International Environmental Modelling and Software Society (iEMSs)
2012 International Congress on Environmental Modelling and Software
Managing Resources of a Limited Planet, Sixth Biennial Meeting, Leipzig, Germany
R. Seppelt, A.A. Voinov, S. Lange, D. Bankamp (Eds.)
http://www.iemss.org/society/index.php/iemss-2012-proceedings

Analysis of simultaneous water end use events using a hybrid combination of filtering and pattern recognition techniques

NGUYEN, A.K, ZHANG, H, STEWART, R.A
Griffith School of Engineering
Griffith University
Gold Coast Campus
AUSTRALIA

Abstract: It is challenging to disaggregate domestic water consumption flow-trace data into end use event categories for urban water management. Currently, domestic end use studies utilise software and analyst experience to disaggregate flow data into end use events (e.g. faucet, dishwasher, toilet, etc.), which often take an excessive amount of time, particularly for separating a combined event (i.e. a group of single events which occur simultaneously) into different isolated events. To tackle this problem, an existing database of end use events for 252 households located in South-east Queensland (SEQ), Australia was utilised with the goal of automating the flow trace analysis process. A newly developed filtering method in conjunction with the Hidden Markov Model (HMM) technique was applied to disaggregate the combined events. The outcome of this practice is a hybrid model which allows great separation accuracy (average of 88%) of a combined event into many different single events. Future work will incorporate this model with existing methods in single event classification to fulfil the development of an automatic system to disaggregate domestic water consumption flow-trace data into end use event categories. Moreover, validation of its accuracy will be examined through independent testing with 20 selected combined samples.

Keywords: Hidden Markov Model; water end use; water micro-component; flow trace disaggregation; water demand management

1 INTRODUCTION

In the current decade, Australia has experienced severe droughts, raising concerns on water security for urban areas. It has been estimated that the influence of climate change and growing urban water demand could halve current supplies (CSIRO, 2010). After decades of inadequate metering of water use, organisations have come to the realisation that it is almost impossible to achieve effective water management schemes without an accurate and appropriate measure of water consumption. This desire to better monitor and analyse water consumption has led to the conceptualisation of a Knowledge Management System (KMS) which is able to collect real-time water consumption data through a smart water metering system, transfer and store the data into a knowledge repository, analyse the data, and produce comprehensive reports which can be accessed on-line by a broad range of users (e.g. consumers, water utilities, government organisations, etc.) (Stewart et al., 2010). Before such an information system can be realised, improved approaches for disaggregating high resolution water consumption data

into end use events are required. Therefore, the key enabler for this KMS is the development of pattern matching algorithms which are able to automatically categorise collected flow trace data points received from wireless data loggers into particular water end use categories. To facilitate this study, a substantial number of samples of end use events were collected from 252 homes and manually separated into nine different types of water pattern, including shower, faucet (tap), dishwasher, clothes washer, full-flush toilet, half-flush toilet, bathtub, irrigation and leak. Water flow data collected directly from water meters includes both single (e.g. shower event occurring alone) and combined events (i.e. an event which comprises of several overlapped single events). With the existing database, the problem of classifying single events from the collected flow trace has been thoroughly clarified through applying a hybrid combination of Hidden Markov Model (HMM) and Dynamic Time Warping Algorithm (DTW) (Nguyen et al., 2012). The remaining challenge to be solved for the fulfillment of the system is related to combined event classification, which is a challenging blind source separation problem. To date, many techniques have been proposed to tackle similar issues, but none of them has comprehensively answered the question (Nguyen & Jutten., 1995). To overcome this difficulty, the present study has suggested an effective solution using the HMM technique in conjunction with a filtering method, which has been developed for this particular study. As HMM is a classic tool that has been used worldwide for decades, the paper only provides a brief introduction to this method and mainly focuses on technical aspects of the new filtering method and the combined event analysis methodology.

HMM is one of the most popular techniques, which has been widely applied in many branches of pattern recognition. Chien and Wang, (1997) presented an adaptation method of speech HMM for telephone speech recognition. The goal of that study was to automatically adapt the HMM parameters so that the adapted HMM parameters can match with the telephone environment. Experiments showed that the proposed approach can be successfully employed for self adaptation as well as supervised adaptation. Cho et al. (1995) applied HMM for the problem of modelling and recognising cursive words. A handwritten word is regarded as a sequence of characters and optional ligatures. With this in mind, an interconnection network of character and ligature HMMs was constructed to model words of indefinite length. Experiments showed that this model can ideally describe any form of handwritten words, including discretely spaced words, pure cursive words and unconstrained words of mixed styles. In this present study, a HMM model was achieved through training the collected flow data from 252 homes having high resolution smart meters (i.e. 0.014 litres/pulse recorded at 5 seconds interval). This technique was utilised as the main classifier for the analysis of both single and combined events. When a random water end use sample is recognised by the existing HMM model, it will be assigned to a particular residential water end use category, including faucet, shower, clothes washer, dishwasher, toilet, irrigation and bathtub, whichever results in the highest possibility.

