FIRST USER BEHAVIOUR MODELS

SmartH2O algorithms for water end-use disaggregation and user modelling

SmartH2O

Project FP7-ICT-619172

Deliverable D3.2 WP3

Deliverable
Version 2.0 – 2 June 2015
Document. ref.:
D32.SUPSI.WP3.V2.1

Programme Name:	ICT
Project Number:	619172
Project Title:	SmartH2O
Partners:	Coordinator: SUPSI Contractors: POLIMI, UoM, SETMOB, EIPCM, TWUL, SES, MOONSUB
Document Number:	smarth2o. D32.SUPSI.WP3.V2.1
Work-Package:	WP3
Deliverable Type:	Document
Contractual Date of Delivery:	31 December 2014
Actual Date of Delivery:	31 December 2014
Title of Document:	First user behaviour models
Author(s):	Dario Piga, Andrea Cominola, Matteo Giuliani, Andrea Castelletti, Andrea Emilio Rizzoli, Alessandro Facchini, Jasminko Novak, Isabel Micheel.
Approval of this report	Submitted for review.
Summary of this report:	This report presents a set of algorithms to derive, from metered water consumption data and socio-psychographic features of the consumers, models describing the users' consumption behavior. Specifically, the deliverable reports: two novel algorithms for decomposing high-resolution water flow data into end use categories, the application of several machine learning and data-mining algorithms to the water user modeling problem. The discussed methodologies are validated against datasets available in the literature.
History:	n/a
Keyword List:	User behavioural models, user profiling, water end use disaggregation.
Availability	This report is restricted

This work is partially funded by the EU under grant ICT-FP7-619172

Document History

Version	Date	Reason	Revised by
1.0	22/12/2014	First submission of the deliverable	A.E. Rizzoli
2.0	2/6/2015	Section 3.6 has been added, reporting on the application of end-use disaggregation on water data at different resolutions.	Andrea Cominola, Matteo Giuliani, Ahmed El Sahaf, Dario Piga, Andrea Castelletti, Andrea E. Rizzoli
2.1	28/7/2015	Broken links fixed, Figure 10 fixed	A.E. Rizzoli

Disclaimer

This document contains confidential information in the form of the SmartH2O project findings, work and products and its use is strictly regulated by the SmartH2O Consortium Agreement and by Contract no. FP7- ICT-619172.

Neither the SmartH2O Consortium nor any of its officers, employees or agents shall be responsible or liable in negligence or otherwise howsoever in respect of any inaccuracy or omission herein.

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7-ICT-2013-11) under grant agreement n° 619172.

The contents of this document are the sole responsibility of the SmartH2O consortium and can in no way be taken to reflect the views of the European Union.

Table of Contents

Executive Summary

The design and the assessment of policies and strategies for urban water demand management at the household level require to build user models that quantitatively describe how water demand is influenced and varies in relation to exogenous determinants (e.g., climate conditions), socio-psychographic features (e.g., age, income, household features), social pressure, water restrictions, water tariffs, and reciprocal influence of these factors. These models should be then used to foresee the consumers' response to different water demand management scenarios and, if models estimating different water demand for different types of users are built, consumer-tailored water demand management strategies (WDMS) can be proposed in order to effectively modify the consumers' attitude for pursuing a water saving behavior.

This deliverable presents a set of algorithms to derive, directly from metered water consumption data and consumers' socio-psychographic data retrieved from the gamified applications, mathematical models describing the users' consumption behavior. Specifically, this deliverable addresses the following three main steps for user behavioral modeling:

- **water end-use characterization**, which aims at decomposing the aggregate (i.e., whole household) high-resolution water flow data collected from a single measurement point into water end use categories (e.g., shower, toilet flush, dishwasher), in order to understand how, when and where water is used. *Two novel disaggregation algorithms, developed in the SmartH2O project, are discussed*.
- **variable selection**, which aims at identifying (based on water consumption data, a set of socio-psychographic features and other external factors) the main drivers influencing water consumption at an individual (household) level. *Several feature extraction algorithms have been used to tackle the variable selection problem*.
- **model learning,** which aims at constructing a model that allows to predict the consumption profile of water users as a function of the determinants identified in the variable selection step. *Bayesian Regressor and Decision Tree classifiers are used to tackle the model learning problem*.

At this stage of the project, the user is modeled as an autonomous entity, thus social interactions and influence/mimicking mechanisms are not considered in the modeling phase. These mechanisms will be considered in Tasks T3.3 and T3.4 of Work Package 3, in order to develop an agent-based model that will be used to simulate whole districts of water users and to understand how some user types (leaders/influencers) can stimulate a behavioural change on other users.

1. Introduction

1.1 Insights on residential water management

Individual and collective behavioural responses to different water conservation policies acting on the demand side of residential water consumption (the so called Water Demand Management Strategies, WDMS) might significantly vary within the same urban context depending on economic drivers as well as socio-psychological determinants. Therefore, in order to design and to assess the effectiveness of alternative WDM policies, it is essential to build models that quantitatively describe how the water demand is influenced and varies in relation to exogenous determinants (e.g., climate conditions), socio-psychographic features (e.g., age, income, household features), social pressure, water restrictions, water tariffs, and reciprocal influence.

High spatial (household) and temporal (up to few seconds) resolution water consumption data gathered by smart meters provide a detailed user consumption profile. This enables an accurate characterization of the water consumption share and patterns of end-uses, which, in turn, constitute the basis for the mathematical modeling of individual and collective user behaviors. In summary, residential water management comprises the sequential phases represented in the flowchart in Figure 1, namely: data gathering, water end-use characterization, user modeling and WDMS design, implementation and assessment in terms of water savings. Within the SmartH2O project, Work Package 3 addresses the second and the third block of the flowchart in Figure 1 (i.e., water end-use characterization and user modeling), which are briefly described in the following:

- The *water end-use characterization phase* aims at decomposing the aggregate (i.e., whole household) water consumption data collected from a single measurement point into water end use categories, to understand how, when and where water is used. Beside using this information for building mathematical models of the user behavior, the generated knowledge can be also directly provided to customers, municipalities and water utilities, so that:
  i. household's components have a detailed knowledge on their water usage. For instance, through the SmartH2O platform, customers can log into a web page to view their hourly consumption, as well as charts on their water end-use patterns across major end-use categories (e.g., washing machine, toilet, shower, irrigation) and they can be alerted of occurring consumption anomalies (e.g., leak events). Furthermore, personalized hints for reducing water consumption can be directly delivered by the municipality and the water utility;
  ii. customers can be informed on potential savings in differing the usage of some water using appliances (e.g., washing machine and dishwasher) to peak-off hours, or in replacing low-efficient appliances into high-efficient ones, and personalized rewards schemes can be then proposed to stimulate customers to adopt water saving actions.
- The *user modeling phase* aims at identifying the drivers influencing the water consumption, and thus at building mathematical models to predict water demand at the individual (household) level based on socio-psychographic features of the consumers and on exogenous variables (e.g., climate conditions, social pressure, awareness campaigns). Thus, the inputs of the *user behavioural models* are the user attributes and the exogenous variables, while the prediction of the household water consumption is the resulting output.
  Depending on the type of information included among the input set and the structure of the model, two groups of user models can be identified among the existing works: (i) the *single-user models,* which describe the user's consumption behavior considering the user as an isolated entity, thus not including social

interactions and influence/mimicking mechanisms in the inputs; and (ii) the *multi-user models* that include dynamic interactions among users. In this deliverable, only the algorithms for modeling the *single-user behaviour* are considered, while the development and implementation of *multi-user models* will be the subject of the next deliverables D3.3 and D3.4.

1.2 Deliverable contribution

In this deliverable, we present two novel algorithms for end use characterization and a two-stage data-mining approach to model single-user consumption behaviors at the household level. The developed disaggregation and user profiling algorithms are tested against data available in the literature or gathered in previous studies focused on water user behavioural modeling. Based on the *single-user models* and on the behavioral data collected from questionnaires and the social gamified platform developed in WP4, *multi-user models* including the dynamic interactions among users will be eventually developed (as part of the next deliverables D3.3 and D3.4) by exploiting agent-based modeling platforms.

It is worth mentioning that the final user models should also be able to describe the future consumers' behavior in face of water price and rewards changes. The latter is the main goal of Work Package 5 ("Saving water by dynamic water pricing"), where econometric models of water demand under new pricing and reward policies will be developed, and eventually integrated with the consumer behavioural models developed in WP3.

1.3 Deliverable outline

The deliverable is organized as follows:

- Section 2 provides a review on the state-of-the-art algorithms for water end-use disaggregation and on previous studies on water user behavioural modeling.
- Section 3 describes two novel algorithms for water end-use characterization developed within the SmartH2O project. The performance of the developed algorithms are tested against high-resolution energy consumption data available in the literature and against water consumption data gathered in the WEEP (Water End Use and Efficiency Project) project [Heinrich07], a study which was conducted in New Zealand in 2005-2007.
- Section 4 describes a novel approach to model the single-user consumption behavior at the household level. The approach is based on a two-step procedure: (*i*) identification of the most relevant determinants of users' consumption profiles; (*ii*) construction of a model that allows predicting the consumption profile of water users as a function of the determinants identified in the previous step. Since a database with water consumption data and the associated users' features is not available yet, the *H2ome Smart* project dataset [Anda12] was used to assess the performance of the proposed user modeling algorithm.

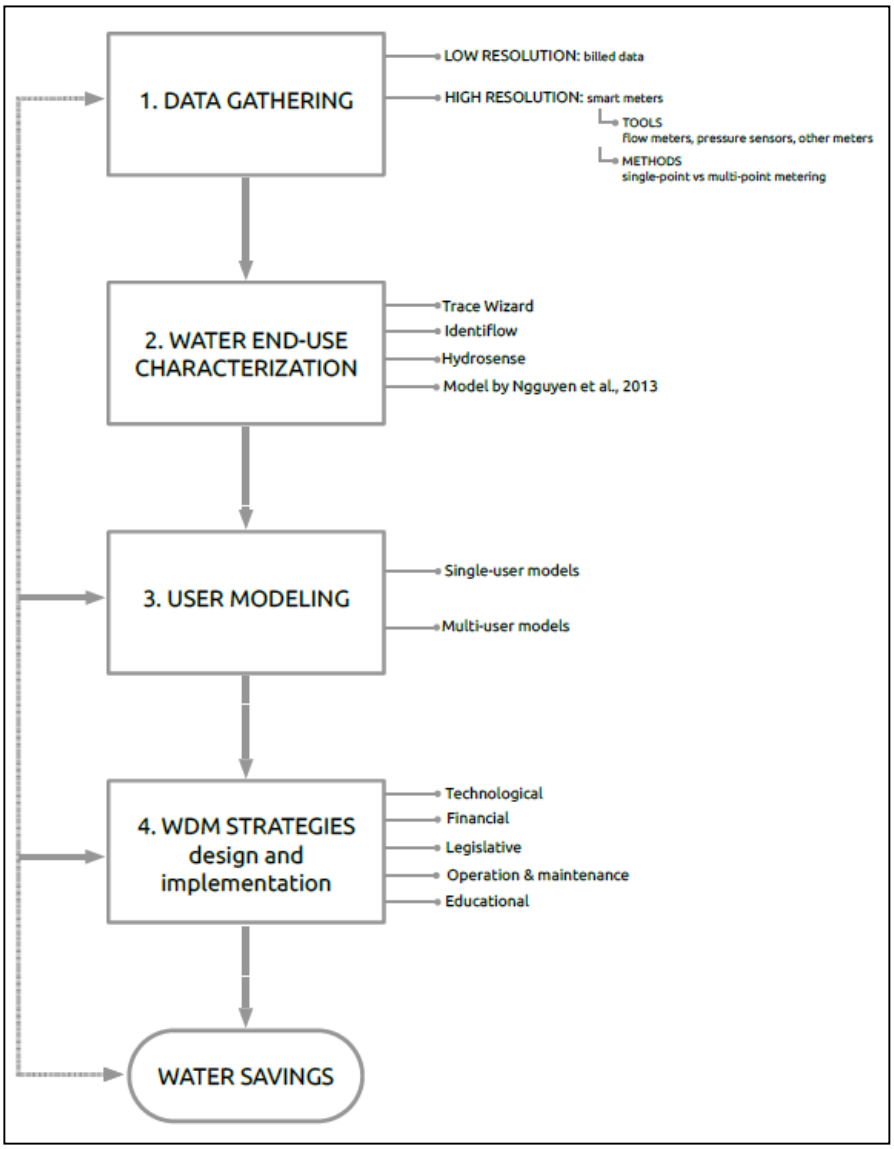

Figure 1: Flowchart presenting the sequential phases of Water Demand Management at the residential scale.

2. State of the art on water end-use characterization and user modeling

2.1 Disaggregation algorithms for end-use breakdown characterization

Several studies aiming at disaggregating water flow data collected from high-resolution smart meters into water end use categories have been conducted in the last two decades (see Table 1 for a summary of recent water end use studies). The summary reported in Table 1 shows that three different algorithms were basically used for water end-use disaggregation in past studies, namely: *Trace Wizard*® (a commercial flow trace analysis toolkit developed by Aquacraft, Inc.); *Identiflow*® (a tool developed by WRc, a research organization based in United Kingdom) and *HydroSense* (a water disaggregation approach originally proposed in [Froehlich09]). Strengths and weaknesses of each approach are briefly described in the next paragraphs.

2.1.1 Trace Wizard®

Trace Wizard applies a decision tree algorithm that interprets data based on some basic flow boundary conditions (e.g., minimum/maximum volume, peak flow rate, duration range, etc.). The disaggregation process requires completing the following tasks:

- Conduct a detailed water appliance/fixture stock inventory audit for each household to determine the efficiency rating of each household appliance/fixture;
- Household's occupants should complete a diary of water use events over a one-week period to gain information on their water use habits;
- Analysts use water audits, diaries, and sample flow trace data for each household to create specific templates that serve to match water end uses patterns based on some basic flow boundary conditions.
- Based on the developed templates, stock survey audit, diary information and analysts' experience, the individual water end uses are disaggregated.

Due to the human resource requirement, the overall process is extremely time and resource intensive, and it relies on the analyst's experience in understanding flow signatures. Furthermore, it has been observed that the prediction accuracy of *Trace Wizard* is significantly reduced when more than two events occur concurrently [Mayer99].

2.1.2 Identiflow®

Similar to Trace Wizard, *Identiflow* is based on a decision tree algorithm to disaggregate the total water consumption into end-use categories. *Identiflow* uses fixed physical features of various water-using devices (e.g., volume, flow rate, duration, etc.) to make different decisions for categorization. Although *Identiflow* has shown better performance than *Trace Wizard*, its classification accuracy strongly depends on the physical features used to describe each fixture/appliance, and two different water events could be placed into the same category if they enjoy similar physical characteristics. Similarly, the accuracy of *Identiflow* significantly decreases if old water-using appliances are replaced by modern ones, whose physical characteristics might be completely different compared to the old ones.

2.1.3 HydroSense

HydroSense is based on a continuous analysis of the pressure within a home water infrastructure. Water end use events are classified based on the unique pressure waves that propagate to the sensors when valves are opened or closed. Specifically, when a valve is

opened or closed (be it a bathroom faucet or a mechanical valve in a dishwasher), a pressure change occurs and a pressure wave is generated in the plumbing system. Based on the pressure wave (which depends on the valve type and its location), water end-use events are classified by using advanced pattern matching algorithms and Bayesian probabilistic models. The main disadvantage of the *HydroSense* system is due to the large number of sensors, which should be connected to each appliance/fixture to build the dataset for model calibration. Therefore, since a distributed sensing network is required for calibration, the approach cannot be considered totally non-intrusive and its portability to a wide urban context would entail costs and privacy issues. Furthermore, the installation of a number of sensors can hardly be accepted by house occupants.

