Available online at www.sciencedirect.com

ScienceDirect

Procedia Engineering 119 (2015) 1419 – 1428

Procedia Engineering

www.elsevier.com/locate/procedia

13th Computer Control for Water Industry Conference, CCWI 2015

Forecasting Domestic Water Consumption from Smart Meter Readings using Statistical Methods and Artificial Neural Networks

David Walker[a,], Enrico Creaco[a], Lydia Vamvakeridou-Lyroudia[a], Raziyeh Farmani[a], Zoran Kapelan[a], Dragan Savić[a]

[a]*Centre for Water Systems, College of Engineering, Mathematics and Physical Sciences, University of Exeter, EX4 4QF, UK.*

Abstract

This paper presents an artificial neural network-based model of domestic water consumption. The model is based on real-world data collected from smart meters, and represents a step toward being able to model real-time smart meter data. A range of input schemas are examined, including real meter readings and summary statistics derived from readings, and it is found that the models can predict some consumption but struggle to accurately match in cases of peak usage.

Peer-review under responsibility of the Scientific Committee of CCWI 2015

Keywords: Smart water meters; domestic water consumption; prediction; artificial neural networks.

1. Introduction

Forecasting domestic water use is of vital importance for water utility companies, while the rising prevalence of smart water meters, collecting high resolution data from individual users, provides a large corpus of data on which predictive models can be based. The iWIDGET [1] project has recently begun collecting large volumes of meter readings, measuring water consumption at 15 minute intervals for domestic properties in locations throughout Europe. In this work, we use data from a pilot case study in Greece to develop a model capable of forecasting water usage at hourly resolution for the households participating in the iWIDGET project. Predictions at this frequency are useful, for example, for leakage detection.

Artificial neural networks (ANNs) are known to be adept at modelling nonlinear relationships, and have been used for predicting domestic water consumption (e.g., [2,3]), as well as being applied in a range of hydroinformatics settings (e.g., [4,5]). ANN training is conducted with an evolutionary algorithm.

An issue with water consumption data at a relatively fine resolution is that it is inherently noisy. An activity such as flusing a toilet or using a washing machine will cause a peak of water usage in that timestep that will not be matched in the preceding or following timesteps. We therefore consider the use of summary statistics to enhance the ability of the ANN models generated herein to generalise to new data despite the presence of such noise. This allows us to capture the history held within days-worth of meter readings without requiring an input for each reading. This greatly

* Corresponding author.
E-mail address: D.J.Walker@exeter.ac.uk

Peer-review under responsibility of the Scientific Committee of CCWI 2015
doi:10.1016/j.proeng.2015.08.1002

reduces the complexity of the ANN training, meaning that a useful model can be achieved with significantly shorter training procedures. As is discussed later in the results section, this is met with some success, although the overall results remain affected by the noise to a certain extent. We discuss possible approaches to ameliorating this issue at the end of the paper.

The remainder of this paper is structured as follows. After some background information is presented in Section 2, we describe the case study and data used to train and test the model (Section 3), before the structure of the ANN model used in this paper is outlined in Section 4. Results and discussion are presented in Section 5. Sections 6 and 7 concludes the paper with a discussion of ongoing and future work.

2. Background

The literature contains a range of studies in which computational intelligence approaches, such as that employed herein, have been used to predict domestic water consumption. This contribution made by this work is to begin to investigate the use of these methods for predicting at the fine resolution that smart water meters operate on.

A common goal is to predict daily water consumption. An example of this is [2], in which the SECM-UA [6] algorithm is used to optimise network weights for ANN modelling; their results demonstrate comparable results to using Bayesian training methods, regression and ANFIS (adaptive-network-based fuzzy inference system) [7]. ANNs trained with EAs were also used in a study that was able to predict up to 24 hours ahead of the current time [3]. Another daily study employed the relevance vector machine [8] to predict water usage for a case study in China [9]. ANNs have been selected in this work as they have been found to offer competitive or superior predictions to other data driven approaches [10].

An example of a study on a comparable timescale to that used in this work is [11], where a stochastic model of water use was presented and demonstrated on a range of timescales including an hour. The model is based on statistical data about water usage, rather than actual water usage measurements as is the case here. Another statistical model was presented by [10], which also demonstrated the ability to predict at the hourly level. Statistical modelling of the variety employed by these papers is beyond the scope of this initial study, however is an avenue that we will explore in the future. Summary statistics describing water usage over periods of time are used in this work as inputs to the ANN in order to reduce the amount of actual meter readings required, simplifying the resulting models.

