E-proceedings of the 36th IAHR World Congress
28 June – 3 July, 2015, The Hague, the Netherlands

MODELLING RESIDENTIAL WATER CONSUMERS' BEHAVIORS BY FEATURE SELECTION AND FEATURE WEIGHTING

ANDREA COMINOLA[(1)], MATTEO GIULIANI[(1)], DARIO PIGA[(2)], ANDREA CASTELLETTI[(1)], ANDREA RIZZOLI[(3)] & MARTIN ANDA[(4)]

[(1)] *Department of Electronics, Information, and Bioengineering – Hydroinformatics Lab, Politecnico di Milano, Milano, Italy, andrea.cominola@polimi.it, matteo.giuliani@polimi.it, andrea.castelletti@polimi.it*

[(2)] *IMT Institute for Advanced Studies Lucca, Lucca, Italy, dario.piga@imtlucca.it*

[(3)] *Istituto Dalle Molle di Studi sull'Intelligenza Artificiale, SUPSI-USI, Lugano, Switzerland, andrea@idsia.ch*

[(4)] *Murdoch University, Perth, Western Australia, m.anda@murdoch.edu.au*

ABSTRACT

Identifying the most relevant determinants of water consuming or saving behaviors at the household level is key to building mathematical models that predict urban water demand variability in space and time and to explore the effects of different Water Demand Management Strategies for the residential sector. This work contributes a novel approach based on feature selection and feature weighting to model the single-user consumption behavior at the household level. A two-step procedure consisting of the extraction of the most relevant determinants of users' consumption and the identification of a predictive model of water consumers' profile is proposed and tested on a real case study. Results show the effectiveness of the proposed method in capturing the influence of candidate determinants on residential water consumption, as well as in attaining sufficiently accurate predictions of users' consumption profiles, which constitutes essential information to support residential water demand management.

Keywords: User Profiling, Residential Water Consumption, Water Demand Management, Feature Extraction

1. INTRODUCTION

Residential water demand nowadays covers a large portion of the public drinking water supply worldwide (Collins *et al.*, 2009, Kenny *et al.*, 2009) and projections show that urbanization and population increase will further boost such a demand (Cosgrove and Cosgrove, 2012). The expansion of supply infrastructures is not always enough to secure demand satisfaction, due to water availability and financial constraints (McDonald *et al.*, 2014). Therefore, Water Demand Management Strategies (WDMS) are key for water utilities to secure reliable water supply at affordable costs (Gleick *et al.*, 2013). In turn, the effectiveness of WDMS strongly relies on our understanding of water consumption drivers (Jorgensen *et al.*, 2009). Revealing the most relevant determinants of water consuming or saving behaviors at the household level is a fundamental step to build predictive models of urban water demand variability in space and time (*e.g.*, Bennett *et al.*, 2013). By capturing the behaviors of water users, these models allow identifying the variety of users' consumption profiles (Gato-Trinidad *et al.*, 2011) as well as exploring the effects of different WDMS for the residential sector (Anda *et al.*, 2013; Fielding *et al.*, 2013), ultimately supporting water utilities and urban planners. The state-of-the-art literature reports a variety of users models, which can be classified as:

- *Descriptive models* (*e.g.*, Gato-Trinidad *et al.*, 2011) that focus on the analysis of water end-use patterns for targeting specific promising areas in designing WDMS (*e.g.*, restriction on irrigation practices in the case where gardening represents the dominant end-use);
- *Predictive models* (*e.g.*, Bennett *et al.*, 2013) that aim to estimate the water demand at the individual (household) level as determined by natural and socio-psychographic factors as well as by the response of water users to different WDMS.

The first class of models allows building users' consumption profiles based on historical trends. This provides the baseline reference for identifying promising areas where water savings and conservation actions may be focused. Yet, they do not quantify the expected impact of demand management actions on water consumption and savings, thus limiting their suitability to support the design of WDMS. In contrast, the second class of models can be employed to effectively predict water consumption at the household level. They are usually composed of two modules: *user profiling*, which regards the identification of the most relevant inputs to explain users' consumption, and *behavioral modeling*, which performs structure identification, parameter calibration and validation of the predictive model. Many state-of-the-art studies (*e.g.*, Makki *et al.*, 2013; Fox *et al.*, 2009; Olmstead *et al.*, 2007) reported the presence of correlations between one or more presumed consumption drivers and the associated consumption profiles, thus accomplishing the user profiling phase. Yet, the number of considered candidate variables is generally limited. In addition, only a few works completed the second phase (*i.e.*, behavioral modeling) and provide a quantitative prediction of the water demand at the household level as a function of the identified drivers and WDMS, thus representing promising decision-aiding tools for water utilities and urban planners.

This work contributes a novel approach based on *feature extraction* techniques (Guyon and Elisseeff, 2003) to model the single-user consumption behavior at the household level. The approach is based on a two-step procedure:

1. **Identify** the most relevant determinants of users' consumption profiles;
2. **Build** a predictive model of water consumption profiles based on the observation of the determinants identified in the previous step.

The use of *feature selection* (*i.e.*, algorithms returning a subset of selected features) and *feature weighting* (*i.e.*, algorithms ranking the features according to their relevance) (Zhao *et al.*, 2010) is motivated by the need to manage a large number of potentially relevant factors influencing water consumers' behaviors along with their redundancy and highly nonlinear relationships, which represent major challenges for standard cross-correlation analyses (Galelli *et al.*, 2014).

In this paper, the proposed approach is applied to a dataset of low-resolution water consumption records associated with a variety of demographic and psychographic users data and household attributes collected in nine towns of the Pilbara and Kimberley Regions of Western Australia throughout the *H2ome Smart* project (Anda *et al.*, 2013).

