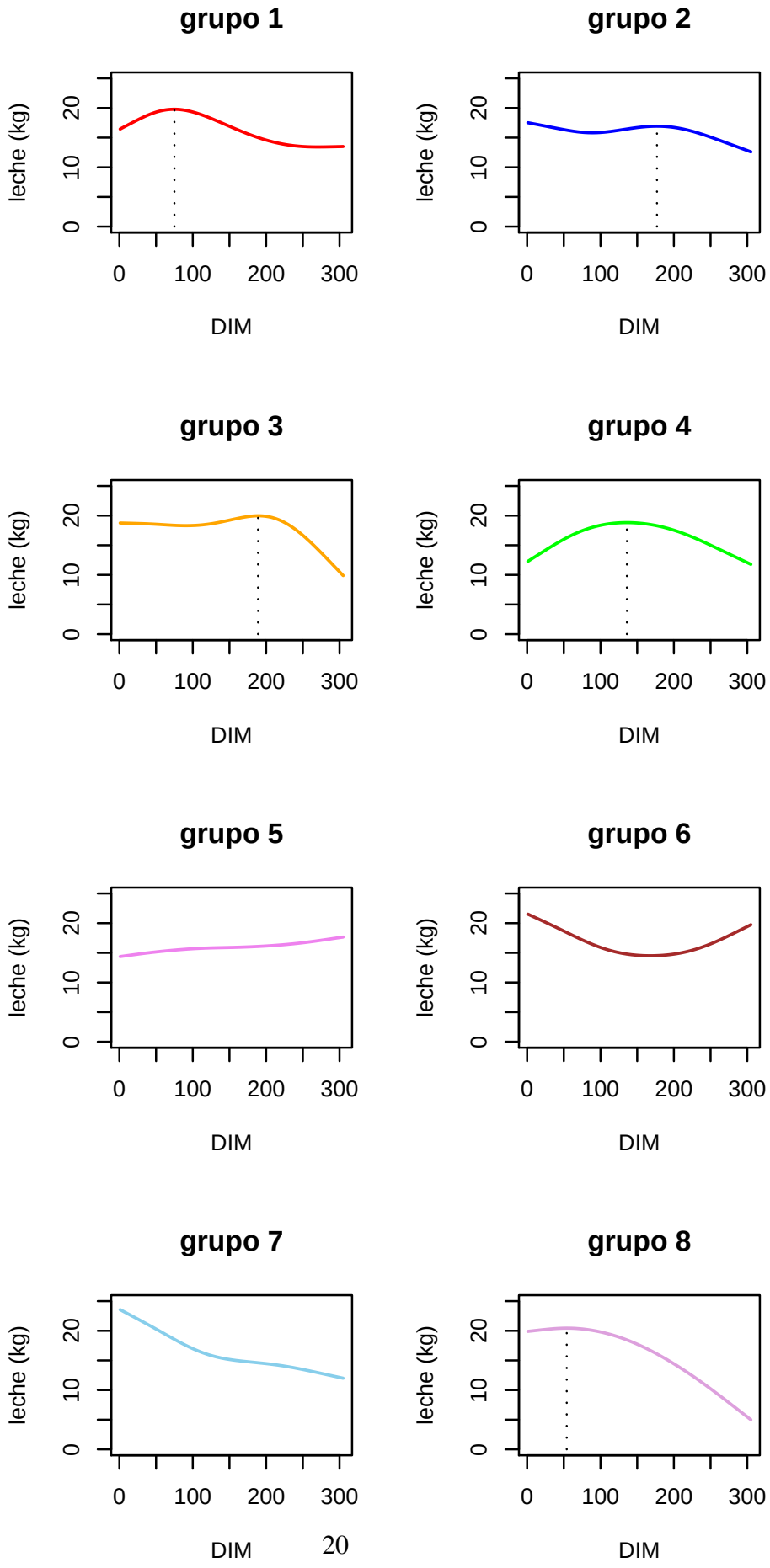


En la Figura 4 también se observa que los grupos 2 y 3 tienen en promedio un comportamiento similar: decrecen o se mantienen constantes hasta un punto, luego crecen y finalmente decrecen nuevamente. Ambos tienen un cambio de concavidad pero que en promedio se da en distintos DIM. Se diferencian en el momento en el cual se empieza a dar el decrecimiento (en el grupo 2 es un poco antes) y en la profundidad de ese decrecimiento (en el grupo 3 es mucho más pronunciado).