2.1.4 SEQREUS algorithm

Another algorithm to disaggregate water flow data into end use categories has been developed within the SEQREUS project [Beal11]. The SEQREUS approach makes use of machine learning tools, i.e., *Hidden Markov Models* (HMMs) and the *Dynamic Time Warping* (DTW) algorithm, and it is basically used to refine the results given by other disaggregation algorithms (e.g., Trace Wizard). Specifically, SEQREUS approach works as follows:

1. Water flow data are first broken down into nine different water end use categories (namely, shower, tap, dishwasher, clothes washer, toilet, bathtub, irrigation, leak and inconclusive) using Trace Wizard*,* and the disaggregated data are used for training Hidden Markov Models (step 2);
2. Based on the training data (obtained at Step 1), compute eighty different *Hidden Markov Models* describing the different water end use categories (excluding the inconclusive event);
3. Physical characteristics for each end use event category are used to refine the estimate given by the HHMs (e.g., any shower event with a volume less than 7 liters or any bathtub event with duration less than 4 minutes is placed in the inconclusive event for future analysis);
4. Use DTW algorithm to determine if any event in the inconclusive dataset is similar to an event in the clothes washer or dishwasher set. DTW is employed because of its ability to search patterns existing in series which have a clearly defined patterns (e.g., the clothes washer and dishwasher cycles);
5. Use time of day probability to assign inconclusive events to an end-use category.

2.1.5 Electric energy disaggregation

Unfortunately, none of the disaggregation algorithms described so far is completely automatic, but all do require some level of interaction with the user (intrusive monitoring).
Nevertheless, in the field of electric energy, there is a rich literature on automatic disaggregation methods (known as *Non Intrusive Appliance Load Monitoring* (NIALM) algorithms) aiming at decomposing the aggregate household energy consumption data collected from a single measurement point into device-level consumption data without requiring a limited interaction with the user. The first algorithm for NIALM was proposed by Hart in 1992 [Hart92]. Hart's approach is based on the segmentation of the aggregate power signal into successive steps, which are then matched to the appliance signatures. However, this method is not able to detect multistate appliances and it is neither able to decompose power signals made of simultaneous on/off events on multiple appliances. Since Hart's contribution, the problem of Nonintrusive Appliance Load Monitoring has been extensively studied in the literature. The survey papers [Zoha12] and [Zeifman11] give a complete review on the state-of-the-art of NILAM methods, which can be classified into two main categories: optimization based and machine learning based approaches. The methods based on sparse coding [Figueiredo13, Dong13] and integer programming [Suzuki08,Camier13] belong the first category, while the approaches discussed in [Srinivasan06, Zia11, Parson12, Johnson13], which make use of Hidden Markov Models and Artificial Neural Networks belong to the second category.

As already mentioned, the energy disaggregation algorithms require a limited interaction with the user (i.e., a monitoring period less intrusive w.r.t. the one required by the methods for water end-use characterization). In order to transfer this property also to water end-use characterization, two novel disaggregation algorithms (an optimization based method and a machine learning based method) have been developed within the SmartH2O project. These algorithms, described in Section 3, can be used to disaggregate both water and energy data and they are able to accurately decompose multiple overlapping device signals, thus overcoming one of the major drawback of the algorithms commonly used for water disaggregation.

Table 1 Summary of residential end use studies conducted in the last 15 years [Nguyen13]

Study	Location	Number of metered household	Sample regime	Reading interval	Tool used for disaggregation	Reference
1998 - REUWS	North America	1188	2 weeks in summer and 2 weeks in winter	10 s	*Trace Wizard*®	[Mayer99]
2003 – Smart metering project in UK	United Kingdom	250	ongoing	1 s	*Identiflow*®	[Kowalski05]
2004 – Yarra Valley Water Residential End Use Study	Yarra Valley, Australia	100	2 weeks in summer and 2 weeks in winter	5 s	*Trace Wizard*®	[Roberts05]
2005-2007 Water End Use and Efficiency Project	Auckland region, New Zealand	51	6 months	10 s	*Trace Wizard*®	[Heinrich07]
2005-2010 California Single Family Water Use Efficiency Study	State of California, USA	732	22 months	10 s	*Trace Wizard*®	[DeOreo11]
2009 Albuquerque Single Family Water Use Efficiency and Retrofit Study	Albuquerque, New Mexico, USA	240	2 weeks	10 s	*Trace Wizard*®	[Aquacraft11]
2009-2011 SEQREUS	South East Queensland region, Australia	252	11 weeks in summer and 4 weeks in winter	5 s	*Trace Wizard*®	[Beal11]
2011 - USA University of Washington	Seattle, USA	5	5 weeks	1 ms (readings from pressure sensors)	*HydroSense*	[Froehlich11]

2.2 Water user modeling studies

The user modeling phase (third block in Figure 1) is composed by two steps: (i) *user profiling*, which consists in the identification and selection of significant inputs for the model (i.e., through *variable selection techniques*) and (ii) *model structure identification, parameter calibration and validation*. The studies limiting their extent to the variable selection phase can qualitatively inform water managers, utilities and decision makers about current users' habits and consumption trends, while the ones completing the second phase provide tools which support policy design and decision making processes, allowing a what-if analysis as well as scenario simulation and analysis. The purpose of this section is to classify the state of the art literature about *single-user behavioural modeling* by distinguishing between those studies stopping to variable selection stage and those completing the user behavioural modeling process.

2.2.1 User profiling

A first level of the existing works on single-user behavioural modeling is given by the studies that stop at the very beginning of the user-profiling phase. Such studies simply try to deepen the understanding of the breakdown structure of water end uses (i.e., the ones disaggregated in the water end-use characterization phase), in order to identify consumption patterns and trends [Loh03, Roberts05], and to build a user consumption profile that constitutes the baseline for identifying the most promising areas where conservation efforts may be polarized [Gato11]. Although these studies do not directly consider drivers for water consumption, and thus they cannot be used to design policies acting on users' behavioral drivers, the interpretation of the water end uses and the profiles constitute an essential basis for estimating the savings achievable by acting on the technical side of the problem, e.g., on the efficiency of water using fixtures [Gato11], or in estimating the difference in daily water consumption due to seasonality [Gato11].

Other single-user modelling studies, in turn, push their aim beyond the analysis of end uses and look for correlations between a set of variables belonging to a *specific domain* (e.g., dwelling features domain, economic domain, social domain) and water consumption, thus approaching user profiling through variable selection and assessment with a limited, pre-defined variable set. In [Fox09], for instance, statistical tools (like univariate analysis, multivariate analysis and ANOVA) were applied to assess the relationship between physical characteristics of the dwelling (e.g., number of rooms, type, presence of garden) and water consumption. Although a water demand forecasting model was not developed in [Fox09], finding the interlinks and dependencies between water consumption and household features constitutes an advantage to forecast water demand for new housing development, where socio-demographic and economic information regarding the (future) inhabitants is therefore not available. Other contributions, like [Olmstead07] and [Olmstead09], look for correlations between purely economic factors and water consumption, as water price and incentives, in order to accurately estimate savings due to price-dependent policies or retrofit efficiency campaigns. Unfortunately, the main shortcoming of [Fox09], [Olmstead07] and [Olmstead09] is that they consider only a limited set of variables influencing residential water consumption and thus they not provide an accurate estimate of the actual most impacting variables.

In contrast, many studies attempted to build user profiles that include variables from *different domains*. Some of them consider a variety of drivers, but focus on specific water uses. For instance, [Syme04] built a structural equation model upon physical, socio-demographic, lifestyle and attitude variables to infer their influence on external water use for gardening. Coherently, [Makki13] found that, as a result of a multi-regression model, household makeup characteristics and devices efficiency are the dominant drivers of water consumption for showering. Some other recent works (e.g., [Suero12], [Willis11], and [Talebpour14]) consider

the problem from a holistic perspective, thus linking a variety of key factors to the total household consumption.

2.2.2 User behavioral modeling for water demand forecasting

To the best of the authors' knowledge, only few pilot studies and experiments have produced prototype models to inform demand management, resulting in the only attempts to complete both the phases of variable selection and behavioral modeling mentioned at the beginning of this section.

In [Gato06], a multi-variable regression approach is used to predict the demand of water end uses from tailored demographic variables. Predictor variables included the number of adults, the number of children less than 12 years of age, and appliance information, such as ownership of a dishwasher, the type of clothes washer and the fraction of dual flush toilets in the household. Significant prediction models were produced for the following water end uses: total internal demand; toilet demand; shower demand; clothes washer demand; dishwasher demand; and tap demand. In [Blokker10], the authors developed a stochastic end-use model based on demographics, end-use category frequency of use, flow duration and event occurrence likelihood to predict water demand patterns at the residential scale and with a high time resolution (1 second), resulting in a tool able to explain large part of the variance for the observed consumption data just based on statistical information on users. More recently, a forecasting model built upon smart metered end-use data gathered during a two-year end-use study in South East Queensland (Australia) has been developed in [Bennet13]. This model couples non-parametric statistical tests and artificial neural networks to: (i) identify key water consumption determinants and (ii) forecast residential water consumption, achieving moderate forecast accuracy levels (R^2 coefficient ranging from 21% to 60% for the diverse water end-use categories).

3. SmartH2O algorithms for end-use disaggregation

This section provides a description of two novel disaggregation algorithms that can be used to decompose both water and energy consumption data into end use categories. The Section is organized as follows: the problem of data disaggregation is formalized in Section 3.1; an optimization based and a machine learning based algorithm are described in Section 3.2 and Section 3.3, respectively. High-resolution energy consumption data available in the literature and water consumption data gathered in the WEEP (Water End Use and Efficiency Project) research [Heinrich07] have been used to assess the performance of the developed disaggregation algorithms (Sections 3.4 - 3.6).

3.1 Problem formulation

Consider the situation where *N* different water-using appliances/fixtures $(\mathrm{L}_1, \ldots, \mathrm{L}_\mathrm{N})$ are available in a house. Each appliance L_i has C_i operating modes and let $\mathrm{B}_\mathrm{i}^{(\mathrm{j})}$ be the water demand of the *i*-th appliance at the *j*-th operating mode (with *j* = 1,…, C_i). The water consumption $\mathrm{y}_\mathrm{i}(\mathrm{t})$ of the *i*-th appliance/fixture at time *t* is then given by:

$$\mathrm{y_i(t)} = \begin{bmatrix} \mathrm{B_i^{(1)}} & \mathrm{B_i^{(2)}} & \ldots & \mathrm{B_i^{(C_i)}} \end{bmatrix} \begin{bmatrix} \mathrm{x_i^{(1)}(t)} \\ \mathrm{x_i^{(2)}(t)} \\ \vdots \\ \mathrm{x_i^{(C_i)}(t)} \end{bmatrix} + e_i(t),$$

with $e_i(t)$ being an error term. The time-varying variables $\mathrm{x_i^{(1)}(t)}, \ldots, \mathrm{x_i^{(C_i)}(t)}$ can be either 0 or 1, and they satisfy the equality constraint $\sum_j^{C_i} \mathrm{x_i^{(j)}(t)} = 1$ (i.e., each water appliance can operate at a single mode at each time instant *t*).

Let $y(t)$ be the aggregate water consumption measured by the smart meter at time *t*, which is given by:

$$\mathrm{y(t)} = \sum_i^N \mathrm{y_i(t)} + e(t),$$

where $e(t)$ is a measurement noise. Given a sequence D_{T_V} of T_V observations of the aggregate water consumption readings $y(t)$ (with *t=1,…,* T_V), our goal is to reconstruct the actual water consumptions $y_i(t)$ (with *t=1,…,* T_V) of each appliance/fixture based on the household aggregate water flow data D_{T_V}.

A training dataset D_{T_E} is assumed to be available. The training set consists of the observations of the water consumption profiles of each appliance/fixture available in the house. An intrusive period is needed to construct the set D_{T_E}. During this period, the patterns of the water consumption of each appliance are observed, and information on time-of-day probability characterizing the usage of each appliance/fixture can be also gathered.

3.2 Optimization based algorithm

The first water disaggregation algorithm developed within the SmartH2O is based on sparse optimization and it is described in this section. The developed algorithm exploits the following assumptions:

- **A1**: A rough knowledge of the water consumption of each appliance/fixture at each operating mode (i.e., the terms $\mathrm{B_i^{(j)}}$) is supposed to be available. For instance, the terms $\mathrm{B_i^{(j)}}$ can be evaluated from the training dataset D_{T_E} through k-means clustering [Likas03].

- **A2**: The water consumption profiles of each appliance/fixture are piecewise constant over time (as it is typical for many residential water-using appliances/fixtures).

The ideas underlying the developed disaggregation algorithms are now described.

3.2.1 Standard least-squares estimate

In order to estimate the water consumption $y_i(t)$ of each appliance/fixture at the time sample *t*, the time varying parameters $x_i^{(j)}(t)$ might be computed by solving the standard least-squares problem:

$$\min_{\substack{x_i^{(1)}(t),\dots,x_i^{(C_i)}(t) \\ t=1,\dots,T_V \\ i=1,\dots,N}} \sum_{t=1}^{T_V} \left(y(t) - \sum_{i=1}^{N} \hat{y}_i(t, x_i) \right)^2, \qquad (1)$$

where $\hat{y}_i(t, x_i)$ denotes the model of the water usage of the *i*-th appliance at time *t*, i.e.,

$$\hat{y}_i(t, x_i) = \begin{bmatrix} \mathrm{B}_\mathrm{i}^{(1)} & \mathrm{B}_\mathrm{i}^{(2)} & \dots & \mathrm{B}_\mathrm{i}^{(\mathrm{C_i})} \end{bmatrix} \begin{bmatrix} \mathrm{x}_\mathrm{i}^{(1)}(\mathrm{t}) \\ \mathrm{x}_\mathrm{i}^{(2)}(\mathrm{t}) \\ \vdots \\ \mathrm{x}_\mathrm{i}^{(\mathrm{C_i})}(\mathrm{t}) \end{bmatrix}$$

Unfortunately, the least-squares optimization problem (1) is an overparametrized problem, since it involves more unknown parameters than measurements. As a consequence, overfitting occurs in computing the time varying parameters $x_i^{(j)}(t)$ through a simple least-squares approach. A possible solution to overcome this problem is to introduce regularization terms (or equivalently penalty terms) terms in (1) in order to:

- enforce each appliance at operating at a single mode at each time instant;
- enforce water usage patterns $\hat{y}_i(t, x_i)$ to be piecewise constant over time, according to assumption **A2**.

3.2.2 Adding regularization

In order to exploit the information that: (i) the parameters $x_i^{(1)}(t), \dots, x_i^{(C_i)}(t)$ can be either *0* or *1*; (ii) each appliance/fixture can only operate at a single mode at each time instant, the following regularized problem can be solved instead of (1):

$$\min_{\substack{x_i^{(1)}(t),\dots,x_i^{(C_i)}(t) \\ t=1,\dots,T_V \\ i=1,\dots,N}} \sum_{t=1}^{T_V} \left(y(t) - \sum_{i=1}^{N} \hat{y}_i(t, x_i) \right)^2 + \gamma_1 \sum_{i=1}^{N} \sum_{t=1}^{T_V} \left\| \begin{bmatrix} x_i^{(1)}(t) \\ x_i^{(2)}(t) \\ \vdots \\ x_i^{(C_i)}(t) \end{bmatrix} \right\|_0, \qquad (2)$$

$$s.t. \;\; x_i^{(j)}(t) \geq 0, \qquad \sum_{j=1}^{C_i} x_i^{(j)}(t) = 1, \quad i = 1, \dots, N; \quad t = 1, \dots, T_V,$$

where $\|\cdot\|_0$ denotes the *0*-norm of a vector (i.e., number of nonzero elements). Note that, on one hand, the second term in the objective function of Problem (2) aims at minimizing the number of nonzero elements in the vector $\left[x_i^{(1)}(t), \dots, x_i^{(C_i)}(t) \right]$. On the other hand, because of the constraint $\sum_{j=1}^{C_i} x_i^{(j)}(t) = 1$, the vector $\left[x_i^{(1)}(t), \dots, x_i^{(C_i)}(t) \right]$ is guaranteed to have at least a nonzero element. The parameter $\gamma_1 \geq 0$ is tuned by the user (for instance through cross validation, see Section 3.2.6) for balancing the tradeoff between minimizing the fitting error (by decreasing the value of γ_1) and minimizing number of the nonzero elements in the vector

$\left[x_i^{(1)}(t), \dots, x_i^{(C_i)}(t)\right]$ (by increasing the value of γ_1). Because of the *0*-norm, Problem (2) is nonconvex, and thus difficult to be solved through numerical optimization solvers available in the literature. Nevertheless, an approximate solution of Problem (2) can be obtained by replacing the 0-norm with the (convex) *1*-norm (i.e., sum of the absolute value of the elements of the vector). Furthermore, the final estimate can be improved by scaling the parameters $\left[x_i^{(1)}(t), \dots, x_i^{(C_i)}(t)\right]$ with nonnegative weights $\left[w_i^{(1)}(t), \dots, w_i^{(C_i)}(t)\right]$. This leads to the following approximation of Problem (2):

$$\min_{\substack{x_i^{(1)}(t),\dots,x_i^{(C_i)}(t) \\ t=1,\dots,T_V \\ i=1,\dots,N}} \sum_{t=1}^{T_V} \left(y(t) - \sum_{i=1}^{N} \hat{y}_i(t, x_i) \right)^2 + \gamma_1 \sum_{i=1}^{N} \sum_{t=1}^{T_V} \left\| \begin{bmatrix} w_i^{(1)}(t) \\ w_i^{(2)}(t) \\ \vdots \\ w_i^{(C_i)}(t) \end{bmatrix} * \begin{bmatrix} x_i^{(1)}(t) \\ x_i^{(2)}(t) \\ \vdots \\ x_i^{(C_i)}(t) \end{bmatrix} \right\|_1, \qquad (3)$$

$$s.t. \;\; x_i^{(j)}(t) \geq 0, \qquad \sum_{j=1}^{C_i} x_i^{(j)}(t) = 1\,, \quad i = 1, \dots, N; \quad t = 1, \dots, T_V,$$

where $*$ denotes the element-wise multiplication. An appropriate choice of the weights $w_i^{(j)}(t)$ is discussed in Section 3.2.4.

