Field implementation of forecasting models for predicting nursery mortality in a Midwestern US swine production system

E. S. Magalhaes[1], D. Zhang[1,2], C. Wang[1,2], P. Thomas[3], C. A. A. Moura[3], D. J. Holtkamp[1], G. Trevisan[1], C. J. Rademacher[1], G. S. Silva[1] and D. C. L. Linhares[1,*]

[1]*Department of Veterinary Diagnostic and Production Animal Medicine, College of Veterinary Medicine, Iowa State University, Ames, IA 50011, USA*

[2]*Department of Statistics, College of Liberal Arts and Sciences, Iowa State University, Ames, IA 50011, USA*

[3]*Iowa Select Farms, Iowa Falls, IA 50126, USA*

[*]Corresponding author: Daniel Linhares, linhares@iastate.edu

Abstract

Swine nursery mortality is highly impacted by the performance of the piglets through the pre-weaning phase of production. Even though the importance of the pre-weaning phase on the downstream post-weaning performance is acknowledged, no predictive modelling is currently being utilized in the swine industry to predict the downstream nursery performance of individual groups of pigs based on their previous pre-weaning phase. One obstacle to building such predictive models is that health, management, and production data for the pre-weaning and post-weaning phases are collected with separate record-keeping programs and stored in unconnected databases. Thus, the objective of this study was to build a master table that automatically integrates dispersed data collected from one swine production system. After that, the performance of 5 forecasting models was investigated for predicting nursery mortality using the master table containing data on 3,242 groups of pigs (~ 13 million animals) and 44 variables, which concerned the pre-weaning phase of production and conditions at placement in growing sites. After training and testing each model's performance through cross-validation, the model with the best overall prediction results was the Support Vector Machine model in terms of Root Mean Squared Error (RMSE=0.406), Mean Absolute Error (MAE=0.284), and Coefficient of Determination (R^2=0.731). Subsequently, the forecasting performance of the SVM model was tested on a new dataset containing 72 new groups, simulating ongoing and near real-time forecasting analysis. Despite a decrease in R^2 values on the new dataset (R^2=0.672), the model demonstrated high accuracy (94.52%) in terms of prediction when the mortality of 72 groups was high (5≥%) or low (5<%) values. This study demonstrated the capability of forecasting models to predict the nursery mortality of commercial groups of pigs using pre-weaning information and stocking conditions variables collected post-placement in nursery sites.

Keywords: swine, mortality, data-wrangling, forecasting, machine-learning.

Introduction

The implementation of precision animal agriculture in livestock is often challenged by the abundance of diverse and large-scale data streams, which requires a multifaceted data-wrangling approach to investigate this complex livestock "big data" (Morota et al., 2018). The use of data management techniques and machine-learning models on this data can overcome its complexity for analytical purposes, such as forecasting. Although forecasting analysis in the livestock realm is acknowledged (Murphy et al., 2014; Nguyen et al., 2020; Zhang et al., 2016), this application has yet not been reported in the swine industry for mortality rate. Swine post-weaning mortality is a key performance indicator (KPIs) utilized to measure the sustainability of swine production system`s (Gebhardt et al., 2020), and is divided in nursery mortality and finisher mortality. Swine nursery mortality refer to the mortality of pigs in the first 5-8 weeks of the overall post-weaning phase (approximately 5,5 months), and accounts for large portion of the overall post-weaning mortality (USDA., 2015).

Information concerning the risk factors for swine mortality is routinely collected, such as health, environment, productivity, and infrastructure. However, difficulty in integrating and merging these data streams prevents its collective utilization for purposes such as forecasting or causal inference, which can be supported through the development of means for data integration and analysis under field conditions, as demonstrated in other risk factor studies (Agostini et al., 2015; Goumon and Faucitano, 2017; Passafaro et al., 2019; Magalhaes et al., 2022). Therefore, the objective of this study was to develop a data-wrangling pipeline within one swine production system to integrate and manage multiple data streams, enabling automated and near real-time data consolidation. Furthermore, the performance of multiple forecasting models was assessed on historical data, and the best model was tested on new data to predict the nursery mortality of prospective closeouts.