The rest of the paper is organized as follows: the next section introduces the proposed feature extraction approach, and Section 3 describes the case study and the experiment setting. Numerical results are reported in Section 4. Section 5 summarizes the limitations of the proposed approach and identifies possible improvements for development.

2. FEATURE EXTRACTION-BASED USER PROFILING

Feature extraction techniques, mostly developed in the data mining and machine learning research communities, represent potentially promising tools to model residential water users behaviors. These methods allow extracting the more relevant determinants in describing the consumption profiles of water users out of a large set of candidate drivers. On the basis of the selected determinants, a behavioral model predicting the water consumption at the household level can be identified.

The general formulation of a water demand predictive model for a generic user i is the following:

$$y_i = f(\boldsymbol{x}_i) \qquad [1]$$

where y_i is the consumption profile of the i-th user and $\boldsymbol{x}_i$ denotes the set of M determinants influencing his behavior, represented by a variety of demographic and psychographic users data (*e.g.*, age, number of house occupants, income level, conservation attitude, etc.), household attributes (*e.g.*, house size, type, garden area, etc.) and exogenous factors (*e.g.*, temperature, and precipitation, water price, etc.). The union of determinants and consumption data yields a sample dataset containing N tuples, one for each user. The i-th tuple (with $i=1,\ldots,N$) is defined as follows:

$$< x_i^1, x_i^2, \ldots, x_i^M, y_i > \qquad [2]$$

The construction of the water demand predictive model defined in Eq. [1] relies on the following two-step procedure:

1. *Feature extraction* to select from the original dataset X of users' data a subset $X' \subseteq X$ of determinants that are relevant to describe the consumption profile Y;
2. *Model learning* to predict the water consumption profile as a function of the selected features X'.

2.1 Feature extraction

Different approaches can be adopted to perform feature extraction as well as for model learning. In particular, feature extraction techniques can be classified in two main categories:

- **Feature selection**, namely algorithms that return a subset of features selected from the original dataset as the most relevant to describe the considered output variable (*i.e.*, consumption profile);
- **Feature weighting**, namely algorithms that rank all the features according to a measure of their relevance, with no actual selection of the most relevant variables, which however are identified as the ones in the first positions of the ranking.

Moreover, depending on their structure, they can be distinguished between *model-free* (or *filter*) algorithms, when they do not include any learning algorithm, or *model-based* in case they explicitly rely on a learning algorithm (Galelli and Castelletti, 2013). Model-based algorithms can be further classified into *wrapper* models, if they include a predetermined learning algorithm, and *embedded*, if the model construction phase includes feature selection (Zhao *et al.*, 2010).
Since no single method is best suited to all datasets and modeling purposes a-priori, we implemented and applied different algorithms for both feature selection and weighting. In particular, we run the following feature selection algorithms[a]:

- Fast Correlation Based Filter (FCBF) (Yu and Liu, 2003);
- Correlation Feature Selection (CFS) (Zhao *et al.*, 2010);
- Bayesian Logistic Regression (BLogReg) embedded method (Guyon *et al.*, 2002);
- Sparse Bayesian Multinomial Logistic Regression (SBMLR) embedded method (Cawley *et al.*, 2007).

We also tested the following feature weighting algorithms:

[a] The 2014 version of the ASU feature selection package downloadable at http://featureselection.asu.edu/ was adopted for this study.

- CHI-square score (Liu and Setiono, 1995);
- Information gain (Cover and Thoma, 2012).

2.2 Model Learning

As far as the model learning is concerned, in principle any data-driven model (regressors or classifiers) can be used (see Maier *et al.*, 2000; Galelli and Castelletti, 2013). In practice, the selected method should have the following desirable features:

1. *Modeling flexibility* to approximate strongly non-linear functions, particularly because the relationships between the candidate inputs (selected features) and the output (consumption profile) is completely unknown a priori;
2. *Computational efficiency* to deal with potentially large data-sets, when considering large number of users;
3. *Scalability* with respect to the number of candidate variables to be analyzed, due to the need of testing several variables with different domains and variability.

In the present experiments, we used two different data-driven models: a Naive Bayes Classifier (Duda and Hart, 1973) and the J48 java implementation of the C4.5 Decision Tree algorithm (Quinlan, 1993).

3. CASE STUDY DESCRIPTION

3.1 The *H2ome Smart* project

The following data, collected within the *H2ome Smart* project, are available:

- **Household water consumption data** from meter readings (measured in m^3). The maximum number of readings per household within the considered period is seven, thus the highest reading resolution is approximately three months;
- **House and occupants attributes**: 26 variables describing different features of the users and their house. Table **1** reports the complete list of data available.

Data were collected between August 2010 and February 2012 for more than 3000 households in the towns of the Pilbara and Kimberley Regions of Western Australia.

Table 1. Water consumers' and household features considered in this study.