Finalmente, los grupos 5 y 6 son los únicos que en promedio crecen al final de la lactancia (el grupo 6 de forma más acentuada). De hecho, el grupo 5 en promedio es siempre creciente, de forma casi recta. En cambio el grupo 6 comienza decreciendo (al igual que el 2 y el 7) y en la mitad de la lactancia empieza a crecer, con concavidad positiva (tiene forma de U).

Esto mismo se puede ver en la Figura 5, en la que se grafica el promedio de la producción de leche (en kg) para cada DIM por grupo (si en vez del promedio se hace la mediana las curvas quedan similares). En la Figura 5, además de corroborar lo analizado a partir de la Figura 4 (y quizás de forma más sencilla), se observa el nivel promedio de la producción de cada grupo (la altura de las curvas). En este sentido, el grupo 3 tiene en promedio un nivel más alto de producción que el grupo 2. Por otro lado, al comienzo se la lactancia en los grupos 6 y 7 se observa una producción promedio más altas que en el resto. El que se muestra con una producción más baja a lo largo de toda la lactancia es el grupo 5 (aunque está bastante parejo con el grupo 2). A su vez, los grupos 3 y 8 son los que en promedio llegan con niveles más bajos de producción al final de la lactancia.

Figura 5: Promedio de producción de leche por DIM dentro de cada grupo



Cuadro 7: Distribución de los meses de parto por grupo (en porcentaje)

	Gr.1	Gr.2	Gr.3	Gr.4	Gr.5	Gr.6	Gr.7	Gr.8	Total
Enero	0,7	1,4	0,7	0,9	5,3	3,0	1,1	0,3	1,5
Febrero	2,4	13,5	15,0	6,9	12,9	3,8	5,7	3,9	8,3
Marzo	6,9	32,5	43,4	24,9	27,9	11,4	12,2	10,3	21,1
Abril	8,9	21,0	22,5	21,8	10,8	5,3	5,8	11,7	13,7
Mayo	8,9	10,8	6,5	17,2	7,4	2,2	4,5	11,5	8,6
Junio	9,2	3,0	3,3	10,6	3,7	1,3	3,4	9,6	5,4
Julio	9,4	2,0	1,1	4,8	2,4	1,2	3,3	8,3	4,0
Agosto	17,7	4,1	2,6	5,4	3,4	3,7	16,5	20,4	9,6
Setiembre	23,7	5,3	1,9	4,9	9,5	24,5	28,0	17,3	14,8
Octubre	7,8	4,2	1,8	1,5	6,1	21,5	13,1	4,5	7,5
Noviembre	3,0	1,7	0,6	0,6	5,4	14,0	4,5	1,6	3,6
Diciembre	1,3	0,5	0,5	0,5	5,4	8,1	1,9	0,6	2,0
Total	100	100	100	100	100	100	100	100	100

El Cuadro 7 busca caracterizar a los grupos obtenidos en el análisis de *cluster* según cómo se distribuye el mes de parto de las vacas a la interna de cada grupo. Los grupos 1 y 8 tienen mayor frecuencia que la marginal en los meses de mayo a agosto y si se mira la Figura 5 se observa que se trata de modelos con un único modo que se da entre los 50 y los 100 días desde el parto. Los grupos 2 y 3 que son los que tienen un modo un poco antes de los 200 DIM (en el caso del grupo 2 también tiene un modo al comienzo de la lactancia), tienen mayor frecuencia en los meses de febrero a abril (el grupo 2 se extiende hasta mayo). El grupo 4 (que comparte con el 1 y el 8 la forma "típica", pero tiene el modo tardío, entre los 100 y los 150 DIM, un poco antes que los grupos 2 y 3), se destaca entre marzo y junio. En cuanto al grupo 7, cuya curva promedio es monótonamente decreciente, tiene mayor presencia entre agosto y octubre. Finalmente, los grupos 5 y 6 que tienen formas "atípicas" (monótonamente creciente y en forma de U respectivamente), tienen una frecuencia más alta que la marginal en varios meses, principalmente de noviembre a enero, meses en los que ninguno de los otros modelos resalta; vale recordar que justo son los meses en los que paren menos vacas en las observaciones disponibles.

4. CONCLUSIONES

En este trabajo se exploraron distintas metodologías para estimar las curvas de producción de leche de vacas Holando en su primera lactancia, de forma paramétrica: Wood, Wilmink y Ali & Shaeffer, y no paramétrica: polinomios de Legendre, *smoothing splines* y *regression splines*.

