

DATA-DRIVEN PREDICTION OF PIGLETS BORN ALIVE PER FARROWING USING MONTE CARLO-OPTIMISED MULTIPLE REGRESSION

Mykola Povod^{1a}, Tomasz Cepowski², Zofia Cepowska³, Igor Voshchenko^{1b}, Oleksandr Mykhalko^{1c}, Oleksandr O. Borshch⁴, Mykhailo Matvieiev⁵, Vitalii Pakholiuk⁶, and Andriy Getya^{7*}

^{1a} Department of Feed Technology and Animal Feeding, Sumy National Agrarian University, Ukraine
<https://orcid.org/0000-0002-2470-4921>

² Faculty of Navigation, Maritime University of Szczecin, Poland
<https://orcid.org/0000-0001-8038-075X>

³ Pomeranian Medical University, Poland
<https://orcid.org/0009-0005-7910-0833>

^{1b} Department of Feed Technology and Animal Feeding, Sumy National Agrarian University, Ukraine
<https://orcid.org/0009-0005-2745-9000>

^{1c} Department of Feed Technology and Animal Feeding, Sumy National Agrarian University, Ukraine
<https://orcid.org/0000-0002-2470-4921>

⁴ Department of Technology Milk and Meat Production, Bila Tserkva National Agrarian University, Ukraine
<https://orcid.org/0000-0002-8450-2109>

⁵ Department of Animal Technology, National University of Life and Environmental Sciences of Ukraine, Ukraine
<https://orcid.org/0000-0003-1281-9032>

⁶ "Ascania Nova" Institute of Animal Breeding in the Steppe Regions named after M.F. Ivanov, National Scientific Center for Sheep Breeding NAAS, Ukraine
<https://orcid.org/0009-0000-7639-3005>

⁷ National University of Life and Environmental Sciences of Ukraine, Ukraine

* Corresponding author: getya@ukr.net

ABSTRACT

In this study, a data-driven nonlinear power regression approach with Monte Carlo-based exponent optimisation was applied to predict the number of piglets born alive per farrowing (PFA). The analysis was based on productivity data from F1 sows obtained by crossbreeding Danish Landrace and Danish Large White breeds and inseminated with semen from Danish Duroc terminal line boars. Data were collected from 114 commercial farms in Denmark and divided into a primary dataset used for model development and an independent test dataset. The Monte Carlo random-search procedure enabled automatic identification of the optimal power exponents, thereby reducing the need for arbitrary variable transformations and improving model accuracy. The final model included a combined nonlinear predictor based on the number of weaned piglets per sow per year (PPSY) and piglet survival rate (PSR%), together with a quadratic term describing the average annual number of farrowings per sow (F). The model achieved high predictive accuracy, with $R^2 = 0.99632$, adjusted $R^2 = 0.99623$, RMSE = 0.04330, RRMSE = 0.23932%, and CCC = 0.99816. The test dataset RMSE was 0.04725, and the test-to-dataset RMSE ratio was 1.09, indicating good predictive stability and no substantial overfitting. The developed model demonstrated high predictive accuracy, statistical robustness, satisfactory diagnostic performance, and good generalisability. The results indicate that

data-driven nonlinear power regression with Monte Carlo-based exponent optimisation may provide a useful approach for modelling biological and zootechnical data, particularly when nonlinear relationships are expected.

Keywords: pigs, breeding, piglet survival rate to weaning, piglets born alive per farrowing, weaning weight of the piglets.

INTRODUCTION

Accurate prediction of pig reproductive performance is important to improving the efficiency of pig farming. Traditionally, sow productivity has been evaluated using statistical methods that quantify the relationships between various factors and reproductive traits (Chen et al., 2025). These approaches remain valuable for analysing production data and identifying statistical patterns (Plaengkaeo et al., 2021). Reliable prediction of sow reproductive performance can improve breeding efficiency, support management decisions, and increase the economic sustainability of pig production systems (Knol et al., 2016). Therefore, the development of more accurate predictive models is highly relevant to modern swine production.

Mathematical modelling is widely used in pig farming to support decision making regarding genetic selection, nutrition, health management, and litter size optimisation. Traditional linear models often fail to describe the complex biological interactions of pigs with many factors (Yu et al., 2022), which has led to growing interest in the use of nonlinear regression models that can help better describe the dynamics of biological processes such as growth, productivity and reproductive performance, offering more accurate predictions and a deeper understanding of complex relationships (Powell et al., 2022; O'Brien and Silcox, 2024).

Until recently, linear approaches were widely used for data analysis in farm animal research (Nayeri et al., 2019). For example, random regression models allow analysing individual growth patterns in piglets over time, providing a useful framework for evaluating the offspring of hyperprolific sows. These models account for both average growth trends and random deviations of each individual, enabling more accurate estimates (Škorput et al., 2023). Similarly, multi-trait models can account for the relationships between many reproductive traits of animals. By jointly incorporating information on litter size, growth, piglet survival, and other traits, these models improve the accuracy of the prediction of productive performance in pigs. This approach is particularly effective for traits with low heritability, such as the number of liveborn piglets (Yang et al., 2024).

The use of nonlinear models to predict reproductive traits in sows remains relatively uncommon. However, according to Zhou et al., (2022) and Atamanyuk et al. (2019), these models offer a number of advantages over traditional approaches, such as linear random regression, multi-trait analyses, Wiener and Kalman methods, and other commonly used statistical techniques. First, nonlinear models are capable of reflecting complex, asymmetric, and asymptotic relationships among biological parameters, which is particularly relevant to the analysis of reproductive cycles, where relationships may deviate substantially from linearity or additivity (Park et al., 2019; Auger-Méthé et al., 2021). Second, they allow modelling time lags, threshold effects, synergism or saturation, which are often ignored in linear multivariable models but are crucial for accurate prediction of prolificacy, culling or productive life of sows (Archontoulis and Miguez 2015; Xu et al., 2024).

Nonlinear regression models may better describe the complex relationships among reproductive traits in sows and, consequently, provide higher predictive accuracy for population-level estimates than traditional linear approaches. Therefore, the objective of the present study was to evaluate the effectiveness of a Monte Carlo-optimised multiple regression model for predicting reproductive traits of sows.

MATERIALS AND METHODS

Farm location and animal material

The study analysed data from F_1 local sows obtained by crossing Danish Landrace and Danish Large White breeds and inseminated with semen from boars of the Danish Duroc terminal line. The data were collected from commercial pig farms located in the western part of the Jutland Peninsula, Denmark. Farm-level information was sourced from the publicly available DB-Tjek pig farm rating analyses for 2021 and 2022, conducted by the consulting firm SvineRådgivningen and provided by members of the Ukrainian Pig Breeding Association in Denmark (Powered by Porcus).

Data collection

The data were collected from the Cloudfarms herd management software. Overall, the database

included average productivity data from 86 farms across 15 traits. These farms housed 133,116 sows of various ages, of which an average of 23.3% were first-parity sows. The following traits were recorded: number of weaned piglets per sow per year (PPSY), average annual number of farrowings per sow (F), piglet survival rate to weaning (PSR%), number of weaned piglets per farrowing (WPF), number of piglets born alive per farrowing (PFA), weaning weight of piglets (WWP), duration of the suckling period (DSP), and average daily weight gain in the suckling period (DWG).

