2020. Journées Recherche Porcine, 52, 59-60.

Comparaison de méthodes pour valider l'estimation par scanner à induction magnétique de la composition de jambons et de poitrines

Gérard DAUMAS (1), Mathieu MONZIOLS (1), Juan Manuel RODRIGUEZ (2), Jacobo ÁLVAREZ-GARCIA (2), David CAUSEUR (3)

(1) IFIP - Institut du porc, BP 35104, 35651 Le Rheu Cedex, France

(2) LENZ Instruments SL, calle Santander 42 nave 36, 08020 Barcelona, Espagne

(3) Agrocampus Ouest, Irmar, UMR 6625 CNRS, 65 rue de St-Brieuc - CS 84215, 35042 Rennes Cedex, France

gerard.daumas@ifip.asso.fr

Comparison of methods to validate magnetic induction scanner estimation of ham and belly composition

Magnetic induction scanning is a promising technology for carcass grading and sorting of cutting parts. The objective of this study was to compare the performance of five prediction methods of the composition of hams and bellies by magnetic induction, based on observations of a calibration sample. The five prediction methods tested were Ordinary Least Squares (OLS), Lasso, Ridge, Partial Least Squares (PLS), and complete selection of sub-models by minimizing Bayesian information criterion (Subset). For each statistical method, R^2 and RMSEP were calculated in a 10-fold cross validation repeated 100 times with random division of the data into 10 segments. Data for two calibration samples were used: one for 100 hams and the other for 80 bellies. Hams and bellies were scanned with a recent commercial device using a low-intensity magnetic field. The four response variables, weights and contents of fat and muscle, were measured by computed tomography. Based on the median values, the PLS gave the best performance for hams. The dispersion of results was lowest with the PLS as well. For bellies, Ridge regression was the most successful, except for fat content, for which Subset was better. Muscle content of hams and fat content of bellies were estimated respectively with a median R^2 of 0.64 and 0.66. The ranking of methods based on their prediction performance depended on the cut. Subset, Ridge and Lasso seemed to show the most stable prediction performance results among the cuts and response variables, always being close or equal to the best performance.

INTRODUCTION

La validation statistique est une étape à ne pas négliger dans le processus de test d'une technologie. Néanmoins, il n'y a pas de consensus sur la méthode à appliquer. Les résultats semblent dépendre de la nature des données. Aussi, il est souvent conseillé de tester plusieurs méthodes.

Ayant estimé la composition de jambons et poitrines par un scanner à induction magnétique sur un échantillon de calibrage (Daumas *et al.*, 2019), les auteurs souhaitaient passer à l'étape de validation. Pour cela, les auteurs ont comparé la performance de cinq méthodes de régression linéaire parmi les plus courantes.

1. MATERIEL ET METHODES

1.1. Matériel

Les données étaient celles des deux échantillons de calibrage, à savoir 100 jambons d'une part, et 80 poitrines d'autre part, qui avaient été scannés avec un appareil commercial récent («HAM-INSPECTOR»), utilisant un champ magnétique de faible intensité. Les quatre variables à prédire, les poids et les teneurs de gras et de muscle, ont été mesurées par tomodensitométrie. Les échantillons et l'estimation de la composition tissulaire ont été présentés par Daumas *et al.* (2019). Les variables explicatives sont nombreuses et présentent de la colinéarité.

1.2. Méthodes

Les cinq méthodes de régression testées sont explicitées ci-dessous, avec mention entre parenthèses des fonctions et packages utilisés du logiciel R (R Core Team, 2017) :

OLS : modèle de régression linéaire estimé par la méthode usuelle des moindres carrés ordinaires (fonction lm du package stats) ;

Lasso : modèle de régression linéaire estimé avec minimisation du critère des moindres carrés ordinaires pénalisé par un terme proportionnel à la somme des valeurs absolues des coefficients de régression, avec choix du paramètre de pénalité par minimisation de l'erreur de prédiction (package glmnet ; Friedman *et al.*, 2010) ;

Ridge : modèle de régression linéaire estimé avec minimisation du critère des moindres carrés ordinaires pénalisé par un terme proportionnel à la somme des carrés des coefficients de régression, avec choix du paramètre de pénalité par minimisation de l'erreur de prédiction (package glmnet ; Friedman *et al.*, 2010) ;

PLS : régression des moindres carrés partiels, avec choix du nombre de composantes par minimisation de l'erreur de prédiction (function selectNcomp avec l'option par défaut du package pls ; Mevik *et al.*, 2019) ;

Subset : minimisation du critère d'information bayésien (BIC) sur tous les sous-modèles possibles (fonction regsubsets du package leaps ; Lumley, 2017).

Pour chaque méthode statistique, le coefficient de détermination (R^2) et l'erreur de prédiction (RMSEP) ont été calculés en validation croisée, en réalisant 100 partitions aléatoires en 10 segments de données. Les méthodes ont été classées selon leur médiane.

2. RESULTATS ET DISCUSSION

La hiérarchie des méthodes étant la même en termes de R^2 et de RMSEP, seules les valeurs de R^2 sont données.

Pour les jambons, pour les quatre variables à prédire, la PLS a donné les meilleures performances et OLS les moins bonnes. Le rang des trois autres méthodes variait. Les boîtes à moustaches de la figure 1 illustrent la comparaison des méthodes pour le pourcentage de muscle. L'étendue des R^2 médians était de 0,14, les extrêmes étant de 0,50 et 0,64. La dispersion des résultats était la plus faible avec la PLS et la plus forte avec OLS et ce pour les quatre variables à prédire.

