Journal of Animal Science, 2020, Vol. 98, No. 4, 1–12

doi:10.1093/jas/skaa089
Advance Access publication March 23, 2020
Received: 26 December 2019 and **Accepted:** 19 March 2020
Housing and Management

HOUSING AND MANAGEMENT

Forecasting beef production and quality using large-scale integrated data from Brazil

Vera Cardoso Ferreira Aiken,[†] Arthur Francisco Araújo Fernandes,[†] Tiago Luciano Passafaro,[†] Juliano Sabella Acedo,[‡] Fábio Guerra Dias,[||] João Ricardo Rebouças Dórea,[$] and Guilherme Jordão de Magalhães Rosa[†,¶,1]

[†]Department of Animal Sciences, University of Wisconsin–Madison, Madison, WI 53706, [‡]DSM Produtos Nutricionais Brasil S.A., São Paulo, Brazil, [||]JBS S.A., São Paulo, Brazil, [$]Department of Dairy Science, University of Wisconsin–Madison, Madison, WI 53706, [¶]Department of Biostatistics and Medical Informatics, University of Wisconsin–Madison, Madison, WI 53706

[1]Corresponding author: grosa@wisc.edu

ORCiD number: 0000-0001-9172-6461 (G. J. M. Rosa).

Abstract

With agriculture rapidly becoming a data-driven field, it is imperative to extract useful information from large data collections to optimize the production systems. We compared the efficacy of regression (linear regression or generalized linear regression [**GLR**] for continuous or categorical outcomes, respectively), random forests (**RF**) and multilayer neural networks (**NN**) to predict beef carcass weight (**CW**), age when finished (**AS**), fat deposition (**FD**), and carcass quality (**CQ**). The data analyzed contained information on over 4 million beef cattle from 5,204 farms, corresponding to 4.3% of Brazil's national production between 2014 and 2016. Explanatory variables were integrated from different data sources and encompassed animal traits, participation in a technical advising program, nutritional products sold to farms, economic variables related to beef production, month when finished, soil fertility, and climate in the location in which animals were raised. The training set was composed of information collected in 2014 and 2015, while the testing set had information recorded in 2016. After parameter tuning for each algorithm, models were used to predict the testing set. The best model to predict CW and AS was RF (CW: predicted root mean square error = 0.65, R^2 = 0.61, and mean absolute error = 0.49; AS: accuracy = 28.7%, Cohen's kappa coefficient [**Kappa**] = 0.08). While the best approach for FD and CQ was GLR (accuracy = 45.7%, Kappa = 0.05, and accuracy = 58.7%, Kappa = 0.09, respectively). Across all models, there was a tendency for better performance with RF and regression and worse with NN. Animal category, nutritional plan, cattle sales price, participation in a technical advising program, and climate and soil in which animals were raised were deemed important for prediction of meat production and quality with regression and RF. The development of strategies for prediction of livestock production using real-world large-scale data will be core to projecting future trends and optimizing the allocation of resources at all levels of the production chain, rendering animal production more sustainable. Despite beef cattle production being a complex system, this analysis shows that by integrating different sources of data it is possible to forecast meat production and quality at the national level with moderate-high levels of accuracy.

Key words: beef, Brazil, forecasting, large scale data, machine learning

Abbreviations

AS	age when finished
CHTC	Center For High Throughput Computing
CQ	carcass quality
CW	carcass weight
FA	feedlot premix with additives
FD	fat deposition
FP	feedlot mineral premix
GLR	generalized linear regression
Kappa	Cohen's kappa coefficient
LR	linear regression
MAE	mean absolute error
mtry	number of explanatory variables included in the model at a time
NN	neural network
PNF	mineral premix for non-feedlot cattle
RF	random forest
$RMSE_p$	predicted root mean square error

Introduction

In the 21st century, agriculture will have to scale up to feed a human population projected to increase 30% by 2050 (FAO, 2009). Nonetheless, it cannot happen at the expense of animal welfare, or an increase in the environmental footprint, or even at a higher cost, which would limit market access for developing countries. Therefore, it is imperative to optimize the whole food chain to overcome such challenge. In this context, large data collections can be a valuable source of information to effectively address the current demand faced by agriculture (Kamilaris et al., 2017; Liakos et al., 2018 and Morota et al., 2018) with data analytics being central to support decision making at the farm level (Morota et al., 2018; Pham and Stack, 2018). The emerging fields of artificial intelligence and machine learning are core to data analytics and present new tools for predicting (i.e., forecasting) outcomes such as yield and quality with large-scale data. By projecting future trends, we can optimize the allocation of resources rendering the whole production chain more efficient and sustainable.

Beef production represents an important sector of animal agriculture as the third most produced meat in the world, after pork and poultry (FAO, 2014). The largest commercial cattle population in the world (213.5 million head of cattle) is located in Brazil (Oliveira, 2019). The country is also the largest exporter in the world with almost 20% of global beef exports (2.1 MM metric tons exported from the 9.9 MM produced) in 2018 (Zia et al., 2019). USDA predicts that Brazilian production will keep increasing in the near future, reaching 23% of the world's total exports by 2028 (Zia et al., 2019), remaining essential in the international agriculture market. Beef cattle production is developed in all Brazilian ecosystems (Oliveira, 2018), with the major producing states being Mato Grosso, Goiás, Minas Gerais, Mato Grosso do Sul, and Pará, which together are responsible for 54.2% of the national production (Oliveira, 2019). A visualization of the geopositioning of those states within Brazil is presented in Figure 1. Brazil has a mature beef cattle industry based on grass-fed cattle (Millen et al., 2011), and cattle spend most of their lives grazing in the pasture. The diversity in environments and conditions in which animals are raised combined with complex animal physiological mechanisms makes the prediction of meat production and quality at a national level a challenging task. In addition, historically the lack of consistent data collection and availability has made forecasting meat production and quality at a national level in Brazil a virtually impossible task.

This paper aimed to evaluate the feasibility and compare different tools for forecasting future beef cattle production and quality, using a large-scale data set integrated from different sectors of industry in Brazil. We compared the efficacy of traditional methods (linear regression [**LR**] or generalized linear regression [**GLR**]) and machine learning approaches (random forests[**RF**], and artificial neural networks [**NN**]) to forecast beef cattle production traits (carcass weight [**CW**], age when finished [**AS**], fat deposition [**FD**], and carcass quality [**CQ**]). Predictor variables included animal traits, farm participation in a technical advising program, nutritional products utilized by the farms, economic variables related to beef production, month when finished, and soil fertility and climate classification in the location in which animals were raised. The data analyzed contain information on over 4 million animals, corresponding to 4.3% of the Brazilian national beef production between 2014 and 2016.

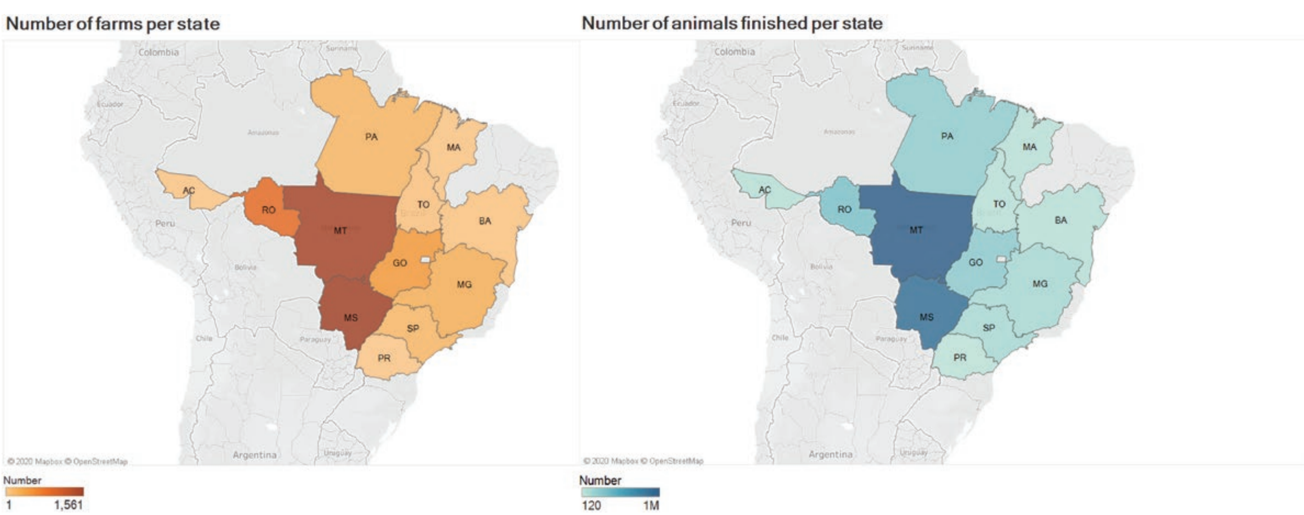

Figure 1. Distribution of farms (in the left) and finished animals (in the right) in the data set per state in Brazil. States are represented as Acre (AC), Bahia (BA), Goiás (GO), Mato Grosso (MT), Mato Grosso do Sul (MS), Maranhão (MA), Minas Gerais (MG), Pará (PA), Paraná (PR), Rondônia (RO), São Paulo (SP), and Tocantins (TO).

