Computers and Electronics in Agriculture 99 (2013) 176–185

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Artificial neural networks vs. Gene Expression Programming for estimating outlet dissolved oxygen in micro-irrigation sand filters fed with effluents

Pau Martí [a,*], Jalal Shiri [b], Miquel Duran-Ros [c], Gerard Arbat [c], Francesc Ramírez de Cartagena [c], Jaume Puig-Bargués [c]

[a] Departament d'Enginyeria Rural i Agroalimentària, Universitat Politècnica de Valencia, c/ Camí de Vera s/n, 46022 València, Spain
[b] Department of Water Engineering, Faculty of Agriculture, University of Tabriz, Iran
[c] Department of Chemical and Agricultural Engineering and Technology, University of Girona, C. de Maria Aurèlia Capmany, 61, 17071 Girona, Spain

ARTICLE INFO

Article history:
Received 1 April 2013
Received in revised form 29 July 2013
Accepted 17 August 2013

Keywords:
Artificial neural networks
Gene Expression Programming
Dissolved oxygen
k-Fold testing
Drip irrigation
Clogging

ABSTRACT

Sand media filters are specially used to avoid emitter clogging when water with large amount of organic pollutants like effluents are used in micro-irrigation systems. Estimation of water quality parameters such as dissolved oxygen at sand filter outlet (DO_o) is of great interest for irrigation engineers. Artificial neural networks (ANN), Gene Expression Programming (GEP) and Multi Linear Regression (MLR) based models were trained for estimating DO_o using data from 769 experimental filtration cycles. Instead of considering a single configuration of the training and test data sets, which is the usual procedure for those applications in agricultural studies, the performance of those models was assessed through *k*-fold testing, ensuring a complete performance evaluation. In general, the GEP model tended to provide the most accurate estimations, followed by ANN and, lastly, by MLR models. After the evaluation of the models, the GEP approach was used to provide a new equation to estimate DO_o based on the complete data set. This procedure revealed that only inlet DO, pH, electrical conductivity and filter head loss were necessary to feed the models. Furthermore, the consideration of leave one out or, at least, *k*-fold assessment should be advisable to perform a suitable evaluation of the model performance. Otherwise, conclusions drawn might be only partially valid.

1. Introduction

The use of effluents for irrigation is a viable alternative in areas where water is scarce or there is intense competition for its use. Micro-irrigation is the best way to apply effluents from the public health and environmental points of view (Bucks et al., 1979). However, emitter clogging, which is an inherent problem of this irrigation system, can be increased when using effluents due to their generally greater concentrations of salts, nutrients, solids and biological materials (Trooien and Hills, 2007). Appropriate filtration is essential for the successful operation of micro-irrigation systems, because it helps to prevent emitter clogging caused by organic and inorganic particulate matter (Ayars et al., 2007). The three common filter types used in micro-irrigation system for preventing emitter clogging are screen, disc and sand media filters. Sand media filters are considered the standard for filtration protection of micro-irrigation systems (Trooien and Hills, 2007), as emitters protected by these filters show less fouling (Capra and Scicolone, 2007; Duran-Ros et al., 2009). However, sand filters are more complex and expensive than the other filter types and they are only appropriate for farms with high technical and professional standards (Capra and Scicolone, 2007). The operational problems of sand filters can be increased if low quality water, like effluents, is used for irrigation, which is a more common practice everywhere.

The usual high content of organic matter in wastewaters allows aerobic microbial growth and reduces dissolved oxygen (DO) concentrations in effluents. So, DO is an indicator of water quality since it can be a limiting factor, especially in intensive agriculture systems (Raviv et al., 2004). Besides, DO can be determined easier and quicker than other related parameters such as chemical oxygen demand (COD) or biochemical oxygen demand (DBO). Sand media filters improve outlet DO (DO_o) moderately, which helps avoiding the agronomic problems caused by hypoxic water (Elbana et al., 2012). A bad sand filter performance reduces DO_o (Elbana et al., 2012) since the increase of available organic nutrients in sand filters promotes microbial activity (Nakayama et al., 2007), which tends to lower the levels of DO within the filter (Bouwer and Zehnder, 1993). Thus, the development of accurate models for estimating DO at filter outlet can be very useful for an

* Corresponding author.
E-mail address: paumarpe@hotmail.es (P. Martí).

0168-1699/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.compag.2013.08.016

appropriate management of both sand filter performance and irrigation water quality.

Genetic Programming (GP) was introduced as a generalization of Genetic Algorithms (GA) (Koza, 1992). The fundamental difference between GP and GA lies in the nature of individuals. In GA individuals are linear strings of fixed length, such as chromosomes; whereas in GP individuals are nonlinear entities of different sizes and shapes, such as parse trees. Gene Expression Programming (GEP) is comparable to GP yet evolves computer programs of different sizes and shapes encoded in linear chromosomes of fixed lengths. The chromosomes are composed of multiple genes, each gene encoding a smaller subprogram. One strength of the GEP approach is that the creation of genetic diversity is extremely simplified, because genetic operators work at the chromosomes level. Another strength of GEP consists of its unique, multigenic nature which allows the evolution of more complex programs composed of several subprograms. The advantages of a system like GEP are clear from nature, but the most important are (Ferreira, 2001b):

- the chromosomes are simple entities: linear, compact, relatively small, easy to manipulate genetically (replicate, mutate, recombine, etc.)
- the expression trees are exclusively the expression of their respective chromosomes; they are entities upon which selection acts, and according to fitness, they are selected to reproduce with modification.