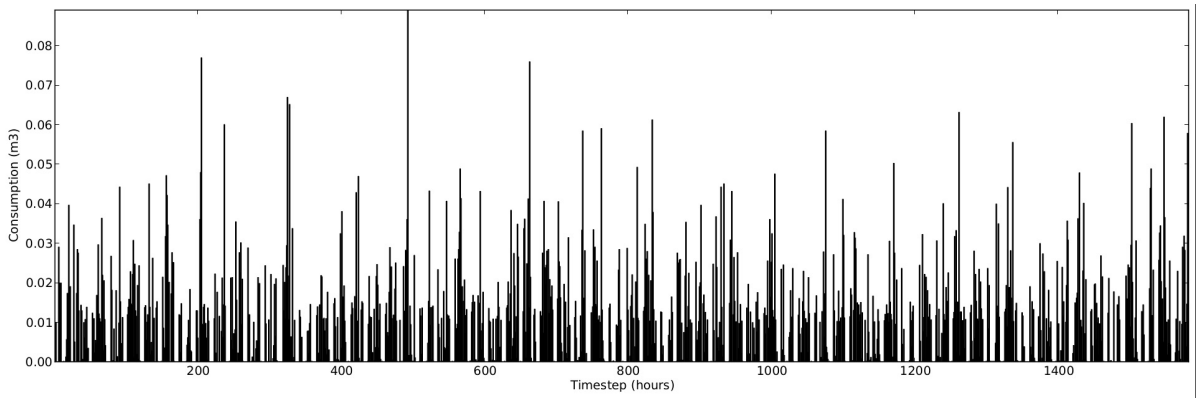
3. Case Study

The data used in this study is drawn from a case study in the iWIDGET project, which has recently begun collecting large volumes of data from three locations in Europe: Greece, Portugal and the United Kingdom. In this work, we use data drawn from the Greek case study, and have selected nine properties for which to construct models of their domestic water usage.

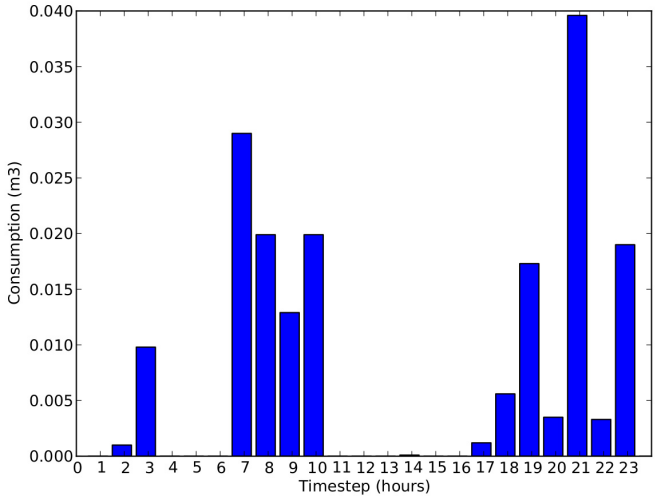
The period with which this study is concerned is 1st February 2015-31st March 2015. The properties were chosen because of the quality of data in this period (some properties in the case have missing data because of meter transmission errors). No other criteria were involved in the selection of properties to model. The time series for one of the properties is shown in Fig. 2, along with a closer view of the first 24 hours of the time series. The model will take as inputs recent measurements from the meter (e.g., the last four hours' readings). However, as can be seen in Figure 1(b), the data is extremely noisy, spiking when water is consumed with frequent periods when there is no water usage within the house. This represents a difficulty for the ANN approach we propose in this paper, so the data relating to recent water usage will be supplemented with statistical information describing historical usage at those times. For this reason, a week of historical data from 25th-31st January 2015 is included at the start of the dataset. The exact amount, and nature of this supplementary statistical information will be discussed and examined later in this paper.

4. Artificial Neural Network

The model presented in this paper comprises two components: an artificial neural network, the design of which was determined experimentally, and an evolutionary algorithm, with which the network weights were trained. These



(a)



(b)

Fig. 1: An example of a user's time series (a) and a close-up view of the same time series for the first twenty four hours (b). The data shown is for the period 25th January 2015 - 31st March.

components are described in the following sections. The general structure of the ANN framework employed is based on that presented in [4].