Statistical analysis

To analyse the relationships between variables, the following nonlinear regression model was applied:

$$Y = a_0 + a_1 \times x_1^{b_1} + a_2 \times x_2^{b_2} + a_3 \times x_3^{b_3} + \dots + a_{m-1} \times x_{n-1}^{b_{m-1}} \times x_n^{b_n} \quad (1)$$

where:

Y – dependent variable,

$x_1, x_2, \dots, x_n$ – independent variables,

$a_0, a_1, \dots, a_m$ – regression coefficients,

$b_1, b_2, \dots, b_n$ – power exponents.

Model construction was based on an iterative data-driven procedure for optimising predictor transformations (Fig. 1). The regression

coefficients a_m were estimated by the least-squares method, whereas the corresponding exponents b_n were adjusted to obtain the best agreement between Eq. (1) and the empirical data. The optimisation of the exponents was carried out using a Monte Carlo random-search algorithm, with the root mean square error (RMSE) used as the objective function. During this procedure, different combinations of exponent values were randomly generated from predefined intervals, typically ranging from -3 to $+3$, using a uniform distribution. For each generated set of exponents, the regression coefficients were recalculated using the least-squares method, and model accuracy was subsequently assessed. The search was performed iteratively without imposing a fixed number of trials and was terminated when subsequent iterations no longer resulted in a reduction in RMSE. In practical computations, stable solutions were generally obtained after approximately 100,000 iterations, requiring approximately 3 h of calculation on an Apple Mac Mini M1 with 16 GB of RAM.

The combination of exponents yielding the lowest RMSE was treated as the optimal solution. This stochastic, data-driven search strategy enabled broad exploration of the parameter space and reduced the likelihood of selecting a solution associated with a local minimum.

Among the generated candidate models, the final model was selected based not only on the

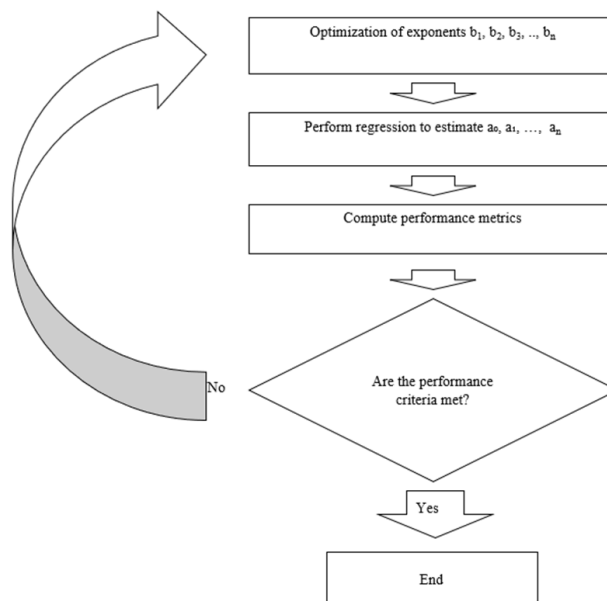


Fig. 1 Flowchart of the regression exponent optimisation algorithm based on model performance criteria.

lowest RMSE but also on additional statistical and diagnostic criteria. The selected model was required to satisfy the following conditions:

- Absence of overfitting, assessed by comparing the prediction error for the test dataset with that for the training dataset; the relative increase in the test error was not allowed to exceed 10%;
- Absence of substantial multicollinearity among predictors, verified using the variance inflation factor, with variance inflation factor (VIF) < 2;
- Statistical significance of the overall regression model, confirmed by the F-test with $p < 0.05$;
- Statistical significance of all individual predictors, confirmed by $p < 0.05$;
- Sensitivity contribution greater than 5% for each predictor, based on sensitivity analysis, with $SA > 5\%$;
- Satisfactory Bland–Altman agreement, indicating the absence of systematic bias and no statistically significant proportional bias, with residual differences remaining within the 95% limits of agreement;
- Approximate normal distribution of residuals;
- Minimal heteroscedasticity of residuals in both the training and test datasets.

Six performance indicators were used to evaluate the predictive ability of the model with respect to statistical parameters: the coefficient of determination (R^2), adjusted R^2 , Pearson correlation coefficient (r), concordance correlation coefficient (CCC), standard error (SE), root mean square error (RMSE), and relative root mean square error (RRMSE). Low values of SE, RMSE, and RRMSE (approaching 0), together with high values of R^2 and r (approaching 1), indicate a strong agreement between predicted and observed values.

In the present study, model evaluation was based on fitting accuracy within the available dataset, without external validation. Following previous studies (Rao et al., 2018; Uncuoglu et al., 2022; Cepowski and Drozd 2023), estimation accuracy based on RRMSE was classified as follows:

- Excellent: $RRMSE < 10\%$,
- Good: $10\% \leq RRMSE < 20\%$,
- Acceptable: $20\% \leq RRMSE < 30\%$,
- Unacceptable: $RRMSE \geq 30\%$.

To further assess the statistical significance and robustness of the models, analysis of variance (ANOVA) and Student's t-tests were performed. ANOVA was used to evaluate the overall

significance of each regression model, while the t-tests assessed the individual significance of each regression coefficient. For each coefficient, a corresponding t-statistic and p-value were calculated. Coefficients with p-values less than 0.05 were considered statistically significant at the 95% confidence level.

In addition to standard goodness-of-fit measures, the concordance correlation coefficient (CCC) and Bland–Altman analysis were used to evaluate the agreement between observed and estimated values. This was particularly important because the analyzed data were biological in nature and therefore subject to natural variability, individual differences, and measurement uncertainty. The CCC was applied to assess both the precision and accuracy of model predictions, providing a more comprehensive measure of agreement than the Pearson correlation coefficient alone. Bland–Altman analysis was used to identify systematic bias and proportional bias between observed and estimated values. The mean difference between observed and estimated values was used to assess systematic bias, while regression analysis of the differences against the mean values was used to evaluate proportional bias. Agreement was considered satisfactory when the mean bias was close to zero, proportional bias was not statistically significant, and most differences were within the 95% limits of agreement.

The performance criteria were calculated using the following equations:

- Pearson correlation coefficient (r):

$$r = \frac{\sum(y_e - \bar{y})(y_o - \bar{y}_o)}{\sqrt{\sum(y - \bar{y})^2 \sum(y_e - \bar{y}_e)^2}} \quad (2)$$

- Concordance correlation coefficient (CCC):

$$CCC = \frac{2 \cdot r \cdot \sigma_y \cdot \sigma_{y_e}}{\sigma_y^2 + \sigma_{y_e}^2 + (\bar{y} - \bar{y}_e)^2} \quad (3)$$

- Determination coefficient (R^2):

$$R^2 = \frac{\sum(y_e - \bar{y})^2}{\sum(y - \bar{y})^2} \quad (4)$$

- Adjusted R^2 (Adj. R^2):

$$Adj. R^2 = 1 - \frac{(1 - R^2)(n - 1)}{n - k - 1} \quad (5)$$

- Standard Error (SE):

$$SE = \sqrt{\frac{\sum(y_e - \bar{y})^2}{n(n - 1)}} \quad (6)$$

- Root Mean Square Error (RMSE):

$$RMSE = \sqrt{\frac{\sum(y - y_e)^2}{n}} \quad (7)$$

- Relative Root Mean Square Error (RRMSE):

$$RRMSE = \frac{RMSE}{\bar{y}_e} \times 100 \quad (8)$$

where:

y – observed value,

y_e – estimated value,

$\bar{y}$, $\bar{y}_e$ – mean values of observed and estimated data, respectively,

σ_y – standard deviation of observed data,

σ_{y_e} – standard deviation of estimated data,

n – number of observations,

k – number of model predictors.

