



Original software publication

WEUSEDTO—Water End USE Dataset and TOols: An open water end use consumption dataset and data analytics tools



Anna Di Mauro^{*}, Salvatore Venticinque, Giovanni Francesco Santonastaso, Armando Di Nardo

Università della Campania Luigi Vanvitelli, Via Roma, 29, 81031, Aversa (CE), Italy

ARTICLE INFO

Article history:

Received 2 April 2021

Received in revised form 31 December 2021

Accepted 22 September 2022

Keywords:

Dataset

Water end-use data

Data modelling

Water management

Users' profiling

ABSTRACT

Globalization, climate changes, innovative technologies and new human habits have increased attention to water conservation and management. Therefore, behavioural studies became a key element to understand how and when water is used in residential environment. Water End USE Dataset and TOols (WEUSEDTO), an open water end use consumption dataset and data analytics tools, has been released to help researchers, water utilities and companies to test models and algorithms on real water consumption data. The dataset combines with some notebook python able to analyse high-resolution water data (data recorded with 1 sample per second) to provide several tools to manage raw data, compute statistical analysis, learn fixture usage and generate synthetic simulation models. In addition, washbasin flow data were used as a test case to illustrate the main features of WEUSEDTO: providing volume and duration of single events, classifying usages and simulating user's behaviour.

© 2022 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Code metadata

Current code version

Permanent link to code/repository used for this code version

Code Ocean compute capsule

Legal Code License

Code versioning system used

Software code languages, tools, and services used

Compilation requirements, operating environments & dependencies

If available Link to developer documentation/manual

Support email for questions

v1.0

<https://codeocean.com/capsule/9225099/tree/v1>

10.24433/CO.3634054.v1 <https://codeocean.com/capsule/9225099>

GPLv3

git

python

matplotlib, pandas, numpy, scipy, sklearn, joblib

<https://water-end-use-dataset-tools.github.io/WEUSEDTO/docs/html/>

anna.dimauro@unicampania.it, salvatore.venticinque@unicampania.it, ing.dimauroanna@gmail.com

1. Motivation and significance

Globalization, urbanization and climate changes increased attention for water scarcity and environmental sustainability issues. Water management has evolved recognizing water consumption data (WCD) as an input for decision-making processes in water distribution systems to achieve water conservation strategies, improve water network efficiency, promote demand management actions and, reduce costs [1,2]. Water conservation in urban areas is an ongoing challenge. Depending on the interest towards the

impact of human behaviour on WCD [3], caused by economical and social changes over the past decades, it emerged an increasing interest on modelling end-use consumption data to investigate how users drive water usages [4]. On this side, the progress in smart water metering technology improved the availability of WCD revealing considerable benefits for water demand modelling [5]. These include, e.g., behavioural studies, customer segmentation, data-driven model, and user-oriented water demand management strategies. Extracting information on water use behaviours from smart meters data allows to understand the way water is used in residential environment. In the literature, several methods of analysing smart water meter data have been proposed to extract water event information, such as flow trace

^{*} Corresponding author.

E-mail address: anna.dimauro@unicampania.it (Anna Di Mauro).

analysis [6], derivative signals [7], classification techniques [8,9] and pressure sensing [10,11].

Despite smart meter evolution, there is a lack of open datasets at household and end-use levels due to the complexity of assessing case studies [12], i.e. intrusiveness of smart meters, privacy issue, difficulty in sensors placement, etc. Water demand studies need data easily available in formats that suit researchers' needs, at relevant spatial and temporal resolutions that are useful to explore water issues.

Furthermore, available data need software and/or statistical tools to extract meaningful information that can be used for more effective decision-making [13]. For this reason, the great interest in gathering and using WCD goes with the spread of the development of software able to manage them.

Many tools have been proposed to analyse water end-use data with different aims: water demand forecasting [14], data filtering [15], statistical tools [16], IoT monitoring systems management [17], integrated urban water modelling [18] and especially water demand disaggregation [7].

These tools use surveys or household-level data to gain further insight into residential water use patterns. As reported in [19], to the authors' knowledge, there are neither relevant open data repository, and nor studies based on data collected real time at fixture level. The lack of this kind of data in the literature [19] represents a barrier for the development of innovative algorithms/techniques for demand-side management and forecasting, and to provide data analytics tools able to develop accurate characterizations of end-use water consumption profiles. Then, the main contribution of this work is providing the first open water end-use dataset, and open-source software for its analysis.