3.2.3 Adding regularization to enforce piecewise constant power consumption profiles

In order to improve the estimate given by (3), we might exploit the additional information that the patterns of water consumption are piece-wise constant over time (Assumption **A2**). In order to enforce the estimated water consumption profiles to be piecewise constant, a new regularization term can be added to Problem (3), i.e.,

$$\min_{\substack{x_i^{(1)}(t),\dots,x_i^{(C_i)}(t) \\ t=1,\dots,T_V \\ i=1,\dots,N}} \sum_{t=1}^{T_V} \left(y(t) - \sum_{i=1}^{N} \hat{y}_i(t, x_i) \right)^2 + \gamma_1 \sum_{i=1}^{N} \sum_{t=1}^{T_V} \left\| \begin{bmatrix} w_i^{(1)}(t) \\ w_i^{(2)}(t) \\ \vdots \\ w_i^{(C_i)}(t) \end{bmatrix} * \begin{bmatrix} x_i^{(1)}(t) \\ x_i^{(2)}(t) \\ \vdots \\ x_i^{(C_i)}(t) \end{bmatrix} \right\|_1 + \qquad (4)$$

$$+\gamma_2 \sum_{i=1}^{N} \sum_{t=2}^{T_V} \left\| k_i \begin{bmatrix} x_i^{(1)}(t) - x_i^{(1)}(t-1) \\ x_i^{(2)}(t) - x_i^{(2)}(t-1) \\ \vdots \\ x_i^{(C_i)}(t) - x_i^{(C_i)}(t-1) \end{bmatrix} \right\|_\infty$$

$$s.t. \;\; x_i^{(j)}(t) \geq 0, \qquad \sum_{j=1}^{C_i} x_i^{(j)}(t) = 1\,, \quad i = 1, \dots, N; \quad t = 1, \dots, T_V,$$

with γ_2 being a tuning parameter playing a role similar to γ_1. The terms k_i (with *i=1,…,N*) are a-priori specified nonnegative weights which can be chosen through the method described in Section 3.2.5. Note that the infinity norm of a vector (i.e., maximum absolute value among the element of the vector) is considered in (4). In this way, if one of the parameters $\left[x_i^{(1)}(t), \dots, x_i^{(C_i)}(t)\right]$ changes from time t-1 to time t, a variation of the other parameters does not change the cost function. Specifically, only the largest time variation among the elements of the vector $\left[x_i^{(1)}(t), \dots, x_i^{(C_i)}(t)\right]$ affects the cost function.

Summarizing, the time-varying parameters $\left[\mathrm{x_i^{(1)}(t),\dots,x_i^{(C_i)}(t)}\right]$ describing the water consumption of each appliance/fixture are computed by solving the regularized (convex) optimization Problem (4).

3.2.4 On the choice of the weights $w_i^{(j)}(t)$

The main idea behind the choice of the weights $w_i^{(1)}(t),\dots,w_i^{(C_i)}(t)$ is the following: if the *i*-th water-using appliance/fixture is likely to operate at mode *j* at time *t*, then the parameter $x_i^{(j)}(t)$ is likely to be equal to *1*, while the other parameters $x_i^{(g)}(t)$ (with $g \neq j$) are likely to be equal to *0*. In terms of the optimization problem (4), the parameters $x_i^{(g)}(t)$ (with $g \neq j$) should be more penalized than $x_i^{(j)}(t)$, or equivalently, the scaling weights $w_i^{(g)}(t)$ (with $g \neq j$) should be higher than $w_i^{(j)}(t)$. The information on time-of-day probability of the usage of each appliance/fixture can be inferred from the training dataset D_{T_E}. Specifically, for given *i* and *t*, the weights $w_i^{(1)}(t),\dots,w_i^{(C_i)}(t)$ can be chosen as follows:

- Given the training dataset D_{T_E}, for each time sample *t* compute the number of times the *i*-th fixture/appliance is operating at mode *j* at the time samples *t+k24h*, where *k=0,1,-1,2,-2,…* Denote the computed number as $q_i^{(j)}(t)$.
- If $q_i^{(j)}(t) \neq 0$, the weight $w_i^{(j)}(t)$ is given by the inverse of $q_i^{(j)}(t)$, i.e., $q_i^{(j)}(t) = \frac{1}{w_i^{(j)}(t)}$. Otherwise, set the parameter $x_i^{(j)}(t)$ equal to *0*.

Note that the weights $w_i^{(j)}(t)$ might be also computed by considering not only the observations at time *t, t+24h, t-24h, t+48h, t-48h, …* but also the observations (possibly weighted) within given time intervals *[t+k24h+Δ, t+k24h-Δ]*.

3.2.5 On the choice of the weights k_i

The weights $\mathrm{k_i}$ (with i=1,…,N) can be chosen as follows: if the i-th appliance/fixture changes its operating mode rarely over the time, than the time variation of the parameters $\mathrm{x_i^{(j)}(t)}$ should be more penalized w.r.t. the time variation of the parameters characterizing an other appliance/fixture which frequently changes its operating mode. The weight $\mathrm{k_i}$ can be then inversely proportional to the number of mode changes observed in the training dataset for the *i*-th appliance.

3.2.6 On the choice of the tuning parameters γ_1 and γ_2

In order to tune the parameters γ_1 and γ_2, a subset $\mathrm{D_{T_C}}$ of length T_C is extracted from the original training dataset $\mathrm{D_{T_E}}$. The $\mathrm{D_{T_C}}$ is referred as calibration dataset. The values of γ_1 and γ_2 are then chosen through a cross-validation procedure, that is by minimizing (with a grid search) the *Total Relative Mean Square Error* (TRMSE) over the calibration dataset $\mathrm{D_{T_C}}$, where the TRMSE is defined as

$$\boldsymbol{TRMSE} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \frac{\sum_{\mathrm{t=1}}^{\mathrm{Tv}} (\mathrm{y_i(t) - \widehat{y}_i(t)})^2}{\sum_{\mathrm{t=1}}^{\mathrm{Tv}} \mathrm{y_i^2\,(t)}}$$

The values of γ_1 and γ_2 leading to the minimum TRMSE are chosen.

3.3 FHMM and iSDTW based algorithm

In addition to the previously explained algorithm, another computationally efficient algorithm for end use disaggregation was developed, as no a priori indications on which category of state-of-the-art literature was identified as best performing. The following novel algorithm is mainly based on *Factorial Hidden Markov Models* [Ghahramani97] and *Subsequence Dynamic Time Warping* and it is developed upon the following assumptions:

- **A3** Each water consuming device can be identified by its specific consumption pattern, i.e., each fixture has a typical "*signature*";
- **A4** The consumption pattern of each water-using fixture can be roughly described with a limited number of states (e.g., state 1: fixture *on/operating; state 2:* fixture *off/not operating*).

Operationally, the algorithm is developed and implemented into three steps:

1. **SIGNATURE IDENTIFICATION**: the purpose of this step is to create a database containing the signature of each fixture contributing to the total consumption, each signature being a representative pattern of each appliance;

2. **FACTORIAL HIDDEN MARKOV MODEL (FHMM) TRAINING and DISAGGREGATION**: the purpose of this step is to disaggregate, as a first rough stage, the aggregate consumption signal into the consumption trajectories of each fixture;

3. **ITERATIVE SUBSEQUENCE DYNAMIC TIME WARPING CORRECTION**: the purpose of this third phase is to refine the disaggregation obtained by FHMM, with the goal of correcting wrong event detections and by increasing the accuracy of the modeled trajectories, based on the signatures identified during step 1.

The description and the implementation details for each of the three steps listed above are given in the next paragraphs.

3.3.1 Signature identification

This preliminary step is performed in order to build a database of *signatures*, i.e., a database containing the typical consumption pattern of each of the fixtures contributing to the total consumption. For instance, Figure 2 provides examples of the signatures for two appliances.

In order to do such an operation and gather all the needed signatures, the current version of the algorithm assumes that a training dataset D_{T_E}, consisting of the observations of the water consumption profiles of each appliance/fixture available in the house, is available (as the algorithm presented in Section 3.2 does).

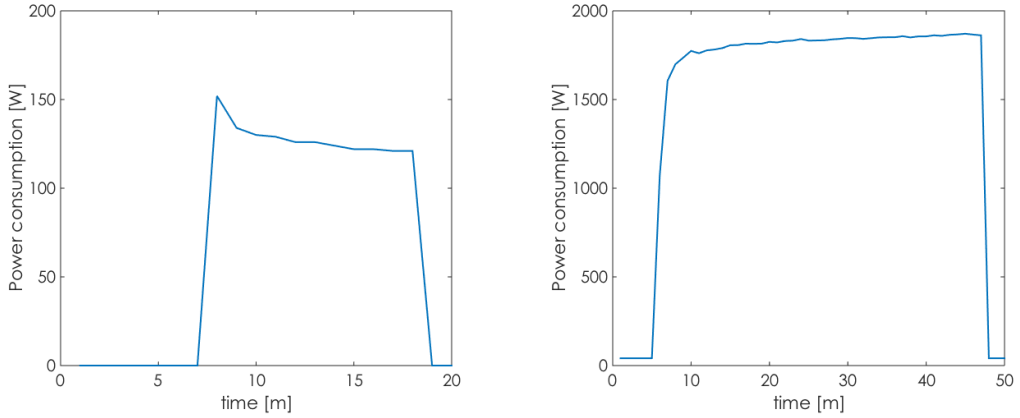

Figure 2: Examples of energy consumption signatures for a fridge (left) and an air conditioner (right).

In practical terms, the identification of signatures operates as follows:

1. For each appliance contributing to the measured total consumption in the household, its consumption trajectory for the full training period is retrieved from training set D_{T_E};
2. The signature of the each considered fixture is built as the set of events for which the consumption trajectory of such a fixture in the training set shows that the fixture is *operating* (on/*open*), i.e., the consumption is larger than 0 or than a given threshold (in order to disregard small measurement noises).

The signature of each appliance i will be from now on defined as s_i, while the database containing all the signatures of the considered appliances will be specified by S.

3.3.2 Factorial Hidden Markov Model training and disaggregation

Factorial Hidden Markov Models (FHMM) [Ghahramani97] are a quite well established technique in machine learning and have already been applied in the field of data disaggregation, mainly within studies of energy disaggregation [Batra14], but a recent study also explored their application for water disaggregation purposes [Nguyen13]. In order to clarify the rationale behind such algorithms, it is relevant to make a couple of remarks regarding their terminology:

- They are called *Hidden* Markov Model as they are used to identify the sequence of states a Markovian Process goes through, just based on the measured output of the system (therefore states are not visible and measurable, i.e., they are hidden).
- Hidden Markov models are called *Factorial* when the considered system is composed of different components, such that the state of the whole system is the combination of the states of each component of the whole system. Figure 3 shows an example where the components of the system act independently from one another. In the case of water disaggregation, the state of the whole system, i.e., the household, is given by the combination of states of each fixture.

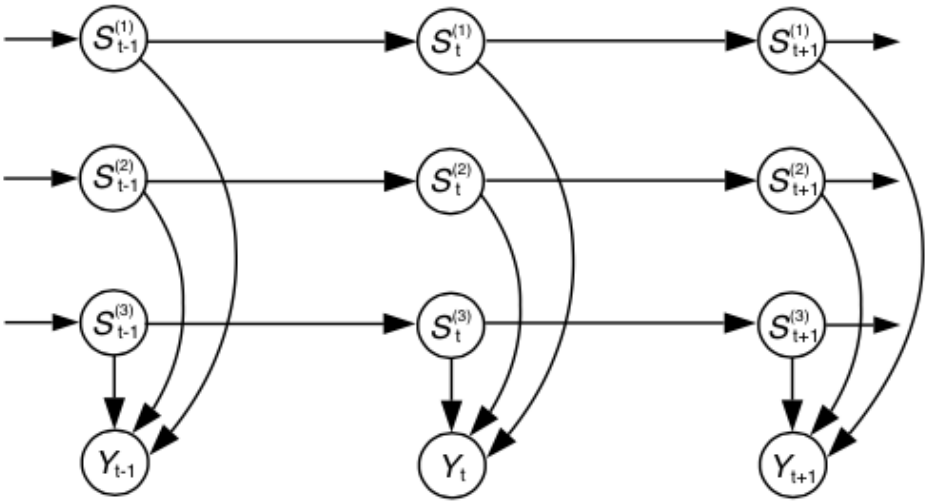

Figure 3. Directed graph representing a system modeled with a Hidden Markov Model.

The FHMM component of the model hereby presents has a two-step operation, composed of a training phase and a testing phase. In detail, the implemented version of the model [Batra14] executes these two phases as follows:

- **TRAINING.** During the training phase, only the data contained in dataset D_{T_E} are considered. The training phase consists in the calibration of the three elements characterizing hidden Markov models, namely:
  - The *initial probability distribution* $P\{X_0^i\}$ for the states of each fixture i. It represents the probability of occurrency of each state (or operating mode j).
  - The *transition probability distribution* $P\{X_t^i \mid X_{t-1}^i\}$ for the states of each appliance i. It represents the transition probability among the different operating modes of each appliance between time *t* and time *t+1*. The output of this phase is the so called *transition matrix.*
  - The *emission probability distribution* $P\{Y_t^i \mid X_t^i\}$ for the states of each appliance i. It represents the probability of observing a particular output of the system depending on its operating state. The output of such a distribution is the so-called *emission matrix*.

  Further information behind the theory of such elements can be found in Ghahramani and Jordan, 1997. The training phase of the FHMM in the current model is performed according to the *Baum-Welch* algorithm [Rabiner89].

- **DISAGGREGATION.** Once the three probability distributions listed above (i.e., prior probabilities, transition matrix and emission matrix) are calibrated, FHMM can be applied to perform a first disaggregation of the aggregate data time series. In short, FHMM solves the following problem:

$$P^*\{X_t^i \mid X_{t-1}^i\}, P^*\{Y_t^i \mid X_t^i\} = argmin_{P\{X_t^i \mid X_{t-1}^i\}, P\{Y_t^i \mid X_t^i\}} \left(\sum_{t=1}^{H} |\overline{Y_t} - \hat{Y}_t|\right)$$

  in order to find the most probable sequence of (hidden) states generating the measured output. In the algorithm here discussed, the *Viterbi* algorithm [Forney73] is used to find such a most probable sequence.

