



Article

Water End Use Disaggregation Based on Soft Computing Techniques

L. Pastor-Jabaloyes , F. J. Arregui *  and R. Cobacho

ITA-Grupo de Ingeniería y Tecnología del Agua, Dpto. de Ingeniería del Agua y Medio Ambiente, Universitat Politècnica de València, Camino de Vera s/n, 46022 València, Spain; laupasja@ita.upv.es (L.P.-J.); rcobacho@ita.upv.es (R.C.)

* Correspondence: farregui@ita.upv.es

Received: 22 November 2017; Accepted: 5 January 2018; Published: 9 January 2018

Abstract: Disaggregating residential water end use events through the available commercial tools needs a great investment in time to manually process smart metering data. Therefore, it is extremely difficult to achieve a homogenous and sufficiently large corpus of classified *single-use* events capable of accurately describe residential water consumption. The main goal of the present paper is to develop an automatic tool that facilitates the disaggregation of the individual water consumptions events from the raw flow trace. The proposed disaggregation methodology is conducted through two actions that are iteratively performed: first, the use of an advanced two-step filter, whose calibration is automatically conducted by the Elitist Non-Dominated Sorting Genetic Algorithm NSGA-II; and second, a cropping algorithm based on the filtered water consumption flow traces. As a secondary goal, yet complementary to the main one, a semiautomatic massive classification process has been developed, so that the resulting single-use events can be easily categorized in the different water end uses in a household. This methodology was tested using water consumption data from two different case studies. The characteristics of the households taken as reference and their occupants were unequivocally dissimilar from each other. In addition, the monitoring equipment used to obtain the consumption flow traces had completely different technical specifications. The results obtained from the processing of the two studies show that the automatic disaggregation is both robust and accurate, and produces significant time saving compared to the standard manual analysis.

Keywords: water end uses; water microcomponents; high frequency smart metering data; residential water flow trace disaggregation; water flow trace filtering

1. Introduction

Since the Brundtland Report [1] was presented, sustainability in the use of water resources has been a steady concern in designing water policies [2–4]. This is a problem with many different faces, from the source (surface or ground water, desalination, reclamation, etc.) to the use (agriculture, residential, industrial, environmental preservation, etc.). All of them are relevant, but bearing in mind that most of the human population lives in cities, urban water management becomes an issue of paramount importance. Therefore, accounting urban water consumption and knowing about end-uses at each customer's household is not only key because of the amount of water resource that is used and/or can be saved, but also because of many other considerations. In this regard, there are well-founded reasons for the research currently being conducted on residential end-uses, such as reduction in treatment costs linked to water consumption; improvement and better effectiveness of conservation measures in the urban environment; conservation of energy linked to water consumption; design optimization of indoor piping systems; improved demand forecasting models; etc.

As an essential tool for enhancing urban water management, the new technologies being implemented today in smart meters are making possible a significant leap forward in recording

and characterizing domestic water consumption. Further than the traditional monthly volume read, new meters may provide hourly consumption time patterns or a volume-flow pattern. They may also send alarms when a leak, a forgotten open faucet or a continuous back flow is detected, and all that information may be immediately sent through an AMR (automatic meter reading) system.

However, the above-mentioned capabilities are only the first tier when considering all the real possibilities current smart meters may yield. Though feasible today, a second, more advanced tier is not fully developed because of its notable complexity. It consists of high frequency monitoring—duration, volume and flow rate—of domestic water consumption, so that every single use in a household (hh) is accurately registered. Then, and after a detailed analysis, all consumption events can be categorized into the different end uses present in the household [5].

As soon as this will be soundly achieved on a large scale, new improvements in efficient water management strategies will be within reach. To name a few, from the consumer's perspective, water conservation measures could be tailored to each individual consumer [6], thus maximizing the saving potential in each case, or the variable term in the tariff could be designed according to consumer's characteristics to guarantee the balance between equity and income [7]. Furthermore, from the utility's view, water demand prediction models could be reliably produced from a more accurate bottom-up approach [8,9].

Nowadays, few commercial tools allow for this water end use analysis exercise—Trace Wizard[®] [10], Identiflow[®] [11] and BuntBrainForEndUses[®] [12]. However, any of these tools involves a great investment in time and human resources, as a significant part of the data processing work requires human intervention. Furthermore, the results from the analysis are unavoidably affected by arbitrary and constantly changing human criteria.

Alternatively, an automatic prototype based on machine learning algorithms was proposed by Nguyen et al. [13–16] to disaggregate and classify water consumption events. Unfortunately, the universal usability and compatibility of the tool is limited by the fact that the algorithms were trained with data originated from a specific water meter/data logger combination. In addition, all data were collected in the same geographical area from consumers sharing very similar water consumption habits and water appliances. Furthermore, the set of data employed for the training of the proposed machine learning tool has been obtained using Trace Wizard[®] software, which has limited capabilities for disaggregating overlapped consumption events [13]. Following a similar approach, Piga et al. [17] proposed an automated water and energy end use disaggregation, which has only been tested against electric energy data. Also, several start-ups claim to have developed software to automatically classify residential water consumption events into various uses [18–20]. Unfortunately, in this case there is not official or public information available about the processing tools and algorithms used, and the real performance achieved in the water end use classification for various types of households.

This paper presents a novel methodology to substitute manual water end use disaggregation and to produce more accurate sets of classified single end use events that can be employed as training sets for automatic recognition algorithms. Figure 1 depicts the general structure of the methodology, comprising two main processes: disaggregation, which is fully automatic and the main objective of this paper, and classification, which is semiautomatic at its current stage.

The proposed disaggregation process focuses on an advanced two-stage filtering based on the algorithm described in Pastor et al. [21], which is calibrated using the Elitist Non-Dominated Sorting Genetic Algorithm NSGA-II [22], and a new cropping algorithm having as input the filtered water consumption flow traces.

The contribution of this methodology can be summarized in two main aspects. The first one is the integration of a universal two-stage filtering algorithm that can be used to simplify, with a minimum loss of information, the flow traces originated in most commercial metering and logging equipment available in the market. The second one is the reduction of human intervention by automatically disaggregating *overlapped* water consumption events (as the one used as an example in Figure 1) into *single-use* events (examples in Appendix A), which are associated with individual uses of water through

different appliances. Both features facilitate and improve the processing of flow traces generated during a long-term metering campaign.

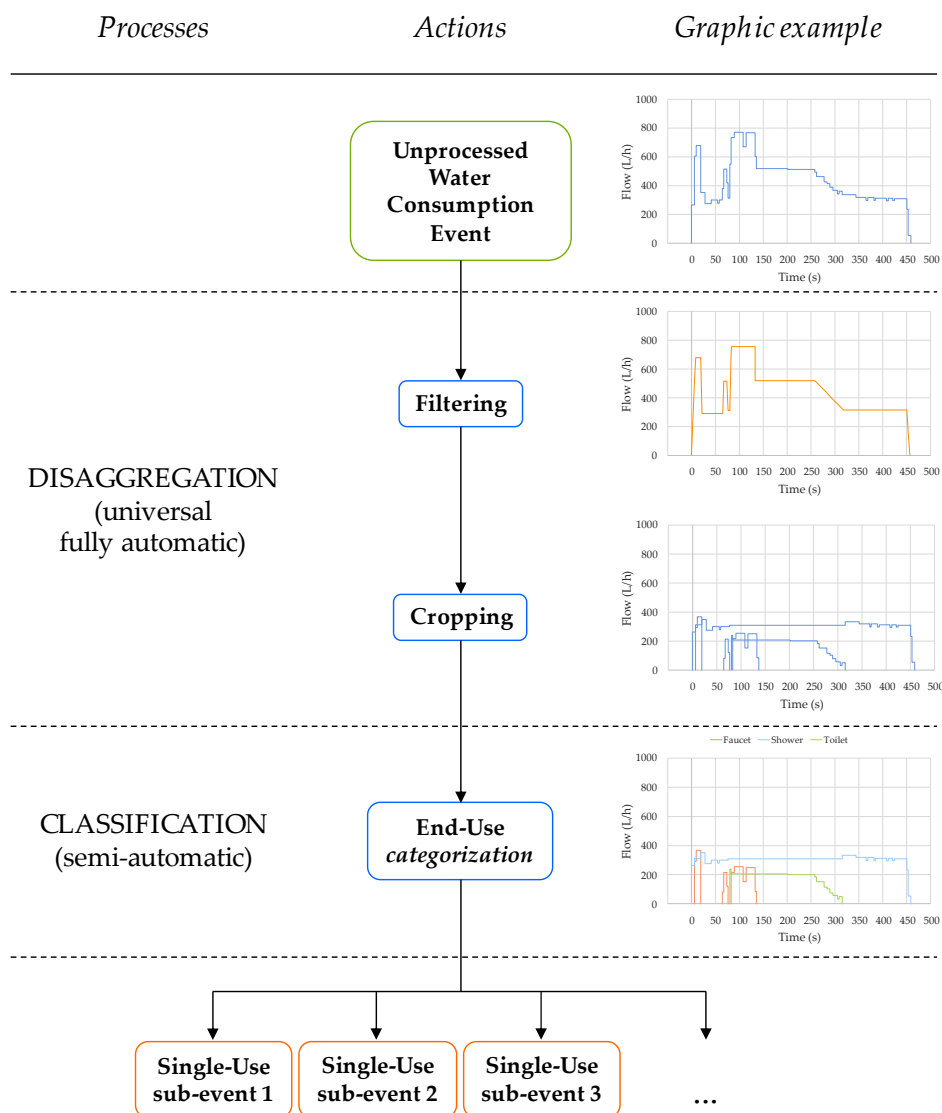


Figure 1. General structure of the proposed methodology.

The classification process corresponds to the use of unsupervised techniques to solve the classification of *single-use* events into the various water end use categories. This step is complementary to the disaggregation process and makes unnecessary the intervention of human analysts, except for validating the results, to classify thousands of automatically obtained *single-use* events. The paper presents a basic fully operational, yet semiautomatic, version to show the capabilities of these techniques for massive classification of consumption events.

As a case study, data originated from two independent water end uses studies conducted in distinct geographical areas were used to test the proposed filtering and disaggregation methodologies. The characteristics of the households and their occupants were unequivocally dissimilar from each other. Moreover, and in order to test the universal applicability of the methodology, the water consumption flow traces analyzed were obtained from monitoring equipment having completely different technical specifications.