Materials and Methods

Data acquisition and integration

The data set utilized in this analysis was integrated from different sources to provide a comprehensive view of the Brazilian beef cattle production context. The animal information utilized was pre-collected by the sources involved, such that procedures involving the use of animals in this study did not have to be approved. The data contained information on 828,292 observations (group of animals) from 5,204 farms comprising a total of 4,022,394 finished beef cattle between 2014 and 2016. It is worth noting that the data analyzed in this study correspond to 4.3% of the Brazilian cattle national production of 94.2 million head of cattle for the 2014 to 2016 period (IBGE, 2018).

The data set contained information on 645 municipalities located in 12 of the 26 Brazilian states (Acre, Bahia, Goiás, Mato Grosso, Mato Grosso do Sul, Maranhão, Minas Gerais, Pará, Paraná, Rondônia, São Paulo, and Tocantins). The distribution of farms and number of finished animals per state is presented in Figure 1.

The integrated data collection included five major sources of data encompassing animal traits, utilization of technology and nutritional products at the farm, economic variables related to beef production, soil fertility, and climate where animals were raised. Each data source and the data integration steps are described in detail below.

The animal and nutrition/technology data sets used for this study were kindly provided by a meatpacking company (JBS S.A., Brazil) and an animal nutrition company (DSM Nutritional products Brazil S.A.), respectively. A farm matching integration procedure, described in detail by Aiken et al. (2019), was implemented to identify which farms in both databases were the same and to connect the information. In this study, results provided by the two best approaches for farm matching highlighted by Aiken et al. (2019), that is, bagged clustering and support vector machines, were overlapped. When the two algorithms disagreed on a matching status, discrepancies were solved by expert clerical review to generate this data set containing 5,204 matched farms.

The animal traits obtained from the meatpacking plant were: CW—measured in kg; AS—obtained by carcass dental evaluation and divided into five categories (up to 20 mo; 20 to 24 mo; 24 to 36 mo; 36 to 48 mo; and above 48 mo old); FD—obtained by visual evaluation of carcass fat coverage, and divided into five categories (absent—lower than 1 mm; low—1 to 3mm; medium—3 to 6mm; high—6 to 10mm; and excessive—above 10mm); CQ—which takes into account CW, AS, FD, gender of the animal, as well as body condition score, and is divided into three major categories (undesirable, acceptable, desirable); and animal category—defined as female, steer, and bull. A distribution of the animal categorical traits is presented in Figure 2. For the continuous trait CW, the average was 252 kg with a standard deviation of 61.6 kg.

The information obtained from the nutritional company comprised two major parts. The first one contained information on farm participation in a technical advising program for improving results, as a binary variable. The second part had information on the amount of nutritional products utilized by farms where animals were raised in the year of slaughter. The amount of product used by farms was divided into three major categories: mineral premix for non-feedlot cattle—**PNF**, feedlot mineral premix—**FP**, and feedlot premix with additives—**FA** (all measured in kilograms). More specifically, PNF contained mainly minerals, while FP included feedlot concentrate and FA had concentrate with the additives of the following classes: essential oils, enzymes, ionophores, buffers, probiotics, and/or yeast. For the three nutritional products previously mentioned, the total amount used (kilograms—kg) was adjusted by the quantity of animals finished per farm in that year, generating a per animal value. Regarding the technical advising program, 921 farms participated in it, 3,835 did not, and the remaining farms had missing information for this variable. The average FP per animal per year was 1.4 kg (SD = 13.8 kg), while PNF was 61.8 kg (SD = 162.4 kg), and FA was 3.1 kg (SD = 28.2 kg).

The information on economic variables related to beef production was extracted from the Agrolink public database (Agrolink, 2019). Two variables were included in this analysis: the finished cattle sales price at the state the farm was located, for the month and year each animal was harvested, and the price for the corn at the state the farm was located 3 mo before the harvesting date. For example, if an animal was harvested in December, the sales price of corn in September for the same state was utilized. All prices were in the Brazilian currency (R$, Reais). The defined time window of 3 mo approximates the average time in Brazil (83 d) that beef animals are finished on feedlots before slaughter (Millen et al., 2011). The average sales price and corn price per state per month are shown in Figure 3.

The soil fertility classification at the farm in which animals were raised was accessed utilizing the interactive geographic mapping platform, available from the Brazilian Institute of Geography and Statistics (Instituto Brasileiro de Geografia e Estatística—IBGE, 2019). The national digital atlas of Brazil for agricultural potential of soils in terms of fertility and characteristics was overlaid to the geopositioning (latitude and longitude) of the farm to determine the soil type where animals were raised. From the 10 soil classifications defined at the atlas, 9 were present in the data set: light green, cream, orange, yellow, purple, dark green, pink, light blue, and gray. A full description of the classification for those soils is presented in Table 1. The number of observations per soil type was as follows: light green = 59,686, cream = 27,206, orange = 527,431, yellow = 370, purple = 10,269, dark green = 9,795, pink = 13,154, light blue = 12,790, and gray = 167,591.

Lastly, we considered the climate at the municipality where cattle were raised. The Köppen's climate classification for Brazil (Álvares et al., 2014) was chosen as it is considered the most widely used classification method across geographical and climatologic societies in the world. This classification utilized historical information on monthly temperature and rainfall to produce a climate map with a high spatial resolution that allows the detection of climatic variations at the landscape level. From the 12 climate classifications identified in Brazil (Álvares et al., 2014), 9 were present at the locations in this data set: Af—tropical zone, without dry season; Am—tropical zone, monsoon; Aw—tropical zone, with dry winter; As—tropical zone, with dry summer; BSh—dry zone, semi-arid, low latitude and longitude; Cfa—humid subtropical zone, oceanic climate without dry season, with hot summer; Cfb—humid subtropical zone, oceanic climate without dry season, with temperate summer; Cwa—humid subtropical zone, with dry winter, and hot summer; and Cwb—humid subtropical zone, with dry winter, and temperate summer. The number of observations for each climate was: Af—29,551; Am—392,180; As—659; Aw—362,603; BSh—417; Cfa—35,349; Cfb—253; Cwa—5,899, and Cwb—1,381.