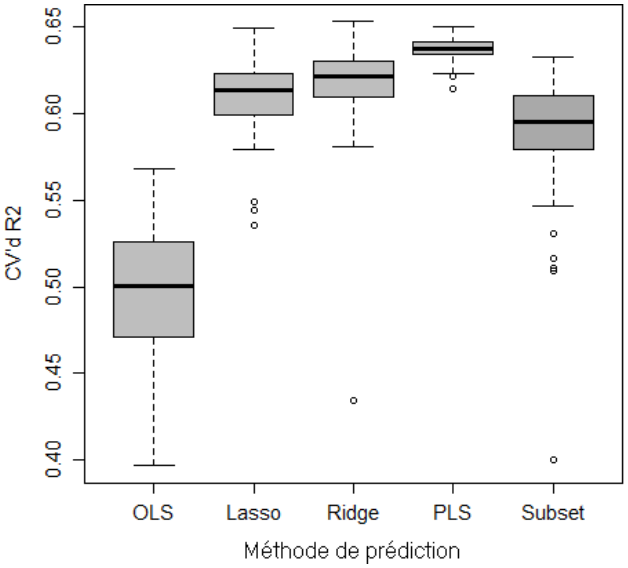

Figure 1 – Distribution du R^2 en validation croisée (CV'd R2) par méthode de prédiction du % de muscle du jambon

Pour les poitrines, la PLS a toujours donné les moins bons résultats. Ridge a été la plus performante, sauf pour le pourcentage de gras, où Subset était meilleure (Figure 2). Concernant la teneur en gras, variable la plus importante pour le tri des poitrines, l'étendue des R^2 médians était de 0,12, les extrêmes étant de 0,55 et 0,67.

La dispersion des résultats était la plus faible avec la PLS pour les teneurs en gras et en muscle.

Lors du calibrage, les modèles avaient été établis par la méthode Subset. Le R^2 médian de 0,66 obtenu ici en validation croisée avec cette même méthode pour la teneur en gras des poitrines est nettement inférieur au R^2 de 0,76 lors du calibrage, conséquence possible d'un surajustement. Pour les jambons où la PLS a donné de meilleurs résultats que Subset, le R^2 en validation croisée était très proche de celui du calibrage (0,65).

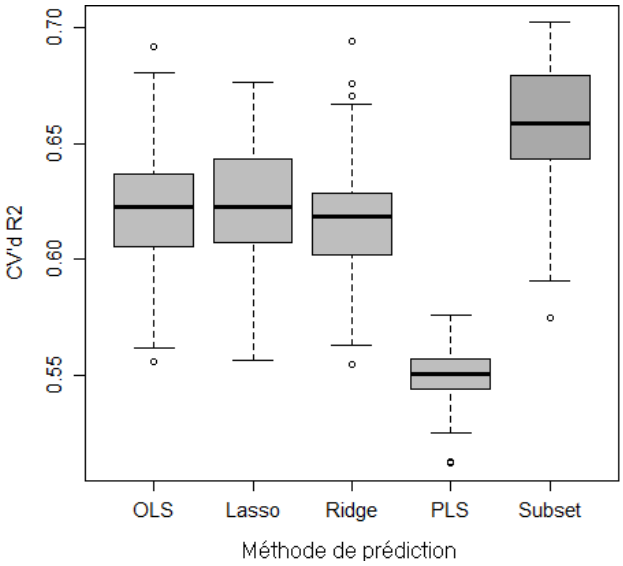

Figure 2 – Distribution du R^2 en validation croisée (CV'd R2) par méthode de prédiction du % de gras de la poitrine

CONCLUSION

Les écarts de R^2 en validation croisée entre les cinq méthodes statistiques testées se sont avérés conséquents. Parmi les cinq méthodes, aucune ne s'est révélée la meilleure sur les deux échantillons, celui de jambons et celui de poitrines. La PLS était la meilleure pour prédire le pourcentage de muscle du jambon, mais la moins bonne pour prédire le pourcentage de gras de la poitrine, sans raison évidente. La hiérarchie des méthodes en termes de performance de prédiction dépendait de la pièce de viande. Les méthodes Subset, Lasso et Ridge seraient à privilégier, car ayant montré des performances de prédiction plus stables pour l'ensemble des deux pièces et des variables à prédire, toujours proches ou égales à la meilleure performance.

REMERCIEMENTS

Les auteurs remercient le CASDAR pour le soutien financier apporté à ce travail dans le cadre du projet HYPER-SCAN. Ils remercient également l'abattoir BERNARD et le groupe Jean Floc'h pour la qualité de leur accueil lors de l'essai.

REFERENCES BIBLIOGRAPHIQUES

- Daumas G., Monziols M., Rodriguez J.M., Álvarez-García J., Causeur D., 2019. Estimation de la composition tissulaire de jambons et poitrines par un scanner à induction magnétique. Journées Rech. Porcine, 51, 339-344.
- Friedman J., Hastie T., Tibshirani R., 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. J. Stat. Softw., 33, 1-22. URL http://www.jstatsoft.org/v33/i01/.
- Lumley T., 2017. leaps: Regression Subset Selection. R package version 3.0. URL https://CRAN.R-project.org/package=leaps
- Mevik B.H., Wehrens R., Liland K.H., 2019. pls: Partial Least Squares and Principal Component Regression. R package version 2.7-1. https://CRAN.R-project.org/package=pls
- R Core Team, 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.