NAME	DESCRIPTION	VARIABLE NATURE	NUMBER OF POSSIBLE CATEGORIES
TOWN	-	Categorical	9
SUBURB	-	Categorical	21
YEARS OF OCCUPANCY	Years since the house is being occupied by the same inhabitants	Integer	-
HOUSE RESPONSIBILITY	Person responsible for paying bills	Categorical	4
NUMBER OF OCCUPANTS	Number of inhabitants in the house	Integer	-
RESIDENT TYPE	Type of resident in the house	Categorical	8
NUMBER OF TOILETS	Number of toilets in the house	Integer	-
LAND USE	Type of land use destination	Categorical	14
HOUSE TYPE	Type of house structure	Categorical	5
WASHING MACHINE TYPE	Type of washing machine	Categorical	3
TOILET TYPE	Type of flush	Categorical	3
SHOWER TYPE	Type of shower	Categorical	3
DISHWASHER PRESENCE	Presence of dishwasher	Binary	-
GARDEN AREA	Area of the house garden [m^2]	Real positive	-
WATERING METHOD	Method used for garden watering	Categorical	4
WATERING TIME	Average weekly watering time	Integer	-
IRRIGATION SYSTEM	Type of irrigation technique	Categorical	3
DRIP IRRIGATION TYPE	Type of irrigation technique	Categorical	3
SURFACE IRRIGATION TYPE	Type of surfasse irrigation	Categorical	3
DRIP IRRIGATION DURATION	Weekly average drip irrigation minutes	Categorical	4
SURFACE IRRIGATION DURATION	Weekly average surfasse irrigation minutes	Categorical	4
MULCH USAGE	Usage of mulch	Binary	-
POOL PRESENCE	Presence of pool	Binary	-
POOL COVER USAGE	Presence of pool chover	Binary	-
SPA PRESENCE	Presence of spa	Binary	-
NATIVE PLANTS PRESENCE	Presence of native plants	Binary	-

3.2 Data pre-processing

3.2.1 Data cleaning

1. Records of users showing data inconsistencies or missing data (*i.e.*, negative consumption rate or no consumption rate measures for any reading period) were removed from the dataset;
2. Empty reading dates fields were filled for as many users as possible with the reading dates of the same accounting reading group;

3. The average daily water consumption rate in [m^3/day] was computed for each water-reading period, for each household, given its water consumption data and reading period dates. This operation was useful to obtain comparable values of water consumption among different houses, since the duration of reading period was heterogeneous in the considered sample;
4. If information about the number of house occupants was present, the per-capita daily water consumption rate in [m^3/day] was computed for each reading period.

The data cleaning process produced the following outputs: a matrix C_{daily} containing six readings of daily average water consumption rate for N = 1624 households and a matrix $C_{pcDaily}$ containing six values of per-capita daily average consumption for N' = 1560 households. Note that N and N' are significantly lower than the initial dimension of the dataset, which included approximately 3000 households, as water consumption readings were partially or totally missing.

3.3 Class label assignment

The real values in C_{daily} and $C_{pcDaily}$ were converted into three classes representing different water consumption profiles: **low-consumers**, **medium-consumer**, and **high-consumers**. *Kmeans* clustering was used to assign consumption data to classes, with k=3 (number of classes) and Squared Euclidean distance settings. It was run over the vectors Y_{daily} and $Y_{pcDaily}$ containing, respectively, the mean of water readings in C_{daily} and $C_{pcDaily}$, for each household.

3.4 Matrix of users' and households features

Two sample datasets X_{daily} and $X_{pcDaily}$ were built, respectively for the users whose consumption is included in Y_{daily} and $Y_{pcDaily}$. Each tuple of the datasets has M = 26 user and house features (see Table **1**) associated to either Y_{daily} or $Y_{pcDaily}$.

The processed datasets X_{daily} and $X_{pcDaily}$ consisted, respectively, of N = and N' = tuples, one for each user satisfying the pre-processing conditions.

4. TESTING AND VALIDATION

4.1 Feature selection and feature weighting

The outputs of the feature selection algorithms are represented in Figure 1, where the user and house features are listed on the y-axis and the color indicates the selection frequency of each feature over different algorithms runs. Each feature extraction algorithm was run 3 times: both X_{daily} and $X_{pcDaily}$ were split into three subsets of equivalent size, and each run considered two thirds of the dataset for feature extraction calibration and the remaining third for predictive modeling validation (Section 4.3). Dark colored features in Figure 1 are the most relevant as they are always selected across the different algorithms runs, while their relevance decreases moving towards gray and white tones. The results of the two figures appear to be quite consistent: the number of household's occupants seems to be the most relevant factor impacting average daily residential water consumption Y_{daily} (left part of the figure), as its selection frequency is higher