This paper presents Water End USE Dataset and TOols (WEUSEDTO), an open water end use dataset and set of tools, able to be used as training dataset&tools to investigate machine learning techniques, characterize end-use water consumption profile, as well as to test, analyse and identify innovative water solutions and management strategies. WEUSEDTO contributes to identifying customers' behaviours essential for user profiling. Moreover, the dataset obtained as reported in [20], it is potentially eligible as numerical benchmarks for training and testing water fixtures signatures, while the software can be used to identify fixtures associated statistics that can be employed for synthetic simulation models as reported in [21]. In addition, the software allows to download/collect time-series, run analysis and provide visualization notebooks.

2. Software description

The presented software allows for the analysis of water demand time-series obtained from raw measures at the fixture level. It has been used to experiment with a methodology that aims at building models of water consumption profiles by the integrated use of statistical parameters and machine learning techniques. The final goal is the exploitation of such models to simulate and predict water consumption at a larger scale (several users, buildings, etc.), but also to design and develop, in future works, disaggregation techniques. The software is available in Zenodo [22] and in this public GitHub repository: <https://github.com/Water-End-Use-Dataset-Tools/WEUSEDTO>. The software can run on any operating system, a Python interpreter (V3.8) has been used to run the current release, and common libraries for data analysis (pandas, scipy, numpy, sklearn, matplotlib, joblib.) are required. The detailed list of required python libraries and other resources can be found at the GitHub repository reported above.

2.1. Software architecture

The software is organized by four Python packages shown in Fig. 1. The *timeseries* package includes the code used to process the raw data to detect relevant sequences of samples or to compute a list of defined features of a time series. The *model* package is used to abstract the objects of the application domain that allow building a consumption profile. Three models are provided to characterize the statistical distribution of water consumption of a certain fixture according to the user's behaviour. Moreover, parametric modelling of the time-series profile is addressed too. The *learning* package uses machine learning techniques for time-series clustering and prediction of fixture usage. Finally, the *simulation* package exploits the models generated and the machine learning techniques to simulate water usage of multiple users, whose behaviour can be assimilated to the one extracted from the measures. Some examples of programs, which use the core packages, are provided to demonstrate how the software can be used from the data analysis to the model generation and for the final simulation.

2.2. Software functionalities

Software functionalities are distributed among the package presented before. The *timeseries* package allows for detecting the occurrence (start and stop) of each fixture usage. A simple splitting function works by comparing the sample with an empirical threshold, which represents the flow rate of a single drop. The threshold value is equal to 6 ml/s and it corresponds to the minimum value of the flow registered by the sensor. Additionally, a more complex algorithm was developed to take into account the delay of sensor transmission and cut usages with a volume lower than a predefined minimum value.

Functions for the computation of significant parameters of each usage are also developed. Finally, filtering of overlays, which correspond to the detection of abnormal usages, is supported. For example, usage whose duration, number of samples or amount of consumed water result below defined values, which can be set by the software, are considered.

The software allows for extracting from the set of detected usages statistical parameters for modelling user's behaviour in terms of usage of that fixture. Three kinds of user's profiles can be generated by the software: global, monthly and weekly usage. The one that best fits the user's behaviour models the probability that a fixture will be used in a day, at a certain hour of the day.

Learning techniques are investigated to understand from data if a different class of water usage, which can be correlated to specific user's activities (e.g., teeth brushing, hand washing, ...), can be distinguished. In particular clustering techniques are used to identify groups of similar time-series and the centroid of each cluster is represented by a spline approximation. The spline representation of clusters complements the statistical one to have a complete model of fixture usage. Machine learning is also used to infer the correlation between the time occurrence of usage (e.g., the day of the week and the hour of the day). The provided implementation uses the Random Forest algorithm.

The last functionality of the software is used to simulate the water consumption of n people who use m fixtures. The number of people n is established a priori and remains constant over time for each simulation. For each of $n \times m$ fixtures the related statistical model is exploited to generate a probable distribution of usages in a day by the users. The Random Forest algorithm allows predicting to which cluster each usage would belong. The spline representation of that cluster is used to reproduce the time-series of the water consumption flow.

