

Establishing applicability limits of regression neural networks using decoders. Application to predict diesel properties using infrared spectra

M. Suliany Rodríguez-Barríos¹, Joan Ferré¹, M. Soledad Larrechi¹,
Enric Ruiz²

Addresses: ¹Universitat Rovira i Virgili, Department of Analytical and Organic Chemistry, Faculty of Chemistry, Campus Sescelades, Carrer Marcel·lí Domingo, 1, Tarragona, 43007, Spain.


²Repsol-Petróleo, Tarragona, Spain

E-mail: mariasuliany.rodriquez@urv.cat

Knowledge of the applicability domain (AD) of regression models based on artificial neural networks (ANNs) is a requisite for applying ANNs in routine analyses using spectroscopic data. The AD limits can be defined from different metrics that condition the confidence of the predictions of the established regression model.

In multivariate regression methods such as the Partial Least Squares (PLS) calibration, the applicability limits are commonly based on Hotelling T^2 and Q statistics [1,2]. These limits allow the detection of samples that are rare compared to those used to establish the model. Other limits based on criteria such as ASTM's RMSSR (Root Mean Square Spectral Residuals) and NND (Nearest Neighbor Distance) [3] have also been used to flag spectra outside the established limits as discordant spectra. Although measures based on the similarity among spectra apply to all types of models, those that consider the specific form of the model, such as Hotelling's T^2 and Q statistics, are preferred, since they are related to how the spectrum is being used by the model. A similar system for defining the limits of applicability of multivariate regression based on ANNs has not been reported yet. This work proposes a methodology to establish the limits of applicability of regression neural networks and shows its implementation for the prediction of a diesel property using infrared spectra.

A test set was created by randomly selecting samples from a data set of 2172 diesel samples. The rest of the samples were split into training and validation sets using the Kennard-Stone algorithm. A feed-forward neural network (FFNN) was trained to predict the density of the diesel samples from their infrared spectra. The activations of the hidden layer of the FFNN were used to train a decoder network to reconstruct the



training spectra. The squared Mahalanobis distance (MD^2) of the hidden layer activations and the spectral residuals (Q residuals) from the decoder network were used to define the applicability limits of the calibration model. The FFNN model provided a high determination coefficient between real and predicted density values ($R^2 = 0.99$), with low prediction errors ($RMSEP = 0.72 \text{ kg/m}^3$) comparable to other reported results [4-6]. The warning limits for the MD^2 and Q residuals were decided from the empirical cumulative distribution function of both metrics. These warning limits define the applicability domain of the FFNN regression model. Beyond these limits, the spectrum of a new sample is flagged as atypical.

References

- [1] J. F. MacGregor and T. Kourti, "Statistical process control of multivariate processes," *Control Eng. Pract.*, 3 (1995) 403–414.
- [2] R. P. Cogdill, C. A. Anderson, and J. K. Drennen, "Process analytical technology case study, part III: Calibration monitoring and transfer," *AAPS PharmSciTech*, 6 (2005) 284-297.
- [3] ASTM E1655-17, 2017. ASTM E1655-17 standard practices for infrared multivariate quantitative analysis. *ASTM Int.* 05, 30
- [4] R. M. Balabin, E. I. Lomakina, and R. Z. Safieva, "Neural network (ANN) approach to biodiesel analysis: Analysis of biodiesel density, kinematic viscosity, methanol, and water contents using near-infrared (NIR) spectroscopy," *Fuel*, 90 (2010) 2007–2015, doi: 10.1016/j.fuel.2010.11.038.
- [5] R. M. Balabin and S. V. Smirnov, "Variable selection in near-infrared spectroscopy: Benchmarking of feature selection methods on biodiesel data," *Anal. Chim. Acta*, 692 (2011) 63–72, doi: 10.1016/j.aca.2011.03.006.
- [6] L. de Fátima Bezerra de Lira, F. V. C. de Vasconcelos, C. F. Pereira, A. P. S. Paim, L. Stragevitch, and M. F. Pimentel, "Prediction of properties of diesel/biodiesel blends by infrared spectroscopy and multivariate calibration," *Fuel*, 89 (2010) 405–409, doi: 10.1016/j.fuel.2009.05.028.