Materials and methods

Overview and study design

This study utilized field data from a large U.S. swine production system located in the Midwestern region. A total of 6 different and disconnected data streams related to 3,242 groups of marketed pigs (over 13 million animals), here referred as closeouts, slaughtered over 3 years were collected for the analyses. The retrospective performance of both the pre-weaning and post-weaning phases of production were imported and integrated into the respective closeouts` information, constructing a dataset (aka., master table) containing breeding-to-market historical information for each closeout. The pre-weaning phase variables available in this master table were utilized as predictors to forecast the downstream post-weaning mortality of each closeout on their initial 60 days in the post-weaning phase (nursery mortality).

Closeouts were defined as the groups of pigs that originated from the company`s breeding herds. The pigs remained in the breeding herd until weaning at approximately 21 days of age. Following weaning, pigs were placed on feed at growing sites where the groups remained for around 5,5 months. The groups were managed all-in-all-out meaning another group of pigs could not start until all of the pigs from the previous groups had been marketed. The mortality of each closeout during the nursery phase was defined as the outcome variable of analysis in this study, and was calculated as the following: (total # pigs at placement – total # pigs 60 days post placement) ÷ total # pigs at placement.

SAS® Version 9.4 (SAS Institute, Inc., Cary, NC) was utilized to build data-wrangling pipeline algorithms, thus, automating the processes of importing, managing, cleaning, and integrating the data streams. The integration of the 6 data streams resulted in a final master table for the 3,242 closeouts that was utilized for comparing the performance of 5 different regression and machine-learning models for forecasting swine nursery mortality. After this step, the model with the best forecasting performance was then utilized on a new dataset to validate the forecasting model on new data, simulating then ongoing near real-time forecasting.

Data-wrangling pipeline

The six different data streams available for the development of the master table were: (1) pre-weaning phase (i.e., breeding herd) productivity and health data; (2) post-weaning phase (i.e., growing phase) productivity data; (3) closeouts` health status reports; (4) pig transportation records; (5) stocking conditions reports; (6) management procedure records. The SAS algorithms developed in this study used a similar methodology described by Magalhaes et al, (2022), where the processes of matching and merging different data streams were conducted based on an identifier (time and location of events) and through the developments of PROC Statements algorithms (PROC MERGE, PROC SET, PROC SQL, PROC SORT, PROC UNIVARIATE, and PROC FREQ). The swine production system provided access to the aforementioned data, where a data workflow was developed using Microsoft Power Automate (Microsoft Corporation, Redmond, WA) and SAS to

automate the data-wrangling processes in this study. Once the master table was built, the dataset contained information for 3,242 closeouts of pigs, originating from 42 breeding herd sources and weaned into 529 different growing sites. The information from each of the 6 data streams was matched and merged to each respective closeout of pigs marketed in this study period (i.e., each closeout historical data from breeding-to-market).

Comparing forecasting models based on training data

The initial step after completing the master table was to select the breeding herd variables from the pre-weaning phase of production and parameters that represent the stocking conditions of the weaned groups into growing sites (i.e., characteristics at placement or day 0 in the post-weaning phase). Among all variables available in the master table, 44 parameters were utilized as predictors in the forecasting analyses (Table 1). The nursery mortality was log-transformed after verifying that its distribution was not normal, thus, utilizing the log-mortality as the response variable.

Table 1: Variables selected from the master table for the forecasting analyses.

