Journal of Agricultural Science and Technology A 14 (2024) 181-186
doi: 10.17265/2161-6256/2024.05.001

Application of Splitting (Cross Validation) Technique and Herzberg Equation in Feed Pellet Quality Prediction Models

Keysuke Muramatsu[1] and Alex Maiorka[2]

1.DSc. DSM Animal Products, São Paulo, São Paulo State, 04543-907, Brazil

2.DSc. Professor at Department of Animal Science, Universidade Federal do Paraná, Curitiba, Paraná State, 80035-050, Brazil

Abstract: The present study had the objectives to apply the splitting technique in feed pellet quality prediction models and validate these models through Herzberg equation. The independent factors (input data) were the particle size (PS: coarse and medium), heat treatment (HT: expanded-pelleted and pelleted), fat addition levels (FA: 15, 25, 35 and 45 g/kg of feed) and moisture addition levels (MA: 0, 7, 14 and 21 g/kg of feed) which were combined in a full factorial design (2 × 2 × 4 × 4), resulting in 64 different treatments with eight replicates each. The intact pellets amount and pellet durability index (PDI) were considered as the model's output data. In the splitting technique the whole data set, composed by 512 observations, were splitted in two data set: (1) model construction set (75% of the total data) and (2) model validation set (25% of the total data set, which were selected randomly from the original data set). Both equations, the one obtained by splitting method and the one obtained by whole data set, had good coefficient of determination and similar residues square means. It was concluded that splitting technique can be successfully applied to fit a prediction equation for feed pelleting process and that Herzberg equation consists in a useful tool to validate the coefficient of determination of those models.

Key words: Modelling, feed, Splitting technique, Herzberg equation, pellets.

1. Introduction

Process modelling consists in obtaining a mathematical equation which adequately represents a phenomenon. According to Teixeira [1], there are empirical modelling techniques based on the introduction of a step or slope of disturbance and the consequent fitting of the response data through a prediction equation. Franke and Rey [2] mentioned several factors that affect feed pellet quality, such as diet formulation, ingredient granulometry, equipment specifications, and process parameters, among others, which can be considered as disturbances introduced into feed pellet manufacturing process.

Once a mathematical model is established for the system under study, it is important to validate the prediction accuracy of the respective model. Model validation involves checking the prediction equation against an independent set of data to verify how well it can predict them, and therefore, an alternative is the technique known as splitting or cross-validation method. According to Snee [3], Steckel and Vanhonacker [4], and Kozak and Kozak [5], in this technique, the collected data (n) are divided into two groups such that: $n = n_m + n_v$, where n_m is the number of observations used in model construction and n_v is the number of observations used for calculating prediction residuals (validation). According to these same authors, the set of data used for validation is randomly designated and usually comprises between 10%-50% of the total data. Snee [3] points out that one of the disadvantages of the splitting technique is the increase in the equation's coefficients variances.

Corresponding author: Keysuke Muramatsu, D.Sc. veterinarian, research fields: feed processing, poultry nutrition and modelling.

In the present research, the pelleting process parameter and pellet quality data obtained in the experiment conducted by Muramatsu et al. [6] were used as the original data set. The input parameters for the prediction equation were constituted by the following parameters: (1) two different particle sizes, (2) two types of heat treatment, (3) four levels of moisture addition in the conditioner, and (4) four levels of fat addition in the mixer. The amounts of intact pellets and the PDI were considered as output data. The objectives of this study were to apply the Splitting technique in pellet quality prediction models and validate the models generated by Herzberg equation.

2. Material and Methods

The production of experimental diets was carried out at a commercial feed mill located in the state of Rio Grande do Sul. The four fat inclusions (15, 25, 35, and 45 g/kg of feed), four levels of moisture addition in the conditioner (0, 7, 14, and 21 g/kg of feed), two particle sizes (medium and coarse), and two types of heat treatment (expanded-pelleted and pelleted) were combined in a 4 × 4 × 2 × 2 factorial scheme, totaling 64 treatments in a completely randomized design. Eight samples (replicates) of each treatment were collected, totaling 512 samples.