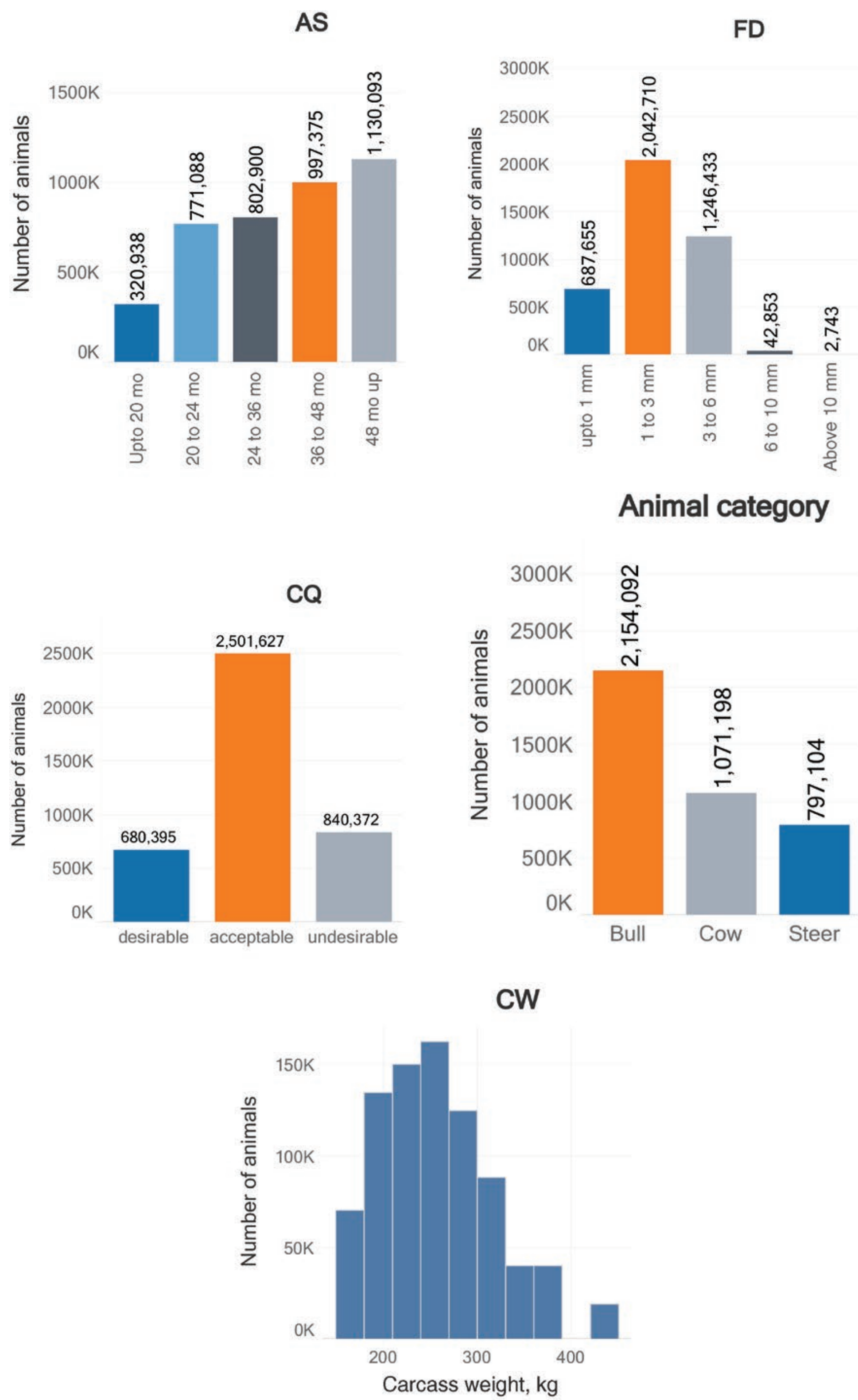

Figure 2. Distribution of animals finished according to AS, FD, CQ, animal category, and CW.

Data analysis

After comprehensive data integration of the previously mentioned sources, the data were preprocessed for the prediction of CW, AS, FD, and CQ. From all variables considered in the models, only the binary variable for participation in the technical advising program had missing information. More specifically, 27% of the farms had missing information (in at least 1 yr). This variable was imputed using bagged trees with

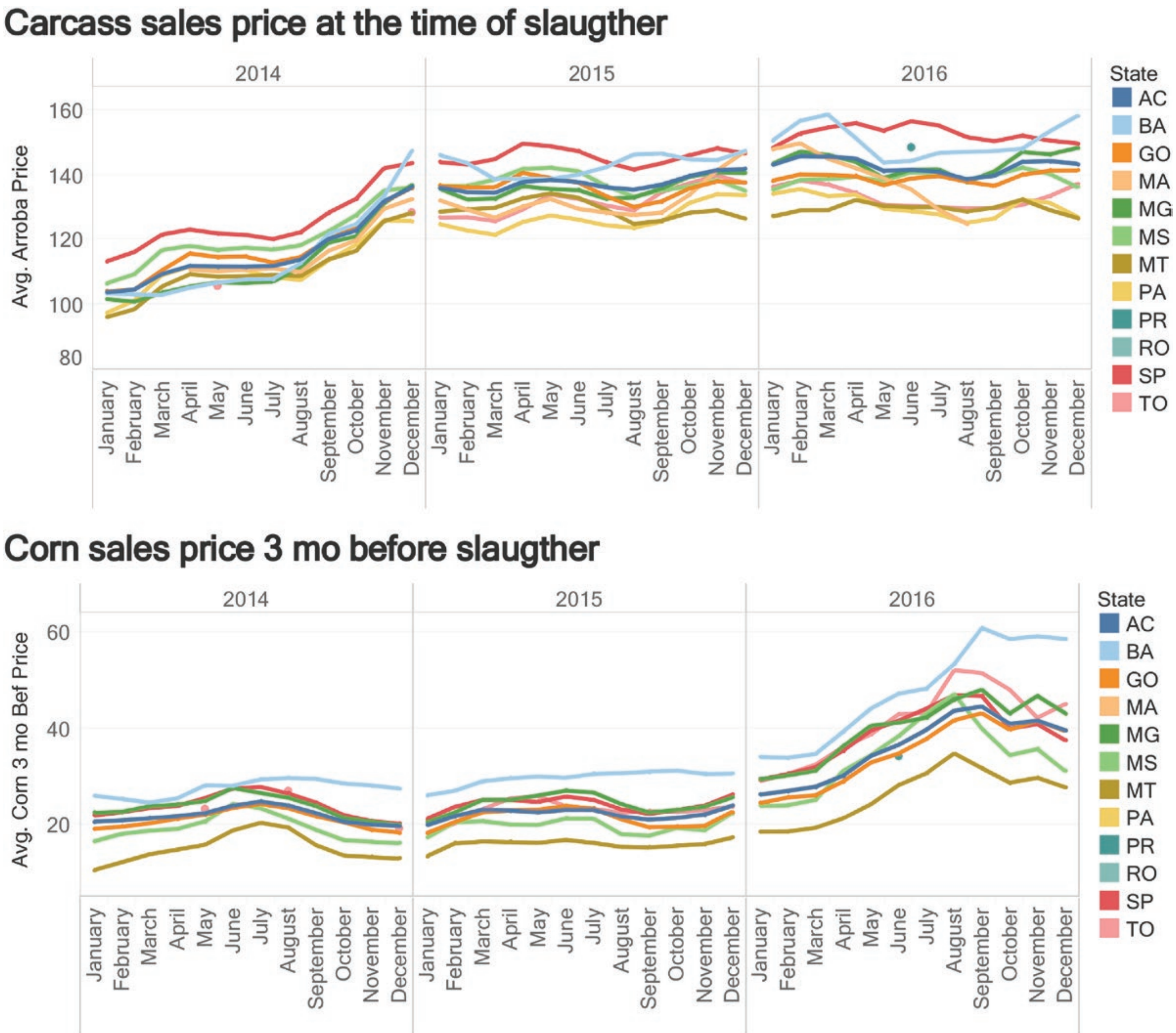

Figure 3. Average cattle sales price per state (top) in Brazilian currency (R$, Reais) and corn sales price per state per month (bottom). Points represent states that contained information on single months. Source: adapted from Agrolink.

Table 1. Classification of soil agricultural potential of Directory of Geosciences, Coordination of natural resources and environmental studies (IBGE, 2019) for the nine soil types in this data

Soil class	Fertility	Attributes	Relief	Major limitations
Light green	High	Good	Flat and slightly undulating	No major limitations
Cream	Mean	Good	Flat and slightly undulating	Medium to low availability of nutrients
Orange	Low	Good	Flat and slightly undulating	Low availability of nutrients, excess of aluminum
Yellow	Low	Regular	Flat and slightly undulating	Low availability of nutrients
Purple	Mean-high	Regular	Flat to undulating	Steep slopes, shallow depth, rough texture
Dark green	Mean-high	Good	Highly undulating	Steep slopes
Pink	Low	Regular	Undulating to mountainous	Steep slopes, restricted drainage, aluminum excess
Light blue	Low	Regular	Flat and slightly undulating	Sodium excess, restricted drainage, flooding risk
Gray	Not recommended to the agricultural activity			

the R package “missForest” (Stekhoven, 2013). Bagged trees were created using all other variables in the training set, such that when a sample had a missing value for a predictor, the bagged model was used to predict this value. The estimated error of the imputation (out-of-bag proportion of falsely classified samples) was very low (0.0074%). The data set presented no major issues with near-zero variance and there was no high correlation among predictors (all correlations below 0.5). All continuous variables were centered and scaled (mean = 0; SD = 1) prior to the analysis.