4.1. Model Structure

The ANN model employed herein is a multi-layer perceptron, a model comprising an input layer, one or more hidden layers, and an output layer. The input layer consists of a set of D inputs, each of which represents an aspect of the data being modelled. In this work we us a mixture of real and statistical information about domestic water usage over time as inputs. The exact configuration of inputs is an area of investigation in this study, as we seek to produce a model with the highest degree of accuracy possible while minimising the number of inputs needed, simplifying the model.

A neural network employs hidden neurons to allow it to model nonlinear functions. The hidden neurons are arranged into one or more hidden layers, the configuration of which depends on the problem at hand, and itself can be cast as an optimisation problem. In this paper we employ a single layer of hidden neurons, and determine the number

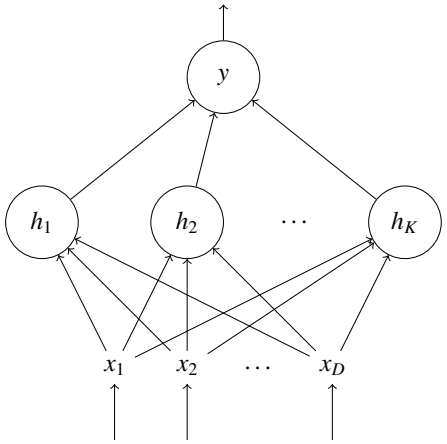


Fig. 2: The general configuration of ANN model used in this study. The model comprises D model inputs, K hidden neurons, and a single output.

of them experimentally. As with model inputs, we seek to minimise the number of hidden neurons in order to reduce the complexity of training the model as far as possible.

The output layer comprises the neuron(s) that produce the actual output of the model. In this case, the task of the model is to predict water consumption at the next timestep, thus we use a single model output. An extension of this work will examine the construction of a model that can predict multiple timesteps ahead. The general configuration of the ANN models examined in this paper is shown in Fig. 2.

The output of a neuron is defined by

$$y_k = a\left(\sum_{d=1}^{D} x_d w_{dk} + w_{k0}\right) \tag{1}$$

where a is the *activation function*. In this study, we employ the sigmoid activation function:

$$a(x) = \frac{1}{1 + e^{-x}}. \tag{2}$$

4.2. Evolutionary Algorithm

An evolutionary algorithm (EA) is employed to optimise the weights of the neural network. EAs are particularly well suited to solving such optimisation problems, and have been used extensively in neural network optimisation. Often they are used to optimise the topology of the neural network, however in this study the optimisation is restricted to the weight values.

A solution is represented as a real-valued vector $\boldsymbol{\theta}$, such that each θ_p represents one of the $P = D \times K + K$ weighted edges between neurons in the ANN. Each weight value is in the range $(-1, 1)$. The fitness ϕ of a solution is computed in terms of the difference between the model's prediction for a given set of inputs in the training data and the true value of those inputs. This is computed with the mean absolute area between the model prediction y_n for inputs $\mathbf{x}_n$ and its corresponding true target value t_n evaluated under the model defined by solution $\boldsymbol{\theta}_i$:

$$\phi_i = \frac{1}{N} \sum_{n=1}^{N} |y_n^i - t_n^i|. \tag{3}$$

The EA is run for a set number of generations, determined experimentally. At each generation, the current parent population is copied to produce a child population. Then, each child solution $\boldsymbol{\theta}'$ is mutated with an additive Gaussian mutation operator that adds a random value drawn from $\mathcal{N}(0, \sigma)$ with probability $1/P$. The standard deviation σ of the Gaussian distribution was 0.1. Once each of the child solutions has been evaluated under the objective function (equation 3) the parent population for the next generation is selected from the union of the current generation's parent and child populations with an elite selection operator; the top ϵ solutions from the combined population are retained.