The data from these 512 samples were used to obtain the regression equation of the model considering the whole data set. For the splitting technique, the 512 sample sets were divided into two groups: 384 observations (75%) were used to fit the prediction equation, and 128 observations (25%) were set aside as validation data. For the models obtained using the splitting technique, the mean squares of the residuals were calculated for both the construction and validation data sets.

The statistical model included replicates, particle size (PS), type of heat treatment (HT), levels of moisture addition (MA), levels of fat addition (FA), and their interactions:

Yijklm = μ + PSi + HTj + FAk + MAl + (PS x HT)ij + (PS x FA)ik + (PS x MA)il +(HT x FA)jk + (HT x MA)jl + (PS x HT x FA)ijk + (PS x HT x MA)ijl + (PS x FA x MA) ikl + (HT x FA x MA)jkl + (PS x HT x FA x MA)ijkl + εijklm

where: Yijklm = observations, μ = average, PSi = Particle Size effect (i = medium or coarse), HTj = Heat Treatment effect (j = Pelleted or Expanded-Pelleted), FAk = Fat Addition effect (k = 15-5 g/kg feed), MAl = Moisture Addition effect (l = 0-21 g/kg feed), and εijklm = residues error.

The general linear model including the effects mentioned previously was used to adjust a regression equation for intact pellet amount and PDI. The statistical software used for this purpose was Statistica version 8.0 (StatSoft, Inc.). Anderson-Darling test, available in Minitab version 16, was used to check the normality of the data regarding intact pellet percentage and PDI and Hartley F-Max test was employed to check the homoscedasticity of the data. The significance of each effect and their interaction in the model was accepted if $p < 0.05$.

In order to validate the prediction equation generated by the splitting technique, the coefficient of determination for the unknown data (or validation data) must be equal to or greater than the coefficient of determination calculated using a technique referenced by Sobol [7], known as the Herzberg equation: $R_c^2 = 1 - \left(\frac{N-1}{N-K-1}\right)\left(\frac{N+K+1}{N}\right)(1 - R^2)$, where R_c^2 is the expected coefficient of determination, N is the total number of validation data points, and K is the number of independent variables in the studied model (including the constant).

3. Results and Discussions

The percentage of pellets and PDI did not follow a normal distribution when submitted to Anderson-Darling test ($p < 0.05$). Consequently, the data were subjected to Johnson’s transformations: (1) Intact pellet amount (g/kg) = -0.498404 + 0.601586*Ln (Intact pellet amount g/kg - 460.565)/(929.921 - Intact pellet amount g/kg); (2) PDI % = -0.499811 + 0.797976*Ln ((PDI % - 47.6941)/(95.8331 – PDI %)).

The whole data set and the splitting data set (model construction data) were independently submitted to regression analysis in order to obtain the prediction equations for PDI and amount of intact pellets (both with a model significance level of $p < 0.001$). Initially, the analyses were performed according to the originally proposed statistical model (complete factorial 2 × 2 × 4 × 4). Subsequently, the Backward elimination method was employed to remove non-significant factors from the prediction equation, limiting the interactions to two by two factor levels.

The regression equations for Intact pellet amount (g/kg) obtained using the whole data and the splitting technique were respectively: (1) Y = -1.7699 + 0.1172*FA (g/kg) - 0.0026*FA(g/kg)^2 - 0.4417*PS + 0.8977*HT + 0.0124*MA (g/kg) + 0.1749*PS*HT + 0.0149*PS*MA (g/kg) + 0.0111*HT*FA (g/kg) + 0.0189*HT*MA (g/kg) + 0.00046*FA (g/kg)*MA (g/kg) and (2) Y = -1.7961 + 0.1192*FA (g/kg) - 0.0026*FA(g/kg)^2 - 0.4474*PS + 0.8682*HT + 0.0126*MA (g/kg) + 0.1858*PS*HT + 0.0152*PS*MA (g/kg) + 0.0116*HT*FA (g/kg) + 0.0199*HT*MA (g/kg) + 0.00047*FA (g/kg)*MA (g/kg). And for PDI the equations were respectively (1) Y = 1.5315 + 0.0680*MA (g/kg) - 0.0011*MA (g/kg)^2 - 0.1203* FA (g/kg) + 0.00073*FA (g/kg)^2 – 1.1420044*PS + 0.829193562*HT + 0.0391*PS*FA (g/kg) + 0.0256*HT*FA (g/kg) and (2) Y = 1.5145 + 0.0690* MA (g/kg) - 0.0012*MA (g/kg)^2 - 0.1188*FA (g/kg) + 0.00070*FA (g/kg)^2 – 1.17472*PS + 0.879146379*HT + 0.0400*PS*FA (g/kg) + 0.0243*HT*FA (g/kg) for the whole data set and the splitting technique set.