2 COMBINED EVENT ANALYSIS

2.1 Combined event dissection process

A combined event in this study is defined as the one which is formed by at least two simultaneous single events. There is no restriction on the starting and finishing time of each end use component, as long as they have an overlapped period on each other. In one combined event, the longest component will be named “base event” (e.g. shower), and all other shorter ones are called “sub event” (e.g. tap) which are located on top of the base one.

The first step in combined event classification is to split the event into several smaller parts for a comprehensive analysis (Figure 1).

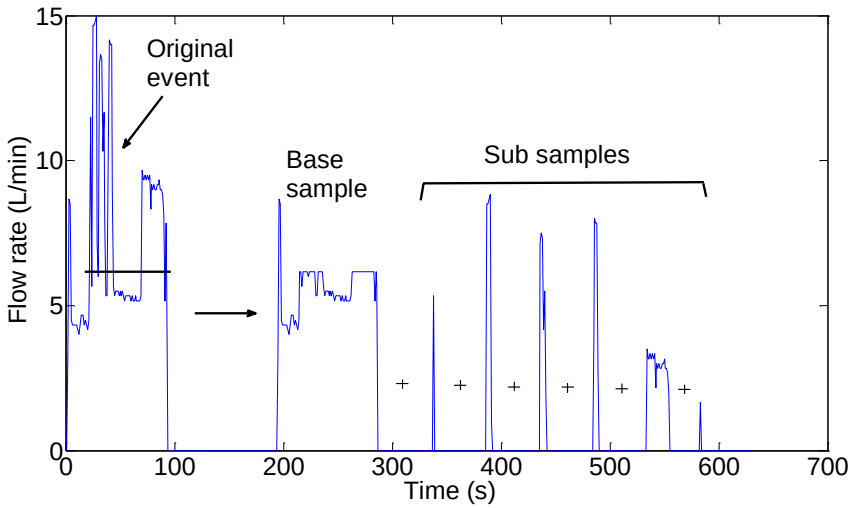

Figure 1 Combined event separation process

To perform this task, a new filtering method has been developed which smooths a combined event to any desired level as displayed in Figure 3. The principle of this technique is based on the examination of gradient change along the sample to make different dissection decisions. The method derivation can be explained as follows:

Given a combined event sample which is demonstrated as vector $\boldsymbol{a} = (a_1, a_2, \dots a_i \dots, a_m)$ of length m, the vector gradient $\boldsymbol{g} = (g_1, g_2, \dots g_i \dots, g_{m-1})$ of $\boldsymbol{a}$ will be determined using the following formula:

$$g_i = \frac{(a_{i+1} - a_i)}{dt} \qquad (1)$$

Where

- i is the sampling index
- dt is the sampling interval ($dt = t_{i+1} - t_i$). In the present study, water flow rate signal is recorded every 5 seconds, then the value of dt will be adopted as 5.
- a_i is the water flow rate expressed in terms of the number of pulses recorded at time t_i.

The main objective is to find the period (i.e. a series of data points) during which the event smoothing would take place. The task will be carried out by setting a criterion for decision making. Given l as a desired filtering level ($l \in N^*$ and dimensionless), a smoothing process is performed on an event section $a_i \rightarrow a_j$, ($i < j$) only if ($\forall\, g_k \in \boldsymbol{g}$) and ($i \le k \le j$), all $|g_k| < l$. The new flow rate for this section ($a'_i \rightarrow a'_j$) is equal to the average flow rate ($\bar{a}$) and can be determined using the formula below:

$$a'_k = \bar{a} = \frac{\sum_i^j a_k}{j - i + 1} \quad , \quad (i \le k \le j) \qquad (2)$$

As can be seen from Figure 2, for example, if l is set as 1, then there are 9 sections where the event smoothing would be identified, and each section is clamped by two points (i.e. absolute gradient values in each particular section are all less than 1). In Figure 3, the filtered patterns of one event from level 1 to 5 were achieved by varying the corresponding value of l from 1 to 5. In the present study, the combined event dissection process was performed by initially smoothing the original event to level 4 ($l = 4$) to attain the base sample, and then eliminating that

base one from the original event to achieve all other sub samples. It should be noted that the limiting value of 4 was selected as it yielded the highest separation accuracy for the overall algorithm over many practices.