In addition to the aforementioned model fit metrics, the regression model was further evaluated based on ANOVA results, the statistical significance of individual predictors, multicollinearity (VIF), heteroskedasticity, and the normality of residuals.

When residuals were not normally distributed, the reliability of individual p-values could be compromised. Therefore, predictor importance was further assessed using sensitivity analysis (SA%), which quantifies the percentage increase in RMSE that occurs when a given predictor is removed from the model. This approach provides a robust, distribution-independent measure of

the practical contribution of each predictor to model accuracy.

All calculations were performed using ndCurveMaster software (Cepowski, 2025).

RESULTS

To predict the dependent variable, the number of piglets born alive per farrowing (PFA), a nonlinear multiple regression model was developed using two transformed predictors: a combined productivity-survival predictor based on the number of weaned piglets per sow per year (PPSY) and piglet survival rate (PSR%), and the average annual number of farrowings per sow (F). During the data-driven model search, several candidate nonlinear regression models were generated; the resulting models are summarized in Table 1. Among the equations evaluated, model no. 57 was selected as the final model because it provided the best balance of predictive accuracy, stability on the test dataset, and overall statistical and diagnostic performance.

The final form of the model is presented in Table 2 and defined by the following equation:

$$PFA = 2.81791 + 46.8487 \left(\frac{PPSY}{PSR\%} \right)^{3/4} - 1.70009 F^2 \quad (9)$$

Table 1 Candidate nonlinear regression models generated during the model selection procedure for predicting the number of piglets born alive per farrowing (PFA)

id	RMSE	Test/Data RMSE	Model
57	0.0432968	1.09123	$PFA = a_0 + a_1 \cdot PPSY^{3/4} \cdot PSR\%^{-3/4} + a_2 \cdot F^2$
55	0.0447359	1.04833	$PFA = a_0 + a_1 \cdot F^{2.4} + a_2 \cdot PPSY^{3/4} \cdot PSR\%^{-3/4}$
54	0.044737	1.04548	$PFA = a_0 + a_1 \cdot F^{2.45} + a_2 \cdot PPSY^{0.8} \cdot PSR\%^{-0.8}$
52	0.0449467	1.04325	$PFA = a_0 + a_1 \cdot F^{2.55} + a_2 \cdot PPSY^{0.95} \cdot PSR\%^{-0.95}$
48	0.0460484	1.0465	$PFA = a_0 + a_1 \cdot F^{2.65} + a_2 \cdot PPSY^{1.15} \cdot PSR\%^{-1.15}$
46	0.045089	1.07705	$PFA = a_0 + a_1 \cdot F^2 + a_2 \cdot PPSY^{1/2} \cdot PSR\%^{-1/2}$
43	0.0435428	1.10649	$PFA = a_0 + a_1 \cdot PPSY \cdot PSR\%^{-1} + a_2 \cdot F^2$
41	0.0472768	1.04198	$PFA = a_0 + a_1 \cdot F^{2.35} + a_2 \cdot PPSY^{0.45} \cdot PSR\%^{-0.45}$
40	0.101956	1.04808	$PFA = a_0 + a_1 \cdot F^{-1.25} + a_2 \cdot PPSY^{-1.85} \cdot PSR\%^{1.85}$
39	0.177554	1.0035	$PFA = a_0 + a_1 \cdot F^{2/3} + a_2 \cdot PPSY^{2.45} \cdot PSR\%^{-1.95}$
37	0.0420326	1.15542	$PFA = a_0 + a_1 \cdot F^{1.4} + a_2 \cdot PPSY^{0.85} \cdot PSR\%^{-0.85}$
36	0.0434458	1.11161	$PFA = a_0 + a_1 \cdot F^{1.95} + a_2 \cdot PPSY \cdot PSR\%^{-1}$
35	0.0461999	1.04092	$PFA = a_0 + a_1 \cdot F^{2.7} + a_2 \cdot PPSY^{1.15} \cdot PSR\%^{-1.15}$
34	0.0524731	1.13889	$PFA = a_0 + a_1 \cdot F^{-0.85} + a_2 \cdot PPSY^{0.65} \cdot PSR\%^{-2/3}$
33	0.0634358	0.974652	$PFA = a_0 + a_1 \cdot F^{2.9} + a_2 \cdot PPSY^{1.75} \cdot PSR\%^{-1.8}$
32	0.063551	0.931718	$PFA = a_0 + a_1 \cdot F^{2.7} + a_2 \cdot PPSY^{1.15} \cdot PSR\%^{-1.2}$
31	0.0790217	1.12304	$PFA = a_0 + a_1 \cdot F^{0.2} + a_2 \cdot PPSY^{2.3} \cdot PSR\%^{-2.35}$
30	0.0899388	1.07147	$PFA = a_0 + a_1 \cdot F^{-0.95} + a_2 \cdot PPSY^{-1.3} \cdot PSR\%^{1.35}$
29	0.0969397	1.06206	$PFA = a_0 + a_1 \cdot F^{-3} + a_2 \cdot PPSY^{1.3} \cdot PSR\%^{-1.4}$
28	0.106896	1.08548	$PFA = a_0 + a_1 \cdot F^{-1.2} + a_2 \cdot PPSY^{0.8} \cdot PSR\%^{-0.7}$
27	0.147382	1.05332	$PFA = a_0 + a_1 \cdot F^{-0.45} + a_2 \cdot PPSY^{1.2} \cdot PSR\%^{-1.4}$
26	0.661817	1.17654	$PFA = a_0 + a_1 \cdot F + a_2 \cdot PPSY \cdot PSR\%$

Table 2. Summary of model performance for prediction of piglets born alive per farrowing.

Observations	86
R ² :	0.99632
Adj. R ² :	0.99623
Pearson correlation coefficient r:	0.99816
Concordance correlation coefficient, CCC	0.99816
Standard error, SE	0.04407
RMSE	0.04330
Test/Dataset RMSE ratio	0.04725
Test/Dataset RMSE ratio	1.09
RRMSE [%]	0.23932

where PFA is the number of piglets born alive per farrowing, PPSY is the number of weaned piglets per sow per year, PSR% is the piglet survival rate from birth to weaning, and F is the average annual number of farrowings per sow.

The Eq.(9) showed an excellent fit to the empirical data (Table 2). The coefficient of determination was $R^2 = 0.99632$, while the adjusted coefficient of determination was Adj. $R^2 = 0.99623$. The Pearson correlation coefficient was $r = 0.99816$, while the concordance correlation coefficient was $CCC = 0.99816$, indicating very strong agreement between observed and predicted PFA values. The root mean square error was $RMSE = 0.04330$, the standard error was $SE = 0.04407$, and the relative root mean square error was only $RRMSE = 0.23932\%$, confirming the high predictive accuracy of the model.