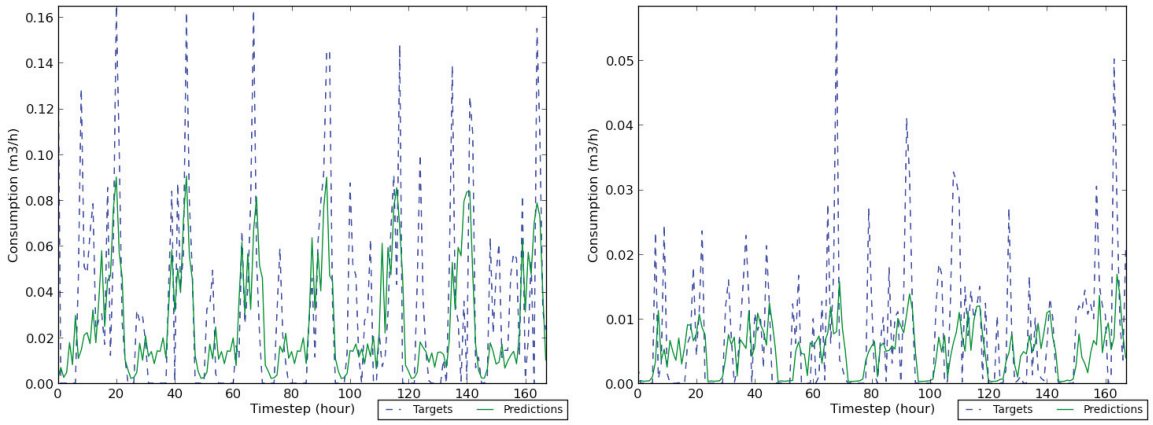


Fig. 3: Test results for the initial ANN configuration on exemplar test cases (weeks one and six from the data). As can be seen, the model is capable of predicting the general shape of the target data, however fails to properly match the magnitude of the larger values.

4.3. Model Training

The ANN model is trained using the supervised approach outlined above. The performance of the approach is evaluated against eight datasets, each of which describes domestic water consumption for one of the properties in iWIDGET's Greek case study. Each of the nine datasets contains 24 measurements for each day (a measurement every hour). We employ a leave-one-out cross validation approach, in which the data is divided into eight folds (one for each week). We then train the data on seven of the folds and retain the eighth for testing the resulting model, which is repeated eight times so that each of the folds is used as test data once.

5. Results

We present results for various arrangements of inputs to examine the difference in predictive quality offered by each. The data is aggregated to the hourly level, and we evaluate the quality of predictions using the correlation between the predictions and target values.

In each configuration, the model is trained using the EA outlined above. The algorithm is executed for 200 generations in each case, with each generation comprising five parent solutions which each generates a child solution.

5.1. Summary Historical Statistics

The initial configuration we examined comprised of three inputs, each of which is used to make a prediction of water consumption at timestep t. The first is the actual water consumption at the previous timestep, $t-1$. The second input is an average of water consumption over the prior seven days for timestep t, and the third is a value representing the current hour of the day.

A test case for this initial configuration is shown in Fig. 3. Clearly, the model is failing to accurately match the peaks in the target data, although it does demonstrate some ability to follow the general trend of the data. For example, the predictions (shown by a solid line) are clearly organised into seven blocks, corresponding to the seven days comprising these test cases. In the left-hand example, usage of up to approximately $0.09m^3$ the model is capable of matching the quantity of water consumed. For water consumption beyond that, the model is unable to properly match the height of the peak in the targets.

In order to address the model's failure to predict peak usage, we increase the data on which the prediction is based by including the standard deviation of the historical values for timestep t. Fig. 4 illustrates the results of training and testing against this input schema. In this case the predictions are smoother, and the model has apparent difficulty predicting the extremely low values, more so than was the case for input schema 1. This indicates that the additional