Three algorithms were contrasted for forecasting CW, AS, FD, and CQ. Those algorithms were: 1) LR for the continuous trait CW or GLR for ordered categorical traits AS, FD, and CQ, 2) RF, and 3) multilayer perceptron NN. The choice of algorithms to analyze the data covers different types of algorithms from methods more traditionally used in animal sciences, such as regression, along with modern machine learning methods, which have been successfully used for the task of forecasting in other fields (Biau and Scornet, 2016; Kuhn and Johnson, 2016). Also, the specific machine learning algorithms were chosen to explore the strengths of the methods for the prediction task. For example, RF is known to be robust to noise in the predictor variables (Biau and Scornet, 2016; Kuhn and Johnson, 2016), which is a likely occurrence with data collected in nonexperimental settings such as the farm data analyzed here. On the other hand, NN is known for its ability to properly model complex nonlinear relationships (Kuhn and Johnson, 2016). The data set utilized in this study aims to use complex relationships between environmental and physiological variables for the task of prediction of meat production and quality. For this reason, NN could be a good algorithm to model such complex relationships. All methods were implemented in the R environment using the "caret" package (Kuhn, 2019). All analyses were performed utilizing the capabilities of the Center for High Throughput Computing (**CHTC**) at the University of Wisconsin, Madison.

The explanatory variables used to predict each outcome (CW, AS, FD, and CQ) with regression, RF, and NN are detailed in Table 2. The training sets had 542,935 observations obtained in 2014 and 2015, while the remaining 285,357 observations from 2016 were used as an independent testing set. A 10-fold cross-validation scheme within the training set was implemented in which for each cross-validation run, the model was trained in 9-folds and the 10th fold was used to validate tuning parameters (for models that required parameter tuning). The model that produced the best results across the 10-fold cross validation was selected and further utilized in the testing set. For GLR, different link functions were tested (logistic, probit, cloglog, loglog, and cauchit) and the one that provided the highest accuracy was chosen. For RF, the only parameter requiring tuning was the number of explanatory variables included in the model at a time (mtry), which was done using exhaustive search (testing 1 to all available explanatory variables) in each model. For NN, three parameters were tuned with a grid search: the number of hidden layers (1 to 3), the number of units per hidden layer (1, 5, 10, 50, or 100), and the rate of weight decay utilized in the training backpropagation procedure (0, 0.0001, or 0.1). A grid search was chosen as a reasonable tuning method due to major constrains of run time and memory related to large size of the data set analyzed and the complexity of analyses. The activation function utilized at each hidden layer was the logistic (i.e., sigmoidal) function. The maximum number of iterations allowed to train the model was 100 (with early stopping criteria), and one hot encoding was applied to all categorical explanatory variables.

Table 2. Models for forecasting CW, AS, FD, and CQ[1]

Outcome	Predictors
CW	AS; animal category; PTAP; PNF, FP, FA; FCSP; CP3B; SOIL; CLIM; and MO
AS	animal category; PTAP; PNF, FP, FA; FCSP; CP3B; SOIL; CLIM; and MO
FD	AS; animal category; PTAP; PNF, FP, FA; FCSP; CP3B; SOIL; CLIM; and MO
CQ	animal category; PTAP; PNF, FP, FA; FCSP; CP3B; SOIL; CLIM; and MO

[1]Explanatory variables to models included: animal category, participation in a technical advising program (PTAP); kg of PNF per beef animal; kg of FP per beef animal; kg of FA products per beef animal; finished cattle sales price (FCSP); corn price 3 mo before finished (CP3B); soil fertility classification (SOIL); climate classification (CLIM); and month when finished (MO).

The predictive ability of each model (regression, RF, and NN) was calculated for each outcome variable. For continuous traits (CW), the predictive ability was assessed in terms of predicted root mean square error (**RMSEp**) of the testing set, coefficient of determination (R^2), and mean absolute error (**MAE**). For categorical traits (AS, FD, and CQ), the predictive ability was evaluated in terms of accuracy and the Cohen's kappa coefficient (**Kappa**). For all outcomes, the respective predictive ability metric is presented with the standard deviation of the resamples.

Among all methodologies tested, two produce a simple and intuitive metric for variable importance: regression and RF. For regression, variable importance was assessed using the absolute value of the *t*-statistic for each explanatory variable used in the best training set. For RF, variable importance was determined by recording the prediction accuracy of the out-of-bag sample when each tree was formed. This was repeated after permuting each of the explanatory variables. The difference between the two accuracies was then averaged across all trees and normalized by the standard error. All measures of importance were scaled to have a maximum value of 100.

Results

Results for the choice of GLR link function for the categorical variables AS, FD, and CQ are presented in Table 3. More specifically, 10-fold cross-validation results (out-of-bag accuracy average and SD) for the training set using different link functions are presented. For each categorical variable, the link function that provided the highest accuracy was chosen and later fitted to the independent test set. The link function that provided the highest accuracy for AS (30.6%) was cloglog, for FD was loglog (45.0%), and for CQ was the cauchit (57.7%). However, the choice of link function seemed largely unimportant.

Results for the parameter tuning of RF using exhaustive search (with number of variables included in the model at a time from 1 to all explanatory variables) are presented in Figure 4. The best results for the 10-fold cross validation performed in the training set, in terms of maximum accuracy for categorical variables and minimum RMSEp for continuous variables were chosen. For the variables CW, AS, FD, and CQ, the best results were with the number of variables included in the model equal to 11, 10, 5, and 9, respectively.

NN results for parameter tuning of number of layers (1 to 3), number of nodes per layer (1, 5, 10, 50, and 100), and rate of decay (0, 0.0001, and 0.1) for the four explanatory variables CW, AS, FD, and CQ are presented in Supplementary Figures S1–S4. Best results for CW (Supplementary Figure S1) were: layer = 3, nodes per layer = 100 in the first, 100 in the second and 100 in the third, and decay = 0. For AS (Supplementary Figure S2) were: layer = 3, nodes per layer = 100 in the first, 100 in the second and 100 in the third, and decay = 0. For FD (Supplementary Figure S3)

Table 3. Accuracy results from 10-fold cross validation for the training set of GLR[1]

	Generalized linear model link function				
Variable	Logistic	Probit	Cloglog	Loglog	Cauchit
AS	0.2924 (±0.0019)	0.2906 (±0.0017)	**0.3056 (±0.0011)**	0.2704 (±0.0011)	0.2991 (±0.0016)
FD	0.4477 (±0.0013)	0.4479 (±0.0013)	0.4301 (±0.0017)	**0.4500 (±0.0022)**	0.4478 (±0.0017)
CQ	0.5768 (±0.0017)	0.5761 (±0.0019)	0.5617 (±0.0012)	0.5735 (±0.0018)	**0.5772 (±0.0037)**

[1]Results are presented as the average accuracy (converted to original scale) across the 10 out-of-bag folds, followed by the ±SD (in parenthesis) for the three categorical variables: AS, carcass FD, and CQ. The highest accuracy across different link functions is highlighted (bold values) for each trait.