The model (9) was further evaluated using an independent test dataset. The RMSE for the primary dataset was 0.04330, whereas the RMSE for the test dataset was 0.04725. The test-to-dataset RMSE ratio was 1.09, which indicates that the prediction error increased by approximately 9% for the test data. This result suggests that the model showed no evidence of substantial overfitting and retained stable predictive performance when applied to independent data.

The ANOVA results presented in Table 3 confirmed the overall statistical significance of the regression model (9). The computed F-statistic was 11234.5, with $p = 5.55 \times 10^{-16}$, indicating that the model explained a statistically significant proportion of the variability in PFA. Both predictors were also statistically significant, with p-values much lower than 0.05. The combined predictor $PPSY^{3/4} \cdot PSR\%^{-3/4}$ had a t-value of 147.42, while F^2 had a t-value of -80.15. These results confirm that both predictors made a statistically significant contribution to the model.

The VIF was 1.85 for both predictors, indicating no evidence of problematic multicollinearity.

The sensitivity analysis further confirmed the practical importance of both predictors. Removing the combined PPSY-PSR% predictor resulted in a 1,521.21% increase in RMSE, whereas removing F^2 resulted in a 785.41% increase. This indicates that the combined productivity-survival predictor had the strongest influence on model performance, although F^2 also made a substantial contribution.

Pairwise Pearson correlations between the transformed predictors and the response variable PFA are summarized in Table 4. The combined predictor $PPSY^{3/4} \cdot PSR\%^{-3/4}$ showed a strong positive correlation with PFA, $r = 0.843$, indicating that it was the dominant individual predictor. In contrast, F^2 showed only a weak positive correlation with PFA ($r = 0.181$). However, its statistical significance and substantial contribution to model sensitivity indicate that F^2 provided additional predictive value when considered jointly with the combined PPSY-PSR% predictor. The two transformed predictors were moderately correlated ($r = 0.679$); however, the VIF values confirmed that this association did not lead to problematic multicollinearity.

The model's predictive performance was confirmed by the observed-versus-predicted plot (Fig. 2). The data points were closely aligned along the identity line, indicating a high level of agreement between observed and predicted PFA values. Importantly, the points representing the test dataset exhibited a similar pattern to those of the primary dataset and were distributed across the full range of predicted values. This distribution indicates that the test observations were not concentrated within a restricted range and provides additional evidence of the predictive stability and generalisability of the model.

The standardised residuals plotted against predicted values presented in Fig. 3 showed no clear systematic trend. The residuals of both the primary dataset and the test dataset were

Table 3. Estimated parameters of the nonlinear regression model for prediction of the number of piglets born alive per farrowing (PFA).

ANOVA	df	SS	MS	F computed	p
Regression	2	43.6431	21.8215	11234.5	5.55×10^{-16}
Errors	83	0.1612	0.00194		
Total	85	43.8043			
PFA =	Value of a	t-Value	P > t 	VIF	SA [%]
a_0	2.81791	23.302	3.59E-38		
$a_1 \cdot \text{PPSY}^{3/4} \cdot \text{PSR}\%^{-3/4}$	46.8487	147.418	3.35E-102	1.85	1521.21
$a_2 \cdot F^2$	-1.70009	-80.149	2.14E-80	1.85	785.41

Table 4. Pearson correlation coefficients between transformed input variables and the number of piglets born alive per farrowing (PFA).

	$\text{PPSY}^{3/4} \cdot \text{PSR}\%^{-3/4}$	F^2	PFA
$\text{PPSY}^{3/4} \text{PSR}\%^{-3/4}$	1	-	-
F^2	0.679	1	-
PFA	0.843	0.181	1

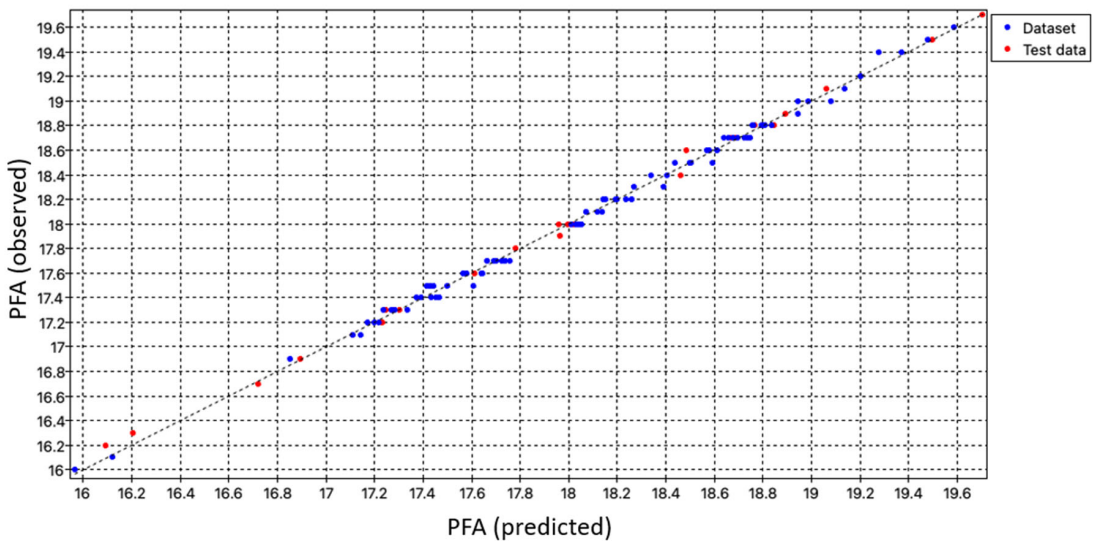


Fig. 2. Plot of observed versus predicted piglets born alive (PFA) per farrowing values.

generally distributed around zero, without a strong funnel-shaped pattern. This indicates limited heteroscedasticity and supports the assumption of approximately constant residual variance. Although several individual residuals were relatively large, they did not form a clear separate pattern for the test dataset.

The Shapiro–Wilk test indicated a normal distribution of residuals, with $W = 0.990$, which was higher than the critical value of 0.971 at $\alpha =$

0.05. The Anderson–Darling test also confirmed normality, with $A^2 = 0.349$ and adjusted $A^2 = 0.353$, both below the critical value of 0.787 at $\alpha = 0.05$. The Q–Q plot further confirmed that the empirical residual quantiles closely followed the theoretical normal distribution, with only minor deviations at the distribution tails (Fig. 4).