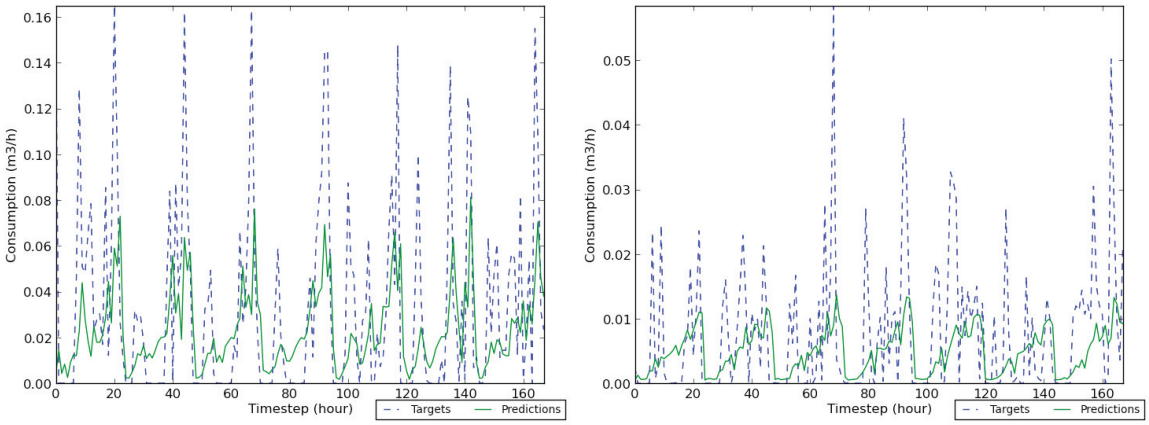


Fig. 4: Test results for the input schema using both the average and standard deviation of historical values for timestep t.

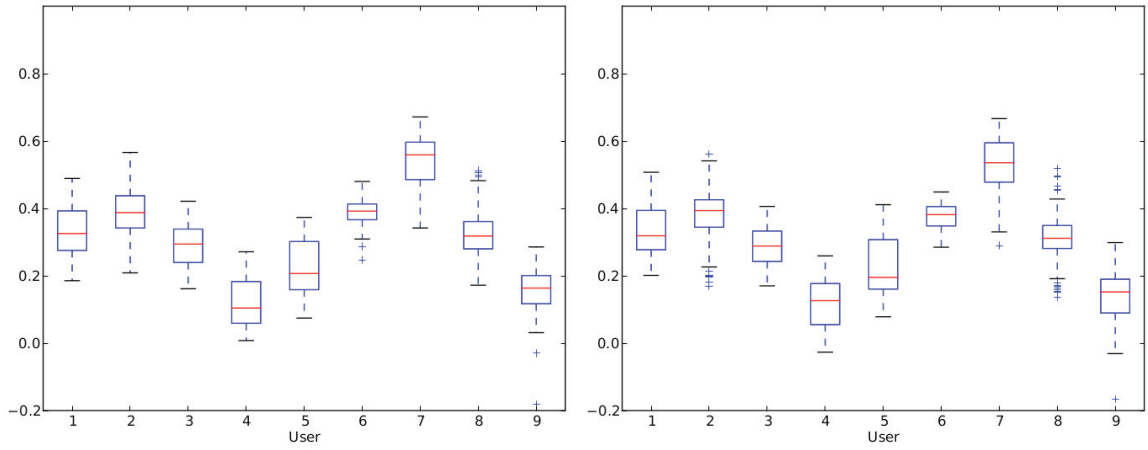


Fig. 5: Correlation between the target values and model output. Each of the nine case study users is shown, with predictions for user 7 showing the greatest correlation between target values and prediction. The left-hand plot shows correlation for models using input configuration 1, while the second shows input configuration 2.

statistical information does not provide any further additional information. That said, from examining Fig. 5 the results are not considerably worse. These show the correlation between the target values and model predictions for the nine case study users examined in this work. An experiment for a user comprises eight weeks, each of which is used as a test case, and was repeated 10 times. Thus, the boxplots represent 80 correlation values per user. Clearly there is a degree of correlation, with values greater than 0.5 being achieved in some cases. Since the model does not properly match the magnitude of the peaks in cases of high water consumption, high correlation values cannot be expected. An interesting feature of the figures is that there is a great deal of variability between users. User 7 showed the best performance in both cases. The results for both input schemas are relatively similar, implying that there is no great improvement given the additional information provided by the standard deviation.