Data streams	Type‡	Variables
(1) Breeding herd productivity & health data*	Rate	Service repeat rate; Abortion rate; Services per inventory; Proportion of gilts bred; Last week weaned sows bred rate; Proportion of sows pregnant at 105 days; Farrowing rate; Stillborn rate; Mummies rate; pre-weaning mortality; pre-natal losses; Sow death rate; Sow culls rate
	Count	Number of services; Number of farrows; Sows inventory
	Average	Wean-to-service interval; Total born; Born alive; Parity at the farrow; Gestation length; Interval between farrows; Pigs weaned/sow; Piglet wean age; Non-productive days; Productive sow days; Litter/female/year; mated inventory; pigs/weaned/female/year
	Category	Breeding herd porcine reproductive and respiratory syndrome (PRRS); Breeding herd *Mycoplasma hyopneumoniae (Mhp)* status
(2) Growing phase productivity†	Rate	Nursery mortality (mortality on the initial 60 days post placement in a growing site)
(3) Closeouts health status*	Category	PRRS status at placement; *Mhp* status at placement
(4) Pig transportation*	Time	Weaning movement year; Weaning movement week
	Count	Number of animals transported
(5) Stocking conditions*	Category	Type of flow; Type of ventilation; Breeding herd origins
	Count	Number of origins; Time to fill the site
(6) Management procedure*	Category	Type of PRRS vaccine; Type of piglet medication at weaning; Breeding herd type of mass medication protocol

†Outcome variable; *Variables utilized as predictors in the forecasting model. ‡Type of variables

To forecast the log-mortality, we investigated five models: multiple linear regression model (MLR), LASSO regression, support vector machine (SVM), neural network (NNet), and random forest (RF). The evaluation criteria for each forecasting model included Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2). Using the R package 'caret' (Kuhn, 2019), and specifically the 'train' function, the optimal parameters of LASSO regression, SVM, and NNet were selected based on the smallest RMSE by doing three repetitions of 5-fold cross-validation, and the optimal parameters of RF were selected based on the smallest out-of-bag (OOB) error. In order to evaluate the prediction performance of each forecasting model, a leave-one-out cross-validation was performed, where, for each record, the training set was the dataset excluding that record. The trained model was then used to predict the log-mortality of the excluded record. The best model was selected based on higher R^2 value and lowest RMSE and MAE values.

Performance of the selected model on independent validation data

After comparing the performance of the different forecasting models on the retrospective dataset of 3,242 groups, which refers to groups stocked into nursery sites between week 29 of 2019 through the week 5 of 2022 (i.e., marketed between January 2020 to August 2022), a new dataset containing 72 new closeouts weaned into nursery sites between weeks 6 and 12 of 2022 (i.e., marketed between August and September of 2022) was obtained through the data-wrangling pipeline. The forecasting model was then utilized on this naïve data to predict the nursery mortality of the groups, and the forecasting performance of the selected model was measured using the same metric of the same step (R^2, RMSE, and MAE). Also, the predicted vs. actual nursery mortality values were classified into relatively high (≥5%) or low (<5%) mortality groups as the company providing the data used the same classification as their target mortality values. The performance of the SVM model on accurately predicting closeouts with high or low mortality was assessed in terms of accuracy (Ac), sensitivity (Se), Specificity (Sp), positive predicted value (PPV), and negative predicted value (NPV), calculated based on the difference between the predicted vs. actual mortality of the 72 groups.

Results and discussion

Data-wrangling pipeline

When assessing data completeness for the 3,242, a total of 93 closeouts (2.87%) were excluded due to a lack of information for all the characteristics included in the master table, resulting in a final dataset composed of 3,149 closeouts and 44 explanatory variables to be used in the forecasting analyses. The algorithms developed for the data-wrangling pipeline allowed the integration of information previously stored independently and without use, now serving as the foundation for forecasting purposes. Also, assuming that the swine production system maintains the data format utilized in this study over time, the algorithms can be utilized to integrate and prepare new incoming information for prospective analyses, including forecasting and causal inference.

Table 2: Performance of the forecasting models on predicting nursery mortality.

Model[1]	Parameters[2]		
	R^2	*RMSE*	*MAE*
MLR	0.385	0.614	0.475
LASSO	0.392	0.611	0.471
RF	0.725	0.421	0. 313
SVM	0.731	0.406	0. 284
NNet	0.533	0.566	0.393

[1]MLR: Multiple Linear Regression; LASSO: LASSO regression; RF: Random Forest; SVM: Support Vector Machine; NNet: Neural Network.

[2]RMSE: Root Mean Square Error; MAE: Mean Absolute Error; R^2: r-square.