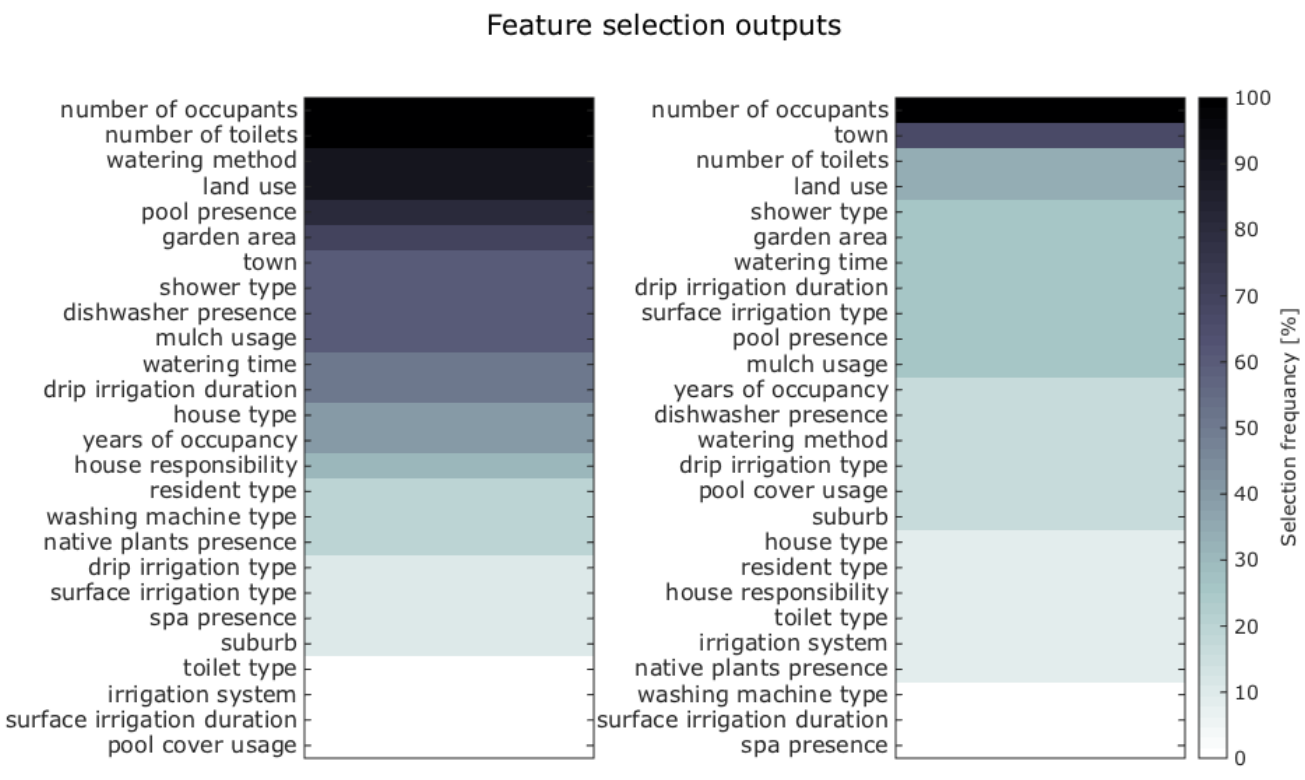

Figure 1. Selection frequency of candidate features over multiple feature selection algorithms runs. Considered predictant: Y_{daily} (left) and $Y_{pcDaily}$ (right).

than 80%; the number of toilets, the method used for irrigation, the presence of a pool and the land use destination are then ranked in the subsequent positions, with a decreasing selection frequency, but still higher than 60%. In addition, the geographical position, expressed by the "town" attribute, is also considered relevant in explaining the average per-capita daily water consumption $Y_{pcDaily}$ (right part of the figure). However, the results obtained considering the average per-capita daily water consumption as predictant enforce the relevancy of the number of house occupants as main driver of water consumption, since its selection frequency is 100%, while all the other candidate variables do not achieve a selection frequency higher than 70%.

Figure 2 shows the results obtained by running the feature weighting algorithms on X_{daily} and $X_{pcDaily}$, respectively. Again, the features are reported on the y-axis, while the x-axis represents the different feature weighting algorithms runs. Colors represent the positions of each feature in the weighting ranking: features with dark color were given higher weights by the algorithms, meaning they are considered relevant in explaining the output variable, while lighter features are associated to lower weights (*i.e.*, less relevant). The two feature weighting algorithms produce consistent results, which are also coherent with the ones obtained by the feature selection algorithms, at least for the majority of the top-ranked features. The results confirm the existence of clear and strong relationships between the extracted features and the corresponding water consumption profiles.