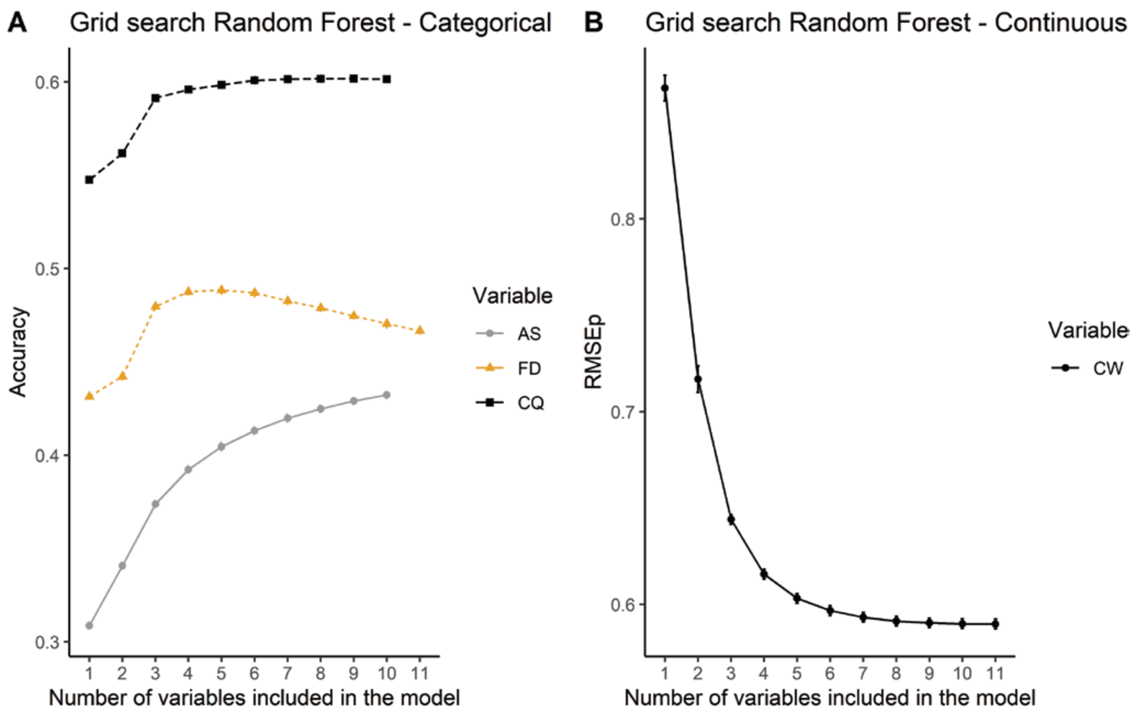

Figure 4. Results for exhaustive grid search, performed with 10-fold cross validation to test for different numbers of explanatory variables included in the RF model at a time. Mean predictive accuracy and SD (vertical line for each point) across the 10 folds are presented for the categorical variables: AS, FD, and CQ in panel A. Mean RMSEp and SD (horizontal line for each point) across the 10 folds are presented for the continuous variable CW in panel B.

were: layer = 3, nodes per layer = 50 in the first, 50 in the second and 50 in the third, and decay = 0. Lastly for CQ (Supplementary Figure S4) were: layer = 3, nodes per layer = 100 in the first, 100 in the second and 100 in the third, and decay = 0.

After parameter tuning of all models using 10-fold cross validation in the training set, the best results in terms of maximum accuracy for categorical variables and minimum RMSEp for continuous variables were chosen to be fitted to the independent test set for each outcome variable. Results for testing set predictive ability in terms of RMSEp, R^2, and MAE, for continuous variables, and accuracy and Kappa for categorical variables are presented in Table 4. The testing set predictive ability measure and respective standard deviation (estimated with 10-fold cross-validation procedure performed in the training set) are presented for CW, AS, FD, and CQ.

The best model for CW was RF (RMSEp = 0.66, R^2 = 0.59, and MAE = 0.50), with results very similar to regression (RMSEp = 0.67, R^2 = 0.60, and MAE = 0.51). The best model for AS was also RF (accuracy = 28.7%, Kappa = 0.08) with very similar performance presented by regression (accuracy = 28.7%, Kappa = 0.07) and slightly lower performance presented by NN (accuracy = 25.3%, Kappa = 0.02). The same pattern was observed for FD (RF: accuracy = 45.7%, Kappa = 0.05 and regression: accuracy = 44.9%, Kappa = 0.05); however, the lower performance of NN (Accuracy = 37.4%, Kappa = 0.05) was more pronounced for this variable. Regarding CQ, the best predictions were obtained using regression (accuracy = 58.7%, Kappa = 0.09), followed by RF (accuracy = 53.9%, Kappa = 0.09) and NN (accuracy = 46.4%, Kappa = 0.07), with a considerable drop in performance of the two models, compared with regression.

Variable importance results (for regression and RF) for CW, AS, FD, and CQ are presented in Figures 5–8, respectively. For CW (Figure 5), sales price for cattle and corn, as well as technical consulting, were important for both RF and regression. However, regression assigned heavier weights to animal category while RF deemed the nutrition given to the animal, month, climate, and soil as important variables. For AS (Figure 6), variable importance was consistent between regression and RF. The most important predictors were animal category, animal nutrition, cattle sales price, and use of technical consulting. Corn price, climate, and soil at the location animals were raised had smaller importance. For FD (Figure 7), results were somewhat consistent

Table 4. Models predictive ability for CW, AS, FD, and quality (CQ)[1]

		Outcome variable		
		Categorical		
Model	Measure	AS	FD	CQ
GLR	Accuracy	0.2867 (±0.0011)	**0.4576 (±0.0022)**	**0.5867 (±0.0019)**
	Kappa	0.0666 (±0.0015)	**0.0476 (±0.0037)**	**0.0862 (±0.0037)**
RF	Accuracy	**0.2871 (±0.0019)**	0.4494 (±0.0020)	0.5390 (±0.0016)
	Kappa	**0.0759 (±0.0026)**	0.0523 (±0.0032)	0.0930 (±0.0032)
Multilayer perceptron NN	Accuracy	0.2536 (±0.0028)	0.3742 (±0.0019)	0.4640 (±0.1999)
	Kappa	0.0237 (±0.0034)	0.0501 (±0.0160)	0.0670 (±0.0017)
		Continuous		
		CW (centered and scaled)	**CW (original scale)**	
LR	RMSEp	0.6765 (±0.0027)	41.2697 kg	
	R^2	0.6017 (±0.0017)	0.6017	
	MAE	0.5097 (±0.0017)	31.0941 kg	
RF	RMSEp	**0.6626 (±0.0025)**	**40.4217 kg**	
	R^2	**0.5920 (±0.0024)**	**0.5920**	
	MAE	**0.5018 (±0.0013)**	**30.6122 kg**	
Multilayer perceptron NN	RMSEp	0.8073 (±0.0030)	49.2491 kg	
	R^2	0.4657 (±0.0037)	0.4657	
	MAE	0.5905 (±0.0045)	36.0233 kg	

[1]For continuous traits (CW), testing set predictive ability was measured in terms of RMSEp, R^2, and MAE. For categorical traits (AS, FD, and CQ), it was assessed in terms of accuracy and Kappa. The testing set predictive ability is presented along with ±SD (in parenthesis) obtained in the training set 10-fold cross validation. Best results for each model are in bold face.