Comparing forecasting models

The overall performance for all forecasting models is reported in Table 2. Notably, the machine learning models performed better than the regression models, where RF and SVM models demonstrated the best overall prediction performance, similarly to other livestock-related studies comparing the performance of multiple forecasting models (Arulmozhi et al., 2021; Nguyen et al., 2020; Semakula et al., 2021). Furthermore, the SVM outperformed the other models measured in terms of R^2 (0.731), and lower errors measured by RMSE (0.406) and MAE (0.284).

Thereafter, the predicted values for each closeout using the SVM model were averaged by week for the data collected in this study (Figure 1), where it was observed that the SVM predicted values were underestimated compared to the actual nursery mortality values of the closeouts. Despite this fact, both the average weekly predicted and actual mortalities followed similar seasonal trends over time, which can be explained by the seasonal activity of major diseases impacting the swine industry (Trevisan et al., 2019).

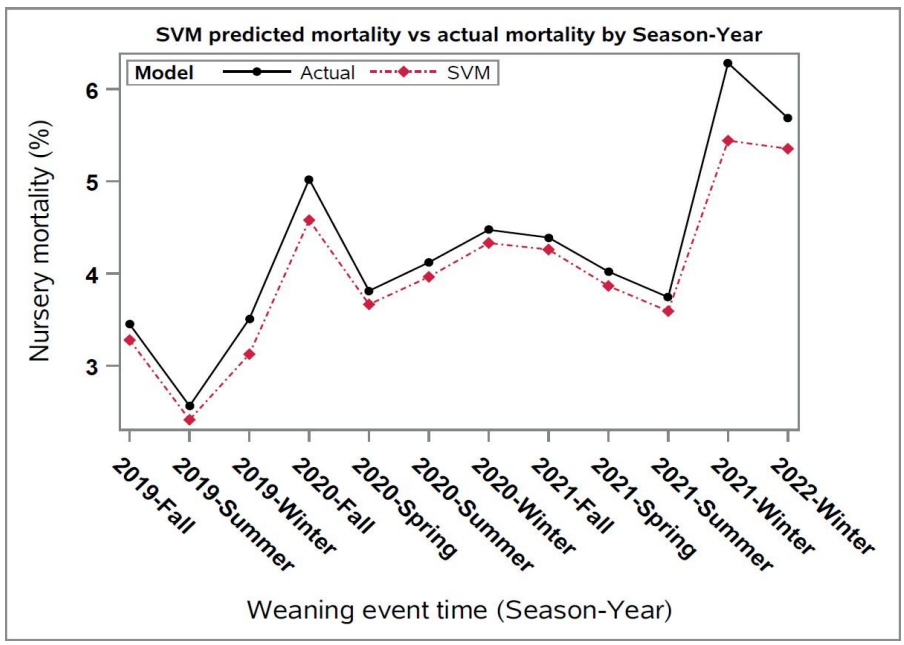

Figure 1: Average predicted versus actual nursery mortality over season-year for the Support Vector Machine (SVM) forecasting model versus the actual mortality.

Performance of the selected model

Identified as the superior model, SVM was prospectively applied to new data consisting of 72 closeouts, representing one month of closeouts, to predict the nursery mortality of the new groups. The overall forecasting performance of the SVM model were lower compared to the performance of the training data based on the cross-validation procedure (R^2 = 0.554 and 0.731, respectively). However, it is important to note that the training step was conducted in a much larger dataset, while the testing of the SVM model was conducted in a small dataset. On the other hand, we observed that most of the groups are located in the positive diagonal axis of the chart, which is the desired area in terms of prediction.

Despite the SVM decreased performance on naïve data, when categorizing both predicted and actual nursery mortality of the 72 closeouts into high (≥5%) or low (<5%), a high accuracy value (78.87%) was observed for the SVM on correctly predicting the closeouts as high or low mortality. Likewise, the values for sensitivity (67.57%), Specificity (91.43%), positive predicted value (89.29%), and negative predicted value (72.73%) demonstrated also an acceptable prediction performance, especially for precisely predicting groups with high

nursery mortality rates (i.e., at high risk). Overall, the SVM model accurately predicted 66.67% of the closeouts with relatively high nursery mortality, and 91.43% of the closeouts with relatively low mortality.