BuntBrainForEndUses[®] was the commercial online software employed for the water end use analysis. This software offers the possibility of exporting raw flow traces and importing them back

after some manipulation has been carried out by the user. This feature is particularly useful for the study conducted as the filters and disaggregating algorithms can be developed with a specialized external analysis software, completely independent from BuntBrainForEndUses[®], and then have the results displayed and corrected in the online application. For the methodology presented, the filtering and disaggregation algorithms were programmed in R statistics [23].

2. Materials and Methods

The methodology proposed is divided into two processes (Figure 1). The first one, disaggregation, works on the original, and generally *overlapped*, consumption events in an iterative way until all the resulting subevents are either *single-use* (the most) or *uncertain* (a few) events (examples in Appendix A). By means of this methodology, the resulting subevents are more homogenous than the ones obtained by manual processing of flow traces through, for example, BuntBrainForEndUses[®]. This software application follows the same analysis procedure and allows human analysts to graphically crop water consumption flow traces into its various individual components.

The second process of the methodology, classification, assigns a specific water end use category to each *single-use* event. In the case study developed, this classification is done by identifying homogeneous subsets of events by means of a non-supervised learning technique and assigning a water end use to each one of them. Whatever the classification technique used, its effectiveness is increased by the fact that the subevents generated by the disaggregating algorithms consistently create subevents using homogeneous and well-defined criteria.

2.1. Disaggregation Process

Figure 2 shows the flow chart of the proposed disaggregation process, which breaks down the unprocessed events defined by the raw flow trace into simpler consumption events, and classifies the resulting subevents as *single-use* or *uncertain*.

The reliability of the process strongly relies on the first analysis stage: filtering of the original flow trace. The filter is controlled by 10 parameters [21], and their calibration is automatically solved per consumption event by the *Elitist Non-Dominated Sorting Genetic Algorithm NSGA-II* [22] (R package *mco*). There are three objective functions to be minimized in this calibration: (a) number of points that describe the filtered flow trace; (b) total accumulated volume difference between raw and filtered flow trace (Figure 3(a2,b2), *Input* and *Output*, respectively); (c) maximum on the curve of accumulated volume difference. The first objective function (FO₁) leads NSGA-II algorithm to solutions that simplify the filtered flow trace, whereas the other two (FO₂ and FO₃) focus on improving its fitting quality respect the original raw flow trace.

The next steps are followed to calculate the curve of accumulated volume difference: (1) given a raw water flow trace, demonstrated as vector $\mathbf{qr} = (qr_1, qr_2, \dots, qr_i, \dots, qr_m)$, and its corresponding filtered water flow trace $\mathbf{qf} = (qf_1, qf_2, \dots, qf_i, \dots, qf_m)$, both expressed in litres per hour (L/h) and recorded at time t_i in seconds (s), two new synchronized time series are generated by linear interpolation, \mathbf{qrs} and \mathbf{qfs} , for the set of unique t_i that belong to \mathbf{qr} and \mathbf{qf} ; (2) vector of time window \mathbf{tw} and vector of reference flow \mathbf{qref} are defined as:

$$\begin{cases} tw_i = 0, & i = 1 \\ tw_i = t_i - t_{i-1}, & 2 \leq i < n \end{cases} \quad (1)$$

$$\begin{cases} qref_i = 0, & i = 1 \\ qref_i = \max(qrs_{i-1}, qfs_{i-1}), & 2 \leq i < n \end{cases} \quad (2)$$

(3) the curve of accumulated volume difference is defined for those components of tw_i greater than 0.001 s (all flow rate jumps that take place in the raw water flow trace have this duration, as it can be seen in Figure 3(a1), *Input*) as follows:

$$\begin{aligned} & \text{if} \left(\left(q_{ref_j} > q_{th} \right) \left| V_{qr} [t \leq (t_j + tw_j)] - V_{qf} [t \leq (t_j + tw_j)] \right| \right) \\ & \quad \text{else} \left| V_{qr} [t \leq (t_j + tw_j)] - V_{qf} [t \leq (t_j + tw_j)] \right| / tw_i \end{aligned} \quad , \quad 2 \leq j < m < n \quad (3)$$

where V_{qr} and V_{qf} are the accumulated volume along the raw and the filtered water flow traces, respectively, until $t = t_j + tw_j$. On the other hand, q_{th} is a user-defined threshold that has been established to properly process the events with a continuous low-flow water leak. In these cases, small differences in leakage flow rate between the raw and the filtered water flow traces can be maintained over a long period of time, resulting in large accumulated volume differences that decrease the representativeness of FO_2 and FO_3 . To avoid this, the accumulated volume difference is divided by tw_i when q_{ref_j} is less than the maximum level of leakage flow rate q_{th} defined by the user. In the same way, the curve of accumulated volume difference is defined for those components of tw_i greater than 0.001 s to reduce the appearance of noise when the flow rate is below q_{th} .

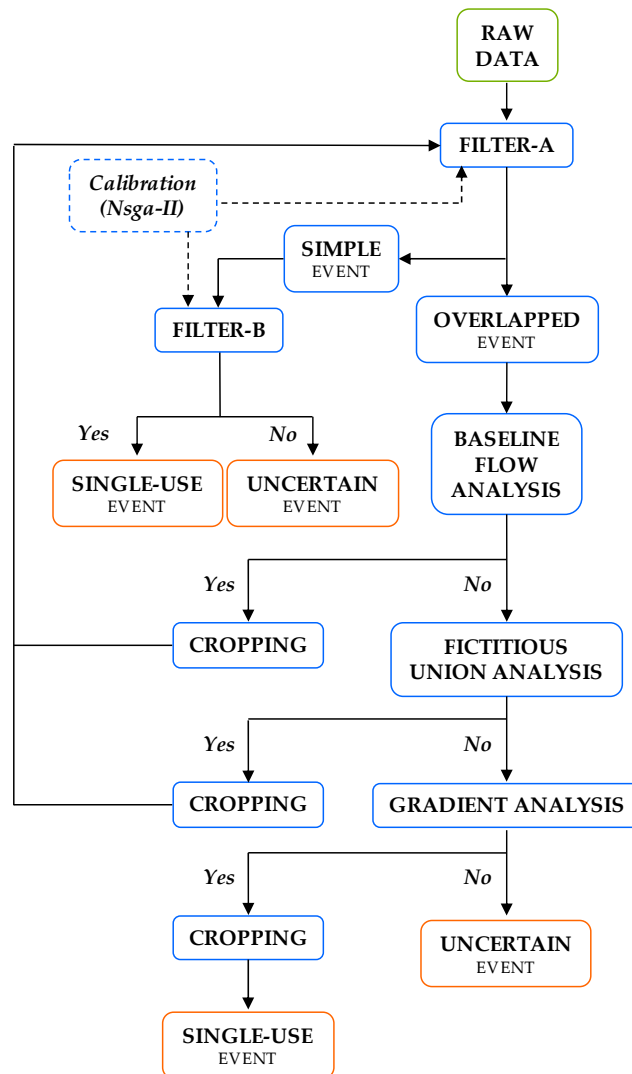


Figure 2. Flow chart of the disaggregation process and sorting of events as *single-use* or *uncertain*.

Regarding NSGA-II parameters, processing time was the most constraining factor to select the population size and the number of generations. The genetic algorithm achieves good results for the most intricate cases—long duration events with a great degree of overlapping, which typically come from households with leaks and high average daily consumption—with 24 individuals and 10 generations in a reasonable computing time. In relation to crossover and mutation probabilities, the default values taken were 0.7 and 0.2, respectively.

The result of this calibration process is a Pareto Front, and the chosen solution is the one for which the following expression is minimized:

$$FO_{\text{select.sol}} = w_1 * \frac{FO_1}{\max(FO_{1-PF})} + w_2 * \frac{FO_2}{\max(FO_{2-PF})} + w_3 * \frac{FO_3}{\max(FO_{3-PF})} ; \quad (4)$$

where the maximum value reached in each objective function within the Pareto Front ($\max(FO_{1-PF})$, $\max(FO_{2-PF})$ y $\max(FO_{3-PF})$) was taken to standardize the corresponding term. A conservative criterion for *Filter-A*, which prioritizes the simplification of the flow traces, establishes the weights for each objective function ($w_1 = 0.8$; $w_2 = 0.1$; $w_3 = 0.1$). This is necessary for subsequent disaggregation processes, which can only be applied if certain requirements are satisfied based on a strong filtering of the raw flow trace. Additionally, the calibration time can be limited by a user-defined threshold. In case a solution is not found within the established time limits (only happening in less than 0.01% of the sample events in the case studies below), the default values for the filter parameters will be used ($p_1 = 150$ (h*ms)/L; $p_2 = 0.16$ L; $p_3 = 80$ L/h; $p_4 = 40$ degrees; $p_5 = 5\%$; $p_6 = 100$ L/h; $p_7 = 10,000$ ms; $p_8 = 6000$ ms, $p_9 = 5\%$, $p_{10} = 10$ degrees). These default values are the result of the authors' experience while developing and applying this process in several projects around the world. In any case, a future sensitivity analysis is already planned to improve the filter performance and improve the speed of the calibration procedure.