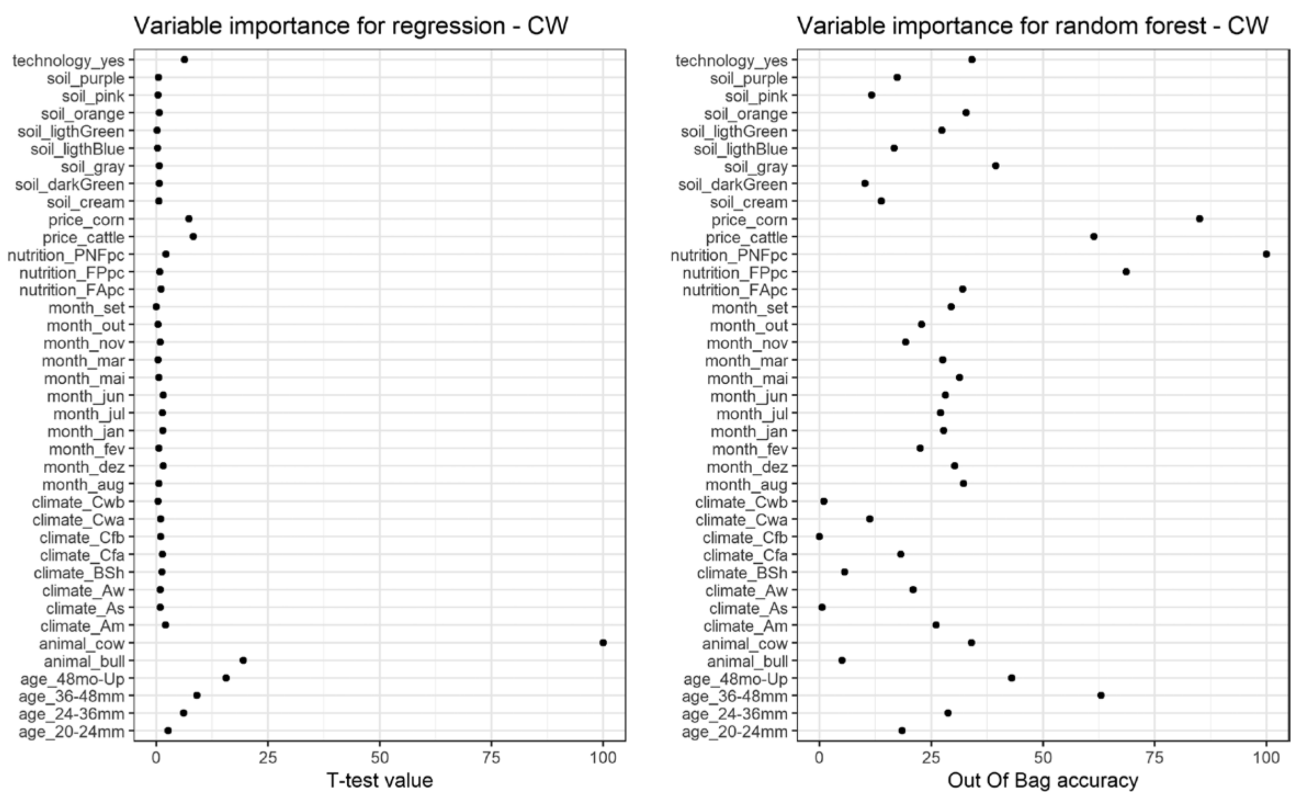

Figure 5. Variable importance results for the prediction of CW with regression and RF. For regression, variable importance was assessed using the *T*-test value of the regression fitted to the test set while for RF it was estimated as the out-of-bag accuracy of permuting each explanatory variable.

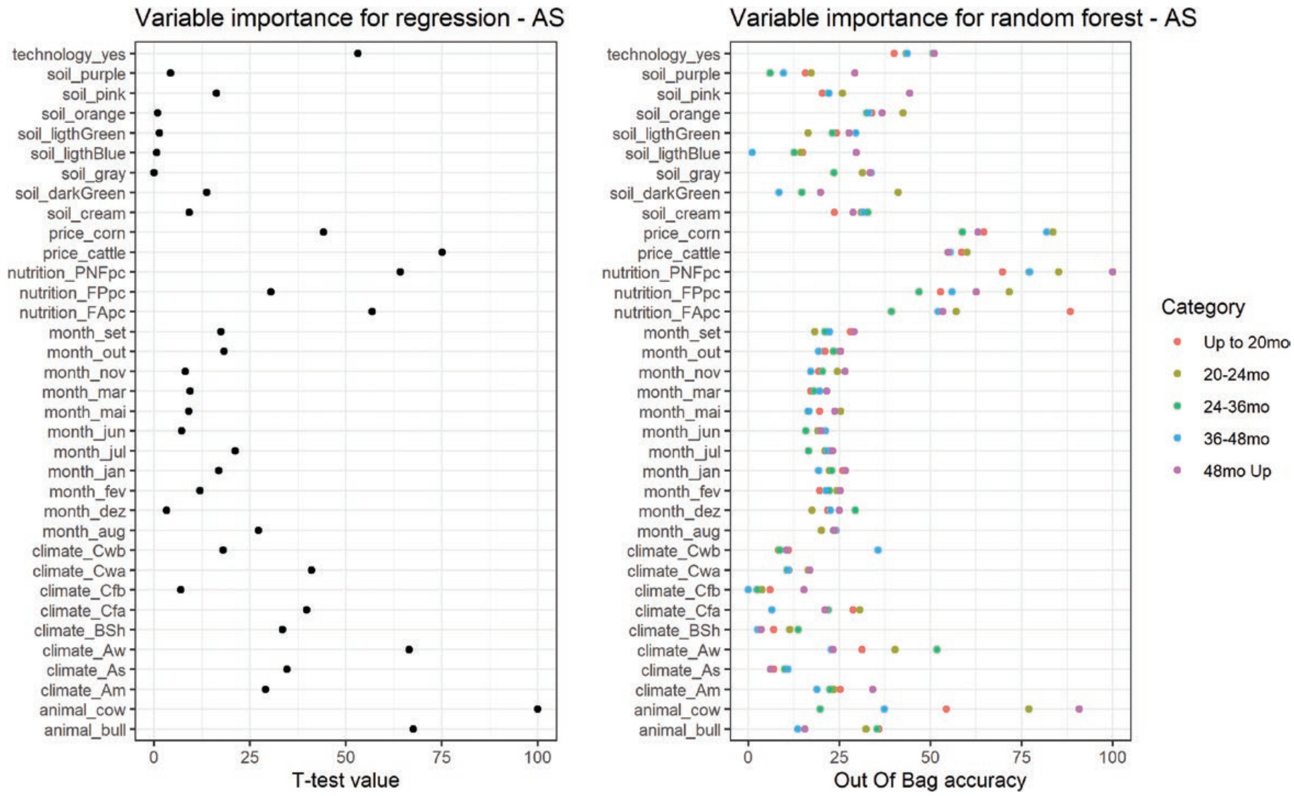

Figure 6. Variable importance results for the prediction of AS with regression and RF. For regression, variable importance was assessed using the T-test value of the regression fitted to the test set while for RF it was estimated as the out-of-bag accuracy of permuting each explanatory variable.

Figure 7. Variable importance results for the prediction of FD with regression and RF. For regression, variable importance was assessed using the T-test value of the regression fitted to the test set while for RF it was estimated as the out-of-bag accuracy of permuting each explanatory variable.

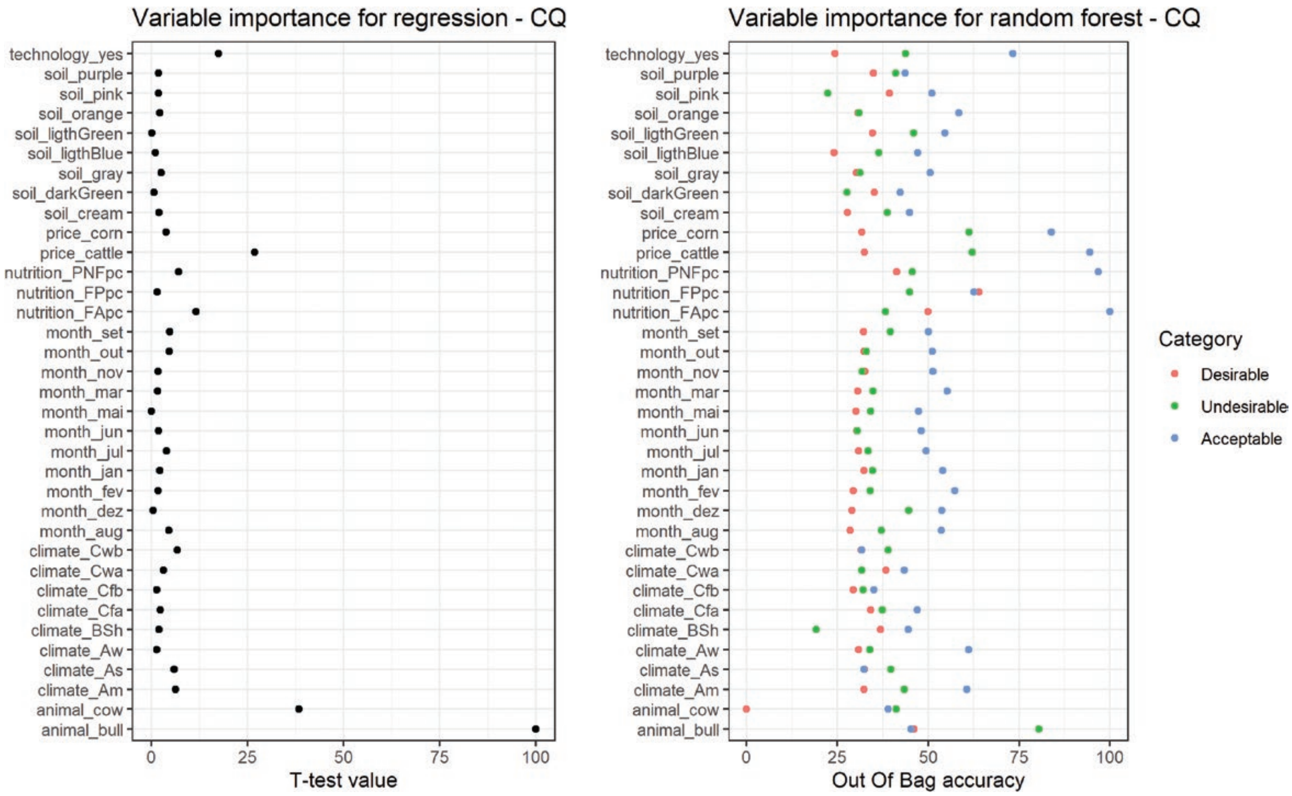

Figure 8. Variable importance results for the prediction of CQ with regression and RF. For regression, variable importance was assessed using the *T*-test value of the regression fitted to the test set while for RF it was estimated as the out-of-bag accuracy of permuting each explanatory variable.