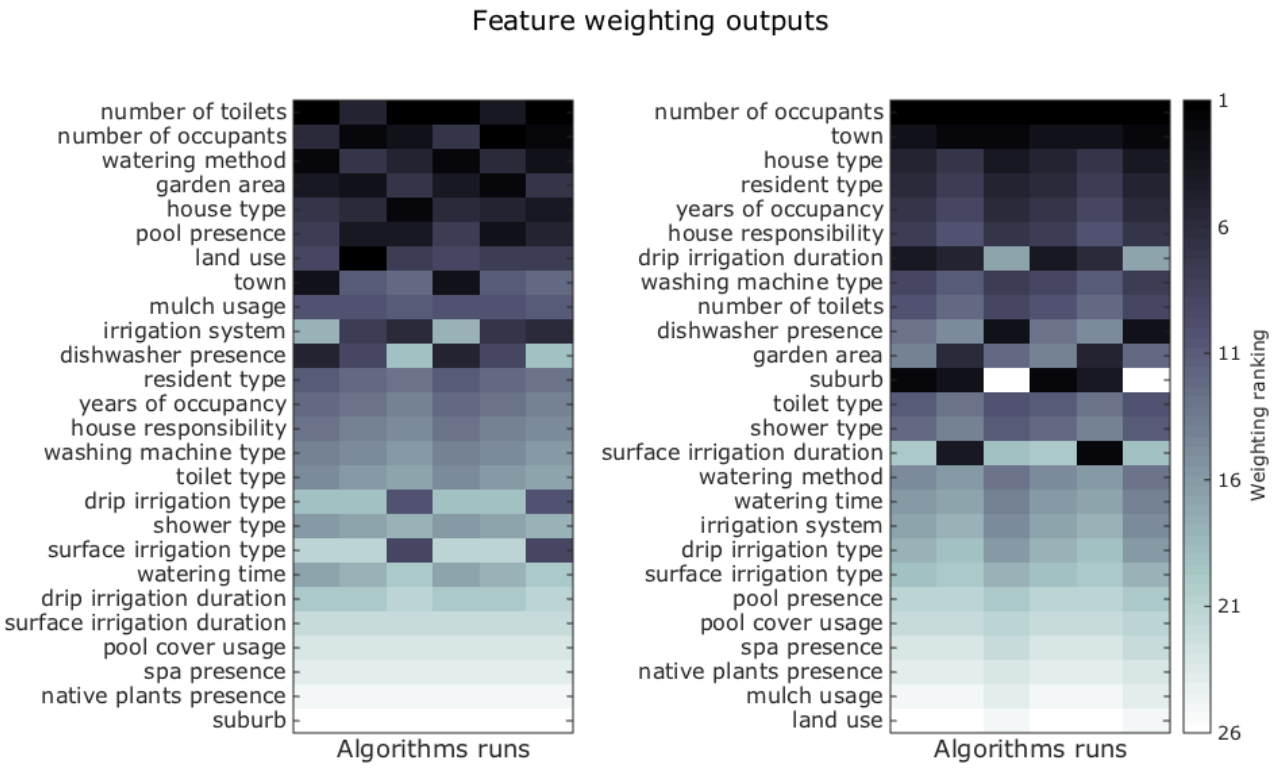

Figure 2. Weighting ranking of candidate features over multiple feature selection algorithms runs. Considered predictant: Y_{daily} (left) and $Y_{pcDaily}$ (right).

4.2 INTERPRETATION OF THE FEATURE EXTRACTION RESULTS

The set of top-ranked features extracted in the previous section is analyzed in this section to better understand the underlying relationships between them and the water consumption profiles.

4.2.1 OCCUPANTS

The first considered feature is the number of occupants of the house, which is always ranked in the first position by all the algorithms. As shown in Figure 3, not surprisingly the median daily water consumption increases with the number of occupants. Yet, the median per-capita consumption decreases with the increasing of the number of occupants. The reason for that can be twofold: the first reason might be that some end-uses represent a sort of fixed-cost, which is shared among the occupants. For example, the water used for irrigation or in a pool is shared among the occupants and,

therefore, the individual cost (*i.e.*, consumption) decreases for increasing number of inhabitants. The second reason might be that when the number of occupants increases, some kind of economies of scale and social pressure develop. As a consequence, water use is better balanced among the inhabitants and wastes are less frequent (Beal *et al.*, 2011).

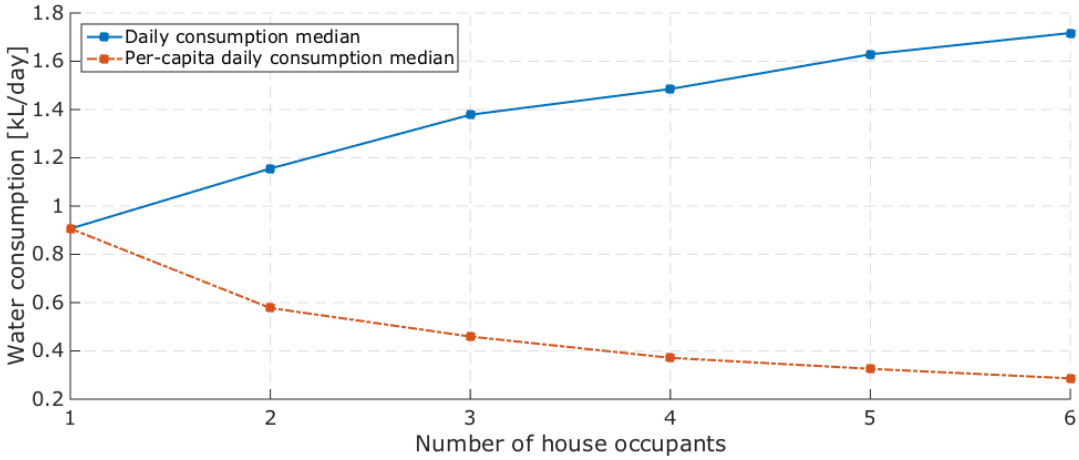

Figure 3. Median daily water consumption and median per-capita daily water consumption for different numbers of house occupants.

4.2.2 TOILET NUMBER

Figure 4 analyzes the number of toilets, where both the median daily and median daily per-capita water consumption level increase with the number of toilets in the house. Since the number of toilets generally increases with the size of the house (and thus with the number of household's occupants), it is reasonable that the daily water consumption increases with the number of toilets. In contrast with the previous case, also the median per-capita consumption increases, probably because with a higher number of toilets there is less "competition" for using the resource (*i.e.*, the toilet).

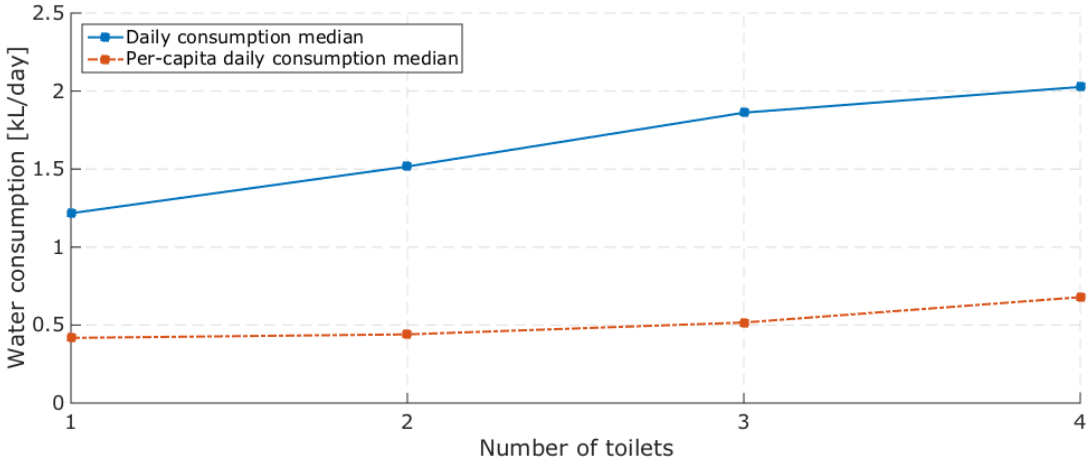

Figure 4. Median daily water consumption and median per-capita daily water consumption for houses with different number of toilets.