5.2. Real Historical Values

To address the smoothing affect, observed above, from the introduction of additional statistical information, we consider the use of real historical meter readings, rather than averaged readings, for the current timestep t. Two new input schemas are employed. In the first, seven inputs are used to incorporate the meter readings for time t on the last

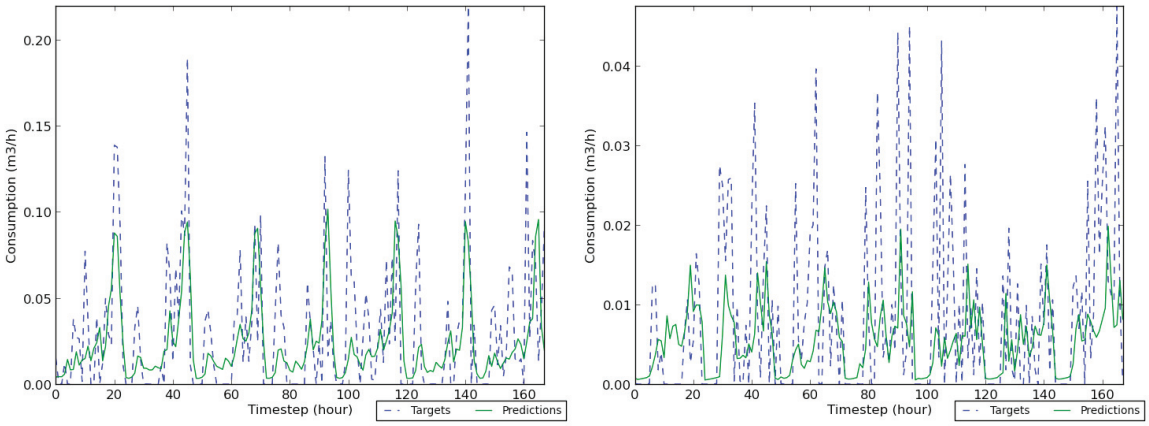


Fig. 6: Test results for the input schema using both the average and standard deviation of historical values for timestep t.

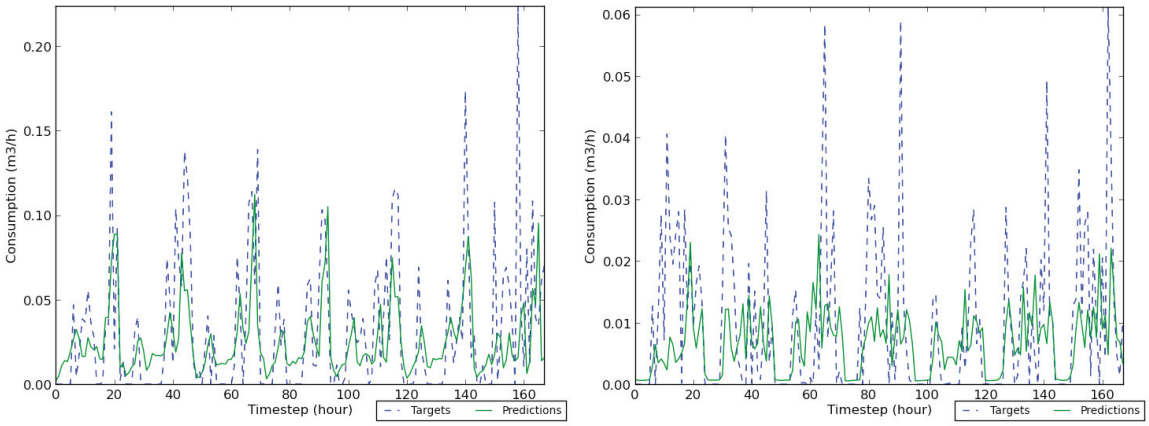


Fig. 7: Test results for the input schema using both the average and standard deviation of historical values for timestep t.

7 days. This is used in combination with the meter reading for $t-1$ (input schema 3) and four readings for ($t-q$, for $q = 1, \ldots 4$). Fig. 6 illustrates two test cases for input schema 3, while Fig. 7 shows results for schema 4. In both cases, the results are comparable to those shown for input schemas 1 and 2. From examining the correlation between targets and predictions in Fig. 8 we can see that there is a similar distribution of results to that of the earlier experiments. Case study user 7 has the best predictions again. Fig. 9 shows the range of correlations for all four experiments. The upper limits, as well as the mean, are generally consistent in all four cases, although there is a slight reduction in mean performance for the experiment using real historical data in lieu of the earlier statistical inputs. This can be attributed to the increased complexity of model training arising from the larger number of inputs.

6. Future Directions

Despite showing some ability to predict domestic water consumption, the models presented in this work suffer from a lack of accuracy. This stems from the amount of noise present in the data; as can be seen in the model output examples shown earlier, cases in which the model fails to predict are typically those where there is significantly greater than normal water consumption. The ANN approach we have employed herein has failed to match the peaks of such instances (e.g., Fig. 3). We therefore consider future directions of work that are ongoing to ameliorate this.