between regression and RF models. Animal category was the most important predictor, followed by cattle sales price and use of technical consulting. However, when compared with regression, RF assigned higher importance to corn price, use of additives in the nutrition, climate, and soil in which the animal was raised. Unlike RF, the regression model assigned higher importance to AS as a predictor of FD. Lastly, for CQ (Figure 8), results were mostly consistent between regression and RF. The most important predictors across models were: animal category, sales price, and use of technology. However, RF assigned higher importance to the nutrition, corn sales price, soil, and climate in which animals were raised, as well as the month of the year.

The run-time and disk space required by all models varied greatly. Regression models were considerably less demanding than the other methods (i.e., RF and NN) with 6 computing hours on 4 CPUs, requiring a total of 40 GB of memory for all outcome variables. For RF analysis, a total of 2,370.5 computing hours on 109 CPUs, and 8 TB total memory were needed, while the NN analysis required 15,482.02 computing hours on 5,580 CPUs, requiring 223.2 TB total memory.

Discussion

When applying statistical and machine learning approaches, a very important step is to choose optimal model hyperparameters. For the GLR, when predicting the class of multi-class ordered variables, the model could be suboptimal if the chosen link function is not appropriate. The selection of a link function usually depends on knowledge of the response distribution (in terms of type and parameters). In cases where the distribution is not known, as in this study, empirical tests can be implemented with the Kuhn (2019) approach to make an educated guess on the best fit available to a data set. As explained in the Materials and Methods section, we compared five different link functions (logistic, probit, cloglog, loglog, and cauchit) to choose the most appropriate one for each of the outcome variables. The best link function was different for each of the categorical variables tested (i.e., cloglog for AS, loglog for FD, and cauchit for CQ). However, the difference in performance across link functions was very small (below 4% accuracy difference for all variables).

Regarding the results for hyperparameter tuning of RF, the best number of explanatory variables available for splitting at each tree node (i.e., mtry) ranged from half of the available number (5 for FD) to all explanatory variables (10 for AS and 11 for CW) (Figure 4). As reviewed by Biau and Scornet (2016), there is no consensus in the literature on the effect of adopting different values of mtry (i.e., number of variables included in the model at a time). Some authors claim that this tuning parameter has little impact on the performance of the method while others recommend using values as large as possible (if possible equal to the number of all explanatory variables). In this analysis, the small number of available explanatory variables allowed us to perform an exhaustive search on mtry. In other words, all possible values of mtry were tested to choose the best one. Results across the analysis of all outcome variables indicate that the model performance benefits from having larger values of mtry, in some cases equal to all available explanatory variables.

For the hyperparameter tuning of NN, the grid search method was used. Grid search is a general approach in which a set of candidate values is defined, then reliable estimates of model performance across candidate values are produced to determine the optimal setting (Kuhn and Johnson, 2016). The candidate values were chosen to span a wide search range, while exploiting the computing capabilities available at a CHTC where the analysis was performed. The NN grid search results tended to favor higher number of hidden layers with higher number of nodes per layer and smaller values of decay (Supplementary Figures S1–S4). Both higher number of hidden units and layers enable expressing more complicated nonlinear functions, extending the classification capability (Liakos et al., 2018). Results may suggest that the saturation of the networks was not yet reached, meaning that fitting more complex networks (in terms of larger number of layers and nodes per layer) could yield better performance. However, due to the large scale of the data set in this study, the effort conduced in the NN analysis already exceeded 15,000 computing hours. Therefore, it would be unfeasible to make further tests for increased complexity (and consequently run-time) with the computational capabilities available. Lastly, it is relevant to understand that even when broader search spaces for the grid are utilized, solutions for best tuning parameters are not guaranteed to be the "global" solution (Kuhn and Johnson, 2016).

Regarding model quality assessment, across all variables, there was a tendency of superior performance of regression and RF methods, while NN tended to present the poorest performance. Results suggest that methods that intrinsically model nonlinear relationships (such as RF and NN) did not perform better for the outcome variables studied (CW, AS, FD, and CQ). It is important to mention that linear models can be adapted to nonlinear trends in the data by manually adding higher-order model terms. However, to do this, one must know the specific nature of the nonlinearity in the data (Kuhn and Johnson, 2016), for example, interaction among specific variables or quadratic effects. The knowledge on such nonlinear relationships was not available for this data set. Inherently nonlinear in nature, models have the advantage that the exact form of the nonlinearity does not need to be known explicitly or specified prior to model training (Kuhn and Johnson, 2016). Lastly, it could be argued that the better performance of RF for CW and AS could be related to the fact of the method being robust to noise in the predictor variables (Biau and Scornet, 2016).

Results for explanatory variable importance for the prediction of CW, AS, FD, and CQ highlighted patterns that interestingly were mostly in agreement with conclusions reached in experimental settings and field observations. One of those patterns is the importance of nutrition used by the farm, and the price of corn 3 mo before animals were finished, followed by soil quality and climate to predict AS. As described by Millen et al. (2011), production cycles carried out solely on grazing systems with only mineral supplementation lead to older animals at the market. This is due to animals putting on weight during the rainy season when the grass quality is higher but losing body weight in the dry season. A big reduction of AS can be achieved when animals are finished in feedlots (Millen et al., 2009). The same nutritional variables also showed importance as predictors of FD, which is in agreement with the observation that feedlot operations are oftentimes utilized just to finish animals and achieve a minimum of 4 mm fat cover as demanded by the Brazilian market (Millen et al., 2009, 2011). Lastly, nutritional variables ranged from moderate to important (for regression and RF) to predict CQ. Even though the effects of nutrition on specific quality parameters, such as FD and CW, are well studied this can be due to the fact that CQ takes into account several other variables, such as AS, gender, and body condition scores.

Participation in a technical advising program showed moderate to high importance for the prediction of all beef production and quality variables. This indicates that expert knowledge is important to aid farmers making management and production decisions that can improve production outcomes. Another variable that showed moderate to high relevance across all outcome variables was the finished cattle sales price. This implies that to a certain degree, the beef market conditions are also relevant for the prediction of meat traits. Lastly, the previously mentioned variable importance results point to an important feature to predict carcass production and quality at the national level: not only physiological variables (such as animal category and AS) are important predictors but also environmental and external variables play an important role as well. Our results highlight that for real-world prediction these factors should not be ignored.

Machine learning approaches are currently being applied for prediction with the objective of optimizing economic efficiency of farm systems in many livestock species (Liakos et al., 2018; Passafaro et al., 2019). Common algorithms in such applications are decision trees and NN. In fact, the use of machine learning techniques to predict beef cattle traits is not new. For example, Alonso et al. (2007) used machine learning to predict beef cattle conformity scores and growth with 91 animals. Additionally, Alonso et al. (2013) applied support vector machines to predict CW in advance to slaughter for the Asturiana de los Valles cattle breed based on zoometric measurement features with 144 animals. The novelty of the application presented here is not only the utilization of machine learning methods to predict beef production and quality but also, more importantly, the capability of utilizing real-world large-scale integrated data, representative of a diverse national context, to do so. Knowledge on which methods and variables are necessary to forecast beef production and quality at the national level can be a valuable tool to predict the future of the Brazilian market. Such projections can be useful in the following years to better allocate resources, with the hope of improving the sustainability of beef production. Lastly, forecasting can also be a tool to aid decision-making, allowing farmers to prepare for changes ahead of time.