In other words, even though the SVM model did not predicted all groups that actually had high nursery mortality as high (false negatives), the model had a high positive predicted value, indicating that 89.29% of the closeouts predicted as high nursery mortality were observed as high nursery mortality.

Although there is an opportunity for improving the prediction of the exact values of nursery mortality (i.e., continuous outcome), there is a trade-off between prediction error and utility of the predicted value when using binary vs. continuous outcome. For example, more relevance was given by the production system in this study to be able to identify relatively high nursery mortality groups instead of predicting their exact mortality values.

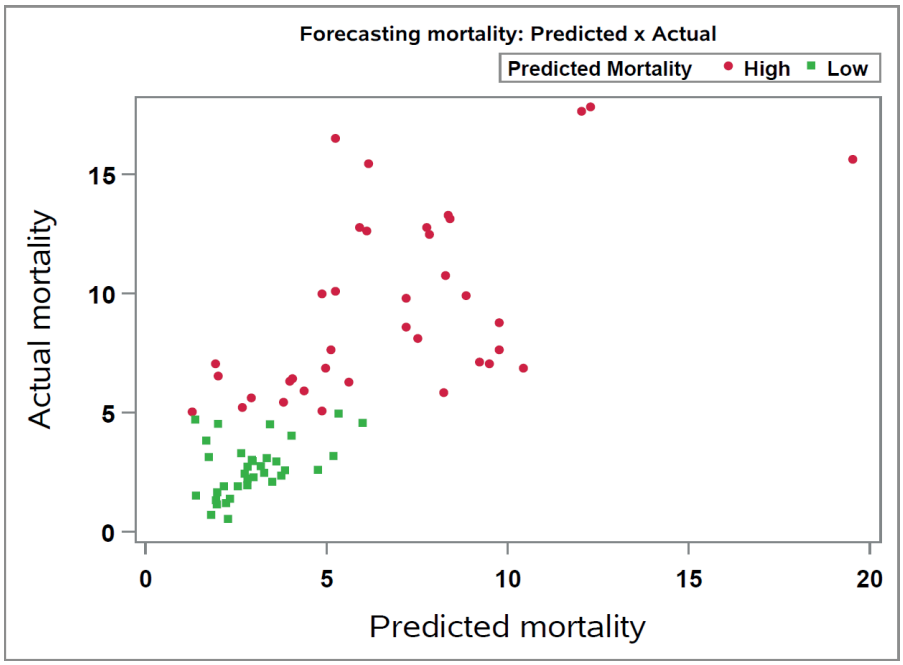

Figure 2: Correlation plot between the observed and predicted nursery mortality using the SVM model on 72 new closeouts. The red squares and green dots refer to groups with high (>5%) and low (<5%) predicted mortality, respectively.

The results of both the data-wrangling pipeline procedure and the forecasting models` comparison allowed training the best model on retrospective data and its further testing on new data, simulating the ongoing application of forecasting models on future data. In other words, utilizing the pre-weaning phase and stocking condition variables to predict the future mortality of closeouts. Also, the algorithms developed in this study can support swine practitioners in their decision-making process to strategically allocate resources (or not) for groups with predicted high nursery mortality. Notably, the predictive performance of the models refers specifically to the dataset collected in this study. The performance may change over time within this company as swine nursery mortality is impacted by multifactorial components that are dynamically interacting over time and period.

Conclusions

Forecasting swine nursery mortality can support decision-makers in allocating resources or interventions to towards precision swine health and productivity management. This study demonstrated the capability of building system-specific algorithms that allows the development of an automated data-wrangling pipeline,

which enables ongoing and near real-time forecasting. Also, this study demonstrated the ability to utilize breeding herd characteristics and data concerning the stocking conditions of weaned pigs placed in nursery sites as predictors for forecasting nursery mortality. Despite the overall acceptable performance for predicting groups at high mortality risk, there is an opportunity for improving the model`s performance by including more predictors and other machine-learning models.