As a result, GEP surpasses the old GP system in 100–10,000 times (Ferreira, 2001a, 2001b). The applications of GEP and GP approaches are relatively scarce in agriculture, and have mainly tackled the estimation of evaporation and reference evapotranspiration (Shiri et al., 2012, 2013).

Haykin (1999) defined Artificial Neural Networks (ANNs) as massively parallel distributed processors consisting of simple processing units, which have a natural propensity for storing experimental knowledge and making it available for use. They are able to perform non-linear mapping of a multidimensional input space onto another multidimensional output space, requiring no detailed information about the system. Synaptic connections between the neurons are adjusted during the training or learning process, and, as a result, knowledge is acquired and stored. They have shown to be efficient and less-consuming in the modeling of complex systems. ANNs have been successfully used to model different target variables in irrigation applications, for instance, water demand (Pulido-Calvo et al., 2003), nitrate distribution (Li et al., 2004), emitter head losses (Martí et al., 2010), reference evapotranspiration (Kumar et al., 2011), and stem water potential (Martí et al., 2013).

So far, the application of ANNs and GEP models in the study of filters for micro-irrigation systems or the emitter clogging when reclaimed wastewater is used is very limited or lacking, respectively. Puig-Bargués et al. (2012) used multilayer perceptrons to estimate DO_o and turbidity at the sand filter outlet, as well as filtered volume, and compared the results with multiple linear regression (MLR). Although the presented results were promising, that work only considered a single data set assignment to evaluate the performance of the neural networks. This is a common practice in ANN and GEP applications, as well as in other data driven approaches. Nevertheless, according to this methodology, results can only be referred to the specific data assignment assumed. If the test set is not representative, the performance assessment of the model might be only partially valid. In order to face this problem, leave one out procedures or *k*-fold testing (Stone, 1974; Shao, 1993) allow for a suitable evaluation of the model performance in the considered data set through a complete testing scan of the data set, considering several train-test stages.

This work presents the application of ANN and GEP based models to estimate DO_o, and evaluates the performance of the models considering *k*-fold testing. Results are compared with those of Puig-Bargués et al. (2012) and with *k*-fold MLR. Finally, a new GEP-based expression is proposed for its estimation.

2. Materials and methods

2.1. Data set and input selection

The experimental setup providing the considered data set is described in Puig-Bargués et al. (2012). According to it, the complete data set consists of 770 data points corresponding to 770 filter runs that lasted 1.6 h on average. After the exclusion of one anomalous pattern, data were arranged as a matrix of 769 patterns (in rows) by 11 variables, namely inlet and outlet dissolved oxygen (DO_i, DO_o) sand media size (d_e), total head loss (ΔH), electrical conductivity (EC), inlet and outlet turbidity ($turb_i$, $turb_o$), flow rate across the filter (Q), filtered volume (V), *pH* and temperature (T). The statistical characterization of these parameters can be found in that paper, too. That study assessed the performance of different input combinations for estimating DO_o, defined based on physical criteria and MLR analysis. Results suggested that the most suitable inputs for that output were d_e, ΔH, DO_i, EC, pH and T. Although there might be other procedures to find out the optimum input combination, the current study assumes that the conclusions of Puig-Bargués et al. (2012) are valid. The discussion of the validity of that input combination is beyond the scope of the present work, focusing on the comparison of 3 modeling approaches relying on the same inputs. Hence, the mentioned inputs were used to feed the current models, too.

2.2. k-Fold assessment

Usually, when dealing with the application of ANN and GEP based models, data are split up, at least into 2 data sets: one for training and one for testing. In contrast to common applications, where only a single definition of these sets is established, this work defines a minimum assumable test size according to criteria relying on computational costs, and performs so many train-test processes, so that the complete data set is scanned considering that test size. The smaller the test size defined the higher the number of stages of the *k*-fold assessment, with a subsequent rise of the operation times. In this case, instead of leaving a single pattern for testing (leave one out), the size of the test set was fixed in 5% of the data set. Smaller sizes would have involved too high computational times. This procedure allows to define smaller test set sizes than in common applications relying on a single assignment, because the train-test process is performed several times and with different configurations. So, this procedure allows for a more effective extraction of the knowledge contained in the data set, given that a higher number of patterns are available for the application of the training algorithm.

For the chosen test size, the complete scan of the data set requires 20 train-test stages (i.e. 20 different training and test sets), 19 stages with 730 and 39 patterns for training and testing, respectively, and one stage with 741 and 28 patterns. Based on the fixed test size percentage, the corresponding number of testing patterns was rounded to a whole number. The last test set is adjusted until completing the total number of patterns.

2.3. Gene Expression Programming

GEP is, like GA and GP, a genetic algorithm as it uses populations of individuals, selects them according to fitness, and