3.3.2.1 On the choice of the number of states for FHMM

The implemented algorithm requires the user to give as an input parameter the number of operating states to consider for each fixture. In particular, the version of the algorithm presented here assumes that the number of states, or operating modes, is the same across all appliances. This is a relevant issue, as the choice of the number of states strongly affects the computational requirements of the algorithm, as the computational complexity grows exponentially with the number of states and the number of appliances.

In turn, it is not easy to a-priori decide which number of states is suitable for describing the consumption pattern of different fixtures, as each fixture has its own consumption pattern and, as preliminary experiments show, the performance in detecting the operational state of an appliance does not monotonically increase with the number of states chosen for the FHMM. Figure 4 shows that the F-score [Batra14], obtained for the disaggregation of electricity consumption over four appliances, does not improve by imposing a higher number of states to the Markov Models.

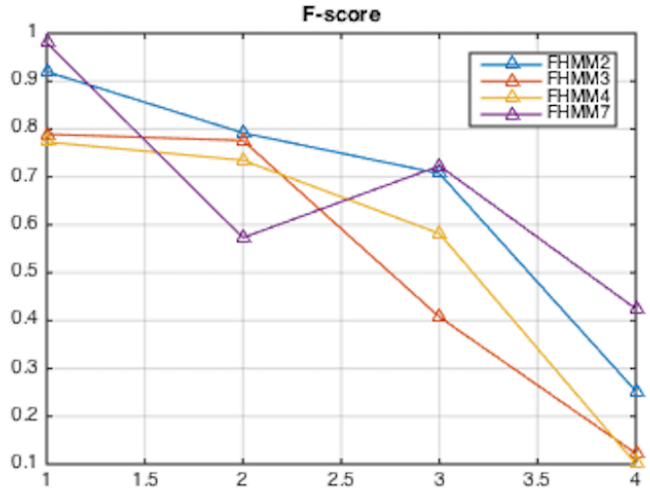

Figure 4: F-score obtained for the disaggregation of the consumption given by 4 appliances (x-axis), with 2,3,4 or 7 Markov states.

As a consequence, on the one hand the user would probably like to limit the number of states, in order save computational resources; on the other hand, limiting the number of states would cause a loss in the accuracy of the model. This means that either the user may choose to have a fast model with limited accuracy, or an accurate model unsustainable from the point of view of computational time (e.g., preliminary experiments show that high levels of accuracy might be reached only with a number of states in the order of 7-10 for each appliance, which means, for instance, that almost two days computational time are required to disaggregate one month of data measured at one minute resolution for a single household with four appliances). In order to increase the performance of traditional FHMM, without compromising their computational sustainability we introduced the iterative use of Subsequence Dynamic Time Warping as explained in the next paragraph.

3.3.3 Iterative Subsequence Dynamic Time Warping correction

Dynamic Time Warping (DTW) and Subsequence Dynamic Time Warping (SDTW) are well known pattern-matching techniques [Sakoe78]. DTW has already been used in combination with FHMM for water consumption data disaggregation [Nguyen13], with the objective of labelling those end uses that FHMM were not able to identify. For doing so, the similarity

between the so-called *unclassified events* with the signatures of a limited number of fixtures was evaluated, and the label relative to the fixture with the closest signature was assigned to the end use FHMM were not able to classify.

In the novel approach proposed here, two main differences with respect to the previously mentioned study characterize the integration of Dynamic Time Warping as an automatic correction:

- The first, and most important one is related to the final goal: DTW is iteratively applied with the goal of directly correcting the trajectories of end use consumption produced as output by FHMM. In fact, while real end use trajectories show pattern characterized by a certain variability, the trajectories produced as output by FHMM are piecewise constant as the number of states considered by FHMM is limited to avoid an increase in the computational burden.
- The second one is related to a fundamental technical aspect: here DTW is applied as a Subsequence Dynamic Time Warping (SDTW) [Muller07], as the total consumption trajectory is first spitted in sequential events that are then compared to the signature of each fixture. Since the length of events is kept much shorter than the total length of the signature, the *best matching sub-sequence* within the signature must be found, thus sub-sequence DTW must be used.

More in details, the Iterative Subsequence Dynamic Time Warping (ISDTW) is integrated into the model works as follows:

1. **EVENT DEFINITION**

   The total consumption trajectory and the single appliances trajectories produced by FHMM are split into events of equal length (10 minutes is the considered event length for the experiment described in the following section).

2. **FIXTURE RANKING**

   For each event, appliances are ranked in descend order according to the values of the 90-th percentile of each of their FHMM trajectories within the event. This ranking gives an idea of the contribution each appliance has to the total event.

3. **ITERATIVE SUBSEQUENCE DTW CORRECTION**

   Subsequence DTW between the total consumption in the considered event and each signature in the database is run. The following cases are then considered:

   a. If the closest signature is the one of the fixture ranked first at step 2, FHMM output is corrected with the values of the closest sequence of such signature:

      $$y_i^{current\ event} = s_i$$

   b. Else, if the closest signature is another, FHMM might have found a false event. In that case:

      i. If the appliance designed by FHMM was already operating in the previous three events (with a higher contribution than the one provided by the appliance with the closest signature), FHMM is not corrected:

         $$y_i^{current\ event} = y_i^{FHMM,current\ event}$$

      ii. Else, the fixture identified by FHMM is switched off for the considered event and the process is repeated from step 2, considering all the signatures but the one of the

fixture just corrected:

$$y_i^{current\ event} = 0$$

This third step is repeated for each fixture, for each event.

3.3.4 Model specifications

The model described in this section is implemented using Python language (v 2.7), starting from the state of the art toolkit (NILMTK) [Batra14] downloadable at (http://nilmtk.github.io/) and exploiting the following packages:

- Anaconda (http://continuum.io/downloads): it is a free Python distribution including over 195 packages for science, math, engineering and data analysis
- Sklearn (http://scikit-learn.org/stable/): it is a simple tool for data mining and data analysis.
- Ucrdtw (https://github.com/klon/ucrdtw): it is a Python extension for highly optimized subsequence search using Dynamic Time Warping.

3.4 Experiment setting

Despite the SmartH2O project is focused on the water sector, the two algorithms presented in the previous paragraphs were initially tested and validated against energy consumption data mainly because (i) the state-of-the art literature on data disaggregation is more advanced in the energy sector, thus allowing with a fair comparisons against benchmark algorithms, and (ii) because a dataset of high-resolution residential water consumption data was not available in the initial phases of the project[1].

3.4.1 Dataset

The AMPds dataset [Makonin13] is used to test the performance of the developed algorithms. The AMPds dataset is available online and it contains the energy consumption readings of a single house located in the Vancouver region in British Columbia, Canada. Specifically, 21 breakers/loads have been sub-metered for an entire year (from April 1, 2012 to March 31, 2013) at one minute read intervals.

For the sake of analysis, we considered only the aggregate power consumption given by the sum of the power consumption readings of the following four electric appliances:

- washing machine
- fridge
- dishwasher
- heat pump

These four appliances share the largest contribution of the total energy consumption both in Summer and in Winter, and they contribute at least for the 5% (Summer period) and 3% (Winter period) of the total energy consumption.

Furthermore, in order to assess the robustness of the disaggregation algorithms w.r.t. a measurement noise which might corrupt the power readings, the aggregate power consumption signal $y(t)$ has been corrupted by an additive zero-mean random Gaussian noise $e(t)$ with standard deviation $\sigma = 4$ W. Note that, because of the added fictitious noise, the aggregate power consumption signal can become negative. At the time samples when this happens, the power consumption signal is set to 0 W.

[1] High-resolution energy consumption datasets, such as the AMPds mentioned in 3.4.1, can be freely downloaded from the internet, while no high-resolution water consumption datasets are available.

The available AMPds dataset has been divided into two disjoint datasets:

- A training dataset D_{T_E} containing the data for the days 16-30 June 2012, which was used for estimating the possible values of the FHMM states and the three FHMM probabilities distribution for the FHMM, described in section 3.3.2. The power readings from October 1, 2012 to November 30, 2012 were used instead as training set for the sparse optimization based algorithm described in Section 3.2. Such a training set is used to estimate the power demand of each appliance at each operating mode (i.e., the terms $\mathrm{B_i^{(j)}}$) as well as the weights $w_i\ (t)$ and k_i through the procedure discussed in Sections 3.2.4 and 3.2.5. Furthermore, in order to tune the parameters γ_1 and γ_2 used in the optimization based algorithm, a calibration dataset D_{T_C} has been extracted from the original training dataset D_{T_E}. Such a calibration dataset consists of the power readings from November 16, 2012 to November 30, 2012. Note that the sub-metered power consumptions of each appliance are supposed to be available in the training and calibration phase.
- The two algorithms were validated on a portion of dataset extracted from the Summer period and a portion from the Winter period, since we expect seasonality to impact on the consumption pattern of the different end uses. In particular, a validation dataset D_{T_V}, which consists of the aggregate power readings from July 1, 2012 to July 31, 2012 (plotted in Figure 5) was considered for the validation of the FHMM+iSDTW algorithm and the power readings from December 1, 2012 to December 31, 2012 (plotted in Figure 6) were taken into account for validating the algorithm based on sparse optimization. In the validation phase, the sub-metered power consumption measurements are not supposed to be available and the aggregate power consumption signal is decomposed into the power consumption of each appliance through the two proposed algorithms. The sub-metered power consumption measurements are only used to assess the performance of the developed algorithms.

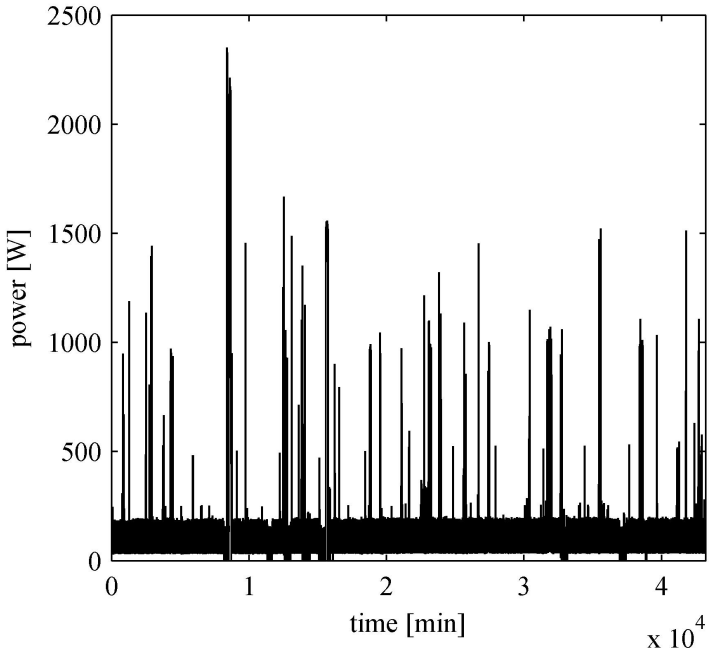

Figure 5: Electric power consumption from July 1, 2012 to July 31, 2012.

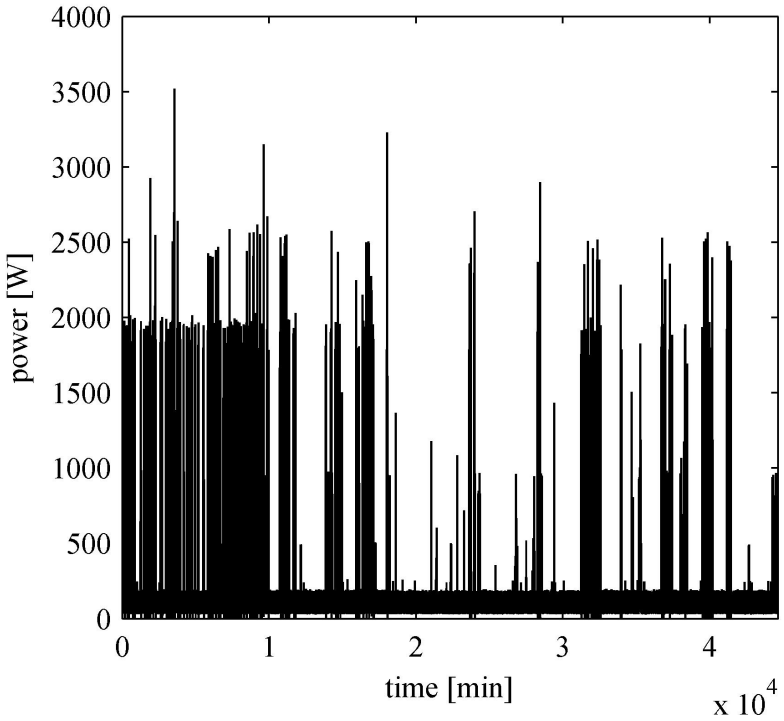

Figure 6: Electric power consumption from December 1, 2012 to December 31, 2012.

3.4.2 Performance metrics

The following metrics have been used to assess the performance of the developed disaggregation tools:

- The *Estimated Energy Fraction Index* (EEFI), defined as:

  $$\hat{h}_i = \frac{\sum_{t=1}^{T_V} \hat{y}_i(t),}{\sum_{i=1}^{N} \sum_{t=1}^{T_V} \hat{y}_i(t)}$$

  The index $\hat{h}_i$ provides the fraction of energy assigned to the *i*-th appliance, and it should be compared to the *Actual Energy Fraction Index* (AEFI), defined as

  $$h_i = \frac{\sum_{t=1}^{T_V} y_i(t),}{\sum_{i=1}^{N} \sum_{t=1}^{T_V} y_i(t)}$$

  which in turn provides the actual fraction of energy consumed by the *i*-th appliance.

- The *Relative Square Error* (RMSE), defined as:

  $$RSE_i = \frac{\sum_{t=1}^{T_V} (y_i(t) - \hat{y}_i(t))^2}{\sum_{t=1}^{T_V} y_i^2\,(t)}$$

  The RSE provides a normalized measure of the difference between the actual and the estimated power consumption of the *i*-th appliance.

- The R^2 coefficient, defined for the *i*-th appliance as:

  $$R_i^2 = 1 - \frac{\sum_{t=1}^{T_V}(y_i(t) - \hat{y}_i(t))^2}{\sum_{t=1}^{T_V}(y_i(t) - \bar{y}_i)^2},$$

  with $\bar{y}_i$ denoting the mean of the power consumption, i.e.,

  $$\bar{y}_i = \frac{1}{T_V}\sum_{t=1}^{T_V} y_i(t)$$

  The R^2 coefficient measures how well the estimated power profiles match the actual power profiles.

3.5 Testing and validation

The FHMM-DTW based algorithm and the sparse optimization based approach have been tested against the validation dataset D_{T_V} (i.e., July 2012 and December 2012). The performance metrics introduced in Section 3.4 and the estimated disaggregate power profiles are computed in order to assess the performance of the algorithms. Specifically:

- Table 2 shows the *Estimated Energy Fraction Index* $\hat{h}_i$ for each appliance, along with the *Actual Energy Fraction Index* h_i*;*
- Table 3 shows the *Relative Square Errors* for each appliance;
- Table 4 shows the R^2 coefficient for each appliance.
- Figure 7 shows the power consumption of each appliance obtained by using the FHMM-DTW based algorithm. For the sake of visualization only the power profiles at July 11, 2012 are plotted.
- Figure 8 shows the power consumption of each appliance (at December 3, 2013, respectively) obtained by using the sparse optimization based approach.

Table 2: Fraction of energy assigned to each appliance (Estimated Energy Fraction Index $\hat{h}_i$) by the sparse optimization based algorithm and by the FHMM-ISDTW based approach, along with the actual fraction of power consumed by each appliance (Actual Energy Fraction Index h_i).

	July 2012		December 2012	
	FHMM-ISDTW	Actual	Sparse optimization	Actual
Washing machine	3.7 %	2.8 %	1.1 %	1.1 %
Fridge	46.6 %	47.7 %	10.3 %	10.6 %
Dishwasher	12.1 %	12.7 %	3.6 %	3.9 %
Heat Pump	37.6 %	36.8 %	85.0%	84.4 %

Table 3: Relative Square Error obtained by the sparse optimization based algorithm and by the FHMM-ISDTW based approach.