Se estudió la frecuencia de las distintas formas que pueden tomar las curvas en Wood y Wilmink y resultó que en ambos casos más del $90\,\%$ eran con forma de U invertida o continuamente decrecientes (en el grupo de las curvas que tienen un R^2 ajustado $> 0{,}75$). Esta distribución no se debe solo al patrón biológico, sino que también resulta del hecho de que el comportamiento real de los datos se fuerza para ajustar alguna de las formas posibles que permiten estos modelos, que ofrecen una única curvatura global (no admiten por ejemplo dos modos). En cambio, los otros modelos (Ali & Shaeffer, polinomios de Legendre, splines) posibilitan otras formas de ajuste, pero esa flexibilidad los hace más sensibles a variaciones locales en la producción de leche. Es decir, son más vulnerables a modificarse por errores de medida, enfermedades de las vacas, etc.

Los mejores modelos según el R^2 ajustado son Ali & Shaeffer, los polinomios de Legendre de cuarto orden y las *smoothing splines*, mientras que los que ajustan peor son el de Wood y el de Wilmink. Además, todos los modelos tienen mejor R^2 ajustado cuando la parición es en invierno y primavera (aproximadamente de junio a octubre) y peor cuando es en verano (enero y febrero).

En cuanto a la proporción de estimaciones diarias negativas, se destacan los modelos de Wood, las *regression splines* y las *smoothing splines* por su buen desempeño; lo contrario sucede con Wilmink y Ali & Shaeffer. Los polinomios de Legendre de tercer orden, las *regression splines* y las *smoothing splines* resaltan por su baja proporción de estimaciones diarias "atípicamente grandes". En cambio Ali & Shaeffer tiene una mala *performance* según este indicador. Con relación a los *regression splines* se vio que con nodos variables para cada vaca disminuyen las proporciones de estimaciones diarias negativas y "atípicamente grandes".

Si bien la elección del mejor modelo depende de los indicadores de bondad de ajuste que se utilicen, tomando en cuenta los resultados obtenidos se recomienda el uso del modelo de Wood, de los polinomios de Legendre de tercer orden, de *smoothing splines* o de *regression splines*.

Por otro lado, dejando de lado al R^2 ajustado, que por su formulación mejora cuando aumenta el

número de observaciones, al aumentar el número de controles de cada vaca hay cada vez menos curvas con una gran proporción de estimaciones diarias negativas o "atípicamente grandes" (salvo en algunos modelos en los que estos indicadores casi siempre son cero).

El R^2 ajustado empeora levemente en todos los modelos si el primer control se da después de los 60 días desde el parto. Lo mismo sucede con la proporción de estimaciones diarias negativas y "atípicamente grandes": dichas proporciones aumentan si el primer control es posterior a los 60 días en los modelos de Wilmink, Ali & Shaeffer y polinomios de Legendre de cuarto orden (en el caso de Wood se da solo para las estimaciones mayores de 50 kg porque, por su formulación, no existen estimaciones negativas).

Finalmente, se realizó un análisis de *cluster* jerárquico con el método de Ward y distancia euclídea, tomando como variables la derivada primera de las *smoothing splines* estimadas evaluada en los diez momentos del tiempo. Se obtuvo una tipología de ocho grupos. Se pudo ver que en las vacas que paren entre noviembre y enero tienen mayor frecuencia de grupos con curva promedio con un patrón "atípico": monótonamente creciente, en forma de U y monótonamente decreciente. Las que paren entre febrero y abril tienen más frecuencia de los grupos con curva promedio bimodal o con modo muy tardío. Si se toman las vacas que paren entre mayo y agosto, los grupos que tienen mayor frecuencia arrojan curvas promedio unimodales. Finalmente, las vacas que paren en setiembre y octubre tienen mayor frecuencia que la marginal en dos de los grupos "atípicos" (con forma de U y monótonamente decreciente) y en uno cuya curva promedio es unimodal "típica".

Los resultados obtenidos tanto en forma descriptiva como en el análisis de *cluster* coinciden con lo mencionado por Urioste et al. (2002) con relación a la existencia de curvas de producción bimodales (aunque en su trabajo se trataba de los partos de otoño y en este de los que se dan entre febrero y abril). Teniendo eso en cuenta se recomienda utilizar un modelo flexible como los polinomios de Legendre de tercer orden, *smoothing splines* o *regression splines*, que fueron recomendados más arriba (especialmente para estimar las curvas de las vacas que paren entre febrero y abril). El modelo de Wood sería adecuado para el resto de las estaciones de parto.

5. REFERENCIAS

ABRAHAM, C.; CORNILLON, P.A.; MATZNER-LOBER, E.; and MOLINARI, N. (2003). Unsupervised curve clustering using B-splines. *Scandinavian Journal of Statistics*, **30**, 581-

595.