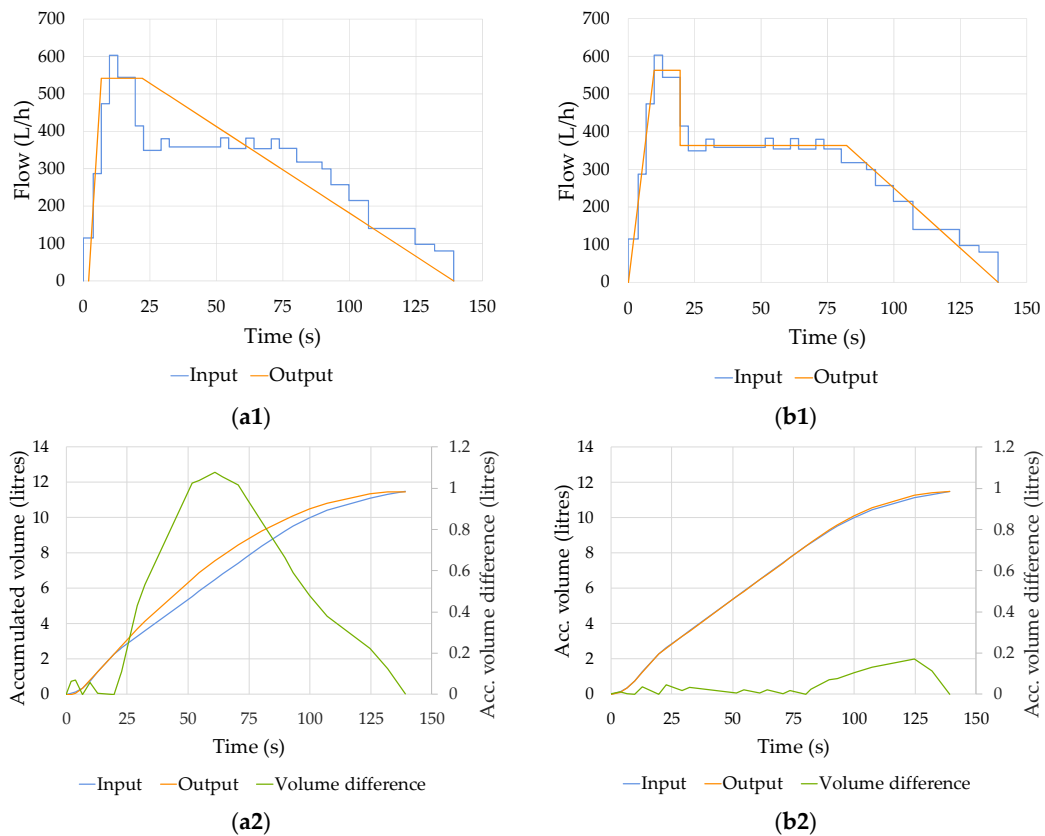


Figure 3. Comparison of raw flow traces vs filtered. (a1,a2) after *Filter-A* and (b1,b2) after *Filter-B*.

Once the consumption event defined by the raw flow trace has been filtered for the first time (*Filter-A* in Figure 2) the event can be classified as *simple*, constituted by only four vertexes, or *overlapped*. *Simple* events are analyzed by an additional filtering process (*Filter-B* in Figure 2). This process prioritises how accurately the filtered flow trace matches the original one. In this case, the solution selected from the Pareto Front should attain a value for the KGE index ([24]; R package *hydroGOF*) higher than 0.8 with a maximum $w1$ weight. Figure 3 compares the resulting filtered flow traces after going through the first and second filtering processes. Only if the final filtered flow trace (after *Filter-B*) is formed by only four vertexes, the event is finally classified as a *single-use* event. Otherwise, it will be classified as an *uncertain* event. The inclusion of a second filter significantly reduces the classification errors of *single-use* events compared to a one-stage filter approach.

An additional step of the disaggregation process is the analysis of the minimum/baseline flow (Q_{base}) in the filtered (*Filter-A*) events that have not been classified as *simple* events. The existence of a base event is considered when more than one horizontal section of the event satisfies the following two conditions: (i) the flow rate falls within a specified range $\{Q_{base} - tolerance, Q_{base} + tolerance\}$, where the tolerance is defined by the user; (ii) the volume associated with the horizontal section of the event is greater than a specified threshold. When these two conditions are met (Figure 4a), the events between horizontal sections are cropped from the base event (Figure 4b). If only the second condition is not satisfied (Figure 4a), the section is processed as a fictitious union, so that the events are separated and the union removed (Figure 4b). Fictitious unions appear when processing the raw flow traces. They do not actually correspond to any real water consumption (they are a distortion in the flow trace caused by the data-acquisition equipment). The end or starting times of the previous and following subevents are then recalculated to account for the volume removed from the fictitious union. Typically, this volume corresponds to one or two pulses from the pulse emitter of the water meter. The resulting subevents generated by these cropping operations are individually analyzed through the whole process again.

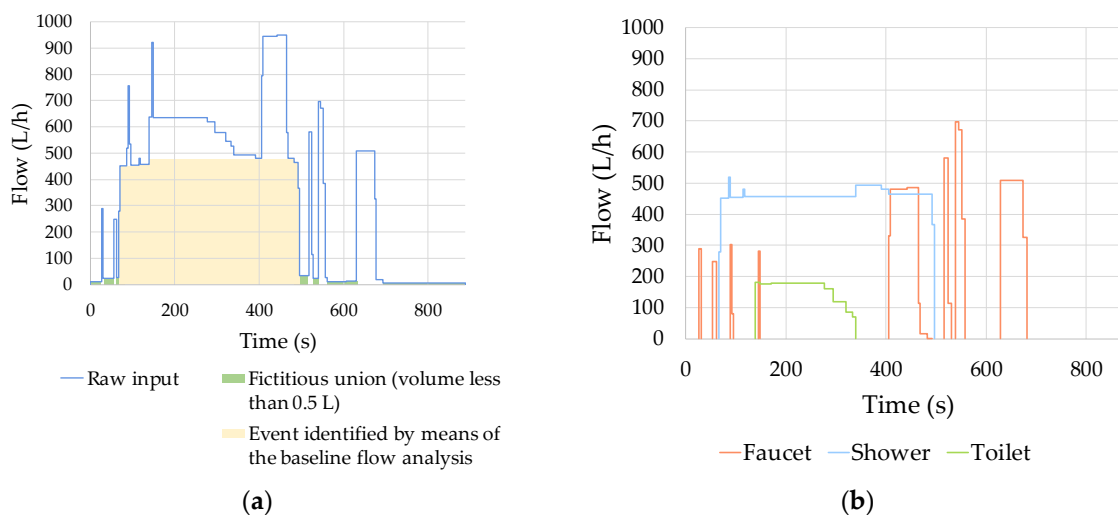


Figure 4. (a) Example of an *overlapped* event. (b) Resulting events obtained after disaggregating fictitious unions and conducting a baseline flow analysis.

Finally, if the event does not fall into any of the previous categories, a gradient analysis of the filtered flow trace is carried out (Figure 2). Only in case that the event has three major slopes, being the first one positive and other two negative or vice versa, it is considered that it is constituted by two or more different events that are *overlapped* in time, which begin or end within the same time range (Figure 5a). In this case, the events are cropped (Figure 5b), and the two new events are classified as *single-use* events. On the contrary, if the number of major slopes in the event is greater than three, it is directly categorized as an *uncertain* event. The key aspect of this separation process is to correctly identify

the start or the end, depending on the case, of the second major slope on the raw flow trace (Figure 5a, point 2) and the flow rate at the analogous instant in the filtered flow trace (Figure 5a, point 1).

It should be highlighted as an important contribution of the proposed methodology that all previous separation processes are implemented on the raw flow trace. Thus, signal smoothing in the filtered flow trace do not generate any loss of information in the resulting subevents obtained. This particular feature can be clearly observed in Figure 5, where the separated consumption events (Figure 5b) maintain the details of the original flow trace (Figure 5a). The presence of these details can be used in a later stage to improve the effectiveness of the automatic classification tools that can be developed.

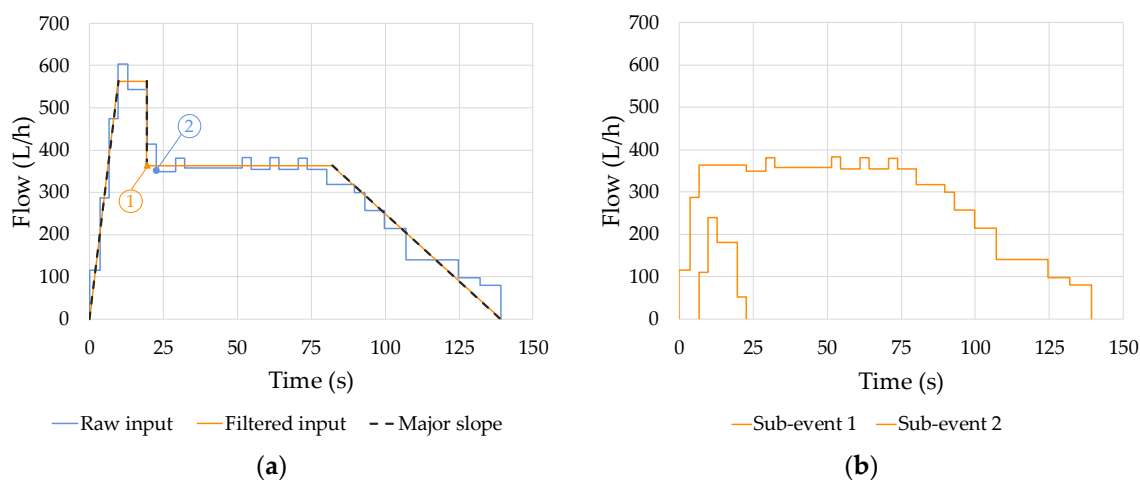


Figure 5. (a) Overlapped event in which the two subevents start at the same time. (b) Two resulting single-use events after the cropping operation.

2.2. Classification Process

At the end of the disaggregation process, all the events have been classified into two categories: *single-use* and *uncertain* events. As discussed below, the first group is the most numerous in any of the households analyzed, typically ranging from 75% to 92% depending on the amount of water consumption in the household. In addition, the amount of *single-use* events generated by the methodology is higher if the *uncertain* events having less than 3 L in volume are assumed to be *single-use* events (in this case the percentage of *single-use* events will range between 85% and 95%). The cropping and classification of the *uncertain* event group, corresponding to intricate events with high flow rate variability, will be the aim of future research. In this sense, it should be noted that high flow rate variability is not always associated to water consumption overlapping from different uses. Occasionally, the so-called *uncertain* events may be originated by pressure fluctuations, or the user changing the opening of a faucet for convenience or to adjust water temperature. Cropping and classification of *uncertain* events is not an easy task when accounting for the previous considerations and the fact that the overlapping of two uses may not produce a consumption flow rate that is the sum of the individual flow rates. The reason for this effect can be found in the pressure losses caused by the plumbing. Depending on the sizing of the pipes, the consumption flow rate of a given use may be significantly reduced by the appearance of additional water usages within the household.

