SESGO EN EVALUACIONES GENÉTICAS Y GENÓMICAS DE MANECH TETE ROUSSE

Legarra[1], A., Astruc[2], J.M. y Reverter[3], A.
[1]INRA, UMR GenPhySE, CS52627, 31326 Castanet Tolosan, Francia. [2]IDELE, CS52627, 31326 Castanet Tolosan, Francia. [3]CSIRO Agriculture & Food; Brisbane, Queensland, Australia
andres.legarra@inra.fr

INTRODUCCIÓN

En especies con generaciones solapadas y diversidad de información como los rumiantes, el BLUP es una herramienta clave para clasificar animales. Sin embargo, para tomar decisiones de selección correctas, es imprescindible la ausencia de sesgo. El ejemplo más típico es que los animales "probados" no deben ser sistemáticamente infra- o sobre-estimados respecto a los animales jóvenes. El uso de toros (y carneros) "genómicos" seleccionados ha mostrado sesgos en la valoración, con consecuencias negativas para seleccionadores y ganaderos (Winkelman, 2015). Sospechamos que el origen último de este problema es que el BLUP no es completamente robusto a selección. Este trabajo evalúa empíricamente la existencia de dicho sesgo.

MATERIAL Y MÉTODOS

El "sesgo" *experimentado* por los utilizadores de animales mejorados se puede describir como "la diferencia entre el nivel esperado promedio y el nivel observado en prueba de progenie". Sin embargo los animales testados en prueba de progenie son los machos genómicos seleccionados, y por tanto hay un proceso de selección. En conjunto, ese sesgo "observado" se puede describir como $\mu_{progenie} - \mu_{genomico}$ y tiene dos componentes: la estimación correcta del promedio de *todos* los machos genómicos, $(\mathbf{1}'\boldsymbol{u} - \mathbf{1}'\widehat{\boldsymbol{u}})/n$ (que es fruto de la selección realizada en sus padres), y la ganancia genética estimada y realmente producida al *retener* (y testar) los *mejores* machos genómicos, $ir\sigma_u - i\sigma_{\hat{u}}$. El primer componente corresponde al error en la estimación del progreso genético: $b_0 = (\mathbf{1}'\boldsymbol{u} - \mathbf{1}'\widehat{\boldsymbol{u}})/n$ (idealmente 0). El segundo componente depende de que $r\sigma_u = \sigma_{\hat{u}}$, lo que sucede si $b_1 = Cov(u,\hat{u})/Var(\hat{u}) = 1$. Si $b_1 < 1$ hay *sobredispersion* de valores genéticos. Los dos estadísticos b_0 y b_1 forman el test Interbull de validación de evaluaciones genómicas (Mantysaari et al., 2010).

Estas ($b_0 = 0$, $b_1 = 1$) son propiedades del BLUP, pero sólo bajo hipótesis muy restrictivas (Henderson, 1982): selección lineal incluida en los datos, modelo correcto, etc. En particular, la utilización de grupos genéticos sesga la estimación del progreso genético (b_0), y la selección no considerada por el BLUP reduce la varianza genética y provoca $b_1 < 1$. Para complicar más el panorama, la confusión existente entre tendencia genética, grupos de contemporáneas y grupos genéticos es básicamente intratable algebraicamente.

Para verificar estas hipótesis, analizamos un juego de datos de cantidad de leche anual en Manech Tête Rousse (o Manex Burugorri), la variedad francesa de la Latxa Cara Rubia. El juego de datos contiene 1.703.515 registros (hasta 2015), 500.626 animales en genealogía, 26 grupos genéticos para considerar un ~20% de padres desconocidos (los últimos de ellos en 2006, 2009 y 2012) y 1.424 carneros genotipados. Las evaluaciones genéticas se realizaron por BLUP o SSGBLUP, en ambos casos utilizando, en lugar de grupos genéticos, metafundadores (Legarra et al., 2015) para garantizar comparabilidad de los valores genéticos. Los valores de cría estimados se refirieron a la base genética del grupo genético "padre desconocido, madre conocida" de 2000. Usamos la heredabilidad estimada por REML con todos los datos (0,34).

Para estimar los sesgos de las evaluaciones usamos el método R (Reverter et al., 1994). A partir de las evaluaciones "total" (u_w) y "parcial" (u_p) y el vector completo de $\widehat{\boldsymbol{u}}$ se calculan $\hat{b}_1 = \widehat{\boldsymbol{u}}_w^{'}\boldsymbol{H}^{-1}\widehat{\boldsymbol{u}}_p / \widehat{\boldsymbol{u}}_w^{'}\boldsymbol{H}^{-1}\widehat{\boldsymbol{u}}_w$ y $\hat{b}_0 = (\mathbf{1}'\boldsymbol{H}^{-1}\widehat{\boldsymbol{u}}_w - \mathbf{1}'\boldsymbol{H}^{-1}\widehat{\boldsymbol{u}}_p)/n$, donde $\boldsymbol{H}$ es la matriz de covarianzas del SSGBLUP.

RESULTADOS Y DISCUSIÓN

Los resultados se muestran en la Tabla 1. Hay que resaltar que estos resultados se refieren a *toda* la población, no solamente a las últimas generaciones. Hay sesgo de tipo $b_0 > 0$, implicando que el progreso genético se *sobreeestima*, y el valor más alto observado (0,12) es equivalente a medio año de progreso genético. El sesgo desaparece a medida que se van introduciendo datos.

La sobredispersión reflejada en el coeficiente b_1 parece muy pequeña (los coeficientes son todos muy cercanos a 1), y por tanto no parece que la selección haya afectado mucho las varianzas genéticas. Sin embargo, los parámetros genéticos se estimaron con toda la base de datos. Por ejemplo, un uso de heredabilidades más altas generaría sobredispersión.

