Computers and Electronics in Agriculture 150 (2018) 74–87

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Original papers

# Machine-learning algorithms for predicting on-farm direct water and electricity consumption on pasture based dairy farms

P. Shine[a], M.D. Murphy[a,*], J. Upton[c], T. Scully[b]

[a] Department of Process, Energy and Transport Engineering, Cork Institute of Technology, Cork, Ireland
[b] Department of Computing, Cork Institute of Technology, Cork, Ireland
[c] Animal and Grassland Research and Innovation Centre, Teagasc Moorepark Fermoy, Co., Cork, Ireland

A B S T R A C T

This study analysed the performance of a range of machine learning algorithms when applied to the prediction of electricity and on-farm direct water consumption on Irish dairy farms. Electricity and water consumption data were attained through the utilisation of a remote monitoring system installed on a study sample of 58 pasture-based, commercial Irish dairy farms between 2014 and 2016. In total, 15 and 20 dairy farm variables were analysed for their predictive power of monthly electricity and water consumption, respectively. These variables were related to milk production, stock numbers, infrastructural equipment, managerial procedures and environmental conditions. A CART decision tree algorithm, a random forest ensemble algorithm, an artificial neural network and a support vector machine algorithm were used to predict both water and electricity consumption. The methodology employed backward sequential variable selection to exclude variables, which added little predictive power. It also applied hyper-parameter tuning with nested cross-validation for calculating the prediction accuracy for each model on unseen data (data not utilised for model development). Electricity consumption was predicted to within 12% (relative prediction error (RPE)) using a support vector machine, while the random forest predicted water consumption to within 38%. Overall, the developed machine-learning models improved the RPE of electricity and water consumption by 54% and 23%, respectively, when compared to results previously obtained using a multiple linear regression approach. Further analysis found that during the January, February, November and December period, the support vector machine overpredicted electricity consumption by 4% (mean percentage error (MPE)) and water consumption by 21% (MPE), on average. However, overprediction was greatly reduced during the March – October period with overprediction of electricity consumption reduced to 1% while the overprediction of water consumption reduced to 8%. This was attributed to a phase shift between farms, where some farms produce milk all year round, some dry off earlier/later than others and some farms begin milking earlier/later resulting in an increased the coefficient of variance of milk production making it more difficult to model electricity and water accurately. Concurrently, large negative correlations were calculated between the number of dairy cows and absolute prediction error for electricity and water, respectively, suggesting improvements in electricity and water prediction accuracy may be achieved with increasing dairy cow numbers. The developed machine learning models may be utilised to provide key decision support information to both dairy farmers and policy makers or as a tool for conducting macro scale environmental analysis.

## 1. Introduction

By 2050, the global consumption of milk and dairy products is expected to increase by 20% (Bruinsma and Alexandratos, 2012). To help meet this increased demand, the European Union milk quota system was abolished in April 2015, thus allowing European dairy farmers to freely expand production. In the long run, the abolishment of milk quotas may significantly increase the monetary value of European exports as the proportion of global milk supplied by European based dairy farms increases. In preparation, the Irish government identified the potential for a 50% increase in milk production by 2020 over 2007–09 levels (DAFM, 2010). Increasing milk production must be met with the sustainable consumption of on-farm electricity and water (E&W) resources to ensure continued sustainable growth of Irelands dairy industry (DAFM, 2016). With increasing milk production, electricity consumption on dairy farms is of key interest, as existing infrastructure

* Corresponding author.
*E-mail address:* michaeld.murphy@cit.ie (M.D. Murphy).

https://doi.org/10.1016/j.compag.2018.03.023
Received 7 December 2017; Received in revised form 9 March 2018; Accepted 17 March 2018

may not be optimally configured for increased production levels. For example, milk cooling electricity cost savings of up to 25% may be achieved through the installation of equipment such as plate heat exchangers for the pre-cooling of milk (Shine et al., 2018b). Concurrently, the water demand (drinking + parlour + miscellaneous) required to meet current milk production targets may increase to unsustainable levels, thus deteriorating groundwater borehole water supplies (O'Connor and Kean, 2014). This may subsequently place additional pressure on the public water supply. Accurate empirical prediction models developed from real-world E&W data may provide: (1) key decision support information regarding both electricity and/or water consumption to both dairy farmers and policy makers. (2) A tool for conducting macro scale environmental analysis for marketing Irish dairy products abroad. (3) A means of calculating the impact of Irish dairy farming on natural resources, (4) a method of benchmarking and improving environmental efficiency and estimating related costs and (5) an aid for developing regulations. Improving the prediction accuracy of dairy farm E&W, validated on unseen data may increase the confidence of key stakeholders for making informed economic and policy related decisions and the confidence of the model's ability to provide decision support for dairy farmers (such as forecasted consumption in hypothetical scenarios). For example, a dairy farmer wishing to upgrade his/her milk cooling system may utilise such an empirical model to calculate the associated change in electricity consumption, which in turn can help assess the viability of the of the investment. Similarly, an empirical model for dairy farm water consumption, in conjunction with borehole volume and replenishment data may be employed to calculate potential stresses on borehole water supplies, which may occur with increasing dairy herd sizes. This data may be particularly useful to water utilities to identify areas of stress, as significant water volumes may be required from an inadequately prepared public water system. For the purpose of dairy farm E&W prediction, multiple linear regression (MLR) has previously been considered (Shine et al., 2018a).