The coefficients of determination for the equations obtained with the total data and with the splitting technique were 0.971 and 0.968 for percentage of pellets, and 0.853 and 0.851 for PDI, respectively (Tables 1 and 2). In this study, different from what was mentioned by Steyeberg et al. [8], both the equation generated by the two data sets (whole data and splitting technique data set) showed a good coefficient of determination for percentage of pellets and PDI. It is likely that the present pelleting process was under control and the main variables that could affect the process were kept constant, except for those used as input data. This is consistent with Fahrenholz [9], who used modeling in the pelleting process to predict the effect of different variation factors on pellet quality and found consistency in his predictions. Moreover, it can be said that the predictive capability of the model was not compromised by the use of the splitting technique.

These equations were used to estimate the predicted values, and from these values the prediction residues in relation to the validation data were calculated (Table 3). The mean squares residues observed between the predicted data and the validation data were 0.0379 for percentage of pellets and 0.1775 for PDI, respectively.

The coefficients of determination obtained by the splitting technique for the validation data set were 0.952 and 0.834 for intact pellets amount and PDI, respectively, and therefore equal to or greater than the coefficients of determination estimated by the Herzberg equation (R_c^2 = 0.952 and 0.826) for these same parameters (Table 3). Thus, as pointed by Oredein et al. [10] the Herzberg equation could be used to validate the prediction quality of the mathematical model derived from the splitting technique using the validation (unknown) set of observations.

The equations generated by the whole data set showed similar mean squares of residues when compared to the equations derived from the splitting technique, both for percentage of pellets (0.0538 versus 0.0596) and for PDI (0.1859 versus 0.1890). In the present study, the set of “unknown” observations used as validation data for the model represented 25% of the total observations. According to Steckel and Vanhonacker [4], when the total number of observations exceeds 100, about 20% of these data can be segregated as validation data.

Table 1 Regression equations for intact pellet amount generated from the whole data set and with the splitting technique data set.

	Prediction equation using whole data set		Prediction equation using splitting technique data set	
Intact pellet amount (Johnson transformation)	Y = -1.7699 + 0.1172*FA (g/kg) - 0.0026* FA (g/kg)^2 - 0.4417*PS + 0.8977*HT + 0.0124*MA (g/kg) + 0.1749* PS*HT + 0.0149*PS* MA (g/kg) + 0.0111*HT*FA (g/kg) + 0.0189*HT*MA (g/kg) + 0.00046*FA (g/kg)*MA (g/kg)		Y = -1.7961 + 0.1192*FA (g/kg) - 0.0026* FA (g/kg)^2 - 0.4474*PS +0.8682*HT + 0.0126*MA (g/kg) + 0.1858*PS*HT + 0.0152*PS*MA (g/kg)+ 0.0116*HT*FA (g/kg) + 0.0199*HT*MA (g/kg) + 0.00047*FA (g/kg)*MA (g/kg)	
R^2	0.971		0.968	
Model *p*-value	< 0.001		< 0.001	
Residue square mean	0.0538		0.0596	
Coefficients	*p*-value	Standard error	*p*-value	Standard error
PS	< 0.001	0.0172	< 0.001	0.0491
HT	< 0.001	0.0326	< 0.001	0.0832
FA	< 0.001	0.0064	< 0.001	0.0078
FA^2	< 0.001	0.0001	< 0.001	0.0001
MA	< 0.001	0.0038	< 0.001	0.0051
PS*HT	< 0.001	0.0103	< 0.001	0.0500
PS*MA	< 0.001	0.001 3	< 0.001	0.0032
HT*FA	< 0.001	0.0009	< 0.001	0.0022
HT*MA	< 0.001	0.0013	< 0.001	0.0032
FA*MA	< 0.001	0.0001	< 0.001	0.0001

PS = 0 if medium, 1 if coarse; HT =0 if pelleted, 1 if pelleted-expanded.