The working hypothesis to categorizing *single-use* events is the following—those events with similar physical characteristics correspond to the same end use. In accordance with this, an initial clustering analysis is conducted by an unsupervised learning machine, Partition Around Medoids (PAM; [25]; R package *cluster*). One advantage of this partition clustering technique relies on the fact that it can work with different similarity measurements other than the Euclidean distance. In this work, the Gower distance has been chosen as the similarity measure, since it scales all the variables considered and allow for the definition of individual weights for the variables. On the other hand,

the characteristics of the events taken into account as input data are the total volume and the average flow rate. Event duration was rejected due to the considerable noise generated by long water uses, which impeded clusters identification. This effect was observed in households with leakage, showing long *single-use* events with low consumption flow rates. However, in these cases the duration of the event is a variable used as a preliminary filter to allow the clustering analysis to solely focus on the bulk of *single-use* events. Once the clustering analysis is finished (Figure 6), the application allows the user to visualize random subsamples of events from each cluster and associate them with an end use according to their physical characteristics.

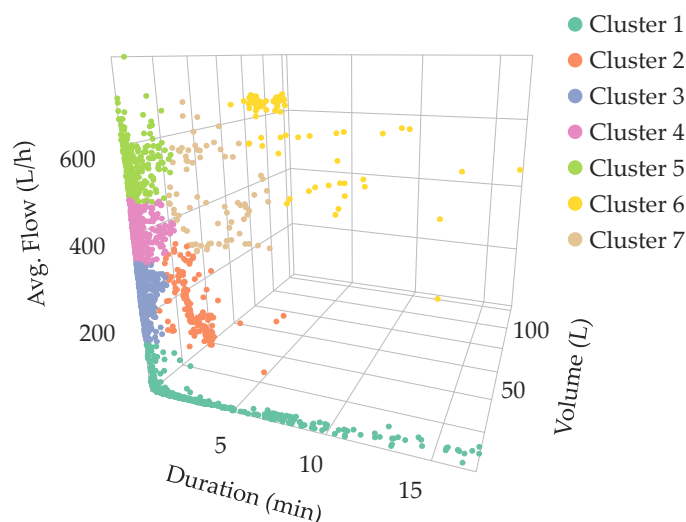


Figure 6. Result of Partition Around Medoids (PAM) algorithm with a number of clusters equal to 7 and similarity matrix based on Gower distance.

The subset of categorized *single-use* events and the subset of *uncertain* events are written in a flat CSV file at the end of the process. This file can be read by BuntBrainForEndUses[®], a web application for manual processing and editing of water end uses. In this way, the user can correct misclassifications and further edit the *uncertain* events that have not been properly analyzed by the algorithms.

3. Case Study

The water consumption data utilized for testing the methodology was sourced from two different water end uses studies conducted in geographically distant regions. One of the main differences between the studies is the type of monitoring equipment employed. In the first study (R1), the smart meters installed for water consumption monitoring were ELSTER Y250 single-jet (maximum flow rate of 5 m³/h) or ELSTER Y250M multi-jet (maximum flow rate of 7 m³/s) depending on the type of residential household. These meters produce a pulse every 0.04 L or 0.06 L of water consumed, respectively. Specially designed data loggers calculated and recorded the average consumption flow rate at approximately 3-s intervals. This recording mode was chosen to optimize the file size while preserving the quality of the flow trace. Files were periodically sent (twice per day) to the server via GPRS/GSM. On the other hand, in the second study (R2) a piston type volumetric water meter was used (Aquadis+, ITRON (WA, United States)), which generates a pulse every 0.1 L. The data logger used, recorded the occurrence time of every pulse with a resolution of 0.02 s.

For this analysis, a selection of significant households of both studies was conducted according to the average daily consumption and the presence or not of continuous leakage. The final selected sample was composed by 20 households—10 from R1 and 10 from R2—for which two-week period of monitoring data were selected. For the first study, the data corresponds to consumption made during autumn 2015, while for the second study the data were collected during autumn 2016. In total

19,858 sampled events were analyzed during the period considered. Figure 7 shows the general characteristics of the selected households and events associated with them.

As shown in Figure 7, there is a considerable difference between the households and events characteristics of these two studies. In the first study, the average daily consumption of the sample was close to 1600 L/hh/day, while in the second it is less than 400 L/hh/day. The number of daily events are also completely different: 110 events/hh/day vs. 35 events/hh/day. The dissimilarities of the households considered in the analysis conducted in this paper emphasize the reliability of the methodology, which can be used independently of the data sources, as long as the quality of the flow traces allows for end use disaggregation.

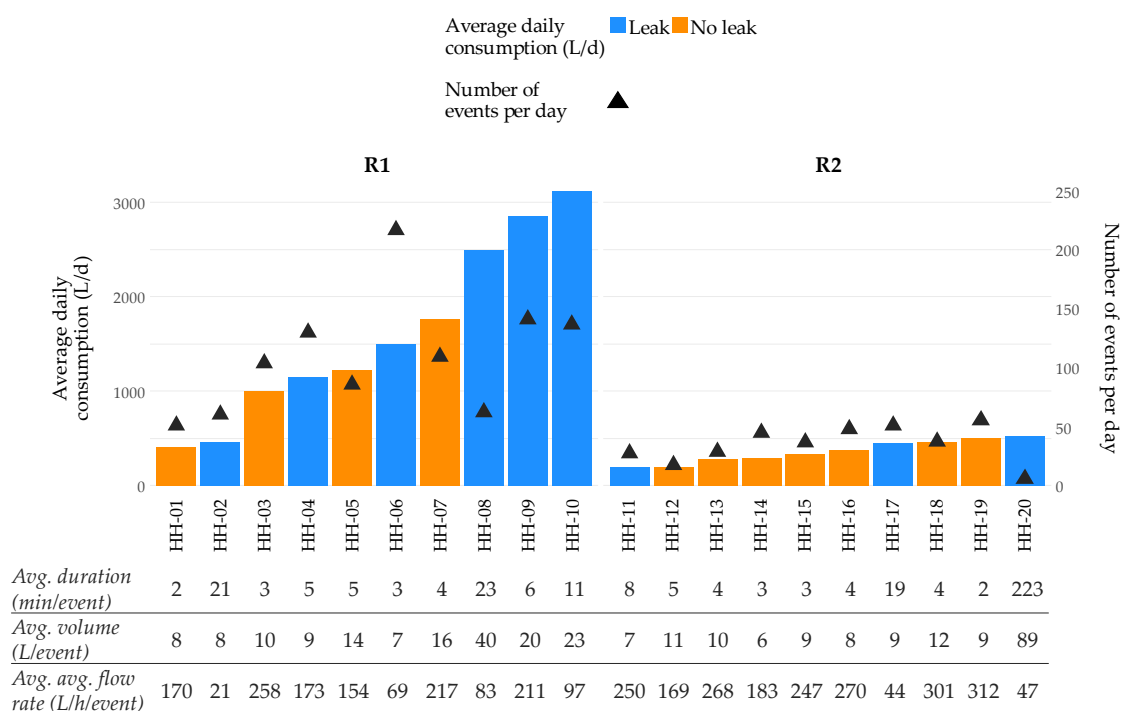


Figure 7. General characteristics of the analyzed households and the events associated with them.

4. Results and Discussion

The proposed methodology has been applied to 19,858 unprocessed water consumption events. After applying the disaggregation process (filtering and cropping) to the flow signal, the total number of events increased to 46,721, being the average number of cropping operations per day equal to 121 and 58 for the studies R1 and R2, respectively (Table 1). The average processing time consumed per each one of these operations is 21.8 s, using an Intel Core i5-4440 processor. Per study, the average cropping time is equal to 18.6 s for R1 and 28.3 s for R2. The calibration of the filtering algorithm is the task requiring more processing time, which increases with the density per unit time of data points in the raw flow trace. For this reason, it takes longer to carry out a cropping operation in the case of an event belonging to the study R2, since in this study flow data were recorded with a lower temporal resolution (0.02 s vs. approx. 3 s). Currently, the research team is working in optimizing the calibration algorithms and reducing the required processing time. The strategies proposed for this optimization are: (1) developing a methodology to calibrate the filter per household rather than per event, without a considerable loss of filtering accuracy; (2) finding the filter parameters through an algorithm that combines heuristic and guided search methods.

Table 1. General statistics about performance of separation process.

	R1 Study	R2 Study	Total
Total number of unprocessed events (10 households)	14,648	5210	19,858
Total number of resulting events (10 households)	32,792	13,929	46,721
Total number of resulting events per household	3279	1393	2336
Average time consumed per cropping operation (s)	18.6	28.3	21.8
Average number of cropping operation per household and day	121	58	90

Analyzing in detail the consumption events selected from study R1 (Figure 8), the result of applying Filter-A to all 14,648 unprocessed events, was that 8768 events were classified as *simple*, whereas the remaining 5880 were classified as *overlapped*. As to the *simple* events, most of them were, as expected, *single-use* events (6394). The remaining 2374 *simple* events were not simple enough and were classified as *uncertain* events. None of these *uncertain* events originated from *simple* events could be cropped or further processed; however, most of them correspond to single-use events with unsteady consumption flows (for example, a faucet that is adjusted to the desired flow rate).

Overlapped events correspond to events that could not be simplified as four-nodes events after Filter-A. These events have a considerable degree of complexity due to the overlapping of water uses. The algorithm proposed is capable of separating into *single-use* events most of these *overlapped* events by accurately cropping the flow traces. The initially identified 5880 *overlapped* events were separated into 24,024 subevents after the disaggregation process. Most of these new subevents were classified as *single-use* events (17,908), and the rest (6116) as *uncertain*. In total, the analysis of the worst case scenario of study R1 produced 24,302 *single-use* events (74.1%) and 8490 *uncertain* events (25.9%). From the *uncertain* events, 4224 had a volume of less than 3 L. These, because of their low volume, could be also be added to the *single-use* group.