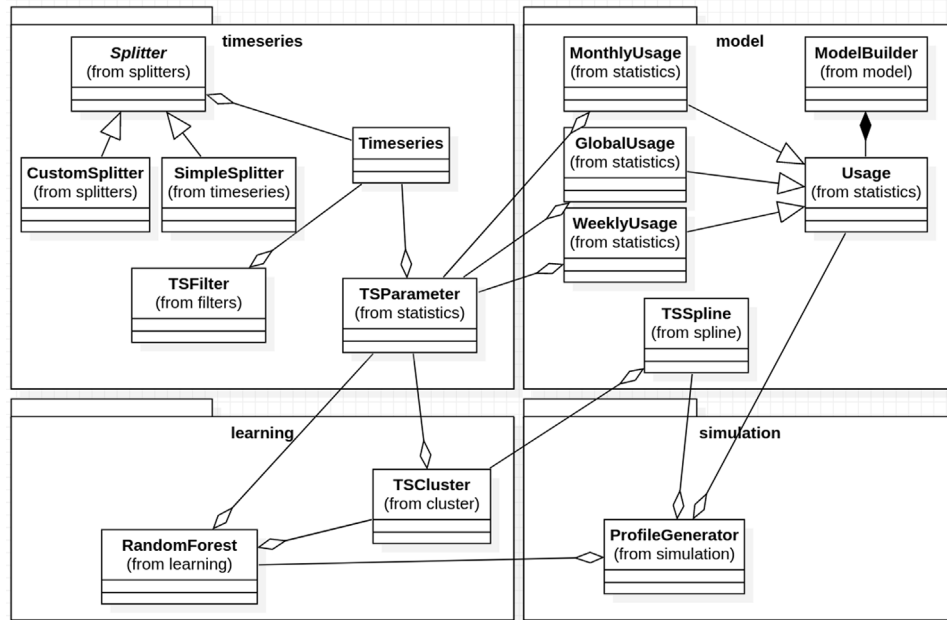


Fig. 1. Unified Modeling Language (UML) class diagram.

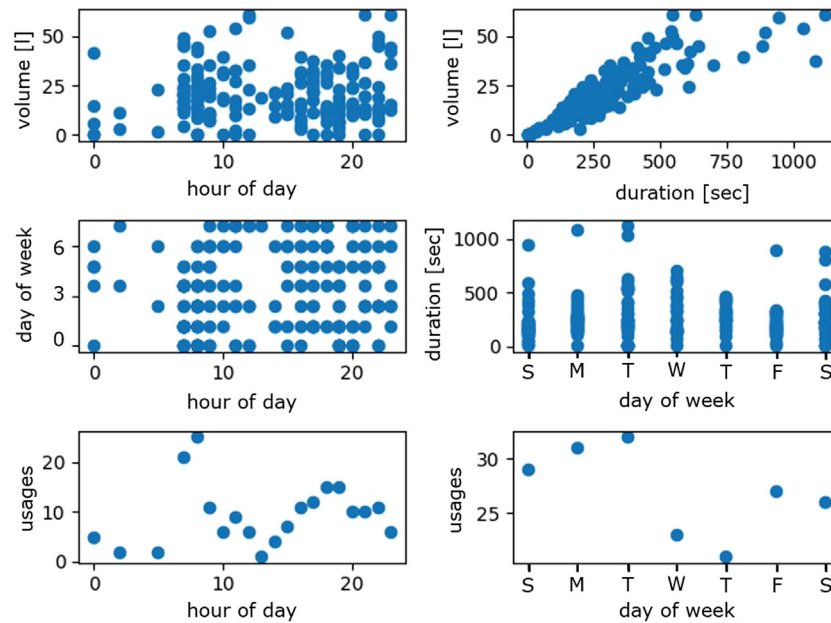


Fig. 2. Water consumption parameters.