	July 2012	**December 2012**
	FHMM-ISDTW	Sparse optimization
Washing machine	9.4 %	4.1 %
Fridge	15.0 %	18.8 %
Dishwasher	9.5 %	9.1 %
Heat Pump	7.8 %	8.1 %

Table 4: R^2 coefficient obtained by the sparse optimization based algorithm and by the FHMM-ISDTW based approach.

	July 2012	**December 2012**
	FHMM-ISDTW	Sparse optimization
Washing machine	5.4 %	95.1 %
Fridge	90.3 %	78.2 %
Dishwasher	80.8 %	90.1 %
Heat Pump	91.2 %	91.1 %

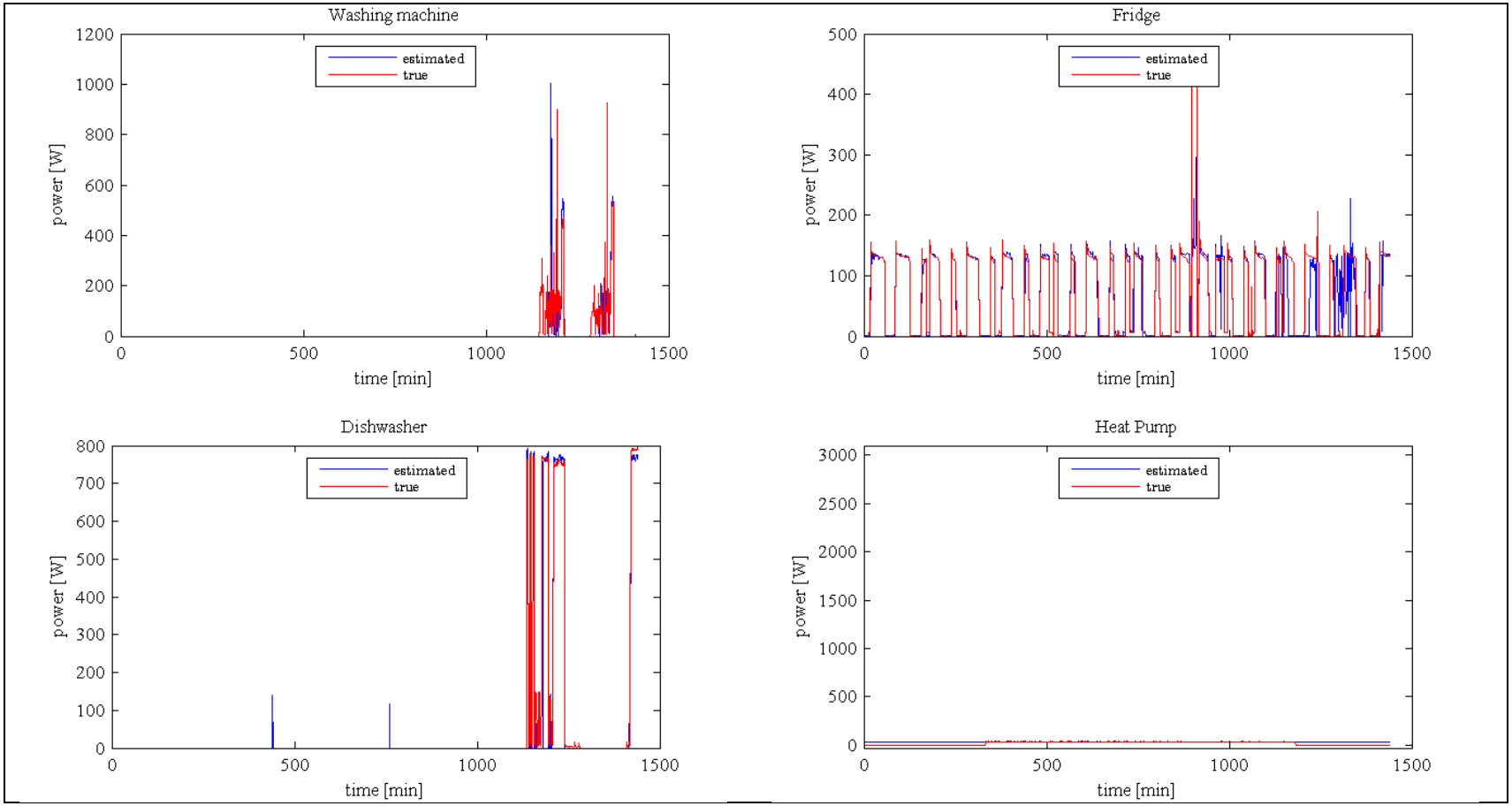

Figure 7: FHMM-DTW based algorithm, July 11, 2012: disaggregated power consumption profiles.

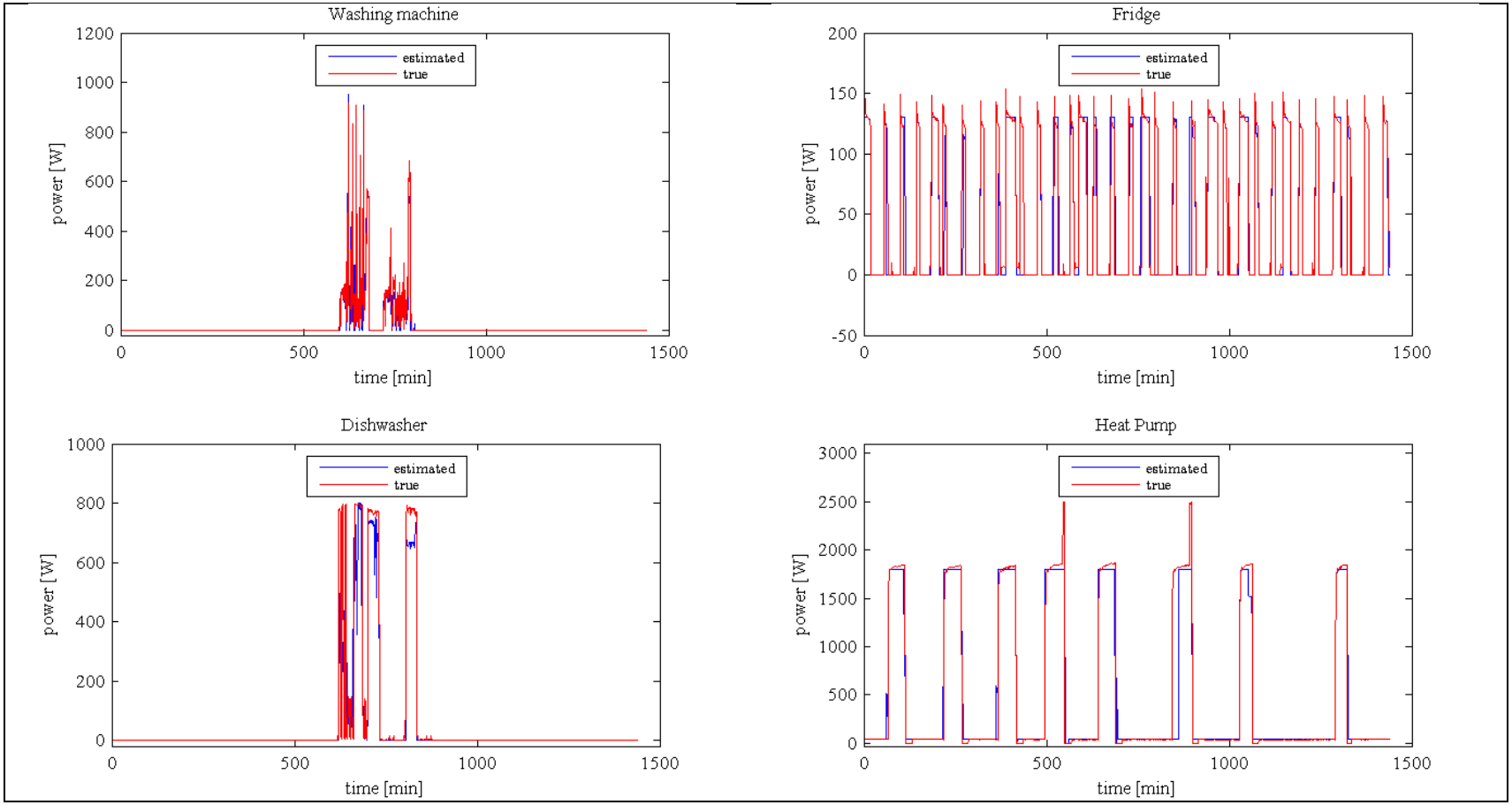

Figure 8: Sparse Optimization based algorithm, December 3, 2012: disaggregated power consumption profiles.

3.6 Applications on water data

Validation on water consumption data was possible only at a second stage of the project, as soon as a dataset containing high-resolution water consumption data for a small set of houses in New Zealand became accessible. The dataset, experiments settings and obtained results are described in the following paragraphs and discussed in comparison with those of experiments on energy data.

3.6.1 Dataset

The dataset considered for the initial testing of disaggregation algorithms on water data contains high-resolution water consumption data for a set of 7 households, which were metered during the 2006 WEEP (Water End Use and Efficiency Project) research [Heinrich07] in New Zealand. For the purpose of performing the first water disaggregation experiments, data from a single house metered for 68 days in the period 27th July – 2nd October 2006 were considered. Data for the following end-uses were available (listed here in descending consumption order):

- toilet
- tap
- shower
- bath
- clothes-washer
- dishwasher
- garden

Differently from the data available for energy disaggregation, the available raw water data required some additional pre-processing, as they did not represent end-use water consumption trajectories metered at a constant time resolution. In contrast, the following information was described by such data:

- starting time of a water consuming event
- ending time of the same event
- cumulative volume of water used during the event
- peak consumption rate within the event.

Based on such data, the following operations were performed, in order to reproduce end-use consumption trajectories suitable to train and test the described disaggregation algorithms:

1. given the starting time, ending time and total volume of consumption events, piecewise constant end-use consumption trajectories at 10-second sampling resolution were generated (i.e., trajectories where the consumption is always 0 but for those time windows in which consumption events happen). Such trajectories are piece-wise constant because consumption events are represented by a constant average consumption rate, which was evaluated dividing the total water volume by the event duration was considered;
2. the generated trajectories were aggregated in time, in order to obtain end-use trajectories sampled at 1 minute resolution and allow for a consistent and fair comparison of the disaggregation algorithms performance between energy and water disaggregation, under same time sampling resolution.

3.6.2 Experiment settings

In order to proceed consistently with the energy disaggregation experiments, the available dataset was split as follows to run the first disaggregation algorithms:

- the training dataset D_{T_E} consisted in 2-week data for the days 27th July – 10th August 2006;
- water consumption readings from August 11th, 2006 to September 9, 2006 were used instead as validation dataset.

In addition, the first experiment run on water consumption data considering and here reported considered the following settings:

- considered appliances: toilet, tap and shower, being the ones most contributing to the total household consumption;
- disaggregation algorithm: experiments were run using both the optimization-based algorithm and the FHMM-iSDTW approach.

All other settings were set in compliance with the experiments on electricity data described in the previous section.

Results from the first experiments are described in the next paragraphs.

3.6.3 Results from disaggregation of high resolution data

Disaggregation results on a 3-appliance experiment were evaluated according to the same performance metrics defined in 3.4.2 are reported and commented here.

Table 5 reports the results in terms of algorithm accuracy in assigning the consumption share of the total to each end-use. As an aggregate consumption result, both of the algorithms show an acceptable performance in estimating the total contribution of each appliance: the maximum estimation error is around 6%. However two drawbacks can be noticed if comparing the result with the one obtained for energy disaggregation. The first is that both algorithms managed to estimate the fraction of energy assigned to each appliance with an error lower than 5%, in the applications on energy data. The second is that, even though the consumption share is estimated with an acceptable error, the ranking of actually most consuming appliances is not accurately detected.

Table 5: Fraction of water assigned to each appliance (Estimated Water Fraction Index $\hat{h}_i$) by the optimization-based and the FHMM-ISDTW based algorithms, along with the actual fraction of water consumed by each appliance (Actual Water Fraction Index h_i).

	optimization-based	FHMM-ISDTW	Actual
Toilet	30.6 %	29.2 %	34.4 %
Tap	36.8 %	35.1 %	35.5 %
Shower	32.5 %	35.7 %	30.1 %

Performance results in terms of Relative Square Error and R^2 score are reported in Table 6 and Table 7. Considering that the disaggregation experiments only considered 3 appliances, the performance significantly decreases in terms of trajectories reproduction accuracy, if compared with the values obtained for electric power disaggregation.

Table 6: Relative Square Errors obtained by the optimization-based and the FHMM-ISDTW algorithms

	optimization-based	FHMM-ISDTW

Toilet	66.5 %	81.1 %
Tap	74.7 %	69.2 %
Shower	4.7 %	7.1 %

Table 7: R^2 coefficients obtained by the optimization-based and the FHMM-ISDTW algorithms

	optimization-based	FHMM-ISDTW
Toilet	32.4 %	-53.2 %
Tap	23.9 %	-63.0 %
Shower	95.3 %	65.5 %

In particular, it is noticeable that the only fixture for which both of the algorithms provide R^2 values larger than 50% is the shower. This gives us important hints on the reasons behind the overall performance decline. If we look at the actual consumption trajectories represented in Figure 9, the following causes can be supposed:

- All appliances operate in a narrow and similar range (in absolute values, the operating range here is 0-30 litres/minute, while energy appliances operated between 0 and few thousand kWh), which already represents a significant limit to appliance identification;
- Tap and toilet show an irregular pattern and operate exactly in the same range. In contrast, shower events can be better distinguished, as they usually show a higher peak and larger durations. This is likely to be the reason why it is the only end-use for which RSE error is low and R^2 higher than 60%. This is further confirmed by an additional performance metric, the *F-score* (defined as in [Batra14]), which provides a measure of the accuracy in detecting the *on/off* status of each appliance. Again, it shows (Table 8) that both of the algorithms achieves an accuracy of around 90% in detecting shower events, while tap and toilets events can hardly be detected.
- Water consumption trajectories for tap and toilet events do not satisfy Assumption **A2** in Section 3.2. In fact, the disaggregated signals are not piecewise constant over a discrete-time scale with sampling time equal to 1 minute. This is the reason why the optimization-based algorithm shows poor performance in reconstructing the consumption trajectories for tap and toilet events.
- Finally, given that the water consumption trajectories do not present a clear signature, the iSDTW module of the FHMM-iSDTW algorithm is likely to be not effective in correcting the FHMM results.

Table 8: F-score obtained by the optimization-based and the FHMM-ISDTW algorithms

	optimization-based	FHMM-ISDTW
Toilet	59.2 %	52.4 %
Tap	58.6 %	58.6 %
Shower	62.2 %	89.9 %

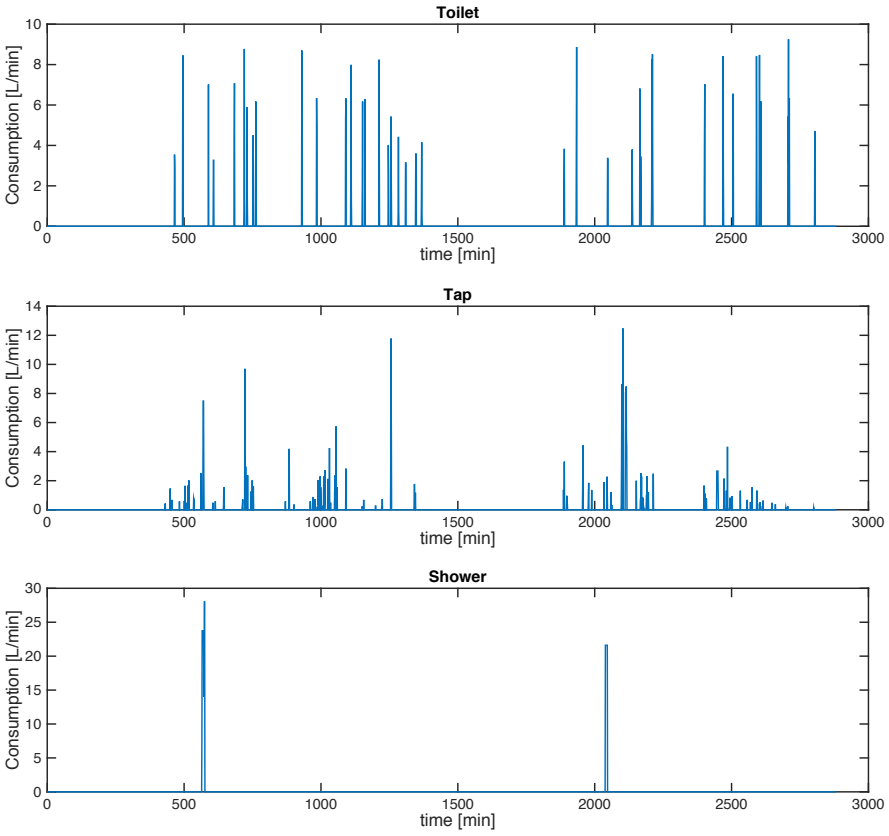

Figure 9. Example of water end-use trajectories for toilet, tap and shower end uses.