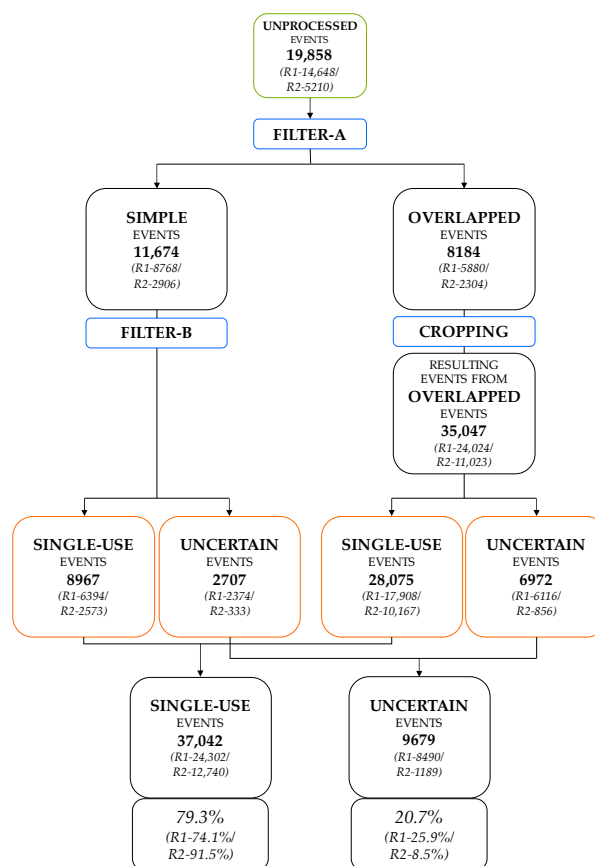


Figure 8. Classification of events after filtering and cropping (disaggregation process).

For study R2, showing simpler flow traces typical of water consumption profiles of a European family, the results are even more positive. As shown in Figure 8, 91.5% of 13,929 resulting events were considered to be *single-use* events and only 8.5% were catalogued as *uncertain* events. The difference between studies is mainly due to the characteristics of the households: the flow traces belonging to the sample R1 are notably more complex, obtained from households with high average daily consumption and frequent overlapping of water uses.

Overall, the methodology generated a number of *single-use* events equal to 37,042 events (79.3% of the 46,721 resulting events), of which 75.8% has been obtained through the proposed disaggregation process. Consequently, human intervention to crop and generate *single-use* events has been significantly minimized, with subsequently large human working time savings. In raw numbers, the total automatic disaggregation process has taken less than 4 days (96 computing hours); whereas, according to the authors' experience, the same work would have required about 45 human-working days (360 h).

The distribution of the physical characteristics of the *single-use* and *uncertain* events for both studies is presented in Figure 9. Some outliers, with a duration of more than 8 min, have been removed to improve the readability of the basic statistics (median, first and third quartile). As expected, the heterogeneity of the *uncertain* events is significantly larger than the one obtained for *single-use* events. In addition, the average duration, volume and flow rate of *uncertain* events are greater than those for *single-use* events. It should also be mentioned that some of the events considered here as *uncertain* correspond to single uses, and their flow rate variability can be caused by adjusting the faucet or variations in the input pressure.

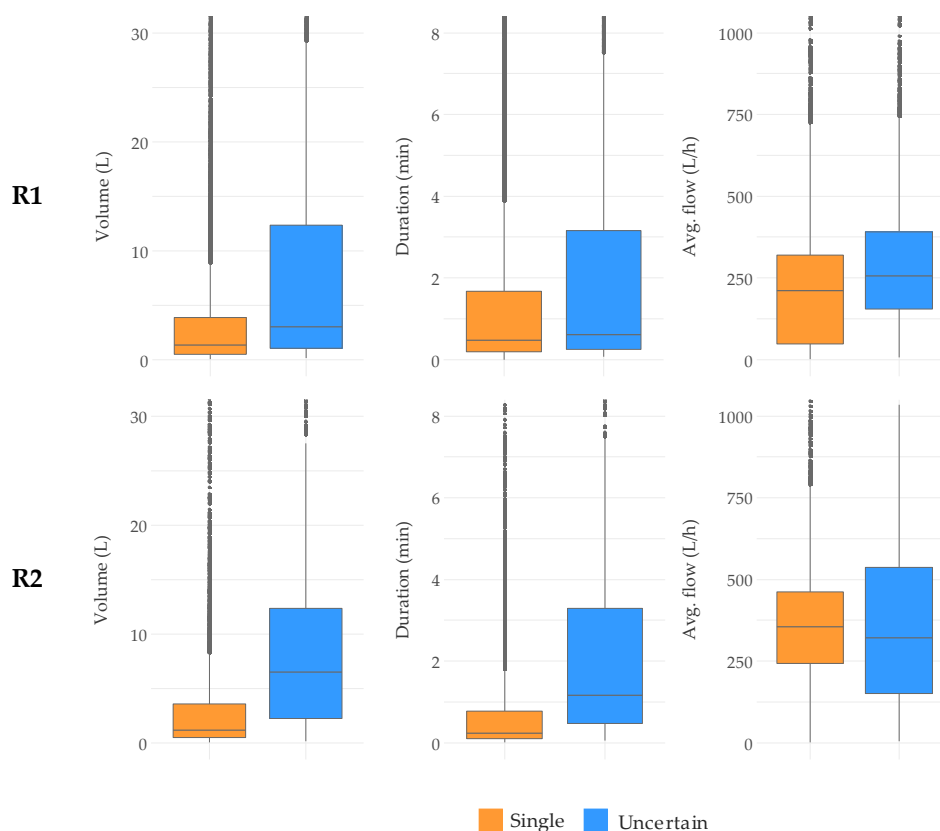


Figure 9. Distribution of the physical characteristics of the events classified as *single-use* and *uncertain* per study.

After the disaggregation process, *single-use* events could be categorized by clustering analysis or any other classification algorithm [15]. Similar methodologies have already been used in other fields, like non-intrusive electric load data disaggregation [26–29]. As an example of the results that can be

achieved by these techniques, Figure 10 shows the findings for one of the most complex households, HH-06 of R1, showing an indoor leak and a high average daily consumption. The unsupervised learning technique used, allows to identify different types of water consumption uses: *Cluster 2* in Figure 10 mostly includes events corresponding to *toilets*, while *Cluster 6* is composed of *washing machine* and *shower* events.

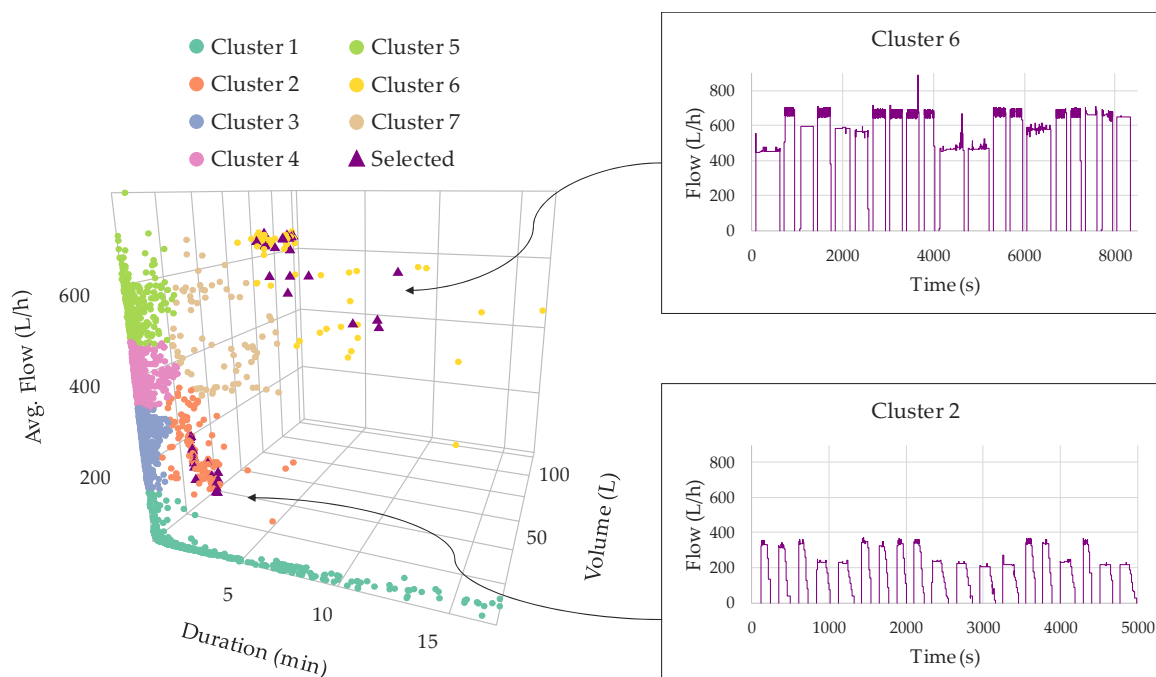


Figure 10. Display of 20 randomly selected individuals from clusters 2 and 6 for household HH-06 that belongs the R1 study.

Figure 11 shows the final results after assigning a water end use to each cluster for the household under study. Given the same monitoring period, the outcome of proposed methodology is compared with the one obtained manually. It can be observed that the physical characteristics of the events in each end use category tend to be similar. Nevertheless, there is a significant deviation, especially with respect to the mean flow rate and duration of the events: the average flow rate for each end use tends to be higher in the presented approach, while the average duration is generally shorter as more cropping operations are conducted through the automatic disaggregation process. In addition, both parameters—volume and duration of *single-use* events—are less dispersed when the flow traces are automatically cropped. This is directly related to the inherent defects of manual editing that can be seen in Figure 12 (additional examples in Appendix B): the automatic algorithms recognize the leakage event by means of a volume check (Figure 12b), whereas the analyst has subjectively decided in this specific case to ignore it (Figure 12a) and add the volume to the *toilet* use. When a human analyst processes the consumption data, the resulting average duration of *faucets* and *toilets* is larger and the average flow rate smaller. Additionally, for the same reason, a greater number of leakage events have been identified and separated from other consumption through the proposed methodology. These findings demonstrate that automatic disaggregation tools can generate a standardized corpus of processed data, which is more homogeneous and reliable because it is obtained as a result of cropping operations based on solid and well established criteria. Therefore, the *single-use* events obtained from the automatic disaggregation algorithms developed in this study are significantly more reliable, in terms of duration, average flow rate and shape than those resulting from a manual processing and cropping the water consumption flow traces. These results have direct implications in the probability functions used to characterize water consumption events frequency, duration and intensity [30–34].