Table 2 Regression equations for PDI generated from the whole data set and with the splitting technique data set.

	Prediction equation using whole data set		Prediction equation using splitting technique data set	
PDI (Johnson transformation)	Y = 1.5315 + 0.0680*MA (g/kg) - 0.0011*MA (g/kg)^2 - 0.1203*FA (g/kg) + 0.00073* FA (g/kg)^2 - 1.1420044*PS + 0.829193562*HT + 0.0391*PS* FA (g/kg) + 0.0256*HT*FA (g/kg)		Y = 1.5145 + 0.0690*MA (g/kg) - 0.0012*MA (g/kg)^2 - 0.1188* FA (g/kg) + 0.00070*FA (g/kg)^2 - 1.17472*PS + 0.879146379*HT + 0.0400*PS*FA (g/kg) + 0.0243*HT*FA (g/kg)	
R^2	0.853		0.851	
Model *p*-value	< 0.001		< 0.001	
Residue square mean	0.1859		0.1890	
Coefficients	*p*-value	Standard error	*p*-value	Standard error
PS	< 0.001	0.1098	< 0.001	0.1276
HT	< 0.001	0.1098	< 0.001	0.1276
FA	< 0.001	0.0118	< 0.001	0.0138
FA^2	0.0001	0.0118	0.0002	0.0200
MA	< 0.001	0.0085	< 0.001	0.0099
MA^2	0.0037	0.0004	0.0103	0.0005
PS*FA	< 0.001	0.0034	0.001	0.0040
HT*FA	< 0.001	0.0034	< 0.001	0.0040

PS = 0 if medium, 1 if coarse; HT =0 if pelleted, 1 if pelleted-expanded.

Table 3 Prediction residues mean squares calculated by the difference between regression estimates values and validation data set.

N	PS	HT	FA (g/kg)	MA (g/kg)	Intact Pellets (Johnson) Obs. Data**	PDI (Johnson) Obs. Data**	Intact Pellets (Johnson) Reg. Estimate*	PDI (Johnson) Reg. Estimate*	Residues Mean Squares Intact Pellets	Residues Mean Squares PDI
1	Coarse	Exp.	15	0	0.2026	0.7504	0.1875	0.5589	0.0002	0.0367
2	Coarse	Pelet.	15	0	-0.985	-0.147	-1.0405	-0.6847	0.0031	0.2892
3	Coarse	Exp.	15	7	0.6315	1.0506	0.5708	0.9831	0.003 7	0.0046
4	Coarse	Pelet.	15	7	-0.8777	0.0424	-0.7966	-0.2605	0.0066	0.0917
5	Coarse	Exp.	15	14	0.7275	1.3524	0.954	1.2897	0.0513	0.0039
6	Coarse	Pelet.	15	14	-0.5492	0.342	-0.5526	0.0461	0	0.0876
7	Coarse	Exp.	15	21	0.9565	1.9126	1.3373	1.4787	0.145	0.1883
8	Coarse	Pelet.	15	21	-0.2906	0.3248	-0.3087	0.2351	0.0003	0.008
9	Medium	Exp.	15	0	0.821	1.097	0.4491	1.1336	0.1383	0.0013
…	…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…	…
119	Medium	Exp.	45	0	-0.0539	-0.3645	-0.3069	-0.4414	0.064	0.0059
120	Medium	Pelet.	45	0	-1.615	-3.1363	-1.6971	-2.414	0.0067	0.5217
121	Medium	Exp.	45	7	0.1316	-0.038	0.0687	-0.0172	0.004	0.0004
122	Medium	Pelet.	45	7	-1.5276	-2.4914	-1.4609	-1.9898	0.0045	0.2516
123	Medium	Exp.	45	14	0.3844	0.196	0.4442	0.2894	0.0036	0.0087
124	Medium	Pelet.	45	14	-1.1905	-2.2473	-1.2246	-1.6832	0.0012	0.3182
125	Medium	Exp.	45	21	0.498	0.4737	0.8198	0.4784	0.1035	0
126	Medium	Pelet.	45	21	-0.744	-1.9725	-0.9884	-1.4942	0.0597	0.2288
127	Medium	Pelet.	15	0	-0.5664	-0.012	-0.5931	-0.11	0.0007	0.0096
128	Coarse	Pelet.	35	21	-0.3189	-1.0263	-0.3273	-0.6409	0.0001	0.1485
Residues square mean									0.0379	0.1775
Sum residues squares									4.851	22. 725
Coefficient of determination									0.952	0.834