As an overall comment to the presented results, it can be concluded that first results on water disaggregation look promising in terms of estimation of end-use contributions to total household consumption, but big improvements are needed to accurately reproduce end-use consumption trajectories. This represents so far an important result for the SmartH2O project, because end-use share information is suitable to understand where major consumptions are, to understand users' consumption profiles and to enforce information and feedback sharing with customers. Yet, the obtained results and relative comments suggest for further investigations on water data disaggregation and intensive testing, in order to possibly achieve high disaggregation accuracy performances as the ones successfully obtained for energy

disaggregation.

3.6.4 Results from disaggregation of 1-hour resolution data

The experiment above was repeated using a resolution of one-hour instead of one minute, using otherwise the same dataset and experiment settings described in sections 3.6.2 and 3.6.3. Due to the relatively short event lengths of the considered water appliances, it is not feasible to retrieve the end-use trajectories through disaggregating at this low resolution. However, it was found that FHMM-ISDTW approach provides better results in terms of estimating the total contribution of each appliance when compared to disaggregating at one-minute sampled data. Moreover, due to the improvement of disaggregation performance with respect to the Estimated Water Fraction Index metric (i.e., the Estimated Energy Fraction Index applied to water data), we are able to disaggregate two more appliances, namely, the bath and the clothes washer. The same considerations do not hold when the optimization-based approach is used. The obtained results are reported in Table 9.

Table 9: Fraction of water assigned to each appliance (Estimated Water Fraction Index $\hat{\mathbf{h}}_i$) by the optimization-based algorithm and the FHMM-ISDTW based approach, along with the actual fraction of water consumed by each appliance using one-hour resolution data (Actual Water Fraction Index $\mathbf{h}_i$).

	optimization-based	FHMM-ISDTW	Actual
Toilet	10.2 %	19.8 %	16.3 %
Tap	14.7 %	19.2 %	17.8 %
Shower	10.3 %	16.8 %	12.4 %
Clothes washer	60.5 %	39.0 %	46.0 %
Bath	4.3 %	5.2 %	7.5 %

These results are also shown graphically in Figure 10, which shows side by side, the percentage

contribution pie charts from the actual data, and the one obtained after disaggregation. The missing slice of the pie chart to the right represents the portion of water consumption that was not accounted for, i.e. not assigned to any of the end-uses, by the disaggregation algorithm. It is equal to approximately 9%. As already mentioned, the Estimated Water Fraction Index metrics obtained by using the optimization-based approach do not accurately match the actual ones, and thus they are not visualized in Figure 10 (but only reported in Table 9).

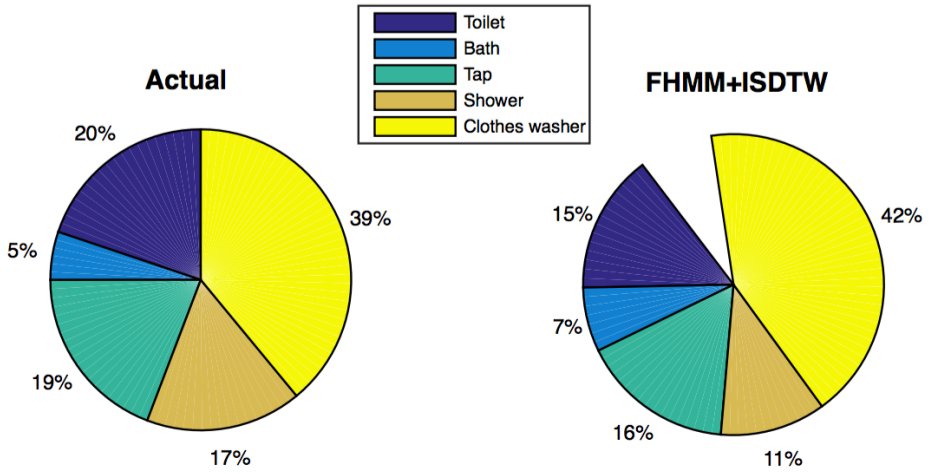

Figure 10. Fraction of water assigned to each appliance (Estimated Water Fraction Index $\hat{\mathbf{h}}_i$) by the FHMM-ISDTW based approach, along with the actual fraction of water consumed by each appliance using one-hour resolution data (Actual Water Fraction Index $\mathbf{h}_i$).

Possible explanations behind the improvement in performance (of the FHMM-ISDTW approach) in contribution error when the resolution is lowered are:

- Downsampling by averaging creates a smoother signal, which leads to better disaggregation.
- Using a lower resolution for the same validation period reduces the size of the dataset, which lessens the likelihood of false positive detections.

On the other hand, the poor performance of the optimization-based algorithm are probably due to fact that:

- the information on time-of-day probability of the usage of each appliance/fixture (used to compute the weighting parameters weight $w_i^{(j)}(t)$ as described in Section 3.2.4) cannot be accurately inferred from low-resolution data.
- the water consumption trajectories of the considered fixtures/appliances do not satisfy Assumption **A2** in Section 3.2 (i.e., the actual disaggregated signals are not piecewise constant over a discrete-time scale with sampling time equal to 1 hour).

3.7 Discussion

The obtained results show that both the algorithms are able to accurately estimate the fraction of energy consumed by each appliance in the household (Table 2), and, most importantly, to extract single power consumption profiles (as shown in TablesTable 3 -Table 4 and Figs. 7-8). The only exception is due to the estimation of power consumption of the washing machine through the FHHM-DTW algorithm (see Section 3.3), which, however, has marginal contribution in the considered period.

In terms of computational complexity, the FHHM-ISDTW approach is less computationally demanding than the optimization based method. As a matter of fact, the time required by the FHHM-ISDTW method to disaggregate a one-month power signal is approximately 15 minutes, while the optimization based method requires approximately 5 hours in 2.40-GHz Intel Pentium IV with 3 GB of RAM.

Waiting for high resolution water data to be available for the SmartH2O project, ongoing research activities are focused on:

- extensive testing of the algorithms' generalization potential:
  (i) w.r.t. new, unseen appliances;

(ii) across different data sampling (i.e., 1s, 15 min, 1 h);

- extensive comparison (in terms of computational complexity and accuracy) between the two developed disaggregation algorithms, along with a comparison with the state-of-the-art disaggregation algorithms;

Related to the application of the disaggregation algorithms to water data, we found that despite the reduction in precision with respect to energy, the algorithms we have developed are performing as well as, and even slightly better, than the state of the art ([Nguyen13]) that considers many appliances, but not overlapping and simultaneous events.

It was also found that, although the optimization-based approach achieves better when high-resolution data (i.e., 1 minute) are available, the FHMM-ISDTW algorithm achieves a good estimate of the breakdown structure of consumption data among end-uses with 1 hour resolution data. This is an important result, especially in the light that most smart meters operate at such a low resolution. The results of this disaggregation will be therefore particularly useful to provide feedback to the users about how they use their water.

4. SmartH2O user modeling algorithms

Understanding 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. By capturing the behavior of water users, these models allow identifying the variety of users' consumption profiles as well as exploring the effects of different Water Demand Management Strategies for the residential sector, thus representing promising decision-aiding tools for water utilities and urban planners.

This section illustrates a novel approach based on *feature extraction* techniques [Guyon03] to model the single-user consumption behavior at the household level. The approach is based on a two-step procedure:

(i) identify the most relevant determinants of users' consumption profiles;
(ii) build a predictive model of water consumption profiles based on the observation of the determinants identified in step *i*.

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) is motivated by the need of managing 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. Many state-of-the-art studies reported about the presence of correlations between one or more presumed consumption drivers and the associated consumption profiles. Yet, the number of considered candidate variables is generally limited. In addition, the subsequent calibration, and validation of a user behavioral model based on the selected input/output is often missing, thus preventing the use of these tools as water demand predictors.

This two-step procedure has been tested and validated on low-resolution (billed) data, as data on the SmartH2O project are not available yet. In particular, we worked on 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 [Anda12].

The Section is organized as follows: the next section introduces the procedure. Section 4.2 describes the case study and Section 4.3 the numerical results. Section 4.4 summarizes the limitations of the proposed approach and identifies possible improvements to be implemented within the smartH2O project.

4.1 Problem formulation

The general formulation of a water demand predictive model for a generic user *i* is given by:

$$y_i = f(\boldsymbol{x_i}),$$

where y_i represents the consumption profile of the *i*-th user and $\boldsymbol{x}_i$ is the set of *M* determinants influencing his/her behavior, represented by a variety of demographic and psychographic user features (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,…,N)* is defined as follows:

$$< x_i^1, x_i^2, \dots, x_i^M, y_i >$$

The construction of the water demand predictive model relies on the following two-step

procedure:

1. *Feature extraction,* to select from the original dataset *X* of user's data a subset $X' \subseteq X$ of determinants that are relevant to describe the consumption profile Y;
2. *Model learning,* that relates the previously generated subset X′ to the water consumption level Y.

4.2 Feature extraction

Feature extraction techniques, mostly developed in the data mining and machine learning research communities, represent promising tools to model residential water user behavior. These techniques 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.

Different approaches can be adopted to perform feature extraction. 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.

Since a-priori no single method is best suited to all datasets and modelling purposes, we implemented and applied different algorithms for both feature selection and weighting. In particular, we run the feature extraction algorithms described in the following paragraphs.

4.2.1 Feature selection algorithms

The following four different feature selection algorithms have been implemented:

- **Fast correlation based filter (FCBF) [Yu13].** This algorithm exploits the formulation of the *Information Gain* algorithm (explained in the next section about feature weighting algorithms), in order to keep into account both the feature-feature and the feature-class correlation, thus considering redundancy issues. The correlation between a feature X_i and a class *C* is computed through the concept of symmetrical uncertainty (*SU*), which is defined as follows:
- $$SU(X_i, C) = 2\frac{IG(X_i, C)}{H(X_i) + H(C)}$$

  where *IG* represents the information gain and *H* is the entropy of a variable (as defined later on).
- **CFS algorithm [Zhao10].** This filter algorithm uses a correlation-based heuristic to determine the relevance of a feature both in terms of feature-class correlation and feature-feature intercorrelation, thus avoiding redundancy issues. Given a subset of *n* features, the algorithm determines the "*worth of the subset"* and then explores different subsets in order to identify the one with the best merit.
- **BLOGREG algorithm [Guyon02].** This an embedded feature selection algorithm, which promotes the sparsity of a logistic model in order to reduce the number of features selected. The BLOGREG algorithm is suitable for managing categorical features.
- **Sparse Bayesian Multinomial Logistic Regression [Cawley07].** This algorithm is an extension of the BLOGREG approach. Making it suitable to be applied to multiclass problems.

4.2.2 Feature weighting algorithms

The following feature weighting algorithms have been implemented:

- **CHI-Square Score [Liu95].** This is a supervised, filter algorithm used to test whether the class labels are independent of a specific feature. In particular, the chi-square score is evaluated as follows:

  $$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{N_{class}} \frac{(n_{ij} - \mu_{ij})^2}{\mu_{ij}}$$

  where:
  - n_{ij} is the number of samples with the *i*-th value for a particular feature in class *j*.
  - $\mu_{ij} = \frac{n_{*j} n_{i*}}{n}$, with n_{i*} being the number of elements with value *i* for a particular feature across all classes and n_{*j} being the number of elements in class *j*.

  The higher chi-square score, the more the class label is dependent on the considered features. The index is suitable to be applied also with categorical (or binary) variables.

- **Information Gain [Cover12].** Information gain is another measure of dependence between a feature and the class labels. Considering a feature X_i and the class labels *C,* the information gain is defined as:

  $$IG(X_i, C) = H(X_i) + H(X_i|C)$$

  Where *H* represents the entropy of a variable, defined as:

  $$H(X_i) = -\sum_{j} P(x_j) log_2 \left(P(x_j)\right)$$

  $$\mathrm{H(X_i|C)} = -\sum_{\mathrm{z}} \mathrm{P(y_z)} \sum_{\mathrm{j}} \mathrm{P(x_j|y_z) log_2 \left(P(x_j|y_z)\right)}$$

  As the maximum value that the information gain can take is 1, the closer the information gain of a feature is to 1, the more relevant the feature is.

It is worth mentioning that both the chi-square score and the information gain algorithms consider each feature separately, thus they do not solve redundancy issues.

4.3 Model learning

As far as the model learning phase is concerned, in principle any data-driven modeling approach (regressor or classifiers) can be used to build a user behavioural model (see [Maier00], [Maier10], [Galelli13]). In practice, the selected method should have the following desirable features:

(i) 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;

(ii) computational efficiency to deal with potentially large datasets, when considering large number of users;

(iii) 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.

The following two data-driven modeling approaches have been implemented:

- **Naive Bayesian Regression**. Bayesian classifiers [Duda and Hart, 1973] learn from training data the conditional probability of each attribute, given the class label. Classification is then performed by computing the probability of each class, given an instance of the attributes and predicting the class with the highest posterior probability.

- **J48 Decision Tree algorithm.** The J48 algorithm used here is an implementation of the C4.5 algorithm used to generate a decision tree [Ross Quinlan, 1993]. It builds a decision tree on the training dataset, where the attributes that most effectively split the set of samples into small subsets, in terms of information gain, are positioned onto nodes.

4.4 Experiment setting

4.4.1 Case study description

The *H2ome Smart* project dataset [Anda12] was then used to assess the performance of the proposed modeling technique. The following data are available, for more than 3000 households in the towns of the Pilbara and Kimberley Regions of Western Australia:

- **Water consumption rate:** household water consumption data from meter readings (measured in m^3), collected between August 2010 and February 2012 (19 months). The maximum number of readings per household, within the considered period, is seven, thus the best reading resolution is approximately three months;
- **House and occupants attributes**: 26 variables describing different features of the users and the house. Table 10 reports the complete list of available data.

Table 10: Customer and household features considered in this study.

Name	Description	Var Type	Number of categories
town	-	categorical	9 possible values
suburb	-	categorical	21 possible values
years of occupancy	years since the house is being occupied by the same inhabitants	integer	-
responsibility	person responsible for paying bills	categorical	4 possible values
number of occupants	number of inhabitants in the house	integer	-
resident type	type of resident in the house	categorical	8 possible values
number of toilets	number of toilets in the house	integer	-
land use	type of land use destination	categorical	14 possible values
house type	type of house structure	categorical	5 possible values
washing machine type	type of washing machine	categorical	3 possible values
toilet type	type of flush	categorical	3 possible values
shower type	type of shower	categorical	3 possible values
dishwasher presence	presence of dishwasher	binary	-
garden area	area of the house garden $[m^2]$	real positive value	-
watering method	method used for garden watering	categorical	4 possible values
watering time	average weekly watering times	integer	-
irrigation system	type of irrigation technique	categorical	3 possible values
drip type	type of drip irrigation	categorical	3 possible values
surf type	type of surface irrigation	categorical	3 possible values
drip duration	weekly average drip irrigation minutes	categorical	4 possible values
surf duration	weekly average surface irrigation minutes	categorical	4 possible values
mulch	presence of mulch	binary	-
pool presence	presence of pool	binary	-
pool cover presence	presence of pool cover	binary	-
spa presence	presence of spa	binary	-
native plants	presence of native plants	binary	-

4.4.2 Data pre-processing

The following data pre-processing steps have been carried out before applying the feature extraction algorithms.