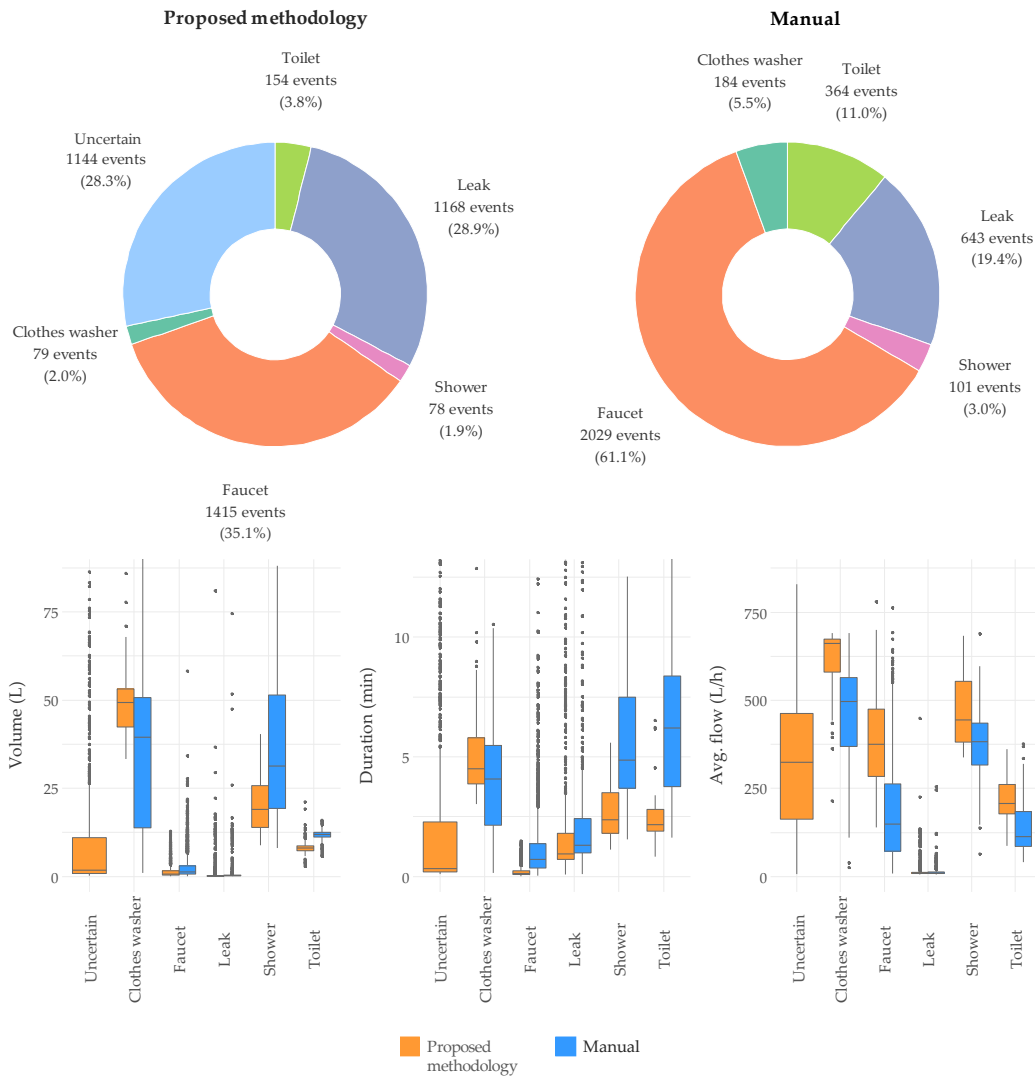


Figure 11. The manual vs proposed methodology final results of complete disaggregation processing for the household HH-06 of the R1 study.

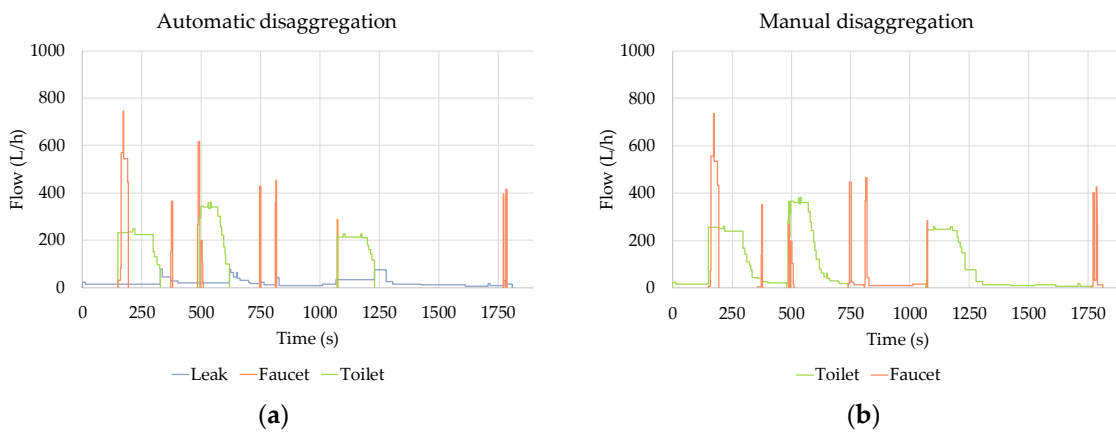


Figure 12. Manual vs automatic disaggregation (example from the household HH-06 of R1).

Obviously, more accurate classification techniques can be developed as processing experience is gained and larger and more reliable data sets are available for training the algorithms. The work

presented should be considered as an important first stage to produce sets of individual events that are built consistent and accurately, which can be used to improve the training of automatic recognition algorithms. Therefore, the main contribution of the proposed methodology is mainly related to the quality of the *single-use* events obtained through an automatic separation technique that can be easily used for developing faster and better performing classification algorithms.

5. Conclusions

The work presented intends to be a step forward to the main objective of understanding in detail how water is consumed through end uses, and the reasons behind it. It proposes a new fully-automatic disaggregation process for water consumption events that is based on a two-step filtering and event cropping algorithm. An additional advantage achieved by the flexibility of the filter is that the whole process can be universally applied to different type of customers and monitoring equipment.

The disaggregation process presented is divided into two main stages:

- (a) The raw water consumption events are filtered and categorized as *simple* or *overlapped* to facilitate subsequent operations. The filtering relies on an advanced algorithm that is automatically calibrated for each water consumption event by means of NSGA-II genetic algorithm. *Simple* events are then characterized as *single-use*, which correspond to actual individual water uses, or *uncertain* events.
- (b) On the other hand, *overlapped* events, originated by simultaneous water uses, are cropped and separated into simpler *single-use* events. All cropping operations are implemented on the raw flow trace, and potential distortions in the filtered signal do not generate any loss of information in the resulting subevents. In other words, all the subevents created maintain the characteristics of the original flow trace. This particular feature increases the amount of information available for the classification algorithms that can be developed in the future, improving their effectiveness.

Finally, as a case study, an example of the way in which the events generated during the previous stages can be easily categorized into various end uses by a semiautomatic algorithm is added to the work presented. *Single-use* events are massively classified into various water end use categories by means of clustering analysis.

Regarding the performance analysis of the first and second stages for the case study presented, the following conclusions were raised: The original raw flow traces, of the 20 households belonging to the studies R1 and R2, covering a monitoring period of 15 days per household, contain 19,858 events. After the filtering and separation process, the number of subevents grew to 46,721, of which 79.3% (37,042 events) are single uses. In other words, the number of water consumption events increased by 130%, and 26,863 new events were created. Up to 75.8% of the *single-use* events that can be classified, have been obtained through the disaggregation process defined in this work. Therefore, the methodology proposed solves most of the cropping operations that need to be performed and reduces significantly the human intervention required to disaggregate the *overlapped* consumption events into *single-use* events, with significant time savings.

Finally, by comparing the manual and the automatically disaggregated events, it was observed that the characteristics of the events originated from the algorithms proposed are more homogeneous and consistent than the ones obtained by manual cropping. This result can be easily justified by the fact that the automatic separation algorithms always apply the same criteria, while a human analyst may change the cropping criteria while conducting the analysis. Furthermore, the inherent subjectivity of manual separation introduces dispersion in the physical characteristics of the events belonging to a specific water end use. More dispersion regarding the physical characteristics of the consumption events associated to an end use, unavoidably lead to poorer performance of whatever automatic classification technique that could be applied. This is why the *single-use* events obtained by the methodology proposed constitute a more reliable corpus for training and developing end use classifications algorithms.

Definitions

Classification: Process by which every *single-use* event is assigned to one of the potential water end uses in a household.

Cropping: Action by which one part of an *overlapped* event that has already been identified as a single water use (and still remains attached to the overlapped event) is effectively removed from it to become a new and independent *single-use* event.

Disaggregation: Process by which an *overlapped* event is effectively separated into all the *single-use* events integrating such event. In this work, the disaggregation process consists in two actions (filtering and cropping), and it is performed through a universal fully-automatic algorithm.

Fictitious union: When a water consumption starts, there is a significant time gap between the previous pulse and the initial pulse recorded. Therefore, the flow rate associated to the first pulse received (calculated as the ratio between the pulse volume and the time gap) is lower than the actual consumption flow rate. With this calculation it is also assumed that during the complete time gap between pulses, the consumption flow rate is constant and equal to the calculated value. Obviously, this calculation does not represent how water was really consumed in the time interval between the two pulses under analysis. Quite frequently, there will be part of the time (most) in which there is no consumption and another part (less) in which there is a consumption at a relative high flow.

Event: Every single water consumption, whatever its volume or duration. An event begins when the flow rate through the meter changes from zero to any positive value, and finishes when the flow rate becomes to zero again.

Unprocessed event: Initial event in the raw flow trace, before any kind of signal processing has been conducted.

Overlapped event: Complex event being the sum of more than one simultaneous single uses of water.

Pulse: Each of the signals sent by the pulse-emitter (attached to or embedded in the water meter) to the data-logger. Each signal corresponds to a fixed volume of water consumed. In the design stage of the monitoring project there are two crucial decisions related to pulse emitters: (i) the consumption volume associated to each pulse, which mostly depends on the water meter design, and (ii) the way pulses are recorded by the logger. Typically, the data loggers may store the number of pulses (volume) received at fixed intervals of time, or it may store the time at which each single pulse is received. Fictitious unions appear when the time of occurrence of the pulses are recorded in the data logger.

