What is the best modelling approach to predict skewed distributed health biomarkers from MIR, a case study with the Haptoglobin in milk?

S. Franceschini[1], M. Calmels[2], C. Grelet[3], J. Leblois[4], Y. Brostaux[1], HappyMoo consortium, N. Gengler[1] & H. Soyeurt[1]

[1] Gembloux Agro-Bio Tech, Université de Liège, Liège, Belgium
Corresponding Author: sfranceschini@uliege.be
[2] Seenovia, Saint-Berthevin, France
[3] Walloon Agricultural Research Center, Gembloux, Belgium
[4] Walloon Breeders Association Group, Ciney, Belgium

Presented in the Session: **Session 7 - Latest tools using MIR-spectra in the ICAR world.**

Keywords: ft-mir health modelling

Abstract

The development of milk Fourier transform mid-infrared (**FT-MIR**) spectrometry allows the collection of information without impairing animal welfare through invasive measures. With the increase in consumer awareness regarding animal welfare, researchers aim to create new decision tools based on FT-MIR predictions of new biomarkers related to animal health. Among them, the concentration of haptoglobin in milk is of interest as an acute-phase protein that indicates inflammation. However, the shape of its distribution, by definition of an acute-phase protein, is very asymmetrical, which implies difficulties in modelling and impacts the prediction performance of the FT-MIR model. Therefore, this study aimed to test different modelling strategies to identify the best methodology to consider these specific distributions.

The dataset used contained 1,361 observations collected from 5 dairy breeds by 11 milk recording organizations located across North-West Europe. We considered 16 models built from different modelling approaches such as linear regression, penalized linear regression, partial least square regression (**PLS**), penalized generalized linear model (**GLM**) and radial and linear support vector machine (**SVM**). The same modelling methodologies were also used after a logarithmic transformation of the concentration of haptoglobin to normalize its distribution. Models were developed with and without somatic cells score (SCS) as a predictive variable in addition to MIR spectra. The model performances were evaluated using cross-validation and external validation. We computed the coefficient of determination (**R^2**), the mean average error (**MAE**) and the root mean square error (**RMSE**). The residuals were also studied to confirm their homoscedasticity which is at risk when modelling asymmetrical distribution.

In this study case, the use of SCS as a predictive variable and logarithmic transformation of Haptoglobin significantly improved the models. From the least good model to the best based on the cross-validation, we have the linear regression (R^2=0.52), the linear SVM (R^2=0.58) and the radial SVM (R^2=0.59) followed by the penalized linear regression (R^2=0.61), the PLS regression (R^2=0.61) and the penalized GLM (R^2=0.70). This last model was obtained using milk MIR spectra and SCS as features with a normal distribution and a logarithmic link function. R^2 of 0.70 and 0.65 as well as MAE of 13.85 µg/L and 10.64 µg/L were observed for this model from a cross-validation

and an external validation, respectively. This regression outperforms the others that use logarithmic transformation by overcoming the assumption of constant variance.

In conclusion, those results show the interest of GLM to predict one of the health biomarkers having a skewed distribution from milk FT-MIR spectra. The regularization was required to prevent overfitting. Therefore, we suggest considering the penalized GLM approach to calibrate health biomarkers of interest with a skewed distribution.