Bland–Altman analysis was used to assess the agreement between predicted and observed PFA values and to identify systematic and

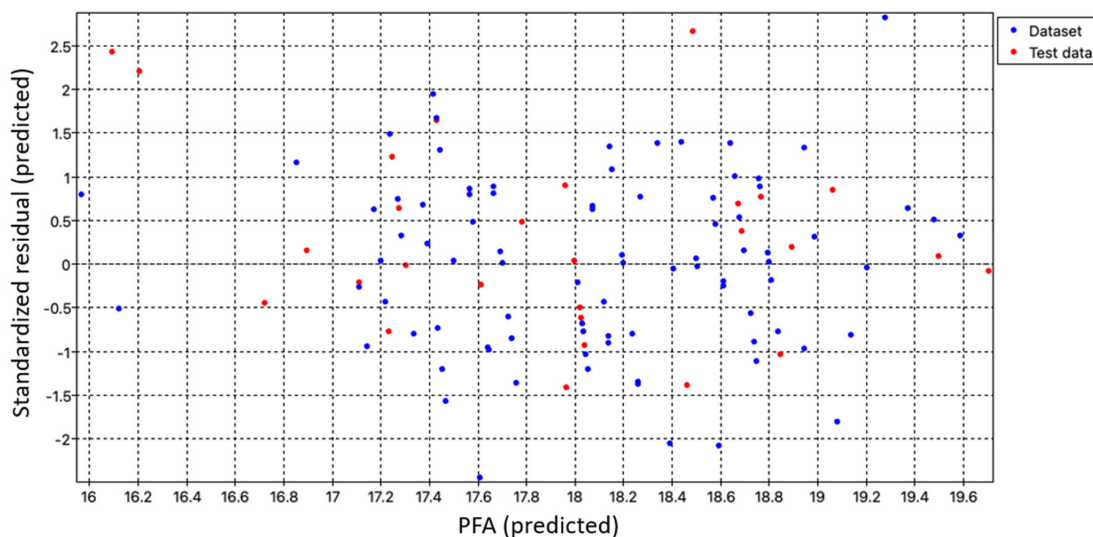


Fig. 3. Plot of standardised residuals against predicted piglets born alive per farrowing values.

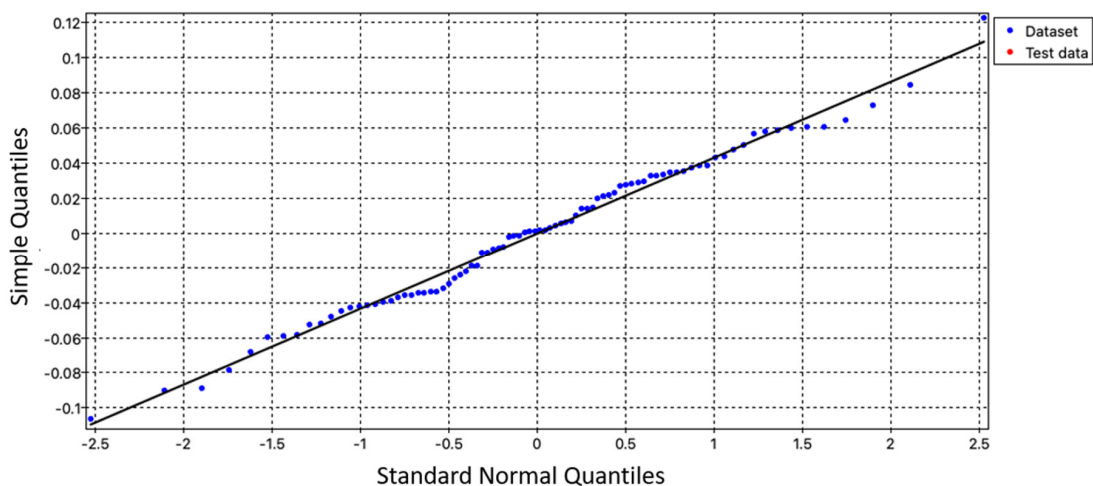


Fig. 4. Q-Q plot of residuals from the piglets born alive per farrowing model with reference line and normality test results.

proportional bias. The results are presented in Fig. 5. For the primary dataset, the regression of differences against mean values was described by the equation $\text{diff} = 0.033 - 0.002 \text{ mean}$, with slope = -0.0018 and $p = 0.813$. This indicates no statistically significant proportional bias. The mean difference was approximately 0.00, suggesting no relevant systematic bias. For the test dataset, the corresponding relationship was $\text{diff} = -0.234 + 0.012 \text{ mean}$, with slope = 0.0124 and $p = 0.213$. Therefore, no statistically significant proportional bias was detected in the test dataset. The mean difference was -0.01 , indicating a very

small average prediction bias. Most differences were within the 95% limits of agreement, confirming satisfactory agreement between the predicted and observed values.

DISCUSSION

The results indicate that the developed data-driven nonlinear multiple regression model achieved high predictive accuracy in estimating the number of piglets born alive per farrowing (PFA). The nonlinear relationships between PFA and the explanatory variables were identified

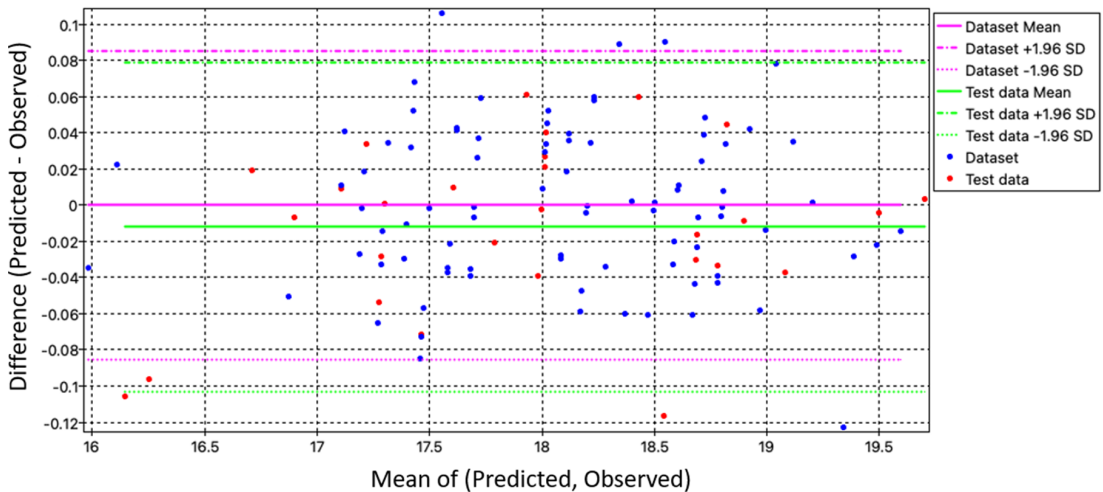


Fig. 5. Bland–Altman plot for agreement between predicted and observed piglets born alive per farrowing values.

through Monte Carlo-based optimisation, in which the power exponents were derived directly from the empirical data, while the regression coefficients were estimated using the least-squares method.

The final model combined the effects of sow productivity and piglet survival in a single nonlinear term, whereas farrowing frequency was represented by a quadratic component. This structure suggests that PFA is determined not only by individual predictors but also by their combined nonlinear contribution. The high coefficient of determination ($R^2 = 0.99632$), adjusted coefficient of determination (Adj. $R^2 = 0.99623$), very low prediction errors (RMSE = 0.04330; RRMSE = 0.23932%), and high concordance correlation coefficient (CCC = 0.99816) confirm both the accuracy of the model and the strong agreement between observed and predicted values. The model also retained good predictive stability on the test dataset. The test RMSE was 0.04725, and the test-to-dataset RMSE ratio was 1.09, corresponding to an approximately 9% increase in prediction error and providing no substantial overfitting. Given the biological nature of the data, agreement between observed and predicted values was further assessed using Bland–Altman analysis. This analysis showed no evidence of relevant systematic and proportional bias in either the primary or test datasets. Thus, the model did not systematically overestimate or underestimate PFA across the analyzed range, which is particularly important for data affected by natural biological variability. Overall, the findings indicate that the proposed data-driven

nonlinear regression approach provides an accurate, stable, and interpretable framework for modelling reproductive performance traits in sows.