Simple event: Processed event, obtained after Filter-A, constituted by four vertexes. Most simple events will become single-use events at the end of the disaggregation process.

Single-use event: Final event obtained after the disaggregation process that corresponds to one specific end use.

Uncertain event: Final event obtained after the disaggregation process that cannot be further cropped into smaller single-use events nor classified as being a single use itself. Additional contextual information is needed to be able to split the event into smaller ones or to categorize it as one single-use event.

Acknowledgments: This study has received funding by the IMPADAPT project /CGL2013-48424-C2-1-R from the Spanish ministry MINECO with European FEDER funds and from the European Union's Seventh Framework Programme (FP7/2007e2013) under grant agreement No. 619172 (SmartH2O: an ICT Platform to leverage on Social Computing for the efficient management of Water Consumption).

Author Contributions: Methodology conception and definition: F.J.A., L.P.-J. and R.C. Data analysis tools and signal filtering algorithms: L.P.-J. End Use software design: F.J.A. Water consumption data gathering and field work: F.J.A. and R.C. All authors contributed to the preparation of the manuscript and approved it.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

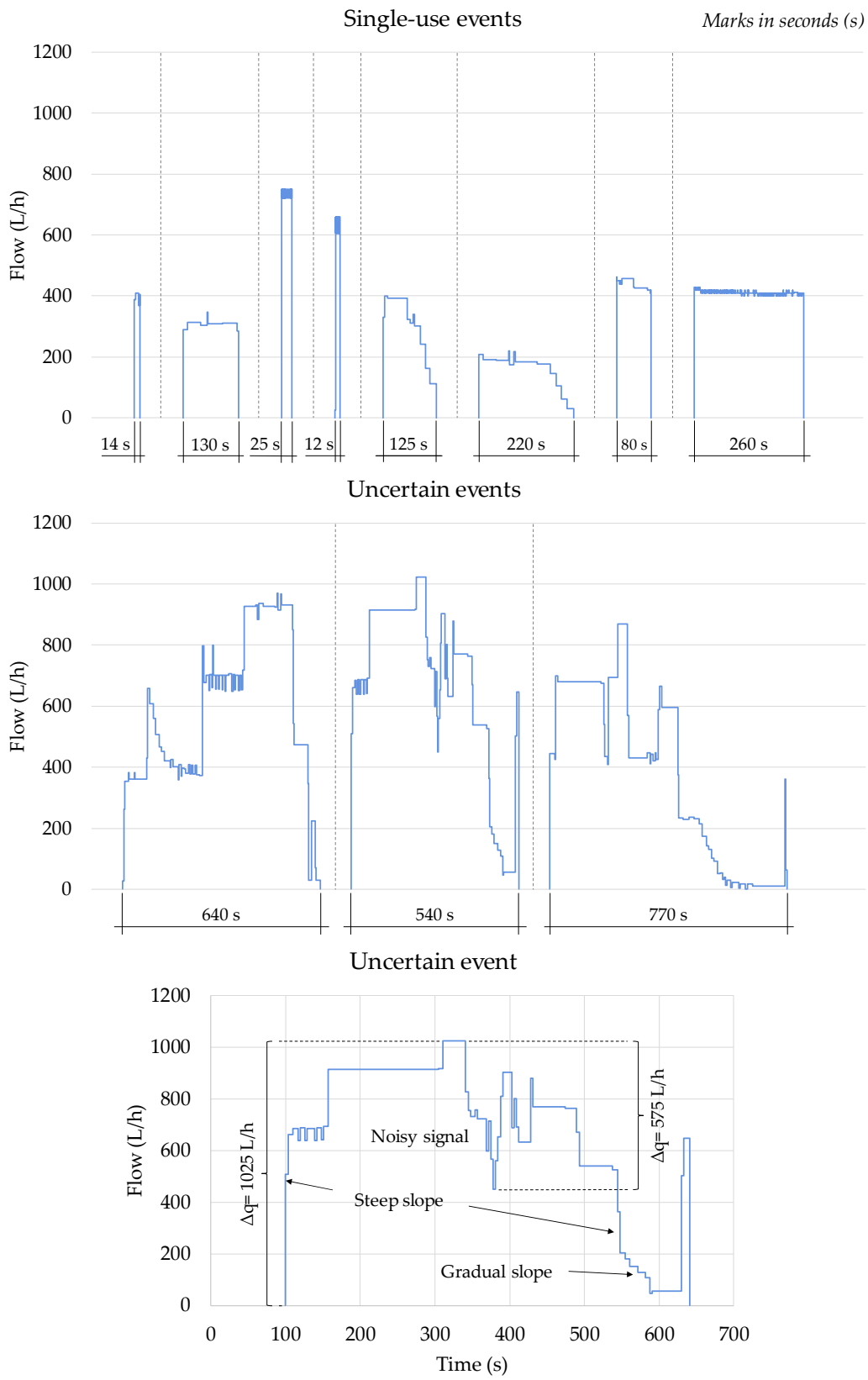


Figure A1. Examples of Single-Use and Uncertain Events.

Appendix B

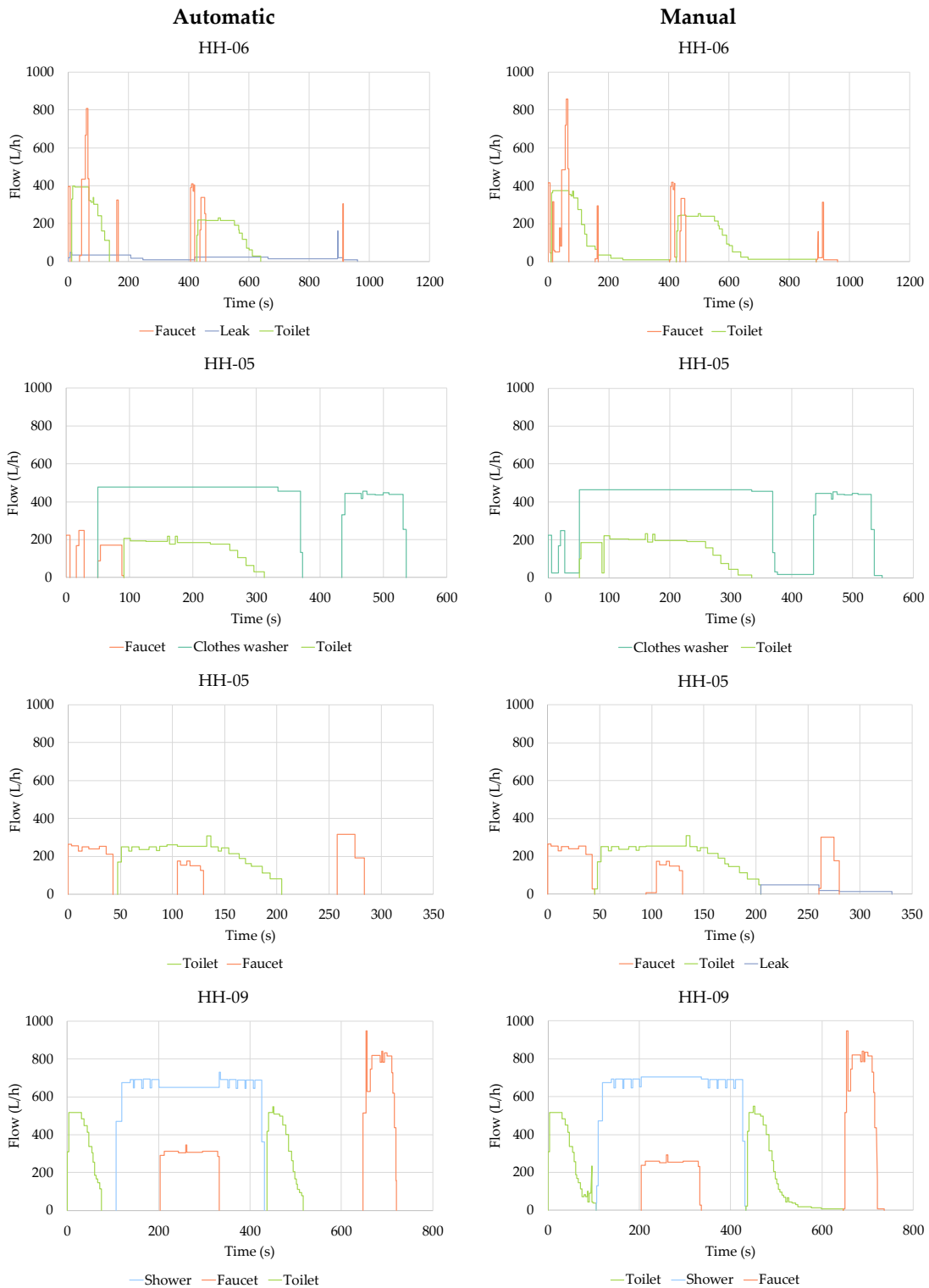


Figure A2. Examples of Manual vs. Automatic Disaggregation from R1 study.

