ANÁLISIS BIG DATA DE REGISTROS PRODUCTIVOS Y REPRODUCTIVOS INDIVIDUALES DE VACAS LECHERAS

López-Suárez, M., Henarejos, L., Calsamiglia, S. y Castillejos, L.
Servei de Nutrició i Benestar Animal (SNiBA), Departament de Ciència Animal i dels Aliments, Facultat de Veterinària, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona; Lorena.Castillejos@uab.cat

INTRODUCCIÓN

Actualmente, debido al grado de tecnificación alcanzado en las explotaciones lecheras, se puede disponer de una gran cantidad de datos, tanto a nivel individual como a nivel de granja. La información contenida en estos datos puede ser de gran ayuda en la toma de decisiones técnico-económicas para mejorar la rentabilidad de las explotaciones. Sin embargo, se requiere de una integración, análisis e interpretación de estos datos para poder extraer información útil y fácilmente interpretable.

En los últimos 20 años, en la era del Big Data, se han llevado a cabo múltiples estudios en el campo de la producción lechera para obtener información útil y aplicable y para definir modelos predictivos (Grezsiak et al, 2003; Murphy et al, 2014; Pinedo & De Vries, 2017 y Hertl et al, 2018).

El objetivo del presente estudio fue analizar una base de datos compuesta de registros individuales de vacas lecheras para identificar indicadores en los datos técnicos de primera lactación que puedan ser utilizados como predictores del rendimiento total esperado de la vida productiva de una vaca.

MATERIAL Y MÉTODOS

En el presente trabajo se utilizaron técnicas de gestión de datos y técnicas estadísticas, para la integración, análisis e interpretación de múltiples registros históricos de más de 800.000 vacas proporcionados por la "Confederación Nacional de la Raza Frisona" (CONAFE). Estos registros contenían información referente a las lactaciones (duración, rendimiento lechero por lactación, contenido de grasa y proteína, etc.), a los controles lecheros oficiales (kg de leche, % de grasa, % de proteína, recuento de células somáticas (RCS), velocidad de ordeño, etc.), al historial de partos (número, fecha, facilidad de parto, etc.), al historial de inseminaciones artificiales (IA) realizadas (número, fecha, toro, diagnóstico de gestación, etc.), a los índices genéticos (ICO, KL, etc.) y a la calificación morfológica (estructura y capacidad, sistema de ordeño, aplomos, etc.).

En un primer paso se integraron los distintos registros anteriormente mencionados en un único archivo utilizando el identificador del animal como nexo. A continuación, se procesaron los datos para eliminar errores de registro. Únicamente se incluyeron aquellos animales con registros completos y con un mínimo de 150 días de vida productiva. A partir de este filtraje, se obtuvo una base de datos de 103.637 vacas con registros de 301.217 lactaciones de 2.035 granjas durante el periodo 2006-2016.

Con el objetivo de estimar o predecir el rendimiento total de la vida productiva de una vaca de forma temprana se utilizaron los datos del pico de la primera lactación, los datos de la primera lactación finalizada y los referentes al rendimiento productivo vitalicio del animal. Para el cálculo de las variables correspondientes al pico de lactación, se consideraron los valores promedio de los controles lecheros comprendidos entre los 50 y 150 días en leche (DEL). En el análisis de la base de datos se incluyeron las siguientes variables: granja y comunidad autónoma (CA) de origen, estación de nacimiento de la vaca, estación del primer parto, edad al primer parto, índices genéticos de producción (KL) y de mérito total (ICO), calificación morfológica del animal, días de vida, días de vida productiva, media de días improductivos por lactación, causa de baja, número de lactaciones, producción vitalicia, producción media por día de vida y por día de vida productiva, índices productivos y reproductivos de la primera lactación (producción total, DEL, producción diaria, intervalo parto-primera IA, número de inseminaciones realizadas y días abiertos), e índices productivos del pico de la primera lactación (producción diaria y contenido de grasa, proteína y RCS). Para la gestión de datos, se utilizó Java para importar los archivos csv originales, transformar los datos y calcular las variables y un sistema SQL (MySQL) para filtrar y almacenar los datos.

Para el análisis de los datos se utilizó el paquete estadístico SAS (v9.4). Primero se realizó la estadística descriptiva de la muestra (PROC MEANS) para conocer la distribución de los datos. A continuación, se estudiaron las relaciones entre variables mediante una tabla de correlaciones (PROC CORR) y por ende se realizaron regresiones lineales (PROC REG). El rendimiento productivo vitalicio se tomó como variable respuesta de los modelos de regresión realizados para explorar la relación de las variables de primera lactación con la producción media por día de vida productiva. En un primer modelo se utilizaron las variables de la primera lactación (edad 1er parto, KL, ICO, calificación morfológica, producción de primera lactación, DEL de la primera lactación, producción media diaria de primera lactación, DEL a primera IA de la primera lactación, número de IA realizadas en la primera lactación y días abiertos de la primera lactación), y en el segundo modelo se incluyeron únicamente aquellas variables conocidas a los 150 DEL de la primera lactación (edad 1er parto, KL, ICO, calificación morfológica, DEL a primera IA de la primera lactación, número de IA realizadas en la primera lactación y producción diaria, % grasa, % proteína y RCS del pico de la primera lactación). La significación estadística se estableció en $P < 0,05$.