En la Figura 1 se ve que la estima de los efectos de los metafundadores (o grupos genéticos) indica una tendencia clara, que sin embargo se rompe en el último. En ovino lechero, los padres desconocidos pertenecen a la población mejorada y por tanto, si hay selección, el grupo genético *n+1* debe ser igual o mejor que el *n*. El error no se debe a falta de datos (la precisión de la estima es 0,95), sino a problemas de modelización. Soluciones razonables son o bien ignorar ese grupo, o bien postular una estructura de autocorrelación entre metafundadores. La estructura de covarianzas de los metafundadores es esencialmente $\boldsymbol{\Gamma} = 0{,}7\boldsymbol{I}$, lo que significa que *a priori* los metafundadores no están correlacionados y tienen valor 0.

Una de las aplicaciones de este tipo de validación retrospectiva es la evaluación de la calidad de las predicciones genómicas frente a las basadas en pedigrí (Olson et al., 2011). Típicamente la fecha de corte es una generación (es decir unos 4 años en vacuno y ovino lechero) Vistos los resultados, parece difícil garantizar siempre el no sesgamiento de las predicciones “parciales” respecto de las “totales”. El promedio de los padres da un estimador no sesgado del valor promedio de sus hijos (es decir $b_0 = 0$), pero no se puede predecir el progreso más allá de una generación. Además, una predicción correcta del progreso genético requiere una estimación correcta de los valores parentales, y los valores genéticos de las hembras son siempre poco precisos.

En conclusión: las valoraciones en su conjunto son sesgadas sobre todo en la predicción del progreso genético, y la utilización de grupos genéticos parece aumentar el sesgo. Sin embargo, las predicciones a corto plazo parecen razonablemente no sesgadas. Se recomienda verificar la falta de sesgo de las valoraciones nacionales.

REFERENCIAS BIBLIOGRÁFICAS

• Henderson, C. R. 1982. Proceedings of the World Congress on Sheep and Beef Cattle Breeding, 1:191–201. Legarra, A., Christensen, O.F., Vitezica, Z. G., Aguilar, I. & Misztal, I. 2015. Genetics 200: 455–68. • Mantysaari, E., Liu, Z. & VanRaden, P. 2010. Interbull Bull 41. • Reverter, A., Golden, B. L., Bourdon, R. M.& Brinks, J. S. 1994. J. Anim. Sci. 72: 34–37. • Olson, K., VanRaden, P., Tooker, M. & Cooper, T. 2011. J. Dairy Sci. 94: 2613–2620. • Winkelman, A. M., Johnson, D. L. & Harris, B. L. 2015. J. Dairy Sci. 98: 659–675. •

Agradecimientos: Estudio financiado por el proyecto Incomings del metaprograma INRA SelGen, y por una acción conjunta INRA-CSIRO. Agradecemos a la empresa y organismo de selección, CDEO y ROLP, por la disponibilidad de los datos.

***Tabla 1**. Sesgo (b0, expresado en desviaciones típicas genéticas) y sobredispersión (b1) de las evaluaciones BLUP y SSGBLUP para todos los animales utilizando diferentes fechas de corte en los datos.*

Fecha de corte	Datos eliminados	Sesgo b0		Sobredispersion b1	
		BLUP	SSGBLUP	BLUP	SSGBLUP
2009	434061	0,10	0,12	0,9905	0,9937
2010	375347	0,11	0,13	0,9925	0,9959
2011	313505	0,06	0,06	0,9922	0,9961
2012	249956	0,03	0,03	0,9962	0,9962
2013	186062	0,03	0,03	0,9948	0,9984
2014	120861	0,02	0,01	0,9979	1,0013
2015	56385	0.01	0.01	0.9984	1.0016
2016	0	0	0	1	1

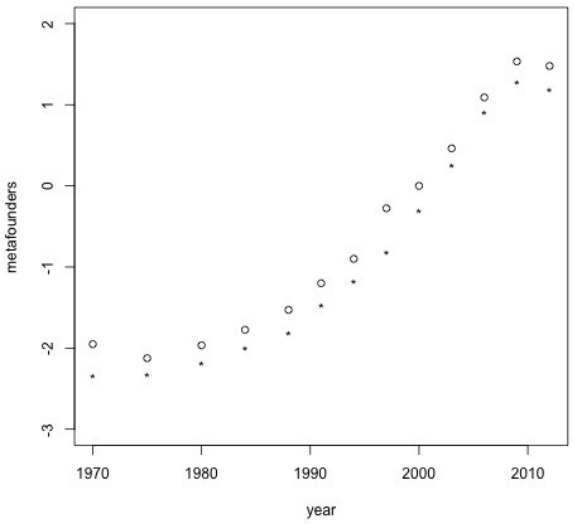

***Figura 1**. Estimas (en desviaciones estándares genéticas; con todos los datos) de los metafundadores (" o ": padre desconocido; " $*$ ": padre y madre desconocidos)*

BIAS IN GENETIC AND GENOMIC EVALUATIONS OF MANECH TETE ROUSSE

ABSTRACT: Bias of genomic selection is due to wrong estimation of genetic trend ($b_0 \neq 0$) and overdispersion ($b_1 \neq 1$) of the candidates' proofs. Unbiasedness only holds under very specific conditions, and therefore national evaluations should be checked for unbiasedness. MethodR provides statistics to assess it using large data sets. Here we analyse Manech Tete Rousse dairy sheep data using different truncation dates from 2011 to 2015. There is bias b_0, overestimation of the genetic trend up to 0.12 standard deviations of the trait, which decreases slowly as data accumulates. Genetic evaluation of this data set is unbiased in the short term but not in the medium term.

Keywords*:* bias, crossvalidation, BLUP, genomic