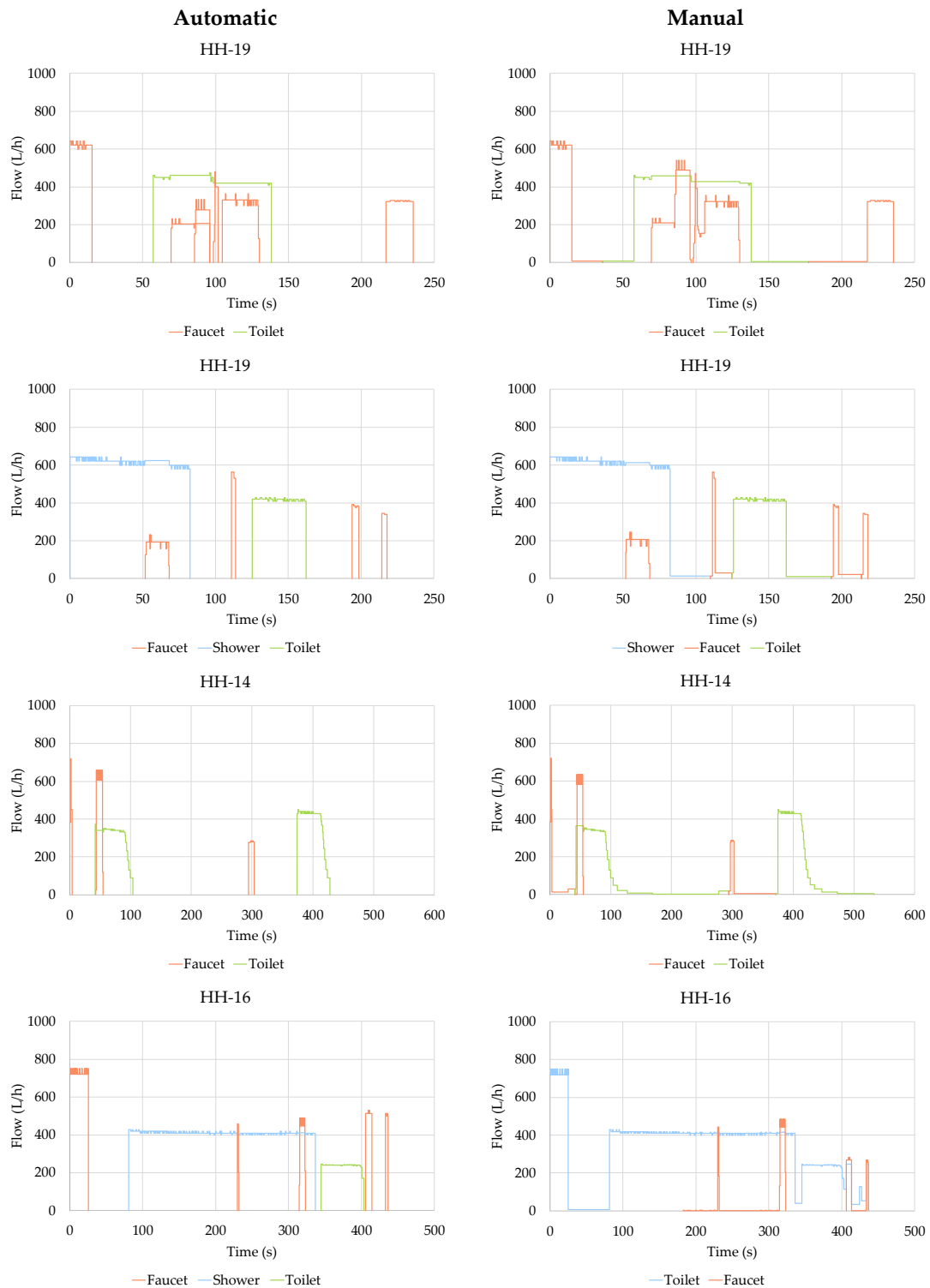


Figure A3. Examples of Manual vs. Automatic Disaggregation from R2 study.

References

1. World Commission on Environment and Development. *Our Common Future*; Oxford University Press: Oxford, UK, 1987; ISBN 019282080X.
2. Hoekstra, A.Y.; Mekonnen, M.M. The water footprint of humanity. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 3232–3237. [[CrossRef](#)] [[PubMed](#)]

3. Jaramillo, F.; Destouni, G. Local flow regulation and irrigation raise global human water consumption and footprint. *Science* **2015**, *350*, 1248–1251. [[CrossRef](#)] [[PubMed](#)]
4. Shiklomanov, I.A. Appraisal and Assessment of World Water Resources. *Water Int.* **2000**, *25*, 11–32. [[CrossRef](#)]
5. Cominola, A.; Giuliani, M.; Piga, D.; Castelletti, A.; Rizzoli, A.E. Benefits and challenges of using smart meters for advancing residential water demand modeling and management: A review. *Environ. Model. Softw.* **2015**, *72*, 198–214. [[CrossRef](#)]
6. Fielding, K.S.; Spinks, A.; Russell, S.; McCrea, R.; Stewart, R.A.; Gardner, J. An experimental test of voluntary strategies to promote urban water demand management. *J. Environ. Manag.* **2013**, *114*, 343–351. [[CrossRef](#)] [[PubMed](#)]
7. Sahin, O.; Bertone, E.; Beal, C.D. A system approach for assessing water conservation potential through demand-based water tariffs. *J. Clean. Prod.* **2017**, *148*, 773–774. [[CrossRef](#)]
8. Makki, A.A.; Stewart, R.A.; Beal, C.D.; Panuwatwanich, K. Novel bottom-up urban water demand forecasting model: Revealing the determinants, drivers and predictors of residential indoor end-use consumption. *Resour. Conserv. Recycl.* **2015**, *95*, 15–37. [[CrossRef](#)]
9. Bennet, C.; Stewart, R.A.; Beal, C.D. ANN-Based residential water end-use demand forecasting model. *Expert Syst. Appl.* **2013**, *40*, 1014–1023. [[CrossRef](#)]
10. DeOreo, W.B.; Heaney, J.P.; Mayer, P.W. Flow trace analysis to assess water use. *Am. Water Works Assoc.* **1996**, *88*, 79–90.
11. Kowalski, M.; Marshallsay, D. *A System for Improved Assessment of Domestic Water Use Components. II International Conference Efficient Use and Management of Urban Water Supply*; International Water Association: Tenerife, Spain, 2003.
12. Arregui, F. New software tool for water end-uses studies. Presented at the 8th IWA International Conference on Water Efficiency and Performance Assessment of Water Services, Cincinnati, OH, USA, 20–24 April 2015.
13. Nguyen, K.A.; Zhang, H.; Stewart, R.A. Development of an intelligent model to categorise residential water end use events. *J. Hydro-Environ. Res.* **2013**, *7*, 182–201. [[CrossRef](#)]
14. Nguyen, K.A.; Stewart, R.A.; Zhang, H. An intelligent pattern recognition model to automate the categorisation of residential water end-use events. *Environ. Model. Softw.* **2013**, *47*, 108–127. [[CrossRef](#)]
15. Nguyen, K.A.; Stewart, R.A.; Zhang, H. An autonomous and intelligent expert system for residential water end-use classification. *Expert Syst. Appl.* **2014**, *41*, 342–356. [[CrossRef](#)]
16. Nguyen, K.A.; Stewart, R.A.; Zhang, H.; Jones, C. Intelligent autonomous system for residential water end use classification: Autoflow. *Appl. Soft Comput.* **2015**, *31*, 118–131. [[CrossRef](#)]
17. Piga, D.; Cominola, A.; Giuliani, M.; Castelletti, A.; Rizzoli, A.E. A convex optimization approach for automated water and energy end use disaggregation. In Proceedings of the 36th IAHR World Congress, Hague, The Netherlands, 28 June–3 July 2015.
18. FLUID—The Learning Water Meter. Available online: <http://www.fluidwatermeter.com> (accessed on 17 November 2017).
19. WaterSmart—Platform Features. Take a Look Under the Hood. Available online: <https://www.watersmart.com> (accessed on 17 November 2017).
20. Aquubiq—Features. A Smart Path for a Greener Lifestyle. Available online: <http://www.aquubiq.com> (accessed on 17 November 2017).
21. Pastor, L.; Arregui, F.; Cobacho, R. Filtering smart metering data to improve detection of water end use events. In Proceedings of the 9th International Conference on Efficient Use and Management of Urban Water, Bath, UK, 18–20 July 2017.
22. Deb, K.; Pratap, A.; Agarwal, S.A. Fast and Elitist Multiobjective Genetic Algorithm: NSGAI. *IEEE Trans. Evolut. Comput.* **2002**, *6*, 182–197. [[CrossRef](#)]
23. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2013.
24. Gupta, H.V.; Kling, H.; Yilmaz, K.K.; Martinez, G.F. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *J. Hydrol.* **2009**, *377*, 80–91. [[CrossRef](#)]
25. Reynolds, A.; Richards, G.; de la Iglesia, B.; Rayward-Smith, V. Clustering rules: A comparison of partitioning and hierarchical clustering algorithms. *J. Math. Model. Algorithms* **1992**, *5*, 475–504. [[CrossRef](#)]

26. Amenta, V.; Tina, G.M. Load demand disaggregation based on simple load signature and user's feedback. *Energy Procedia* **2015**, *83*, 380–388. [[CrossRef](#)]
27. Elhamifar, E.; Sastry, S. Energy disaggregation via learning powerlets and sparse coding. In Proceedings of the National Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015; pp. 629–635.
28. Bonfigli, R.; Squartini, S.; Fagiani, M.; Piazza, F. Unsupervised algorithms for nonintrusive load monitoring: An up-to-date overview. In Proceedings of the 15th International Conference on Environment and Electrical Engineering (EEEIC), Rome, Italy, 10–13 June 2015; pp. 1175–1180.
29. Cominola, A.; Giuliani, M.; Piga, D.; Castelletti, A.; Rizzoli, A.E. A Hybrid Signature-based Iterative Disaggregation algorithm for Non-Intrusive Load Monitoring. *Appl. Energy* **2017**, *185*, 331–344. [[CrossRef](#)]
30. Guercio, R.; Magini, R.; Pallavicini, I. Instantaneous residential water demand as stochastic point process. *WIT Trans. Ecol. Environ.* **2001**, *48*, 129–138. [[CrossRef](#)]
31. Alvisi, S.; Franchini, M.; Marinelli, A. A stochastic model for representing drinking water demand at residential level. *Water Resour. Manag.* **2003**, *17*, 197–222. [[CrossRef](#)]
32. Garcia, V.J.; Garcia-Bartual, R.; Cabrera, E.; Arregui, F.; Garcia-Serra, J. Stochastic model to evaluate residential water demands. *J. Water Resour. Plan. Manag.* **2004**, *130*, 386–394. [[CrossRef](#)]
33. Blokker, E.J.M.; Vreeburg, J.H.G.; van Dijk, J.C. Simulating residential water demand with a stochastic end-use model. *J. Water Resour. Plan. Manag.* **2010**, *136*, 375–382. [[CrossRef](#)]
34. Creaco, E.; Alvisi, S.; Farmani, R.; Vamvakieridou-Lyroudia, L.; Franchini, M.; Kapelan, Z.; Savic, D. Methods for preserving duration-intensity correlation on synthetically generated water-demand pulses. *J. Water Resour. Plan. Manag.* **2016**, *142*. [[CrossRef](#)]



© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).