BRONSHTEIN, I. and SEMENDIAEV, K. (1993). *Manual de matemáticas para ingenieros y estudiantes*. Editorial Mir. Madrid.

DEJEAN, S., MARTIN, P.G.P., BACCINI, A. and BESSE, P. (2007). Clustering time-series gene expression data using smoothing spline derivatives. *EURASIP Journal on Bioinformatics and Systems Biology*, **2007**, 1-10.

D'URSO, P. (2000). *Classificazione fuzzy per matrici a tre vie temporali*. Tesi di Dottorato di Ricerca in Statistica Metodologica, XII Ciclo. Dipartamento di Statistica, Probabilità e Statistiche Applicate, Università La Sapienza. Roma.

FARAWAY, J. (2002). *Practical Regression and Anova using R* [En línea]. Disponible en: `http://cran.r-project.org/doc/contrib/Faraway-PRA.pdf` [Consulta: 10/7/2011]

FARAWAY, J. (2006). *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. Chapman & Hall/CRC

Fox, J. (2002). *Nonparametric Regression. Appendix to An R and S-PLUS Companion to Applied Regression* [En línea] Disponible en: `http://cran.r-project.org/doc/contrib/Fox-Companion/appendix-nonparametric-regression.pdf` [Consulta 10/7/2011]

GARCIA-MUÑIZ, J.G. et al. (2008). Comparación de ecuaciones para ajustar curvas de lactancia en bovinos. *Revista Científica*, FCV-LUZ, **XVIII**, No. 2, 160-169.

GREEN, P.J and SILVERMAN, B.W. (1993). *Nonparametric regression and generalized linear models*. Chapman & Hall/CRC

GROSSMAN, M and KOOPS, W.J. (1988). Multiphasic Analysis of Lactation Curves in Dairy Cattle. *Journal of Dairy Science*, **71**, 1598-1608.

GYORFI, L; KOHLER, M.; KRZYIAK, A. and WALK, H. (2002). *A distribution-free theory of nonparametric regression*. Springer-Verlag. Nueva York.

HASTIE, T. (1993). Generalized Additive Models. CHAMBERS, J. amd HASTIE, T.: *Statistical models in S*. Chapman & Hall. Londres, 249-308.

LIN, X. and CARROLL, R.J. (2008). Non-parametric and semi-parametric regression methods: Introduction and overview. FITZMAURICE et al. (2008) *Longitudinal Data Analysis*. Chapman & Hall/ CRC Press

MA, P.; CASTILLO-DAVIS, C.; ZHONG, W. and LIU, J. (2005). Curve Clustering to Discover Patterns in Time Course Gene Expression Data. *Gene*, **617**, 1-32

MACCIOTTA, N.P.P.; DIMAURO, C.; CATILLO, G.; COLETTA, A. and CAPPIO-BORLINO, A. (2006). Factor affecting individual lactation curve shape in Italian river buffaloes. *Livestock Science*, **104**, 33-37.

MACCIOTTA, N.P.P; VICARIO, D. and CAPPIO-BORLINO, A. (2005). Detection of Different Shapes of Lactation Curve for Milk Yield in Dairy Cattle by Empirical Mathematical Models. *Journal of Dairy Sciencie*, **88**, 1166-1177.

R Development Core Team (2009). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

SILVESTRE, A.M; PETIM-BATISTA, F. and COLAÇO, J. (2006). The Accuracy of Seven Mathematical Functions in Modeling Dairy Cattle Lactation Curves Based on Test-Day Records From Varying Sample Schemes. *Journal of Dairy Science*, **89**, 1813-1821.

SILVESTRE, A.M. et al. (2009). Lactation curves for milk, fat and protein in dairy cows: a full approach. *Livestock Science*, **122**, 308-313.

URIOSTE, J.; NAYA, H. and CHILBROSTE, P. (2002). *Evaluación cuantitativa de curvas de lactancia de vacas holando en Uruguay*. 25o. Congreso Argentino de Producción Animal. Tres resúmenes: 1) descripción de la población, 2) ajuste de un modelo bifásico, 3) implicancias biológicas de las curvas de producción multifásica. [En línea] Disponibles en: `http://www.aapa.org.ar/congresos/2002/SpPdf/sp42.pdf` `http://www.aapa.org.ar/congresos/2002/SpPdf/sp43.pdf` `http://www.aapa.org.ar/congresos/2002/SpPdf/sp44.pdf` [Consulta: 12/3/2011]