It is arguable that the analysis performed here could be improved. For instance, the accuracy of prediction obtained with AS and FD was rather low and even the moderate-high predictive accuracy for CQ could be enhanced. Similarly, the size of the MAE obtained for CW could be decreased. This could have happened for different reasons. For example, the data set might be missing important explanatory variables for the prediction of meat production and quality, such as the genetic merit, breed composition, and health of the animals. Additionally, higher accuracies would likely be achieved if different or even more phenotypes were available for animals. Unfortunately, no other variables were available from the sources used. Recording those variables in different sectors of the market, such that they can be included in future applications, could increase the accuracy of prediction models. With the advent of drones and sensors, we expect that collections of such phenotypes become more feasible. Another important point is that despite the fact that this analysis aims to provide a national snapshot of the Brazilian production, the production system is quite heterogeneous (Millen et al., 2009; Oliveira, 2018). This means that regional variations not accounted for in this analysis are possible, and this could be explored in future studies. Lastly, we acknowledge that the objective of this analysis was solely prediction of future trends, in other words, accurately projecting the chances that something will (or not) happen. The focus of this type of method is to optimize prediction accuracy (Kuhn and Johnson, 2016). Therefore, no causal claim can be made from these results (Rosa and Valente, 2013; Bello et al., 2018).

In the years to come, it will be essential to address the current challenge of augmenting production to nourish a growing human population without increasing the environmental footprint. With greater awareness of the need to preserve natural resources, methods with sustainable perspectives become more appealing (Millen et al., 2011). Understanding how to predict the future of livestock production using large-scale data will be core to projecting future trends and optimizing the allocation of resources at all levels of the production chain, rendering animal production more sustainable. In this analysis, we were capable of predicting future beef production and quality with information on over 4 million head of cattle, corresponding to 4.3% of the Brazilian national production. Despite beef cattle production being a complex system, many times influenced by the farmer's personal interests, meat market regulators, and sanitary issues (such as spread of diseases), this analysis shows that by integrating different sources of data, it is possible to forecast meat production and quality at the national level with moderate-high levels of accuracy.

Acknowledgments

This research was performed using the computing resources and assistance of the UW–Madison Center for High Throughput Computing (CHTC) in the Department of Computer Sciences. The CHTC is supported by UW–Madison, the Advanced Computing Initiative, the Wisconsin Alumni Research Foundation, the Wisconsin Institutes for Discovery, and the National Science Foundation, and is an active member of the Open Science Grid, which is supported by the National Science Foundation and the U.S. Department of Energy's Office of Science.

Conflict of interest statement

There are no potential conflicts of interest that may affect our ability to objectively present or review the research data here submitted.

Literature Cited

Agrolink. 2019. Available from https://www.agrolink.com.br/cotacoes [accessed October 20, 2019].

Aiken, V. C. F., J. R. R. Dórea, J. S. Acedo, F. G. Sousa, F. G. Dias, and G. J. M. Rosa. 2019. Record linkage for farm-level data analytics : comparison of deterministic, stochastic and machine learning methods. *Comput. Eletron. Agr*. **163**:104857. doi:10.1016/j.compag.2019.104857

Alonso, J., A. Bahamonde, A. Villa, and A. R. Castañón. 2007. Morphological assessment of beef cattle according to carcass value. *Livest. Sci*. **107**:265–273. doi:10.1016/h.livsci.2006.09.027

Alonso, J., A. R. Castañón, and A. Bahamonde. 2013. Support vector regression to predict carcass weight in beef cattle in advance of the slaughter. *Comput. Eletron. Agr*. **91**:116–120. doi:10.1016/j.compag.2012.08.009

Álvares, C. A., J. L. Stape, P. C. Sentelhas, J. L. de Moraes Gonçalves, and S. Gerd. 2014. Köppen's climate classification map for Brazil. *Meteorol. Z*. **22**(6):711–728. doi:10.1127/0941-2948/2013/0507

Bello, N. M., V. C. Ferreira, D. Gianola, and G. J. M. Rosa. 2018. Conceptual framework for investigating causal effects from observational data in livestock. *J. Anim. Sci*. **96**:4045–4062. doi:10.1093/jas/sky277

Biau, G., and E. Scornet. 2016. A random forest guided tour. *Test* **25**(2):197–227. doi:10.1007/s11749-016-0481-7

FAO. 2009. *How to feed the world in 2050*. Rome: Food and Agriculture Organization of the United Nations.

FAO. 2014. Meat consumption. Available from http://www.fao.org/ag/againfo/themes/en/meat/background.html [accessed November 2, 2019].

Institutatística—IBGE. 2018. Indicadores IBGE: estatística da produção pecuária.. Available from ftp://ftp.ibge.gov.br/Producao_Pecuaria/Fasciculo_Indicadores_IBGE/abate-leite-couro-ovos_201802caderno.pdf. Instituto Brasileiro de Geografia e Est

Instituto Brasileiro de Geografia e Estatística—IBGE Atlas Nacional. 2019. Available from https://www.ibge.gov.br/apps/atlas_nacional/ [accessed September 10, 2019].

Kamilaris, A., A. Kartakoullis, and F. X. Prenafeta-boldú. 2017. A review on the practice of big data analysis in agriculture. *Comput. Eletron. Agr*. **143**:23–37. doi:10.1016/j.compag.2017.09.037

Kuhn, M. 2019. CARET: Classification And REgression Training. *R package version* **6**:0–84. Available from http://caret.r-forge.r-project.org [Accessed December 1, 2019].

Kuhn, M. and K. Johnson. 2016. *Applied predictive modeling*. 1st ed. Saline (MI): Springer.

Liakos, K. G., P. Busato, D. Moshou, S. Pearson, and D. Bochtis. 2018. Machine learning in agriculture: a review. *Sensors* **18**(2674):1–29. doi:10.3390/s18082674

Millen, D. D., R. D. Pacheco, M. D. Arrigoni, M. L. Galyean, and J. T. Vasconcelos. 2009. A snapshot of management practices and nutritional recommendations used by feedlot nutritionists in Brazil. *J. Anim. Sci*. **87**:3427–3439. doi:10.2527/jas.2009-1880

Millen, D. D., R. D. L Pacheco, P. M. Meyer, P. H. M Rodrigues, and M. D. B. Arrigoni. 2011. Current outlook and future perspectives of beef production in Brazil. *Anim. Front*. **1**(2):46–52. doi:10.2527/af.2011-0017

Morota, G., R. V. Ventura, F. F. Silva, M. Koyama, and S. C. Fernando. 2018. Machine learning and data mining advance predictive big data analysis in precision animal agriculture. *J. Anim. Sci*. **96**:1540–50. doi: 10.1093/jas/sky014

Oliveira, M. 2018. Contributions of Brazilian cattle. São Paulo (Brazil): *Pesquisa FAPESP*. Available from https://biblioteca.ibge.gov.br/visualizacao/periodicos/84/ppm_2018_v46_br_informativo.pdf [Accessed December, 20, 2019].

Oliveira, M. 2019. Produção da pecuária municipal 2018. *Catalog of the Instituto Brasileiro de Geografia e Estatística* **84**(01014234):1–8. https://biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?view=detalhes&id=784 [Accessed November 15, 2019].

Passafaro, T. L., D. Van de Stroet, N. M. Bello, N. H. Williams, and G. J. M. Rosa. 2019. Generalized additive mixed model on the analysis of total transport losses of market-weight pigs. *J. Anim. Sci*. **97**:2025–2034. doi:10.1093/jas/skz087

Pham, X., and M. Stack. 2018. How data analytics is transforming agriculture. *Bus. Horiz*. **61**(1):125–33. doi:10.1016/j.bushor.2017.09.011

Rosa, G. J., and B. D. Valente. 2013. Breeding and Genetics Symposium: inferring causal effects from observational data in livestock. *J. Anim. Sci*. **91**:553–564. doi:10.2527/jas.2012-5840

Stekhoven, D. J. 2013. missForest: nonparametric missing value imputation using random forest. R package version 1.4. Available from https://cran.r-project.org/web/packages/missForest/index.html [Accessed October 29, 2019].

Zia, M., J. Hansen, K. Hjort, and C. Valdes. 2019. Brazil once again becomes the world's largest beef exporter. Washington, DC: United States Department of Agriculture - Economic Research Service. Available from https://www.ers.usda.gov/amber-waves/2019/july/brazil-once-again-becomes-the-world-s-largest-beef-exporter/ [accessed October 21, 2019].