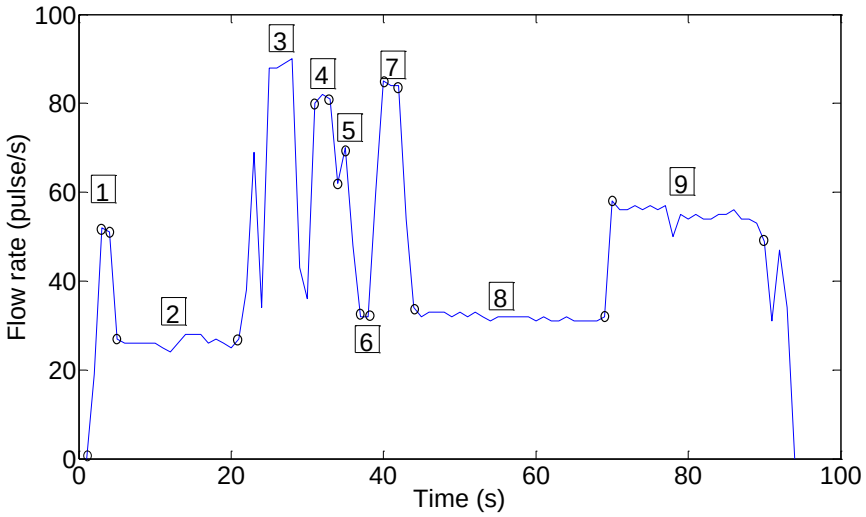

Figure 2 9 sections for leveling purpose when $l = 1$

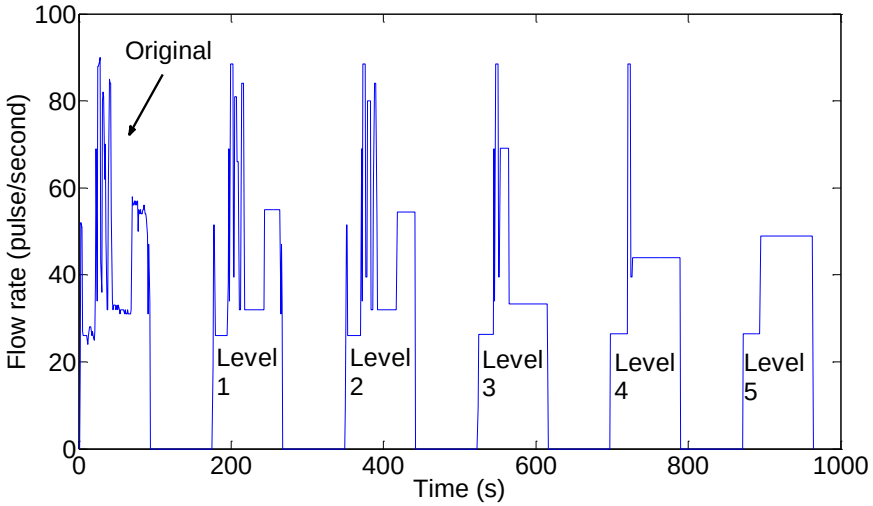

Figure 3 Different levelled patterns of the original water end use event (l varies from 1 to 5, which corresponds to the dissecting levels of 1 to 5)

2.2 Sub event analysis

In the first stage of this study, a combined event will be disaggregated into one base sample, which is actually the smoothed pattern of the original event at l of 4, and several sub samples using a new filtering method. It should be noted that the term “sample” is used to refer to all products achieved from the dissection process. Once these samples have been assigned to proper categories, they will be called “events”. The main task of this section is to employ the existing HMM to place these sub samples into appropriate classes. The output of this process on each sample can be presented in terms of vector $\boldsymbol{c} = (c_1, c_2, c_3, c_4, c_5, c_6, c_7, c_8)$**,** which shows the likelihood of that sample to be recognised as shower (c_1), faucet (c_2),

clothes washer (c_3), dishwasher (c_4), full flush toilet (c_5), half flush toilet (c_6), irrigation (c_7), and bathtub (c_8). The sample will be assigned to the end use event category (i.e. c_1 to c_8) which has the highest probability value.