Although the PFA model showed a very high goodness of fit, its interpretation requires caution because PPSY, F, and PSR% are biologically and partly mathematically related to PFA. PPSY is a composite productivity indicator and may indirectly capture information associated with the number of piglets born alive per farrowing, the annual number of farrowings, and piglet survival rate. Therefore, an additional Pearson correlation analysis was performed for the original variables, and the results are presented in Table 5.

The correlation matrix showed that none of the individual input variables was a direct linear substitute for PFA. The correlation between PPSY and PFA was moderate ($r = 0.585$), whereas the correlation between F and PFA was weak ($r = 0.183$), and the correlation between PSR% and PFA was negligible ($r = -0.038$). This indicates that the very high fit of the nonlinear PFA model cannot be attributed to a simple bivariate linear relationship between PFA and any single predictor.

Nevertheless, because PPSY, F, and PSR% are components of the same biological and production system, their combined information may partially reflect the numerical structure of PFA. Accordingly, the PFA model should be interpreted primarily as an empirical estimation equation describing relationships among key production indicators, rather than as a fully independent predictive model based on external

Table 5. Pearson correlation matrix for the original variables included in PFA model.

	PPSY	F	PSR%	PFA
PPSY	1	-	-	-
F	0.637	1	-	-
PSR%	0,698	0,241	1	-
PFA	0,585	0,183	-0,038	1

explanatory variables. Truly independent prediction of PFA would require the inclusion of predictors that are not directly embedded in the same productivity structure, such as genetic, nutritional, environmental, health-related, or management-related factors.

These findings are consistent with previous studies (Teleken et al., 2017; Besteiro et al., 2024), which demonstrated that nonlinear models can describe asymptotic processes of growth and reproduction more accurately than classical linear approaches.

Previous studies have shown that nonlinear models, such as Gompertz, von Bertalanffy, Richards, or Logistic models, can adequately reflect the S-shaped (sigmoid) growth dynamics of animals and allow the identification of inflection points at which growth rates are maximal (Afrouziyeh et al., 2021; Ceron et al., 2020). Consistent with these findings, our results show that polynomial transformations of the predictors provided sufficient flexibility while maintaining stable quality of data approximation by the developed model, capable of predicting complex nonlinear patterns in sow reproduction. Unlike models with a predefined inflection point, such as the Richards function, our models exhibited improved convergence and showed no evidence of the systematic errors reported in previous studies (Wang et al., 2023).

The results obtained are consistent with the findings from previous studies (Chen et al., 2025; Plaengkaeo et al., 2021), which have highlighted the limitations of the linear model in capturing complex relationships among productivity-related factors. In the present study, the use of nonlinear regressions enabled the model to capture not only direct associations but also synergistic and threshold effects that are often ignored by traditional linear models. These findings are further consistent with the conclusions of Archontoulis and Miguez (2015) and O'Brien and Silcox (2024), who emphasised the utility of nonlinear approaches for describing saturating and asymmetric dependencies in biological systems. Furthermore, compared to other approaches involving linear random-regression models (Škorput et al., 2023), our

approach focuses on population patterns rather than on individual ones. This may have a limitation in terms of the influence of inter-individual variation, but it ensures simplicity of interpretation and stability of the model at the herd level.

The models developed in the present study should primarily be regarded as tools for describing and predicting current herd-level reproductive trends based on available reproductive data. The absence of genetic, environmental, nutritional, and health-related factors is acknowledged as a limitation in the Discussion. Nevertheless, herd-level has considerable practical relevance to modern pig production, as it enables producers to evaluate overall herd productivity, optimise replacement and breeding strategies, and improve long-term production planning (Koketsu et al., 2017). Herd-level models are particularly valuable for detecting population trends and supporting management decisions under commercial production conditions, where individual variability may be less informative than the overall reproductive performance of the herd (Koketsu et al., 2017). However, it should be noted that the present prediction models were developed exclusively using contemporary reproductive parameters available in the dataset. Consequently, the obtained predictions mainly reflect current herd dynamics and do not account for additional biological or environmental factors such as genetics, nutrition, housing conditions, health status, or climate effects, which may also influence sow reproductive performance (Solà-Oriol and Gasa, 2017). Future studies should therefore incorporate a broader range of explanatory variables and independent validation datasets to improve the robustness and generalisability of the models. As noted by Bono et al. (2014), mixed nonlinear models that include random effects can provide more accurate estimations for large samples and repeated measurements. Therefore, further development of our approach may involve incorporating random effects or hierarchical levels into the model. Previous studies have shown the possibility of using machine learning methods for forecasting in animal production even under conditions of

incorrect data distribution (Matvieiev et al., 2025; Önder et al., 2025), which is often encountered in data processing.

Teleken et al. (2017) emphasised the importance of evaluating model fit criteria, such as the Akaike information criterion (AIC) and Bayesian information criterion (BIC), together with cross-validation, to reduce the risk of overfitting. In the present study, the high predictive accuracy observed across the evaluated datasets provide evidence of a good model performance for the available data. However, this should be further assessed using independent samples or bootstrap validation. This consideration is consistent with the recommendations of Colby and Bair (2013), who pointed out that conventional accuracy assessment methods may be insufficient for nonlinear models due to the structural dependence of the data.

Compared with previous studies employing machine-learning algorithms, including Support Vector Regression (Guan et al., 2022) and Gradient Boosting (Cepowski, 2021; Zhou et al., 2022), the predictive performance of the models developed in the present study was comparable to, and in some cases exceeded, that reported for these approaches. However, unlike the “black boxes” of artificial intelligence, the proposed models offer greater interpretability of the parameters.

The nonlinear interaction identified between sow productivity traits and piglet survival may reflect the biological trade-off between prolificacy and offspring viability. Modern hyperprolific sows are genetically selected for increased litter size; however, larger litters are frequently associated with prolonged farrowing duration, lower average birth weight, reduced colostrum intake, and increased pre-weaning mortality (Björkman and Grahofner, 2021; Koketsu et al., 2017). Therefore, the combined nonlinear contribution of productivity and survival parameters observed in the present model likely represents the physiological balance between reproductive intensity and piglet viability under commercial production conditions. Such relationships are biologically plausible because reproductive efficiency in sows is determined not only by the number of piglets born alive but also by the ability of piglets to survive during the early postnatal period.

The quadratic effect observed for farrowing frequency may indicate the existence of an optimal reproductive intensity threshold. Moderate increases in reproductive frequency can improve annual sow productivity, whereas excessive reproductive pressure may negatively affect body condition recovery, metabolic status, and subsequent fertility performance

(Koketsu et al., 2017). In commercial herds, intensive reproductive schedules may increase physiological stress and reduce the adaptive capacity of sows, especially under conditions of high litter size and repeated lactation cycles. Consequently, the nonlinear response identified by the model may reflect the biological limitation of reproductive compensation mechanisms in highly productive females.