4.2.3 IRRIGATION

The relationship between water consumption and the type of irrigation is shown in Figure 5: households where irrigation is performed by hand consume (on average) less water than those houses where irrigation is performed with automatic irrigation systems or both by hand and automatically. This evidence can be explained by relating the water consumption levels to the area of the garden to be irrigated (right y-axis). Houses equipped with automatic irrigation systems generally have a wide garden and high water consumption for irrigation. On the contrary, small gardens are irrigated by hand, resulting in a lower consumption. Reasonably, in houses with a medium-size garden and medium consumption levels, irrigation can be either manual or automatic.

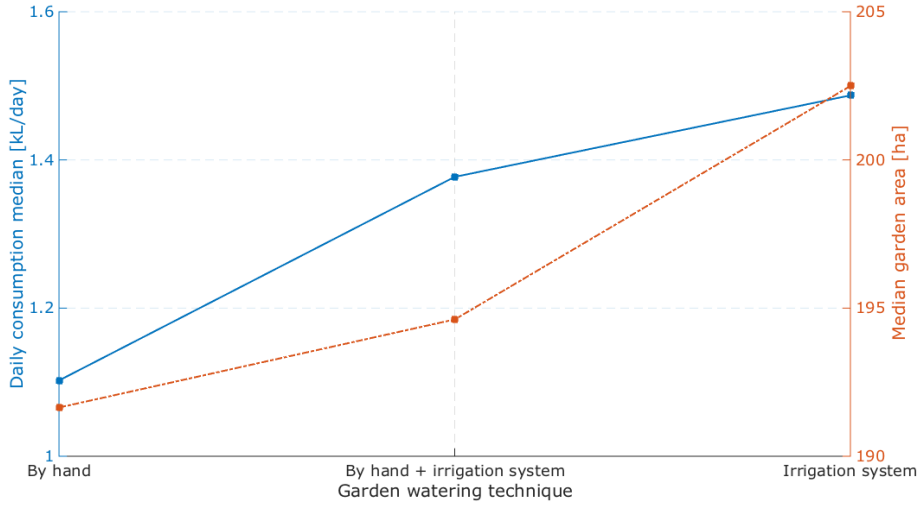

Figure 5. Median daily water consumption and median garden area for households adopting different irrigation techniques.

4.2.4 TYPE OF HOUSE

Figure 6 shows how the consumption level increases with the size of the house. This phenomenon can be probably explained as bigger houses generally are occupied by a higher number of inhabitants and, also, they have a higher number of toilets or very likely larger gardens. In turn, the per-capita water consumption flattens for the same reasons previously discussed about the relationship between the number of occupants and their associated per-capita consumption.

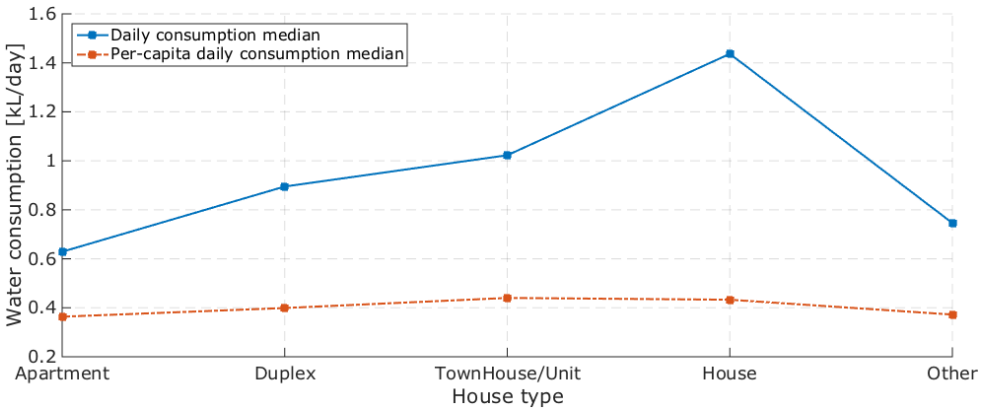

Figure 6. Median daily water consumption and median garden area for different types of house

4.3 FORECASTING USERS' WATER CONSUMPTION PROFILE

The second step of our procedure aims at identifying a model having the features extracted in the previous section as input, and the predicted water consumption profile of the users as output. This second step is fundamental in order to properly support the design of Water Demand Management Strategies, as well as assess their effectiveness. Indeed, the first step of feature extraction provides indications of potentially relevant water consumption drivers, thus supporting the empirical design of WDMS with a description of the status quo of users' behaviors. In turn, a predictive model able to forecast consumers' profile based on relevant attributes enables quantifying changes in household water consumption due to modifications in natural and socio-demo-psycho-graphic drivers, thus supporting utilities and planners by anticipating the effect of WDMS.

