J. Dairy Sci. 97:3352–3363
http://dx.doi.org/10.3168/jds.2013-7451

Comparison of modeling techniques for milk-production forecasting

M. D. Murphy,*† M. J. O'Mahony,† L. Shalloo,* P. French,* and J. Upton*[1]
*Animal and Grassland Research Innovation Centre, Teagasc Moorepark, Co. Cork, Ireland
†Department of Mechanical Engineering, Cork Institute of Technology, Co. Cork, Ireland

ABSTRACT

The objective of this study was to assess the suitability of 3 different modeling techniques for the prediction of total daily herd milk yield from a herd of 140 lactating pasture-based dairy cows over varying forecast horizons. A nonlinear auto-regressive model with exogenous input, a static artificial neural network, and a multiple linear regression model were developed using 3 yr of historical milk-production data. The models predicted the total daily herd milk yield over a full season using a 305-d forecast horizon and 50-, 30-, and 10-d moving piecewise horizons to test the accuracy of the models over long- and short-term periods. All 3 models predicted the daily production levels for a full lactation of 305 d with a percentage root mean square error (RMSE) of ≤12.03%. However, the nonlinear auto-regressive model with exogenous input was capable of increasing its prediction accuracy as the horizon was shortened from 305 to 50, 30, and 10 d [RMSE (%) = 8.59, 8.1, 6.77, 5.84], whereas the static artificial neural network [RMSE (%) = 12.03, 12.15, 11.74, 10.7] and the multiple linear regression model [RMSE (%) = 10.62, 10.68, 10.62, 10.54] were not able to reduce their forecast error over the same horizons to the same extent. For this particular application the nonlinear auto-regressive model with exogenous input can be presented as a more accurate alternative to conventional regression modeling techniques, especially for short-term milk-yield predictions.
Key words: dairy production, milk-production forecasting, modeling

INTRODUCTION

Milk production from pasture-based dairy cows is susceptible to variation due to seasonality of pasture production (Adediran et al., 2012), grazing conditions (Baudracco et al., 2012), disease (Collard et al., 2000), nutritional interventions (Kolver and Muller, 1998), and other disturbances (Olori et al., 1999; Tekerli et al., 2000). The ability to forecast herd milk yield days, weeks, and months in advance provides benefits for management at processor and farm level as total daily milk production strongly influences energy consumption, plant utilization, and farm income. The usefulness of a milk-yield prediction system depends upon how accurately it can predict daily milking patterns and its ability to adjust to factors affecting supply. Milk yield prediction models have proven useful for genetic analysis (Ptak and Schaeffer, 1993) and for bio-economic modeling (Shalloo et al., 2004).

Studies have been undertaken by Wood (1967), Ali and Schaeffer (1987), Wilmink (1987), and Guo (1995), who all developed algebraic equations for the purpose of fitting a lactation curve to empirical data. Jones (1997) stressed the need for increased flexibility and adaptation among curve-fitting techniques and introduced an empirical Bayes method for fitting Wood's lactation curve (incomplete gamma function; Wood, 1967). Macciotta et al. (2002) and Vasconcelos et al. (2004) employed auto-regressive models to predict individual lactations using limited numbers of test days throughout the lactation cycle. Other attempts to forecast milk yields have involved large regression models such as artificial neural networks (**ANN**) and multiple linear regression (**MLR**) models (Lacroix et al., 1995; Salehi et al., 1998; Sharma et al., 2006; Sharma et al., 2007). These models proved to be very successful; however, they require large amounts of detailed information for each specific cow. The ANN model developed by Sharma et al. (2007) requires 12 individual traits of each cow (genetic group, season of birth, period of birth, birth weight, age at maturity, weight at maturity, season of calving, period of calving, age at calving, weight at calving, peak yield, and days to attain peak yield); likewise the model tested by Lacroix et al. (1995) required 16 network inputs including information such as logarithm of somatic cell count, energy fed on test day, protein fed on test day, DM fed on test day, and so on. Brun-Lafleur et al. (2010) modeled variation in milk yield with respect to energy and protein supply, but acquiring even this information for an entire pasture-based herd is not practical. A balance is required between the availability of detailed

Received September 2, 2013.
Accepted February 11, 2014.
[1]Corresponding author: john.upton@teagasc.ie

information at farm level and the prediction accuracy of the milk-supply model. The high levels of detailed data required to construct these milk-yield predictors inhibit their practical implementation on commercial dairy farms.