Data cleaning

1. records of users showing data inconsistencies or missing data (i.e., negative consumption rate or no consumption rate measures) were removed from the dataset;
2. empty reading date 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 household from water consumption data and reading dates, since the number of water consumption readings and the length of reading period was very heterogeneous among different households;
4. if the information about the number of house occupants is present, the per-capita daily water consumption rate in [m^3/day] is computed.

The data cleaning process produced the following outputs: a set Y_1 containing the daily average water consumption rate for N = 3325 households (users) and a set Y_2 containing the per-capita daily average consumption for N' = 3197 households (users).

Class label assignment

The real values in Y_1 and Y_2 were converted into the following classes representing different consumption profiles:

- low-consumers when the user consumption is lower (or equal) than the 25th percentile value;
- high-consumers when it is higher than the 75th percentile value;
- medium-consumers for the ones between 25th and 75th percentiles.

Matrix of user features

Two sample datasets X_1 and X_2 were built, respectively for the users whose consumption is included in Y_1 (daily average water consumption) and Y_2 (daily average per-capita water consumption). Each tuple of the datasets has M = 26 user and house features (see Table 10) associated to either Y_1 or Y_2.

4.5 Testing and validation

4.5.1 Feature selection and feature weighting

The outputs from the feature selection algorithms are represented in Figure 11 (daily average water consumption) and Figure 12 (per-capita daily average water consumption), where the user and house features are represented on the y-axis and the color that indicates the selection frequency of each feature: white colored features are the most relevant as they always selected across the different algorithms runs, while their relevance decreases moving towards gray and black tones.

The results of Figure 11 and Figure 12 appear to be quite consistent: the number of household's occupants seems to be the most important factor of residential water consumption; the number of toilets, the method used for irrigation, the presence of pool and the type of house are then ranked in the subsequent positions with high frequency (i.e., 80%); the town is also considered relevant in explaining the per-capita daily water consumption. However, the selection frequencies in this second experiment are lower than 70%, except for the number of occupants, which is always selected in the first position.

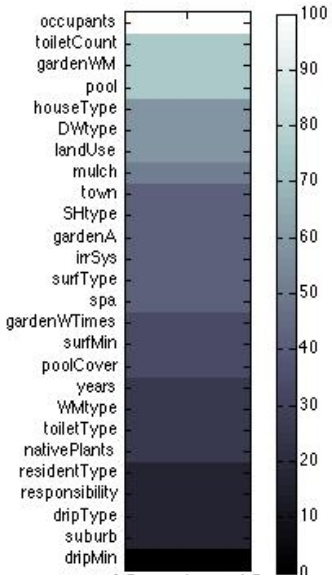

Figure 11: Selection frequency obtained considering as output the daily average water consumption.

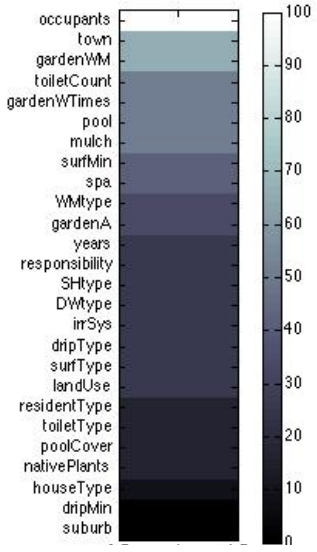

Figure 12: Selection frequency obtained considering as output the per-capita daily average consumption.

Figure 13 and Figure 14 show the results obtained running the feature weighting algorithms on X1 and X2, respectively. Again, the features are reported on the y-axis, while the x-axis represents different algorithm runs. Colors represent the positions of each feature in the weighting ranking: features with white color were given higher weights by the algorithms, meaning they are considered as relevant in explaining the output variable, while darker features are associated to lower weights (i.e., less relevant).

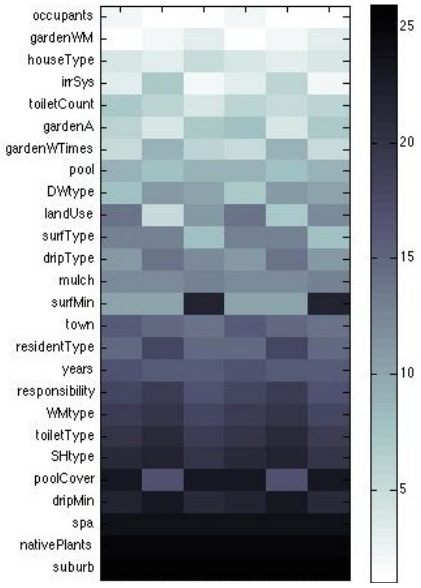

Figure 13: Weighting ranking considering as output the daily average consumption.

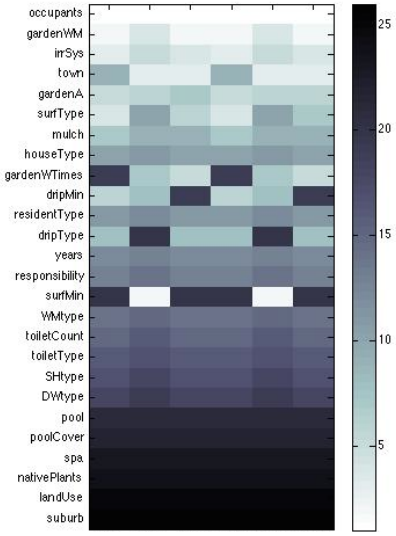

Figure 14: Weighting ranking considering as output the per-capita daily average consumption.

The two feature weighting algorithms produce consistent results, which are also consistent with the ones obtained by the feature selection algorithms, thus suggesting clear and strong relationships between the extracted features and the corresponding water consumption profiles.

4.5.2 Interpretation of the feature extraction results

The set of features extracted in the previous section has been analyzed to better understand

the underlying relationships between them and the water consumption profiles.

OCCUPANTS

The first considered feature is the number of occupants of the house, as it was always ranked in the first position by all the implemented feature extraction algorithms. As expected, Figure 15 shows that the daily water consumption increases with the number of occupants. Yet, the per-capita consumption decreases as the number of household's occupants increases. The reason for this behavior can be double:

1) 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;
2) when the number of household's occupants increases, some kind of economies of scale and social pressure are developed. As a consequence, water use is better balanced among the inhabitants and wastes are less frequent.

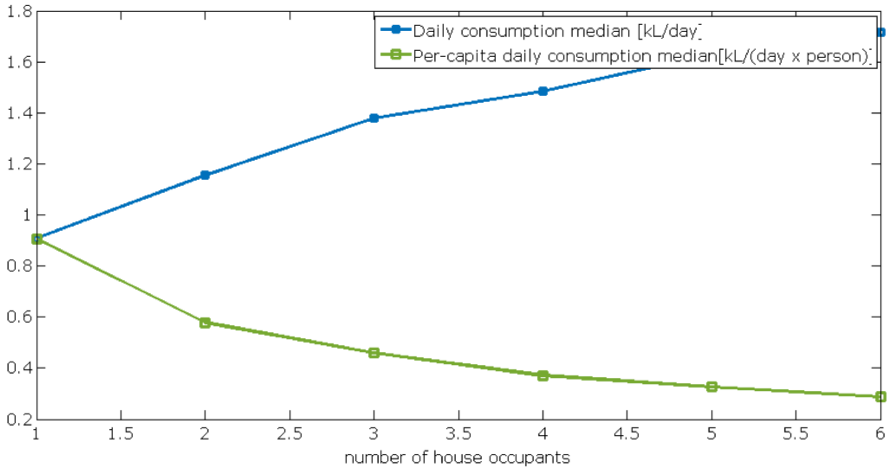

Figure 15: Median daily water consumption and median per-capita daily water consumption for houses with different number of occupants.

TOILET NUMBER

Considering now the number of toilets (see Figure 16) both the average daily and average daily per-capita water consumption level increase with the number of toilets in the house. Since the number of toilets generally increases with the number of household's occupants, it is reasonable that the daily water consumption increases with the number of toilets as well. In contrast with the previous case, in this case the per-capita consumption increases, probably because with a higher number of toilets there is less "competition" for using the resources (i.e., the toilet).

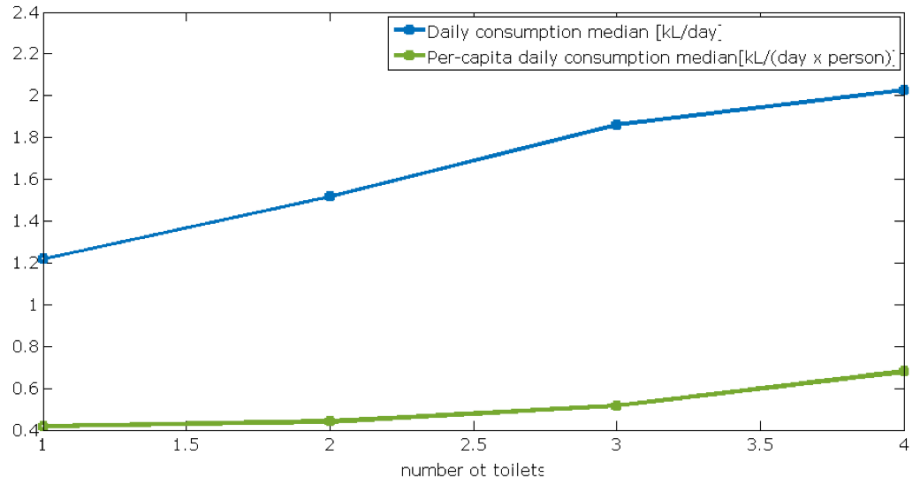

Figure 16: Median daily water consumption and median per-capita daily water consumption for houses with different number of toilets.

HOUSE TYPE

Figure 17 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. The per-capita water consumption flattens for the reasons previously discussed about number of occupants and associated consumption.

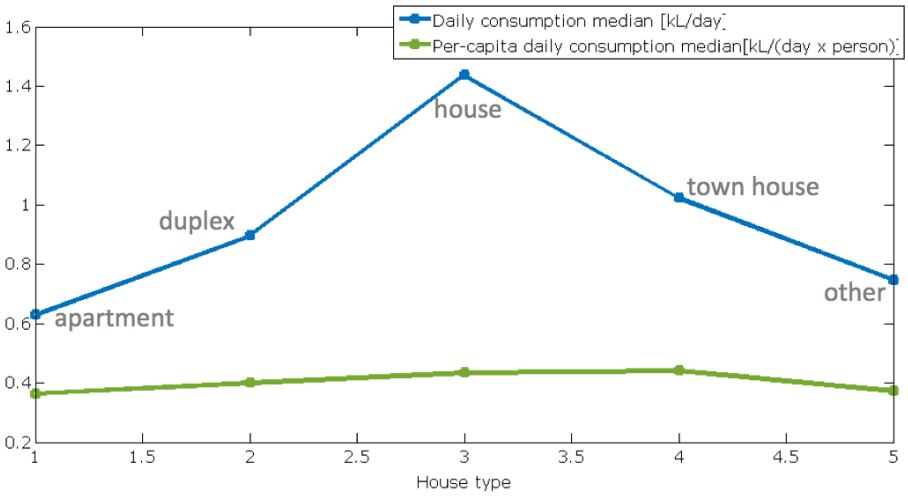

Figure 17: Median daily water consumption and median per-capita daily water consumption for different types of house.

IRRIGATION

The relationship between water consumption and the type of irrigation is shown in Figure 18. 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 (bottom part of the figure). 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.

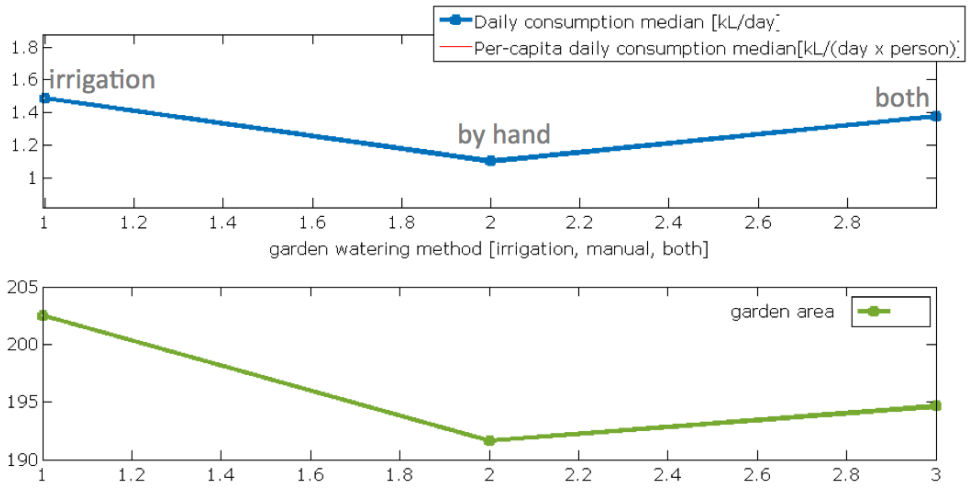

Figure 18: Median daily water consumption and median per-capita daily water consumption for different types of irrigation systems.

4.5.3 Forecasting user 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. Such a model allows water utilities and municipalities to quantitatively assess the effectiveness of future water demand management strategies.

Working on low-resolution consumption data, our model allows classifying users to the three consumption profiles introduced in Section 4.2.2, namely low-, medium-, and high-consumers.Among the available data-driven modeling methods, we employed naive *Bayesian Regressor* and *Decision Tree* classifiers, 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 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. Figures 17-20 report the resulting average model 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. The proposed method 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.

The final model accuracy might improve when moving from low-resolution billed data on water consumption to high-resolution smart-metered data, as they would allow the definition of more detailed user profiles on the basis of the disaggregated end-use patterns. However, the underlying relationships between users' features and end-use patterns are likely more

complex and, consequently, the effectiveness of the proposed approach should be further tested and validated. Finally, since the users' psychographics 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 a direct interaction with the users for data gathering.

Table 11: Classifier models accuracy on daily household water demand prediction.

Feature selection algorithm	Bayes Naive accuracy (mean and standard deviation) over three runs [%]	J48 tree accuracy (mean and standard deviation) over three runs [%]
FCBF	0.51±0.03	0.51±0.03
CFS	0.51±0.01	0.49±0.02
BLOGREG	0.50±0.03	0.48±0.03
SBMLR	0.51±0.02	0.49±0.00

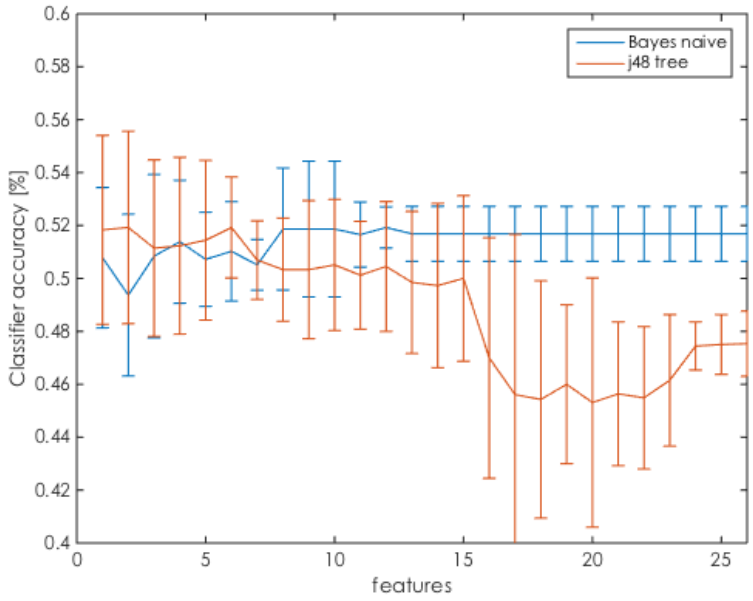

Figure 19: Classifier models accuracy on daily household water demand prediction, considering Chi2 feature weighting algorithm.