2.3. Original data

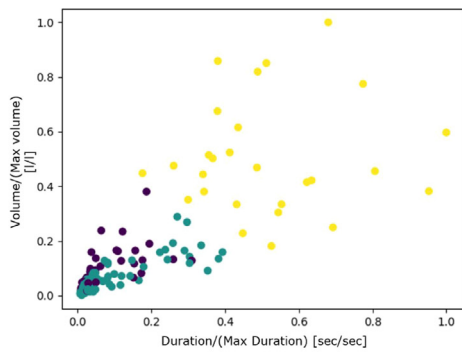
The software repository contains 9 time-series of raw measures exported as csv files. They correspond to the fixtures monitored in the apartment used as a case study and, to the whole-house aggregate consumption. Data gathered have 1s resolution for disaggregate time series and 10 s for the aggregate measurements, spanning 1 year from March 2019–October 2020. WCD at fixture level are collected using an Internet of things (IoT) water end use monitoring system, as reported in [20]. WCD at the household level are gathered using an ultrasonic water meter based on Long Range (LoRa) wireless transmission technology. Data specification and WEUSEDTO time series are available in this public GitHub repository: <https://github.com/AnnaDiMauro/>

WEUSEDTO-Data. The data are released with the Creative Commons Attribution 4.0 International License CC-BY-4.0.

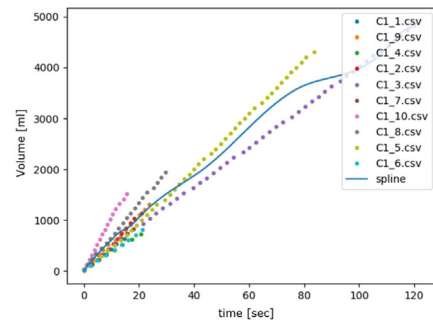
3. Illustrative examples

To explain the main functionalities of WEUSEDTO the outputs of the following packages are reported: *models*, *learning* and *simulations*. Washbasin was used as a test case to show the results of the software. In Fig. 2, the usages of washbasin fixture, detected by *models* package, are represented, specifically volume [L], duration [s] and the number of usage are displayed with two temporal scales: the hour of day and day of the week.

The outputs of the *learning* package are shown in Fig. 3. The clustering results for the washbasin usages are reported in terms of duration of usage against the amount of consumed

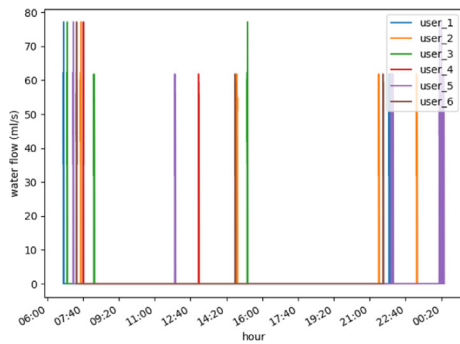


(a) Duration against volume of clustered usages.

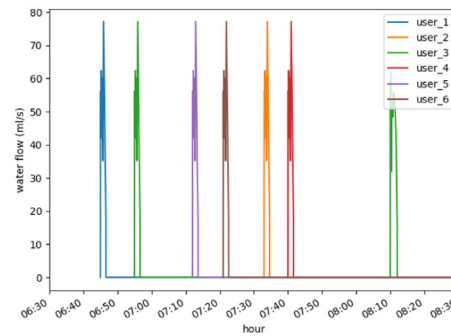


(b) Spline representation of a cluster.

Fig. 3. Modelling time-series usage by clustering and spline approximation.



(a) All day visualization



(b) Zoom-in visualization in early hours.

Fig. 4. Simulation results of Washbasin usage by six users in a random day of August.

water (Fig. 3(a)), three clusters were identified: C1, C2 and C3. Additionally, Fig. 3(b) shows ten time series belonging to cluster C1 and the spline which represents the mean profile of water consumption event of the cluster. Specifically, the regression function spline is computed using the methodology reported in [23].

To simulate 6 users who exhibit the same behaviour we exploited the statistical model to generate, for each user, the time distribution of fixture usages during the day. Using the random forest algorithm a trained model is used to associate to each occurrence of washbasin usage the corresponding cluster. Then the water profile of different usages have been generated from the corresponding splines. Then, the *simulation* package results are shown in Fig. 4(a). The same time-series is magnified in Fig. 4(b). Preliminary estimation of the accuracy of the software, and its test with another similar dataset have been performed and provided admissible results. However, further investigations will be carried out in future work. The interested readers can visit the GitHub repository for further and updated information about all the elaborations carried out with the original dataset and to the water data belonging to the Almanac of Minutely Power dataset (AMPdS) [24].