In addition, the polynomial transformations derived through Monte Carlo optimisation should not be interpreted solely as mathematical adjustments. Rather, they may represent threshold and saturation effects that are typical for biological systems. Reproductive traits in pigs are rarely linearly associated across the entire physiological range because endocrine regulation, uterine capacity, maternal energy reserves, and piglet competition interact dynamically during gestation and lactation. Similar nonlinear biological dependencies have been reported previously in studies of sow prolificacy, piglet survival, and reproductive dynamics in commercial pig production systems (Nystén et al., 2023). Therefore, the proposed nonlinear approach may better reflect the actual biological mechanisms underlying sow reproductive performance than conventional linear models.

The validity of our approach is further confirmed by the consistency of our results with those of Thiengpimol et al. (2024), who reported that a third-degree polynomial model yielded the lowest prediction variance for reproductive performance in sows under tropical conditions. Similarly, in the present study, nonlinear optimisation of polynomial coefficients resulted in a stable model convergence and high predictive performance while maintaining the relatively simple equation structure.

It is also worth noting that our results corroborate the findings of Park et al. (2019) and Auger-Méthé et al. (2021), who emphasised the importance of modelling time lags and considering nonlinear dependencies in animal behaviour and reproductive cycles. The incorporation of nonlinear and threshold effects enables the models to better capture the underlying dynamics of reproduction, which is particularly important for herd management in intensive pig production systems.

Overall, the findings of this study support the conclusions of recent research indicating that nonlinear regression models can provide more accurate predictions of pig reproductive performance than conventional linear models (Powell et al., 2022; Yang et al., 2024; Archontoulis and Miguez 2015). Nonlinear regression models

can not only capture complex biological patterns but also provide biologically interpreted parameters, making them useful for practical applications. Nevertheless, as noted by Feng et al. (2021), a key area for future research is the development of hybrid approaches that combine the accuracy of nonlinear algorithms with the adaptability of modern machine-learning methods.

It should be acknowledged that there are certain limitations to the study related to the dataset and modelling assumptions. Seasonal effects and management differences were not explicitly included as covariates, which could affect overall reproductive capacity. In addition, the model focused on population-level trends rather than individual-level variability, and an independent dataset was not available for external validation. Although nonlinear regression provides flexible and accurate approximations, it can be sensitive to initial parameter settings and data distribution.

Future studies should incorporate multi-species datasets, independent samples for external validation, and potential gene-environment interactions to further assess and improve the robustness and predictive reliability of the proposed models.

CONCLUSIONS

From a mathematical and statistical perspective, the data-driven nonlinear power regression approach, combined with Monte Carlo-based exponent optimisation, proved effective for modelling the number of piglets born alive per farrowing (PFA). The final model achieved high predictive accuracy while maintaining a relatively simple and interpretable structure.

The model was statistically significant, demonstrated strong agreement between observed and predicted values, and satisfied the main diagnostic requirements, including the absence of relevant multicollinearity, an approximately normal distribution of residuals, limited heteroscedasticity, and no statistically significant proportional bias. Its stable performance on the test dataset provided no evidence of substantial overfitting and generalised well to independent data.

The combined nonlinear predictor based on sow productivity and piglet survival was the most influential component of the model, whereas farrowing frequency provided an additional statistically and practically significant contribution. This indicates that PFA is shaped not only by individual productivity parameters but also by their nonlinear combined effects.

Overall, the proposed approach may be considered a useful and robust modelling tool for zootechnical and biological data, particularly when nonlinear relationships between explanatory variables and the response variable are expected.

ACKNOWLEDGMENTS

The authors would like to thank the staff of the Ukrainian Pig Breeders Association in Denmark (Powered by Porcus) and the consulting firm Svine Rådgivningen for their assistance in completing the database.

Author contributions

Conceptualization, Mykhailo Matvieiev, Tomasz Cepowski. Methodology, Tomasz Cepowski and Zofia Cepowska. Formal analysis, Mykhailo Matvieiev and Zofia Cepowska. Investigation, Oleksandr Mykhalko. Resources, Mykola Povod, Igor Voshchenko and Oleksandr O. Borshch. Data curation, Mykola Povod and Tomasz Cepowski. Writing – original draft preparation, Tomasz Cepowski, Oleksandr Mykhalko and Igor Voshchenko. Writing – review and editing, Mykhailo Matvieiev, Andriy Getya and Vitalii Pakholiuk. Visualization, Tomasz Cepowski and Zofia Cepowska. Supervision, Mykola Povod and Mykhailo Matvieiev. Project administration, Mykhailo Matvieiev and Andriy Getya. All authors have read and agreed to the published version of the manuscript.

Declaration on the use of generative AI

No generative AI was used to create the text, tables or figures in this article.

LITERATURE CITED

- Afrouziyeh, M., R. P. Kwakkel, and M. J. Zuidhof. 2021. Improving a nonlinear Gompertz growth model using bird-specific random coefficients in two heritage chicken lines. *Poultry Science* 100(5): 101059. <https://doi.org/10.1016/j.psj.2021.101059>
- Archontoulis, S.V., and F.E. Miguez. 2015. Nonlinear Regression Models and Applications in Agricultural Research. *Agronomy Journal* 107: 786-798. <https://doi.org/10.2134/agronj2012.0506>
- Atamanyuk, I. P., V. Y. Kondratenko, A. S. Kramarenko, O. E. Novikov, and V. Y. Lykhach. 2020. Method of individual forecasting of sow reproductive performance on the basis of a non-linear canonical model of a random sequence. *Biosystems Diversity* 27(4): 309–313. <https://doi.org/10.15421/011940>