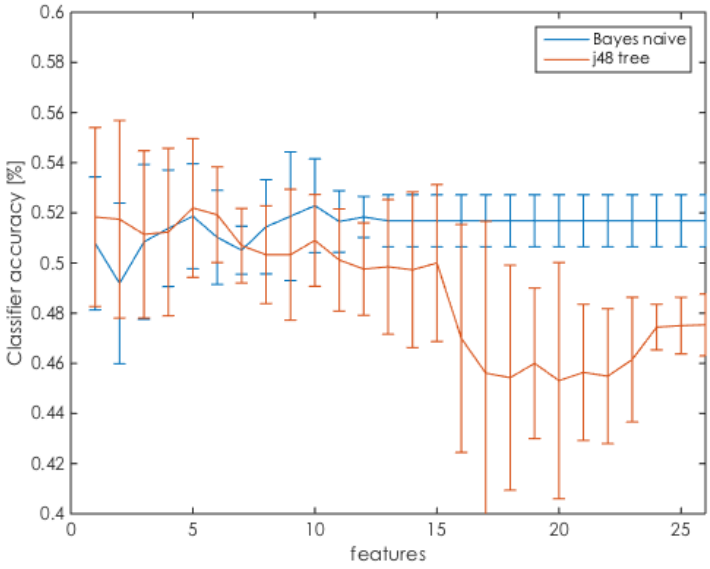

Figure 20: Classifier models accuracy on daily household water demand prediction, considering Infogain feature weighting algorithm.

Table 12: Classifier models accuracy on per-capita daily household water demand prediction.

Feature selection algorithm	Bayes Naive accuracy (mean and standard deviation) over three runs [%]	J48 tree accuracy (mean and standard deviation) over three runs [%]
FCBF	0.56±0.02	0.54±0.00
CFS	0.56±0.02	0.54±0.00
BLOGREG	0.55±0.02	0.53±0.00
SBMLR	0.56±0.02	0.52±0.01

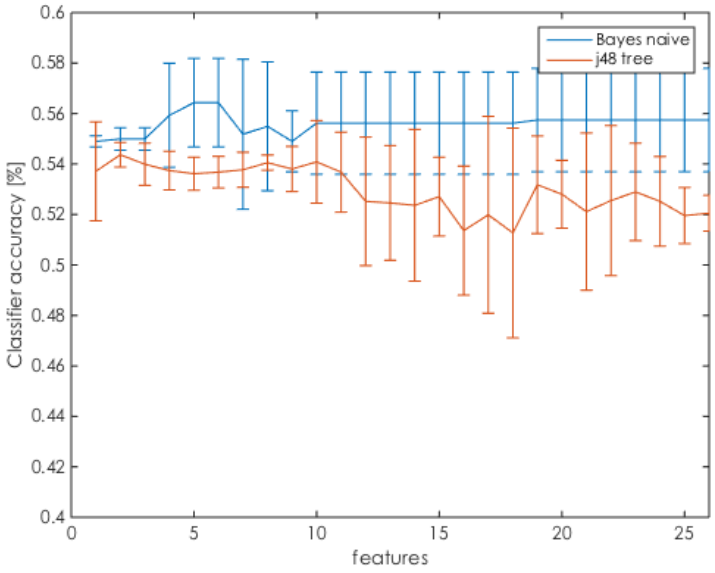

Figure 21: Classifier models accuracy on per-capita daily household water demand prediction, considering Chi2 feature weighting outputs.

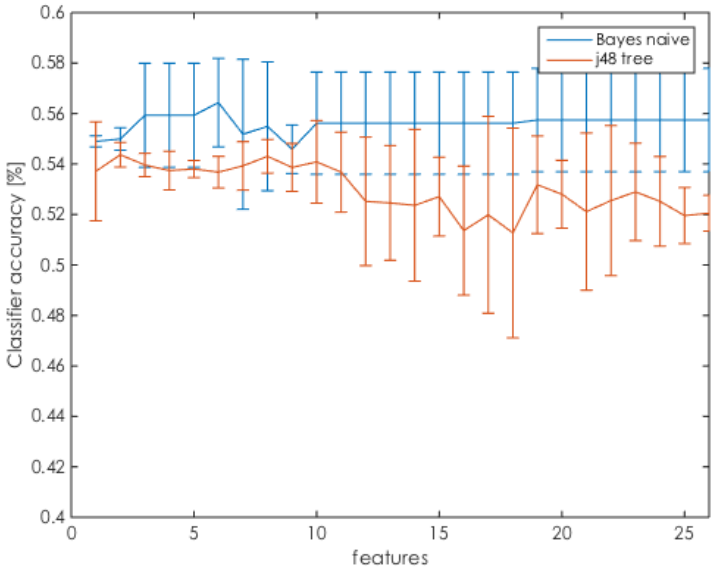

Figure 22: Classifier models accuracy on per-capita daily household water demand prediction, considering Infogain feature weighting outputs.

5. Conclusions and follow-up

In this deliverable, a set of algorithms for data-driven modelling of the user behaviour have been presented. Specifically:

- Two novel algorithms for water end use characterization developed in the SmartH2O project have been described;
- Several machine learning and data-mining algorithms have been applied to build user behavioural models from low-resolution water consumption data.

The models of user behaviour derived through the proposed algorithm will be implemented in the SmartH2O platform. Through the SmartH2O platform, the water utility can visualize the water consumption of each customer at a fixture/appliance level, in order to identify consumption patterns and trends, and thus identifying the most promising areas where conservation efforts may be polarized. Furthermore, the water utility can foresee the consumer behavior in front of exogenous variables (climate), social awareness campaigns, social pressure, water restrictions, etc.

Next steps within WP3 will be:

- Testing the water end use characterization algorithms and the user modelling algorithms against high resolution water consumption data provided by the water utilities taking part at the SmartH2O project (i.e., Thames Water and SES);
- The development of an agent based model that, by combing the single user behavioural models and a set of rules describing the social interaction among the consumers, allows the water utility to simulate whole districts of users and to understand how some user types (leaders/influencers) can stimulate a behavioural change on other users.

6. References

- [Anda12] M. Anda, J. Brennan and E. Paskett, "*Behaviour change programs for water efficiency: Findings from North West and Metropolitan Residential Programs in Western Australia*". In: IWA World Water Congress & Exhibition, September, Busan, Korea, 2012.
- [Aquacraft11] "*Albuquerque Single-family Water Use Efficiency and Retrofit Study*", Aquacraft Inc., 2011. Available online at: http://www.aquacraft.com/node/71
- [Batra14] N. Batra, J. Kelly, O. Parson, H. Dutta, W. Knottenbelt, A. Rogers, A. Singh, M. Srivastava, "*NILMTK: An open source Toolkit for Non-Intrusive Load Monitoring*", in Fifth International Conference on Future Energy Systems (ACM e-Energy), Cambridge, UK, 2014. arXiv:1404.3878 DOI:10.1145/2602044.2602051.
- [Beal11] C. Beal and R. Stewart, "*South East Queensland Residential End Use Study-Final Report*", 2011. Available online at: http://www.urbanwateralliance.org.au/publications/UWSRA-tr47.pdf
- [Bennett13] C. Bennett, R. A. Stewart, and C. D. Beal, "*Ann-based residential water end-use demand forecasting model*", Expert Systems with Applications, vol. 40, 2013.
- [Blokker10] E. Blokker, J. Vreeburg, and J. van Dijk, "*Simulating residential water demand with a stochastic end-use model*", Journal of Water Resources, Planning and Management, vol. 136, 2010.
- [Camier13] T. R. Camier, S. Giroux, B. Bouchard, and A. Bouzouane, "*Designing a NIALM in smart homes for cognitive assistance*," Procedia Computer Science, vol. 19, pp. 524–532, 2013.
- [Cawley07] G. C. Cawley, N. L. Talbot, M. Girolami, "*Sparse multinomial logistic regression via bayesian l1 regularisation*", Advances in neural information processing systems, vol. 19, 2007.
- [Cover12] T. M. Cover, and T. A. Thomas, "*Elements of information theory*", John Wiley & Sons, 2012.
- [DeOreo11] B. DeOreo, "*California Single-Family Water Use Efficiency Study*", Aquacraft Inc., 2011. Available online at: http://www.aquacraft.com/node/63
- [Dong13] R. Dong, L. Ratliff, H. Ohlsson, and S. Sastry, "*A dynamical systems approach to energy disaggregation*," in IEEE 52nd Annual Conference on Decision and Control, 2013.
- [Figueiredo13] M. Figueiredo, B. Ribeiro, and A. de Almeida, "*On the regularization parameter selection for sparse code learning in electrical source separation*," in Adaptive and Natural Computing Algorithms. Springer, 2013.
- [Forney73] G.D. Forney, "*The Viterbi algorithm*", Proceedings of the IEEE, vol. 61, 1973.
- [Fox09] C. Fox, B. McIntosh, and P. Jefirey, "*Classifying households for water demand forecasting using physical property characteristics"*, Land Use Policy, vol. 26, 2003.
- [Froehlich09] Froehlich *et al.*, "*HydroSense: infrastructure-mediated single-point sensing of whole-home water activity*", In Proceedings of the 11th international Conference on Ubiquitous Computing, Orlando, Florida, 2009
- [Froehlich11] J. Froehlich *et al.*, "*A Longitudinal Study of Pressure Sensing to Infer Real-World Water Usage Events in the Home*", In Proceedings of Pervasive, 2011.
- [Galelli13] S. Galelli, and A. Castelletti, "Tree-based iterative input variable selection for hydrological modeling". Water Resour. Res. 49, 42954310, 2013.
- [Gato06] S. Gato, N. Jayasuriya, and P. Roberts, "*Forecasting urban residential water demand*", Ph. D. thesis, RMIT University, 2006.
- [Gato11] S. Gato, N. Jayasuriya, P. Roberts, "*Understanding urban residential end uses of water*", Water Science & Technology 64 (1), 2011.

- [Guyon02] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, "*Gene selection for cancer classification using support vector machines",* Machine learning, vol. 46 (1-3), 2002.
- [Guyon03] I. Guyon, and A. Elisseeff, "*An introduction to variable and feature selection",* Journal of Machine learning Research, vol. 3, 2003.
- [Hart92] G. Hart, "*Nonintrusive appliance load monitoring*," Proceedings of the IEEE, vol. 80, no. 12, 1992.
- [Heinrich07] M. Heinrich, "*Water End Use and Efficiency Project*", 2007. Available online at: http://www.branz.co.nz/cms_show_download.php?id=9bf916e031023c9323d5abe093a02a0b0741cc9e
- [Johnson13] M. Johnson and A. Willsky, "Bayesian nonparametric hidden semi Markov models," The Journal of Machine Learning Research, vol. 14, no. 1, 2013.
- [Kolter07] J. Kolter and T. Jaakkola, "*Approximate inference in additive Factorial HMMs with application to energy disaggregation*," in International Conference on Artificial Intelligence and Statistics, 2012.
- [Kowalski05] M. Kowalski and D. Marshallsay, "*Using measured microcomponent data to model the impact of water conservation strategies on the diurnal consumption profile*", Water Supply, 2005.
- [Likas03] A. Likas, N. Vlassis, and J. Verbeek, "*The global k-means clustering algorithm*," Pattern recognition, vol. 36, no. 2, 2003.
- [Liu95] H. Liu, and R. Setiono, "*Chi2: Feature selection and discretization of numeric attributes*", In: IEEE 24th International Conference on Tools with Artificial Intelligence. IEEE Computer Society, 1995.
- [Loh03] M. Loh, and P. Coghlanm, "*Domestic water use study: In Perth, Western Australia, 1998-2001*", Water Corporation, 2003.
- [Makki13] A. A. Makki, R. A. Stewart, K. Panuwatwanich, and C. Beal, "*Revealing the determinants of shower water end use consumption: enabling better targeted urban water conservation strategies*", Journal of Cleaner Production, vol. 60, 2013.
- [Makonin13] S. Makonin, F. Popowich, L. Bartram, B. Gill, and I. Bajic, "*AMPds: a public dataset for load disaggregation and eco-feedback research*," in Electrical Power and Energy Conference (EPEC), 2013 IEEE, 2013.
- [Maier00] Maier, H.R., Dandy, G.C. "Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications". Environ. Model. Softw. 15, 101e124, 2000.
- [Maier10] Maier, H.R., Jain, A., Dandy, G.C., Sudheer, K.P. "Methods used for the development of neural networks for the prediction of water resource variables in river systems: current status and future directions". Environ. Model. Softw. 25, 891e909, 2010.
- [Mayer99] P.W. Mayer and W. B. DeOreo, "*Residential end uses of Water"*, AWWA Research Foundation and American Water Works Association, 1999. Available online at: http://www.aquacraft.com/node/56
- [Nguyen13] K.A. Nguyen, H. Zhang, and R. Stewart, "*Development of an intelligent model to categorise residential water end use events*", Journal of Hydro-environment Research, vol. 7, 2013.
- [Olmstead07] S. M. Olmstead, M. W. Hanemann, and R. N. Stavins, "*Water demand under alternative price structures*", Journal of Environmental Economics and Management, vol. 54, 2007.
- [Olmstead09] S. M. Olmstead, and R. N. Stavins, "*Comparing price and non price approaches to urban water conservation*", Water Resources Research, vol. 45, 2009.
- [Parson12] O. Parson, S. Ghosh, M. Weal, and A. Rogers, "*Non-Intrusive Load Monitoring using prior models of general appliance types*", in AAAI, 2012.
- [Rabiner89] L. Rabiner, "*A Tutorial on Hidden Markov Models and Selected Applications*

in Speech Recognition", Proceedings of the IEEE, vol. 77, 1989.

- [Roberts05] P. Roberts, "Yarra Valley *Water 2004 Residential End Use Measurement Study"*, 2005. Available online at: https://www.yvw.com.au/yvw/groups/public/documents/document/yvw1001680.pdf
- [Sakoe78] H. Sakoe and S. Chiba, "*Dynamic programming algorithm optimization for spoken word recognition*", IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 26, 1978.
- [Srinivasan06] D. Srinivasan, W. Ng, and A. Liew, "*Neural-network-based signature recognition for harmonic source identification*," IEEE Transactions on Power Delivery, vol. 21, no. 1, 2006.
- [Suero12] F. Suero, P. Mayer, D. Rosenberg, "*Estimating and verifying united states households' potential to conserve water*", Journal of Water Resources Planning and Management, vol. 3, 2012.
- [Suzuki08] K. Suzuki, S. Inagaki, T. Suzuki, H. Nakamura, and K. Ito, "Nonintrusive appliance load monitoring based on integer programming," in SICE Annual Conference, 2008
- [Syme04] G. J. Syme, Q. Shao, M. Po, and E. Campbell, "Predicting and understanding home garden water use. Landscape and Urban Planning, vol. 68, 2004. [Talebpour14] M. Talebpour, O. Sahin, R. Siems, and R. A. Stewart, "*Water and energy nexus of residential rain water tanks at an end use level: Case of Australia*". Energy and Buildings, 2014.
- [Willis11] R. M. Willis, R. A. Stewart, D. P. Giurco, M. R. Talebpour, and A. Mousavinejad, "End use water consumption in households: impact of sociodemographic factors and efficient devices", Journal of Cleaner Production, vol. 60, 2011.
- [Yu03] L. Yu, and H. Liu, "*Feature selection for high-dimensional data: A fast correlation-based filter solution*". In: ICML. vol. 3, 2003.
- [Zeifman11] M. Zeifman and K. Roth, "*Nonintrusive appliance load monitoring: Review and outlook*," IEEE Transactions on Consumer Electronics, vol. 57, no. 1, 2011.
- [Zhao10] Z. Zhao, F. Morstatter, S. Sharma, S. Alelyani, A. Anand, and H. Liu, "*Advancing feature selection research*", ASU Feature Selection Repository, 2010.
- [Zia11] T. Zia, D. Bruckner, and A. Zaidi, "*A Hidden Markov Model based procedure for identifying household electric loads*," in IECON 2011-37th Annual Conference on IEEE Industrial Electronics Society, 2011.
- [Zoha12] A. Zoha, A. Gluhak, M. Imran, and S. Rajasegarar, "*Non-intrusive load monitoring approaches for disaggregated energy sensing: A survey*", Sensors, vol. 12, no. 12, 2012.