In particular, working on low-resolution consumption data, our model allows classifying users to the three consumption profiles introduced in Section 3.3, namely low-, medium-, high-consumers. Among the available data-driven models, we employed Naive Bayes Classifier and Decision Tree algorithm (see Section 2.2) which are particularly suitable for these classification experiments. In order to minimize the risk of overfitting the model over the calibration data, we run a k-fold cross-validation, with k=3, by randomly splitting the dataset into k mutually exclusive subsets of equivalent size. Each time the predictive model is validated on one of the k folds and calibrated using the remaining k–1 folds, on which the feature extraction algorithms are run. Table **2**, Table **3** and Figure 7 report the resulting average models accuracy across the k-fold cross-validation, measured in terms of percentage of correct assignments of users on the basis of their features to their actual consumption profile. Results show that both the models allow attaining a sufficiently good accuracy in predicting the consumption profiles of the users, both when users are classified according to the total consumption of their house or the per-capita consumption level. Moreover, although Figure 7 shows that the prediction accuracy slightly varies when the number of features considered in the model increases, feature extraction algorithms succeeded in identifying the smallest subset of most relevant features, allowing for a sufficient level of prediction accuracy. The proposed method hence shows the potential to effectively capture urban water demand variability with respect to users psychographics and house characteristics data, thus representing promising decision-aiding tools for water utilities and urban planners.

Table 2. Naive Bayes Classifier prediction accuracy based on feature selection (FS) algorithms outputs

FS algorithm	AVERAGE NAIVE BAYES CLASSIFIER ACCURACY ON Y_{daily} [%]	AVERAGE NAIVE BAYES CLASSIFIER ACCURACY ON $Y_{pcDaily}$ [%]
FCBF	62.11 ± 2.80	80.45 ± 7.75
CFS	63.22 ± 2.91	80.45 ± 7.75
BLOGREG	62.48 ± 3.83	80.45 ± 7.75
SBMLR	63.03 ± 3.11	80.45 ± 7.75

Table 3. J48 Decision Tree prediction accuracy based on feature selection (FS) algorithms outputs.

FS algorithm	AVERAGE J48 DECISION TREE ACCURACY ON Y_{daily} [%]	AVERAGE J48 DECISION TREE ACCURACY ON $Y_{pcDaily}$ [%]
FCBF	63.46 ± 1.99	80.64 ± 7.89
CFS	64.94 ± 2.71	80.64 ± 7.89
BLOGREG	61.92 ± 0.85	80.00 ± 7.35
SBMLR	62.91 ± 1,72	81.15 ± 7.84

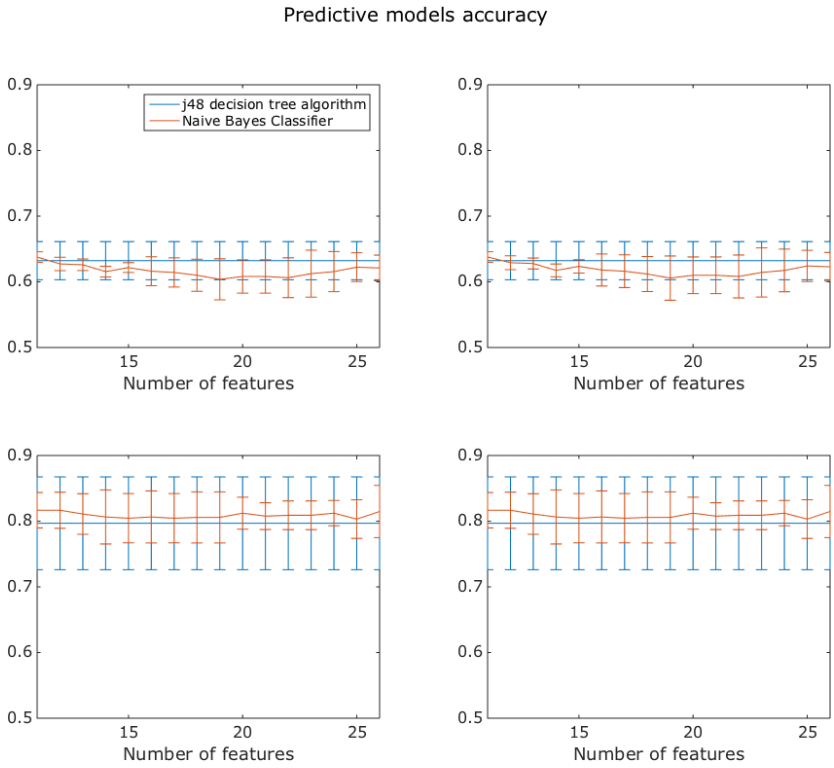

Figure 7. Predictive models accuracy based on feature weighting (FW) algorithms outputs. The following FW algorithm – predictant are represented: Y_{daily} and Information gain FW (top-left), $Y_{pcDaily}$ and Information gain FW (top-right), Y_{daily} and Chi-square score FW (bottom-left), $Y_{pcDaily}$ and Chi-square score FW (bottom-right).

5. CONCLUSIONS

A novel approach based on *feature extraction* techniques to model the single-user consumption behavior at the household level has been presented in this paper. A two-step procedure consisting of the extraction of the most relevant determinants of users' consumption profiles and the identification of a predictive model of water consumers' profile was proposed and tested against a dataset containing low-resolution water consumption records associated with a variety of demographic and psychographic users' data collected within the *H2ome Smart* project, in Western Australia.

Results show overall consistency among the *feature extraction* techniques applied. The analysis of the results allows understanding the relationships between the selected features and the consumption profiles, demonstrating the suitability of such techniques as tools for capturing the influence of candidate determinants on residential water consumption. The development of predictive models of users' behavior attains sufficiently high accuracy in predicting the household water consumption as a function of the user features, which constitutes essential information to support residential water demand management strategies.

Further analysis will focus on assessing the robustness of these results and test the influence of the different steps of the proposed method on the overall quality. For instance, preliminary tests show that the clustering technique used for the construction of the users' consumption profiles impacts on the final results of the predictive model. Moreover, we will assess how the overall procedure accuracy might vary when moving from low-resolution billed data on water consumption to high-resolution smart-metered data, which would allow the definition of more detailed user profiles on the basis of disaggregated end-use patterns. Finally, since the psychographic users data and the house characteristics were collected via survey with no guarantees that all the relevant determinants of users' behaviors are observed, the entire user profiling process would benefit from the use of alternative methods for direct interaction with the users for data gathering as well as for providing personalized feedbacks.

ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 619172 (SmartH2O: an ICT Platform to leverage on Social Computing for the efficient management of Water Consumption).

REFERENCES

Anda, M., Le Gay Brereton, F., Brennan, J. and Paskett, E. (2013). Smart metering infrastructure for residential water efficiency: Results of a trial in a behavioural change program in Perth, Western Australia. In: *Proceedings of the First International Conference on Information and Communication Technologies for Sustainability*, 14 - 16 February, Zurich, Switzerland.

Beal, C., Stewart, R. A., Huang, T., and Rey, E. (2011). SEQ residential end use study. *Journal of the Australian Water Association*, *38*(1), 80-84.

Bennett, C., Stewart, R. A., & Beal, C. D. (2013). ANN-based residential water end-use demand forecasting model. *Expert Systems with Applications*, *40*(4), 1014-1023.

Cawley, G. C., Talbot, N. L., and Girolami, M. (2007). Sparse multinomial logistic regression via bayesian l1 regularisation. *Advances in neural information processing systems*, *19*, 209.

Collins, R., Kristensen, P., and Thyssen, N. (2009). Water resources across Europe-confronting water scarcity and drought, Office for Official Publications of the European Communities.

Cosgrove, C. E., and Cosgrove, W. J. (2012). The Dynamics of Global Water Futures: Driving Forces 2011-2050, The United Nations World Water Development Report, vol. 2, UNESCO.

Cover, T. M., and Thomas, J. A. (2012). *Elements of information theory*. John Wiley and Sons.

Duda, R. O., and Hart, P. E. (1973). Pattern classification and scene analysis. *A Wiley Interscience Publication, John Wiley and Sons, Inc*.

Fielding, K. S., Spinks, A., Russell, S., McCrea, R., Stewart, R., and Gardner, J. (2013). An experimental test of voluntary strategies to promote urban water demand management. *Journal of environmental management*, *114*, 343-351.

Fox, C., McIntosh, B. S., and Jeffrey, P. (2009). Classifying households for water demand forecasting using physical property characteristics. *Land Use Policy*, *26*(3), 558-568.

Galelli, S., Humphrey, G., Maier, H., Castelletti, A., Dandy, G., and Gibbs, M. (2014). An evaluation framework for input variable selection algorithms for environmental data-driven models. Environmental Modelling & Software, 62, 33–51.

Galelli, S., and Castelletti, A. (2013). Tree-based iterative input variable selection for hydrological modeling. *Water Resources Research*, *49*(7), 4295-4310.

Gato-Trinidad, S., Jayasuriya, N., and Roberts, P. (2011). Understanding urban residential end uses of water. *Water Science and Technology*, *64*(1), 36-42.

Elements of information theory, P., D. Haasz, C. Henges-Jeck, V. Srinivas, G. Wolff, K. Cushing, and A. Mann (2003). Waste not, want not: The potential for urban water conservation in California, Pacific Institute for Studies in Development, Environment, and Security Oakland, CA.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, *46*(1-3), 389-422.

Guyon, I., and Elisseeff, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research*, *3*, 1157-1182.

Jorgensen, B., Graymore, M., and O'Toole, K. (2009). Household water use behavior: An integrated model. *Journal of environmental management*,*91*(1), 227-236.

Kenny, J. F., Barber, N. L., Hutson, S. S., Linsey, K. S., Lovelace, J. K., and Maupin, M. A. (2009). Estimated use of water in the United States in 2005, US Geological Survey Reston, VA.

Liu, H., and Setiono, R. (1995). Chi2: Feature selection and discretization of numeric attributes. In *2012 IEEE 24th International Conference on Tools with Artificial Intelligence* (pp. 388-388). IEEE Computer Society.

Maier, H. R., and Dandy, G. C. (2000). Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. *Environmental modelling and software*, *15*(1), 101-124.

Makki, A. A., Stewart, R. A., Panuwatwanich, K., and Beal, C. (2013). Revealing the determinants of shower water end use consumption: enabling better targeted urban water conservation strategies. *Journal of Cleaner Production*, 60, 129-146.

McDonald, R. I., Weber, K., Padowski, J., Flörke, M., Schneider, C., Green, P. A., Gleeson, T., Eckman, S., Lehner, B., Balk, D., and Montgomery, M. (2014). Water on an urban planet: Urbanization and the reach of urban water infrastructure. *Global Environmental Change*, *27*, 96-105.

Olmstead, S. M., Michael Hanemann, W., and Stavins, R. N. (2007). Water demand under alternative price structures. *Journal of Environmental Economics and Management*, *54*(2), 181-198.

Quinlan, J. R. (1993). *C4. 5: programs for machine learning* (Vol. 1). Morgan Kaufmann.

Wichman, C. J. (2014). Perceived price in residential water demand: Evidence from a natural experiment. *Journal of Economic Behavior and Organization*.

Yu, L. and Liu, H. (2003). Feature selection for high-dimensional data: A fast correlation-based filter solution. In *T. Fawcett and N. Mishra, editors, Proceedings of the 20th International Conference on Machine Learning (ICML-03),* August 21-24, 2003, pages 856–863, Washington, D.C., 2003. Morgan Kaufmann.

Zhao, Z., Morstatter, F., Sharma, S., Alelyani, S., Anand, A., and Liu, H. (2010). Advancing feature selection research. *ASU Feature Selection Repository*.