- Auger-Méthé, M., K. Newman, D. Cole, F. Empacher, R. Gryba, A. A. King, V. Leos-Barajas, J. Mills Flemming, A. Nielsen, G. Petris, and L. Thomas. 2021. A guide to state-space modeling of ecological time series. *Ecological Monographs*. <https://doi.org/10.1002/ecm.1470>
- Besteiro, R., T. Arango, M. R. Rodríguez, and M. D. Fernández. 2024. Linear and Nonlinear Mixed Models to Determine the Growth Curves of Weaned Piglets and the Effect of Sex on Growth. *Agriculture* 14(1): 79. <https://doi.org/10.3390/agriculture14010079>
- Björkman, S., and A. Grahofner. 2021. Tools and protocols for managing hyperprolific sows at parturition: Optimizing piglet survival and sows' reproductive health. *IntechOpen*. <https://doi.org/10.5772/intechopen.91337>
- Cepowski, T. 2021. An estimation of motor yacht light displacement based on design parameters using computational intelligence techniques. *Ocean Engineering* 231: 109086. <https://doi.org/10.1016/j.oceaneng.2021.109086>
- Cepowski, T. 2025. ndCurveMaster (8.7) [Computer software]. SigmaLab. <https://www.ndcurvemaster.com>
- Cepowski, T., and A. Drozd. 2023. Measurement-based relationships between container ship operating parameters and fuel consumption. *Applied Energy* 347: 121315. <https://doi.org/10.1016/j.apenergy.2023.121315>
- Ceron, M. S., V. de Oliveira, N. A. N. N. Pieve, N. C. Dias e Silva, C. A. R. Rossi, B. N. Fraga, H. da C. M. Muniz, and A. de M. Kessler. 2020. Nonlinear equations to determine the growth curve of immunocastrated pigs. *Pesquisa Agropecuária Brasileira* 55: e01184. <https://doi.org/10.1590/S1678-3921.pab2019.v55.01184>
- Chen, W., A. Zhao, J. Pan, K. Tan, Z. Zhu, L. Zhang, F. Yu, R. Liu, L. Zhong and J. Huang. 2025. Statistical analysis of reproductive traits in Jinwu Pig and identification of genome-wide association Loci. *Genes* 16(5): 550. <https://doi.org/10.3390/genes16050550>
- Feng, Y., Q. Liu, Q. Yao, and G. Zhao. 2021. Model averaging for nonlinear regression models. *Journal of Business and Economic Statistics* 40(2): 785–798. <https://doi.org/10.1080/0735015.2020.1870477>
- Guan, R., X. Zhou, H. Cai, X. Qian, X. Xin, and X. Li. 2022. Study on the influence of different production factors on PSY and its correlation. *Porcine Health Management* 8(1): <https://doi.org/10.1186/s40813-022-00253-2>
- Knol, E. F., B. Nielsen, and P. W. Knap. 2016. Genomic selection in commercial pig breeding. *Animal Frontiers* 6(1): 15–22. <https://doi.org/10.2527/af.2016-0003>
- Koketsu, Y., S. Tani, and R. Iida, R. 2017. Factors for improving reproductive performance of sows and herd productivity in commercial breeding herds. *Porcine Health Management* 3(1): 1. <https://doi.org/10.1186/s40813-016-0049-7>
- López, S., and J. France. 2022. A holistic approach to evaluating linear and non-linear mixed models to predict phosphorus retention in growing and finishing pigs. *Animals* 12(13): 1611. <https://doi.org/10.3390/ani12131611>
- Matvieiev, M., L. Bila, A. Ugnivenko, D. Nosevych, A. Getya, and T. L. Tyasi. 2025. Application of data mining algorithms for estimating live body weight from linear body measurements of Ukrainian beef cattle breed. *Applied Ecology and Environmental Research* 23(2): 1853–1864. https://doi.org/10.15666/aeer/2302_18531864
- Nayeri, S., M. Sargolzaei, and D. Tulpan. 2019. A review of traditional and machine learning methods applied to animal breeding. *Animal Health Research Reviews* 20(1): 31–46. <https://doi.org/10.1017/s1466252319000148>
- Nystén, M., Y. Jinhyeon, H. Shah, B. Stefan, V. Anna, S. Nicoline, F. Chantal and P. Olli. 2023. Sow nest-building behavior in communal farrowing relates to productivity and litter size. *Applied Animal Behaviour Science* 266: 106117. <https://doi.org/10.1016/j.applanim.2023.106117>
- O'Brien, T. E., and J. W. Silcox. 2024. Nonlinear regression modelling: A primer with applications and caveats. *Bulletin of mathematical biology* 86(4): 40. <https://doi.org/10.1007/s11538-024-01274-4>
- Önder, H., C. Tirink, T. Yakubets, A. Getya, M. Matvieiev, R. Kononenko, U. Şen, Ç. Ö. Özkan, T. Tolun, and F. Kaya .2025. Predicting live weight for female rabbits of meat crosses from body measurements using LightGBM, XGBoost and support vector machine algorithms. *Veterinary Medicine and Science* 11(1): e70149. <https://doi.org/10.1002/vms3.70149>
- Park, S. Y., C. Li, S. M. Mendoza Benavides, E. van Heugten, and A. M. Staicu. 2019. Conditional analysis for mixed covariates, with application to feed intake of lactating sows. *Journal of Probability and Statistics* 2019: 1–14. <https://doi.org/10.1155/2019/3743762>

- Plaengkaeo, S., M. Duangjinda, and K. J. Stalder. 2021. Identifying early indicator traits for sow longevity using a linear-threshold model in Thai Large White and Landrace females. *Animal Bioscience* 34(1): 20–25. <https://doi.org/10.5713/ajas.19.0855>
- Powell, C. D., J. L. Ellis, R. S. Dias, S. López, and J. France. 2022. A holistic approach to evaluating linear and non-linear mixed models to predict phosphorus retention in growing and finishing pigs. *Animals* 12(13): 1611. <https://doi.org/10.3390/ani12131611>
- Rao K, D.V.S.K., M. Premalatha, and C. Naveen. 2018. Analysis of different combinations of meteorological parameters in predicting the horizontal global solar radiation with ANN approach: A case study. *Renewable and Sustainable Energy Reviews* 91: 248–258. <https://doi.org/10.1016/j.rser.2018.03.096>
- Škorput, D., N. Jančo, D. Karolyi, A. Kaić, and Z. Luković. 2023. Analysis of early growth of piglets from hyperprolific sows using random regression coefficient. *Animals* 13(18): 2888. <https://doi.org/10.3390/ani13182888>
- Solà-Oriol, D., and J. Gasa. (2017). Feeding strategies in pig production: Sows and their piglets. *Animal Feed Science and Technology* 233: 34–52. <https://doi.org/10.1016/j.anifeedsci.2016.07.018>
- Teleken, J. T., A. C. Galvão, and W. D. S. Robazza. 2017. Comparing non-linear mathematical models to describe growth of different animals. *Acta Scientiarum. Animal Sciences* 39(1): 73–81. <https://doi.org/10.4025/actascianimsci.v39i1.31366>
- Thiengpimol, P., S. Koonawootrittriron, and T. Suwanasopee. 2024. Assessing reproductive performance and predictive models for litter size in Landrace sows under tropical conditions. *Animal Bioscience* 37(8):1333-1344. <https://doi.org/10.5713/ab.23.0406>
- Uncuoglu, E., H. Citakoglu, L. Latifoglu, S. Bayram, M. Laman, M. Ilkentapar, and A. A. Oner. 2022. Comparison of neural network, Gaussian regression, support vector machine, long short-term memory, multi-gene genetic programming, and M5 Trees methods for solving civil engineering problems. *Applied Soft Computing* 129: 109623
- Wang, L., H. Shi, Q. Hu, W. Gao, L. Wang, C. Lai, and S. Zhang. 2022. Modeling net energy partition patterns of growing-finishing pigs using non-linear regression and artificial neural networks. *Journal of Animal Science* 101 (2023): skac405, <https://doi.org/10.1093/jas/skac405>
- Xu, M., X. Lai, Y. Zhang, Z. Li, B. Ouyang, J. Shen, and S. Deng. 2024. An integrated hog supply forecasting framework incorporating the time-lagged piglet feature: Sustainable insights from the Hog Industry in China. *Sustainability* 16(19): 8398. <https://doi.org/10.3390/su1619839>
- Yang, H., L. Yang, J. Qian, L. Xu, L. Lin, and G. Su. 2024. Information of growth traits is helpful for genetic evaluation of litter size in pigs. *Animals* 14(18): 2669. <https://doi.org/10.3390/ani14182669>
- Yu, G., C. Wang, and Y. Wang. 2022. Genetic parameter analysis of reproductive traits in Large White pigs. *Animal Bioscience* 35(11): 1649–1655. <https://doi.org/10.5713/ab.22.0119>
- Zhou, X., R. Guan, H. Cai, P. Wang, Y. Yang, X. Wang, X. Li, and H. Song. 2022. Machine learning based personalized promotion strategy of piglets weaned per sow per year in large-scale pig farms. *Porcine Health Management* 8(37). <https://doi.org/10.1186/s40813-022-00280-z>