Acknowledgments

This study was funded by the U.S. Department of Agriculture - National Institute of Food and Agriculture (USDA-NIFA) grant #022-68014-36668, and the C. R. Henderson Fund for Excellence in Predictive Inference and Its Applications.

References

Agostini, P.d.S., Manzanilla, E.G., De Blas, C., Fahey, A.G., Da Silva, C.A., and Gasa, J. (2015) Managing variability in decision making in swine growing-finishing units. *Irish Veterinary Journal* 68, 1-13.

Arulmozhi, E., Basak, J.K., Sihalath, T., Park, J., Kim, H.T., and Moon, B.E. (2021) Machine learning-based microclimate model for indoor air temperature and relative humidity prediction in a swine building. *Animals* 11(1), 222.

Gebhardt, J.T., Tokach, M.D., Dritz, S.S., DeRouchey, J.M., Woodworth, J.C., Goodband, R.D., and Henry, S.C. (2020) Postweaning mortality in commercial swine production. I: Review of non-infectious contributing factors. *Translational Animal Science* 4(2), 462-484.

Goumon, S., and Faucitano, L. (2017) Influence of loading handling and facilities on the subsequent response to pre-slaughter stress in pigs. *Livestock Science* 200, 6-13.

Kuhn, M. (2012) Contributions from wing j, weston s, williams a, keefer c, engelhardt a. *Caret: Classification and Regression Training*, 6-0.

Magalhães, E.S., Zimmerman, J.J., Thomas, P., Moura, C.A.A., Trevisan, G., Holtkamp, D.J., Wang, C., Rademacher, C., Silva, G.S., and Linhares, D.C.L. (2022) Whole-herd risk factors associated with wean-to-finish mortality under the conditions of a midwestern USA swine production system. *Preventive Veterinary Medicine* 198, 105545.

Morota, G., Ventura, R.V., Silva, F.F., Koyama, M., and Fernando, S.C. (2018) Big data analytics and precision animal agriculture symposium: Machine learning and data mining advance predictive big data analysis in precision animal agriculture. *Journal of Animal Science* 96(4), 1540-1550.

Murphy, M.D., O'Mahony, M.J., Shalloo, L., French, P., and Upton, J. (2014) Comparison of modelling techniques for milk-production forecasting. *Journal of Dairy Science* 97(6), 3352-3363.

Nguyen, Q.T., Fouchereau, R., Frenod, E., Gerard, C., and Sincholle, V. (2020) Comparison of forecast models of production of dairy cows combining animal and diet parameters. *Computers and Electronics in Agriculture* 170, 105258.

Passafaro, T.L., Van de Stroet, D., Bello, N.M., Williams, N.H., and Rosa, G.J.M. (2019) Generalized additive mixed model on the analysis of total transport losses of market-weight pigs. *Journal of Animal Science* 97(5), 2025-2034.

Semakula, J., Corner-Thomas, R.A., Morris, S.T., Blair, H.T., and Kenyon, P.R. (2021) Application of machine learning algorithms to predict body condition score from liveweight records of mature romney ewes. *Agriculture* 11(2), 162.

Swine, U. (2015) *Part I: Baseline reference of swine health and management in the united states, 2012*. USDA: Fort Collins, CO, USA.

Trevisan, G., Linhares, L.C.M., Crim, B., Dubey, P., Schwartz, K.J., Burrough, E.R., Main, R.G., Sundberg, P., Thurn, M., and Lages, P.T.F. (2019) Macroepidemiological aspects of porcine reproductive and respiratory syndrome virus detection by major united states veterinary diagnostic laboratories over time, age group, and specimen. *PLoS One* 14(10), e0223544.

Zhang, F., Murphy, M.D., Shalloo, L., Ruelle, E., and Upton, J. (2016) An automatic model configuration and optimization system for milk production forecasting. *Computers and Electronics in Agriculture* 128, 100-111.