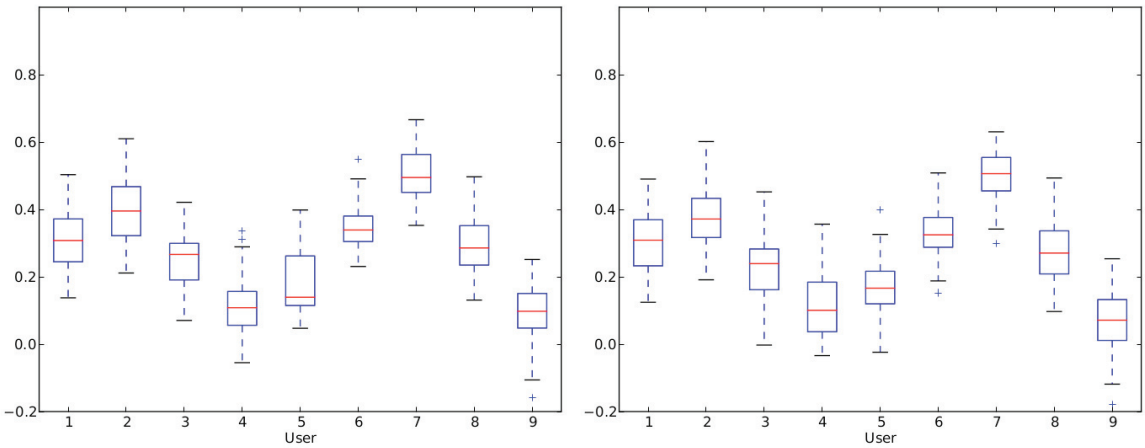


Fig. 8: The correlation between target and model outputs for the nine case study users. The pattern of results here is similar to those obtained for experiments with input schemas one and two.

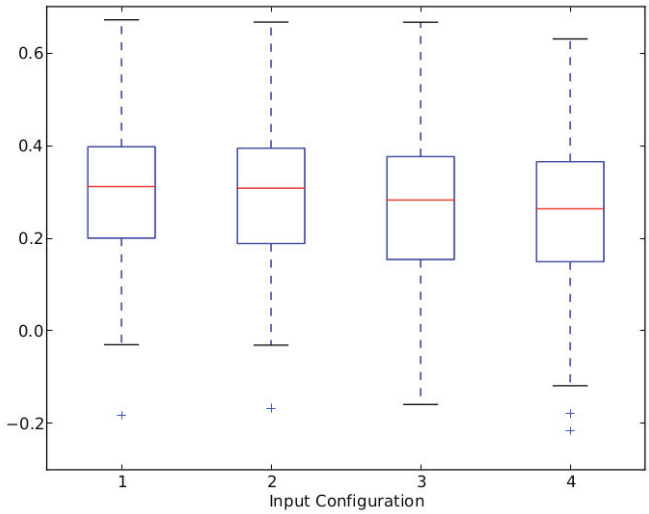


Fig. 9: Correlation between targets and model predictions shown across the four experiments conducted in this study. The best and mean performance is generally consistent throughout, however the poorer worst performance in the case of input schemas three and four indicates there is little to be gained from using real input values instead of summary statistics in these cases.

One important avenue of investigation is to determine where the fault for this inability to predict peak usage lies. We are currently comparing the predictive quality of the ANN model presented here against that of other data driven approaches. This will identify whether an alternative configuration of this model produces better results, or if other approaches also struggle to properly model data of this type and frequency.

We intend to refine the ANN approach in order to better model the available data. As the iWIDGET project continues a greater corpus of training and testing data will be collected, which will enable more sophisticated learning strategies to be developed, benefiting from the larger amount of data that can be retained for testing purposes, without reducing the amount of training data. Increased data will also enable us to build better statistics for use as inputs to the model. A simple example is the averaging over more historical timesteps to provide more accurate predictions.