The aim of this study was to assess the suitability of a static neural network (**SANN**), a MLR model, and a nonlinear auto regressive model with exogenous input (**NARX**) for the prediction of total daily herd milk yield (**DHMY**) over varying forecast horizons. The most successful model was selected according to its abilities to generate the most accurate forecast using very limited training data in low volumes over a long- (305 d), medium- (30 to 50 d), and short-term (10 d) horizon.

MATERIALS AND METHODS

Data Collection

Data were collected from a research farm in the south of Ireland for a period of 4 yr (2006–2010). Daily herd milk yield (liters) and number of cows milked (**NCM**) on that corresponding herd DIM was collected because this is the most accessible data for commercial farmers to obtain. Milk yields were recorded from a conventional herringbone swingover milking parlor using ICAR (International Committee for Animal Recording)-approved milk meters. The model was set at herd level and evaluated by comparing daily milk yields across a herd of 140 pasture-based Holstein-Friesian (**HF**) cattle (North American HF and New Zealand HF genetic strains). The milking season of 2010 was selected as the target prediction horizon and the previous 3 yr of data were used to train the model.

Model Inputs

In previous studies certain variables were found to have an influence on milk production: season of calving (Wood, 1967), climatic conditions (Smith, 1968), number of DIM (Grzesiak et al., 2006), and stocking rate (McCarthy et al., 2011). In this study the farm grazing area remained static, whereas the number of cows grazing varied throughout the year. Similarly, the season of calving (spring) was kept constant in the herd over several years. Hence the total herd milk production behaves in a cyclical pattern (assuming no catastrophic external factors). This pattern is influenced by the herd size at any one time, the DIM of the herd, and other factors such as atmospheric conditions (ambient temperature, irradiance, and precipitation). This information is readily available on commercial dairy farms; hence, DIM and NCM were selected as model inputs. The localized prediction of atmospheric conditions was deemed outside the scope of this study.

The models were trained with basic information (DIM and NCM) and used to predict DHMY over specified time horizons. The total DHMY can be viewed as a time series that is being primarily driven by a handful of factors. The number of cows coming in and out of lactation can be factored in by recording the NCM on each milking day. The DIM is factored in by chronologically applying a day number relative to the beginning of lactation for the entire herd. All 3 model predictions were trialed over several prediction horizons: 305, 50, 30, and 10 d. For the horizons less than 305 d, the models repeatedly projected over the specific horizon in a moving piecewise manner until the end of the series. After every horizon step the previous DHMY data were added to the models training set before the next prediction, updating the model state. (All 3 models were developed using the software package MATLAB R2012a; Mathworks, Natick, MA.) The statistics toolbox was used to create the MLR model, and the neural networks toolbox box was used to create the neural network models. For detailed information regarding the data processing, structure, and training of these models, please refer to Demuth et al. (2010).

Neural Networks

An ANN is a mathematical model whose operating principle is based on biological neural networks (Haykin, 1999). The ANN architecture comprises a series of interconnected layered neurons through which inputs are processed. These inputs values are multiplied by the synaptic weights, which represent the strength of the neural connections. Figure 1 shows a typical feedforward ANN structure containing an input, hidden, and output layer. This configuration is very popular for function approximation in systems where no time-dependent relationship exists among the network inputs. Increasing the size of the hidden layer allows for more intricate function fitting of nonlinear processes; however, overfitting of training data is undesirable when good generalization abilities are needed (Demuth et al., 2010). Many methods exist for improving generalization such as data filtering, feedback elements, regularization, and network reduction. Reducing the number or neurons in the hidden layer is an effective method of improving generalization because small networks do not have the capability of overfitting the training data. The synaptic weights are configured during back propagation training (Hecht-Nielsen, 1989). Once trained, a SANN has no feedback elements and contains no delays.

Dynamic artificial neural networks are also known as recurrent neural networks because of their dynamic