MLR is a relatively basic modelling technique, which has traditionally proven to be the most popular method for predicting residential and industrial electricity consumption (Al-Ghandoor et al., 2008; Samhouri et al., 2009) as well as urban and residential water consumption (Domene and Saurí, 2006; Jorgensen et al., 2013). Previous work applied MLR modelling to predict monthly electricity and water consumption for pasture based Irish dairy farms. Utilising data from 58 pasture based dairy farms, the MLR models were validated on unseen data using *k*-fold cross-validation and were found capable of predicting monthly dairy farm electricity consumption to within 26% (relative prediction error (RPE)), while water consumption was poorly predicted to within 49% (Shine et al., 2018a). A MLR model offers a unique advantage regarding its simplicity and ease of deployment for government bodies and dairy farmers for predicting macro level consumption figures. However, increased prediction accuracy may be achieved through utilising machine-learning (ML) algorithms that are capable of handling non-linear relationships and interactions between input variables.

ML models have already been applied to a wide range of applications within the agricultural industry. These include milk production forecasting using artificial neural networks (Grzesiak et al., 2006; Murphy et al., 2014; Zhang et al., 2016), comparing the performance of random forests, support vector machines and extremely randomised trees for discriminating between grassland types in Ireland (Barrett et al., 2014) and identifying clinical mastitis on dairy cows using decision tree induction (Kamphuis et al., 2010). Similarly, the yield and protein content of brown rice was predicted through support vector machine algorithms (Saruta et al., 2013) while the random forest algorithm has proved effective for global and regional crop yield predictions, improving the accuracy of global wheat predictions by 76% over a standard MLR model (Jeong et al., 2016).

Literature related to the application of ML algorithms for predicting dairy farm E&W consumption is limited, instead primarily focusing on MLR modelling (Higham et al., 2017; Murphy et al., 2017; Todde et al., 2017) and mechanistic modelling of electricity consumption (Upton et al., 2014). However, an adaptive neural-fuzzy inference system was developed for predicting the energy output of 50 Iranian dairy farms using variable inputs related to electricity consumption and fossil fuel usage with their respective calorific values. The adaptive neural-fuzzy inference system was shown to reduce the root mean squared error by 50% compared to a MLR model (Sefeedpari et al., 2014). To the authors' knowledge, no ML models exist for predicting dairy farm electricity or water consumption and as such, a comparative analysis of ML modelling techniques applied to dairy farm E&W consumption prediction may offer a valuable understanding of the potential improvements in prediction accuracy over standard methods.

Comprehensive research methodologies for the development of ML models require the tuning of certain parameters (hyper-parameters) and the calculation of model accuracy (Alonso et al., 2013; Saruta et al., 2013). Hyper-parameters control the internal behaviour of a ML model, effecting their generalisation capability. The selection of optimum hyper-parameter combinations for particular ML algorithms is important for good predictions. However, over-fitting of a ML algorithm (through the tuning of hyper-parameters) may overestimate the prediction accuracy and thus, must be minimised. One technique to minimise model over-fitting is through employing a stratified nested cross-validation methodology (Cawley and Talbot, 2010; Statnikov et al., 2005; Stone, 1974).

In addition to the tuning of hyper-parameters for specific ML algorithms, the selection of input variables may also have a considerable impact on model prediction capability. Multiple variable selection methods exist which aim to extract high prediction yielding variables, as the inclusion of too many variables is likely to introduce inherent noise reducing model prediction accuracy. Conversely, the inclusion of too few variables may not incorporate sufficient information to make useful predictions. Bermúdez-Chacón et al. (2015) recommended the utilisation of multiple variable selection techniques as a pre-processing step while considering multiple ML algorithms to determine which variable subset and algorithm is most suited to the data. Such variable selection methods, as described by Guyon and Elisseeff (2003), include (but are not limited to): (1) variable ranking methods such as the use of the random forest algorithm for quantifying variable importance, as employed by Barrett et al. (2014) for classifying different types of grasslands, and Jeong et al. (2016) for predicting global and regional crop yields. (2) Correlation analysis methods, as employed by Shine et al. (2018a), Todde et al. (2017) and Higham et al. (2017), for predicting dairy farm electricity and/or water consumption using MLR modelling. (3) Variable subset selection methods such as the utilisation of forward and backward variable elimination as employed by Ahmed et al. (2012) for identifying the optimal variables for classifying different types of crops and weeds from digital images.

This work utilised E&W consumption data collected from 58 Irish commercial dairy farms and corresponding data related to milk production, stock, farm infrastructure, managerial processes and environmental conditions as previously utilised for developing MLR models (Shine et al., 2018a). The primary objective of this paper was to assess the prediction accuracy of four ML algorithms for their ability to improve the prediction accuracy of monthly E&W consumption on Irish dairy farms over MLR modelling. The prediction accuracies of a CART decision tree algorithm (CDT), a random forest ensemble algorithm (RF), an artificial neural network (ANN) and a support vector machine regression (SVR) algorithm were assessed. The methodology employed in this paper utilised a recommendation from Bermúdez-Chacón et al. (2015) to consider multiple ML algorithms and variable selection techniques to increase the probability that the prediction performance of the final model (ML model with the greatest accuracy) is maximised. Concurrently, the methodology employed in this research article followed that of Statnikov et al. (2005), whereby a stratified nested cross-validation methodology was employed for unbiased model selection