4. Impact

Water utilities, stakeholders, and researchers struggling to find available high-resolution datasets and tools valuable for research application to investigate and validate innovative solutions to improve water demand management, estimate demand peak timing, identify demand pattern, characterize users' behaviour, define itemized billing and tailor disaggregation techniques [19]. On this side, understanding how water consumption is shared

among individual fixtures (i.e., shower, toilet, tap, etc.) represents a key element to develop disaggregation algorithms able to disaggregate WCD metered at the household level into single end-use categories avoiding sensors placement at fixtures level. WCD gathered for each fixture of a residential household with high-resolution time sampling include water signatures corresponding to a single event of use that needs to be addressed using new data analytics tools able to correlate water demand to users' habits.

Due to sensors intrusiveness, water uses are frequently defined via surveys, audits and water event diaries and disaggregation is approached as a problem of blind identification [5]. In the water research literature, emerged the need to identify unsupervised or semi-supervised learning methods [25] that avoiding data acquisition allow to recognize typical consumption behaviours and patterns valuable to design personalized demand management strategies. Combining a software with high-resolution data at fixture level, WEUSEDTO addresses the need of both open water end-use consumption data and data analytics tools able to improve water disaggregation and investigate new machine learning techniques.

5. Conclusions

WEUSEDTO is an open water end use dataset and software based on real measurement of household fixtures with 1s time resolution. The software, organized by four Python packages *time-series*, *model*, *learning* and *simulation*, has been used to test a methodology to build a water consumption profiles model based on statistical analysis of time-series and machine learning techniques. The model obtained is used to understand how high-resolution data can be used to simulate and predict water consumption at different scale. The dataset coupled with software

constitutes a great opportunity for researchers and water utilities because it represents the starting point to develop tailored machine learning for water data, validate disaggregation algorithms and provide innovative management strategies. In fact, a detailed demand profiling and forecasting allow to act an advanced water supply aimed to provide more resource only when it is really needed, with several advantages in terms of available pressure, costs, and water losses reduction.

Funding

The research was conducted as part of the activities financed with the awarding of the V:ALERE: 2019 project of the University of Campania Luigi Vanvitelli.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