As well as exploiting the increased amount of available data, we also intend to optimise the learning process by which the model is constructed. This involves optimising the structural aspects of the model, for example the number

of hidden neurons, as well as optimising the EA used to train the network. The experimental approach taken in this paper involved selecting a set of values likely to produce good results based on experience (e.g., [4]) and the literature, however this itself can be cast as an optimisation task and optimised with a self-adaptive EA. A problem with this particular avenue is the computational expense of the network training problem, measured on the order of minutes. In order to develop a self-adaptive EA designed to optimise a good training procedure the network must be trained many more times than has been done in this paper. Here, the training required a runtime of the order of three to five minutes, so steps must be taken to speed up the execution if this is to be a viable approach.

An alternative is to consider a statistical approach to modelling the data. A type of model that is used for examples such as this, whereby the data follows a certain trend but contains outliers (here we consider consumption peaks to be outliers) is to use a Student-t distribution to model the data [12]. That distribution contains heavy tails, increasing the likelihood that the outliers will be incorporated into the model as is desirable here. Our proposed approach is to use the expectation-maximisation algorithm to learn a parametrisation of the Student-t distribution that allows us to properly model the data used in order to make more accurate predictions.

7. Conclusion

This paper has presented a preliminary study into forecasting domestic water usage with data collected during the ongoing iWIDGET project. The study employed an artificial neural network to predict the next timestep's water usage for nine users from the project's Greek case study. Models were trained using an evolutionary algorithm, with parametrisations determined experimentally. The models resulting from the work used a range of input schemas, contrasting the use of historical input readings with inputs based on summary statistics constructed from those readings. While the peak and mean performance of the different schemas was similar, the worst case of the schemas using real historical values was lower than those using the statistical inputs. This can be attributed to the complexity of optimising network weights for greater numbers of inputs.

In all cases, the models failed to accurately predict the magnitude of consumption in peak cases. Future work will explore avenues to address this issue, exploring the extensions to the neural network approach described above, while also examining the efficacy of statistic modelling approaches.

8. Acknowledgements

This study was carried out as part of the ongoing project iWIDGET (2012-2015), which is being funded by the European Commission within the 7th Framework Programme (Grant Agreement No 318272).

References

[1] D. Savić, L. Vamvamkeridou-Lyroudia, Z. Kapelan, Smart Meters, Smart Water, Smart Societies: The iWIDGET Project, Procedia Engineering 89 (2014) 1105–1112.
[2] P. Cutore, A. Campisano, Z. Kapelan, C. Modica, D. Savić, Probabilistic prediction of urban water consumption using the scem-ua algorithm, Urban Water Journal 5 (2008) 125–132.
[3] M. Romano, Z. Kapelan, Adaptive water demand forecasting for near real-time management of smart water distribution systems, Environmental Modelling & Software 60 (2014) 265 – 276.
[4] D. Walker, E. Keedwell, D. Savić, R. Kellagher, An artificial neural network-based rainfall runoff model for improved drainage network modelling, in: Proc. International Conference on Hydroinformatics (HIC2014), 2014.
[5] M. Firat, M. Yurdusev, M. Turan, Evaluation of artificial neural network techniques for municipal water consumption modeling, Water Resources Management 23 (2009) 617–632.
[6] J. A. Vrugt, H. V. Gupta, W. Bouten, S. Sorooshian, A shuffled complex evolution metropolis algorithm for optimization and uncertainty assessment of hydrological model parameters, Water Resources Research 39 (2003) 1201.
[7] J.-S. Jang, Anfis: adaptive-network-based fuzzy inference system, IEEE Transactions on Systems, Man and Cybernetics 23 (1993) 665–685.
[8] M. Tipping, Sparse bayesian learning and the relevance vector machine, Journal of Machine Learning Research 1 (2001) 211–244.
[9] Y. Bai, P. Wang, C. Li, J. Xie, Y. Wang, A multi-scale relevance vector regression approach for daily urban water demand forecasting, Journal of Hydrology 517 (2014) 236–245.
[10] C. J. Sutton, Z. Kapelan, A probabilistic methodology for quantifying, diagnosing and reducing model structural and predictive errors in short term water deman forecasting, Environmental Modelling & Software 66 (2015) 87–97.

[11] E. J. M. Blokker, J. H. G. Vreeburg, J. C. van Dijk, Simulating residential water demand with a stochastic end-use model, Journal of Water Resources Planning and Management 136 (2010) 19–26.
[12] J. Christmas, R. Everson, Robust autoregression: Student-t innovations using variational Bayes, IEEE Transactions on Signal Processing 59 (2011) 48–57.