RESULTADOS Y DISCUSIÓN

Como se puede observar en la *Tabla 1*, los valores promedio de edad al primer parto, número de lactaciones, producción vitalicia, producción de la primera lactación y producción diaria del pico de lactación fueron de 806 días (26,8 meses), 2,91 lactaciones, 29.346 kg, 10.196 kg (8.520 kg/305 días) y 31,2 kg/día, respectivamente. Cabe comentar que existe una ligera desviación en los valores medios de la base de datos respecto a la media española, debido a que las granjas que aportan información a CONAFE están en control lechero y normalmente tienen buenos índices productivos y de calidad de leche.

***Tabla 1**. Resultados de la estadística descriptiva de las variables incluidas en el estudio (media, desviación estándar, mínimo, máximo).*

Variable	Promedio	Desv.Est.	Min.	Max.
Edad 1er parto (d)	806,1	91,5	570	1.265
KL	30,4	505,74	-2.123	2.202
ICO	1.455	544,08	-845	3.537
Calificación morfológica	78,5	2,92	63	91
Longevidad (d)	1.943	622,40	794	4.990
Vida productiva (d)	1.089	607,52	150	4.194
Días improductivos/lactación (d)	65,6	23,27	0	150
Número de Lactaciones	2,91	1,507	1	11
Producción vitalicia (kg)	29.346	16.464,	2.053	149.177
Producción media por día de vida (kg/d)	14,2	4,542	1,7	34,6
Producción media por día de vida productiva (kg/d)	27,3	5,07	4,8	53,7
Primera lactación				
Producción (kg)	10.196	2.998	1.693	31.535
DEL (d)	361,9	84,16	150	800
Producción media diaria (kg/d)	28,2	5,27	8,9	53,7
DEL a primera IA (d)	86,3	34,69	30	220
Número de IA realizadas	2,35	1,759	1	16
Días abiertos (d)	131,8	67,92	30	350
Pico (50-150 DEL) de primera lactación				
Producción media diaria (Kg/d)	31,2	5,96	9,0	59,9
% Grasa	3,52	0,612	1,18	8,28
% Proteína	3,11	0,223	2,17	8,42
RCS (x1.000 CS/ml)	154,2	439,6	1,50	19.634

El coeficiente de determinación obtenido en el primer modelo fue de 0,75, donde la variable de producción media diaria de la primera lactación por si sola obtuvo un R^2 parcial de 0,72. Estos resultados indican que la cantidad de leche producida diariamente por una vaca en su vida productiva estuvo directamente relacionada con su producción diaria en la primera lactación.
Con respecto al segundo modelo realizado, se obtuvo un coeficiente de determinación inferior (0,69) y, en este caso, la variable de producción media del pico de la primera lactación presentó un coeficiente parcial de 0,63 y, al añadir la variable ICO, el coeficiente parcial incrementó a 0,67. Por lo tanto, el modelo a los 150 días de la primera lactación con estas dos variables, puede explicar el 67% de la variabilidad existente en el rendimiento productivo vitalicio de una vaca.
Los resultados presentados son un ejemplo de cómo la integración y el análisis de datos, a pesar de la complejidad actual del manejo del Big Data, permiten identificar indicadores útiles para la toma de decisiones a nivel de granja.

REFERENCIAS BIBLIOGRÁFICAS

Murphy, M.D., O'Mahony, M.J., Shalloo, L., French, P., & Upton, J. Comparison of modelling techniques for milk-production forecasting. 2014. J. Dairy Sci. 9:3352-3363 • Grzesiak, W., Lacroix, R., Wójcik, J., & Blaszczyk, P. A comparison of neural network and multiple regression predictions for 305-day lactation yield using partial lactation records. 2003. Can. J. Anim. Sci. 83: 307-310 • Hertl, J.A., Schukken, Y.H., Tauer, L.W., Welcome, F.L. & Gröhn, Y.T. Does clinical mastitis in the first 100 days of lactation 1 predict increased mastitis occurrence and shorter herd life in dairy cows? 2018. J Dairy Sci. 101: 2309-2323. • Pinedo, P.J. & De Vries, A. 2017. Season of conception is associated with future survival, fertility, and milk yield of Holstein cows. J. Dairy Sci. 100: 6631-6639.

Agradecimientos: Este proyecto ha sido financiado por el MINECO/FEDER (AGL2015-67409-C2-1-R). La base de datos original fue proporcionada por CONAFE.

BIG DATA ANALYSIS OF DAIRY COWS' PRODUCTIVE AND REPRODUCTIVE RECORDS

ABSTRACT: Dairy Big Data analysis may facilitate decision making to maintain or increase herd productivity. The objective of this study was to analyze dairy cow recordings to find indicators to estimate the expected lifetime performance of a cow. A database containing more than 800,000 Holstein-Frisian cows' recordings, provided by CONAFE, was cleaned and filtered obtaining a new dataset with 301,217 lactations (103,637 cows). Genetic, productive, and reproductive indexes, among other variables, were included in the study. Two regression models were performed with the average daily yield per productive lifetime as dependent variable. The first model using first lactation variables, obtained a determination coefficient of 0.75, where first lactation daily production explained, by itself, the response variable (partial R^2 = 0.72). The second model, including variables from the first 150 days of the first lactation, obtained a coefficient of determination of 0.69. In this case, the average daily peak milk production obtained a partial R^2 of 0.63, which increased to 0.67 when variable ICO was added. Therefore, this second model explained 67% of the variability existing in the outcome variable. These results suggest that dairy data analysis is a useful tool providing indicators for predictive models which can facilitate dairy farm management.

Keywords*:* Big Data, predictive model, dairy production, first lactation