** Obs. Data: Validation data set values collected in the experiment.

* Reg. Estimate: values estimated by the regression equation.

4. Conclusions

When there is a numerous data set (> 100 observations) the splitting technique can be applied for process modelling validation without compromising the residues mean squares; and the comparison of coefficient of determination obtained by Herzberg equation and by splitting technique can be successfully used as a tool to validate the proposed mathematical model.

References

[1] Teixeira, G. M. 2010. "Application of diferential fractional equation to control absorption columns." Master thesis, Technology Area of Universidade Federal do Paraná (in Portuguese)

[2] Franke, M., and Rey, A. 2006. "Improving Pellet Quality and Efficiency." *Feed Tech.* 10 (3): 12-5. http://www.allaboutfeed.net/PageFiles/10790/001_boerderij-download-AAF10415D01.pdf.

[3] Snee, R. D. 1977. "Validation of Regression Models: Methods and Examples." *Technometrics* 19 (4): 415-28. http://links.jstor.org/sici?sici=0040-1706%28197711%2919%3A4%3C415%3AVORMMA%3E2.0.CO%3B2-6.

[4] Steckel, J. H., and Vanhonacker, W. R. 1993. "Cross-Validating Regression Models in Marketing Research." *Marketing Science* 12 (4): 415-27. http://www.jstor.org/stable/183938.

[5] Kozak, A., and Kozak, R. 2003. "Does Cross Validation Provide Additional Information in the Evaluation of Regression Models?" *Can. J. For. Res.* 33: 976-87.

[6] Muramatsu, K., Maiorka, A., Vaccari, I. C. M., Reis, R. N., Dahlke, F., Pinto, A. A., Orlando, U. A. D., Bueno, M., and Imagawa, M. 2013. "Impact of Particle Size, Thermal Processing, Fat Inclusion and Moisture Addition on Pellet Quality and Protein Solubility of Broiler Feeds." *J Agri. Sci. Tech. A* 17: 1017-28.

[7] Sobol, M. G. 1991. "Validation Strategies for Multiple

Regression Analysis: Using the Coefficient of Determination." *Interfaces* 21 (6): 106-20. http://interfaces.journal.informs.org/content/21/6/106.full.pdf+html.

[8] Steyeberg, E. W., Harrel, J. R. F. E., Boorsboom, G. J .J. M., Eijkemans, M. J. C., Vergouwe, Y., and Habbema, J. D. F. 2001. "Internal Validation of Predictive Models: Efficiency of Some Procedures for Logistic Regression Analysis." *Journal of Clinical Epidemiology* 54: 774-81. http://www.jclinepi.com/article/S0895-4356%2801%290 0341-9/pdf. doi.: S0895-4356(01)00341-9.

[9] Fahrenholz, A. C. 2012. "Evaluating Factors Affecting Pellet Durability and Energy Consumption in a Pilot Feed Mill and Comparing Methods for Evaluating Pellet Durability." Ph.D., thesis, Kansas University.

[10] Oredein, A. I., Olatayo, T. O., and Loyinmi, A. C. 2011. "On Validating Regression Models with Bootstraps and Data Splitting Techniques." *Global Journal of Science Frontier Research* 11 (6): 1-6. https://globaljournals.org/GJSFR_Volume11/1-On-Validating-Regression-Models-with-Bootstraps.pdf.