- [1] Inman D, Jeffrey P. A review of residential water conservation tool performance and influences on implementation effectiveness. *Urban Water J* 2006;3(3):127–43.
- [2] Sharma SK, Vairavamoorthy K. Urban water demand management: Prospects and challenges for the developing countries. *Water Environ J* 2009;23(3):210–8.
- [3] Willis RM, Stewart RA, Panuwatwanich K, Williams PR, Hollingsworth AL. Quantifying the influence of environmental and water conservation attitudes on household end use water consumption. *J Environ Manag* 2011;92(8):1996–2009.
- [4] Koop SH, Van Dorssen AJ, Brouwer S. Enhancing domestic water conservation behaviour: A review of empirical studies on influencing tactics. *J Environ Manag* 2019;247:867–76.
- [5] Cominola A, Giuliani M, Piga D, Castelletti A, Rizzoli AE. Benefits and challenges of using smart meters for advancing residential water demand modeling and management: A review. *Environ Model Softw* 2015;72:198–214.
- [6] DeOreo WB, Heaney JP, Mayer PW. Flow trace analysis to assess water use: Analyzing flow traces from residential water meters enabled researchers to collect precise data about water use by individual fixtures. *J Am Water Works Assoc* 1996;88(1):79–90. <http://dx.doi.org/10.1002/j.1551-8833.1996.tb06487.x>.
- [7] Bethke GM, Cohen AR, Stillwell AS. Emerging investigator series: Disaggregating residential sector high-resolution smart water meter data into appliance end-uses with unsupervised machine learning. *Environ Sci: Water Res Technol* 2021;7(3):487–503. <http://dx.doi.org/10.1039/d0ew00724b>.
- [8] Vitter JS, Webber M. Water event categorization using sub-metered water and coincident electricity data. *Water (Switzerland)* 2018;10(6).
- [9] Meyer BE, Jacobs HE, Ilembade A. Extracting household water use event characteristics from rudimentary data. *J Water Supply: Res Technol - AQUA* 2020;69(4):387–97.
- [10] Larson E, Froehlich J, Campbell T, Haggerty C, Atlas L, Fogarty J, et al. HydroSense : Disaggregated water sensing from a single , non-intrusive pressure-based sensor. *Pervasive Mob Comput* 2010. <http://dx.doi.org/10.1016/j.pmcj.2010.08.008>.
- [11] Froehlich J, Larson E, Saba E, Campbell T, Atlas L, Fogarty J, et al. A longitudinal study of pressure sensing to infer real-world water usage events in the home. In: *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)*, Vol. 6696 LNCS. 2011, p. 50–69. http://dx.doi.org/10.1007/978-3-642-21726-5_4.
- [12] Fagiani M, Squartini S, Gabrielli L, Spinsante S, Piazza F. A review of datasets and load forecasting techniques for smart natural gas and water grids: Analysis and experiments. *Neurocomputing* 2015;170:448–65. <http://dx.doi.org/10.1016/j.neucom.2015.04.098>.
- [13] Grover P, Kar AK. Big data analytics: A review on theoretical contributions and tools used in literature. *Glob J Flexible Syst Manag* 2017.
- [14] Pesantez JE, Berglund EZ, Kaza N. Smart meters data for modeling and forecasting water demand at the user-level. *Environ Model Softw* 2020;125.
- [15] Pastor-Jabaloyes L, Arregui FJ, Cobacho R. A filtering algorithm for high-resolution flow traces to improve water end-use analysis. *Water Sci Technol: Water Supply* 2019;19(2).
- [16] Kossieris P, Makropoulos C. Exploring the statistical and distributional properties of residential water demand at fine time scales. *Water (Switzerland)* 2018;10(10).
- [17] Pacheco CJ, Horsburgh JS, Tracy JR. A low-cost, open source monitoring system for collecting high temporal resolution water use data on magnetically driven residential water meters. *Sensors (Switzerland)* 2020;20(13):1–30.
- [18] Mostafavi N, Shojaei HR, Beheshtian A, Hoque S. Residential water consumption modeling in the integrated urban metabolism analysis tool (IUMAT). *Resour Conserv Recy* 2018;131:64–74.
- [19] Di Mauro A, Cominola A, Castelletti A, Di Nardo A. Urban water consumption at multiple spatial and temporal scales. a review of existing datasets. *Water (Switzerland)* 2021;13(1):1–31. <http://dx.doi.org/10.3390/w13010036>.
- [20] Di Mauro A, Di Nardo A, Santonastaso GF, Venticinque S. An IoT system for monitoring and data collection of residential water end-use consumption. In: *Proceedings - international conference on computer communications and networks*, Vol. 2019-July. IEEE; 2019, p. 1–6.
- [21] Di Mauro A, Di Nardo A, Santonastaso GF, Venticinque S. Development of an IoT system for the generation of a database of residential water end-use consumption time series. *Environ Sci Proc* 2020;2(1):20.
- [22] Di Mauro A, Venticinque S, Santonastaso G, Di Nardo A. WEUSEDTO-water end USE dataset and tools: an open water end use consumption dataset and dataanalytics tools. 2021, Available online: <http://dx.doi.org/10.5281/zenodo.4651443>.
- [23] Horn G, Venticinque S, Amato A. Inferring appliance load profiles from measurements. In: Di Fatta G, Fortino G, Li W, Pathan M, Stahl F, Guerrieri A, editors. *Internet and distributed computing systems*. Cham: Springer International Publishing; 2015, p. 118–30.
- [24] Makonin S, Ellert B, Bajić IV, Popowich F. Electricity, water, and natural gas consumption of a residential house in Canada from 2012 to 2014. *Sci Data* 2016;3. <http://dx.doi.org/10.1038/sdata.2016.37>.
- [25] Carboni D, Gluhak A, McCann JA, Beach TH. Contextualising water use in residential settings: A survey of non-intrusive techniques and approaches. *Sensors (Switzerland)* 2016;16(5). <http://dx.doi.org/10.3390/s16050738>.