However, given that the base sample from this analysis process were achieved by removing spiky sections from the original combined event over a subjective dissection process; there still remains some uncertainties as to whether these extracted sub samples are the actual concurrently occurring events or just fluctuating parts of the base event. To address this issue, a boundary condition, which is actually a physical parameter for this particular study, was set for decision making and was represented by vector $\boldsymbol{b} = (b_1, b_2, b_3, b_4, b_5, b_6, b_7, b_8)$ with each element standing for the threshold value of shower (b_1), faucet (b_2), clothes washer (b_3), dishwasher (b_4), full flush toilet (b_5), half flush toilet (b_6), irrigation (b_7), and bathtub (b_8). The achievement of these values was performed through an intensive examination of the existing database; and over many practices, the threshold values which resulted in the highest classification accuracy were adopted for the study. By incorporating this boundary condition for each end use category, if one sample was already assigned to an appropriate category based on the HMM method, but, its likelihood is less than the threshold value for that particular class then that sample still remains unclassified and will be put aside for further analysis. It was also found through many trials that one sub sample achieved from the separation process will be most likely belonging to the more prevalent end use categories, such as: clothes washer, dishwasher, faucet, full flush toilet, and half flush toilet as they are all short events. Therefore, within this sub event analysis process, the end use type of one sub sample is limited to the above mentioned categories. In summary, the final likelihood for one sub sample is displayed in Table 1 below.

Table 1 Final likelihood of one sample for sub event analysis

Category	Original HMM likelihood	Threshold value
Faucet	c_2	b_2
Clothes washer	c_3	b_3
Dishwasher	c_4	b_4
Full flush toilet	c_5	b_5
Half flush toilet	c_6	b_6
Shower	0	N/A
Irrigation	0	N/A
Bathtub	0	N/A

With the attained likelihood values shown in Table 1, the subjected sub sample will be assigned to the category which has the highest probability and also exceeds the threshold value. The sample which does not exceed the limiting threshold discussed prior still remains undetermined, and will fall into one of the following cases :

a. A part of the original base event.
b. Combination of many single events.
c. Combination of many single events with parts of the original base event.

The next analysis stage aims to classify an undetermined sub sample into appropriate end use category. The first task to be undertaken in this stage is to break down the subjected undetermined sample into its respective smaller sample elements, which are named “secondary sub samples”, using the proposed filtering method. All products achieved from this separation process will be categorised using the HMM method along with the subsequent threshold value criterion. The outcome achieved from this classification can fall into one of the following cases:

i. Some of the secondary sub samples are assigned to particular end use categories and some still remain uncertain due to the failure to meet the threshold values.
ii. All secondary sub samples are successfully assigned to their appropriate water end use categories.
iii. All secondary sub samples still remain undetermined.

The first case means that the original undecided sub sample is actually a combination of some single events with parts of the original base event. This conclusion can be explained based on the extraction of classified single events in the subjected sub sample, and undetermined secondary sub samples which are parts of the actual base event. In this case, all remaining uncertain secondary sub samples will be returned to the original base sample at the same time index as it was before the separation process. In the second case where all secondary sub samples are assigned to different end use categories, then it could be inferred that the undetermined sub sample is actually a combination of many other single events. In the third case, if all separated secondary sub samples fail to be classified into any end use category, it is likely that they are all parts of the actual base event, and they will be returned to the base sample for further analysis.

2.3 Base event analysis

Once all sub events have been fully classified, the final step in this combined event study is to analyse the base sample. This base sample analysis occurs after the returning of all undetermined secondary sub samples from the sub event analysis process to the original base one. Manually completed end use analysis on the collected data revealed that a majority of base events are formed by only one or a combination of the following end use categories: shower; full flush toilet, bathtub and long irrigation events. The recognition of the base sample is similar to that for the sub sample which employs the existing HMM model with the output restricted to shower, toilet, bathtub and irrigation. The likelihood vector of one base sample is therefore in form of $\boldsymbol{c} = (c_1, 0,0,0, c_5, 0, c_7, c_8)$**.** However, unlike all of the above described recognition tasks, the process of checking against the threshold value was not required and the score achieved from the HMM technique was adopted as the final likelihood for decision making.

3 MODEL VERIFICATION

The developed combined end use event disaggregation method was verified using 20 independent combined events extracted from the existing database mentioned prior. These events were divided into three types with an increasing level of complexity. Type 1 combined events included two events which occur simultaneously. This is the simplest event combination and 5 samples of this type were utilised to facilitate the testing process. Type 2 combined events were comprised of a group of concurrent events, in which, the largest event was the base one and all other smaller events lied on top of this larger event. This is the most common type of event combination and 10 samples were obtained for verifying. The last and also most complex type of combined event was comprised of two layers of simultaneous events. Specifically, one combined event included three components; the longest event occurred across the bottom serving as the base event. Simultaneous events occurred above the base event, which was considered as the first overlapping layer. In the second overlapping layer, some other events occurred on top of the sub events in the first layer. There were 5 samples of this type selected for the present testing.

The efficiency of the proposed technique was thoroughly verified using three accuracy indices. These indices enabled an assessment on the effectiveness of the method for analysing each combined event, as well as the degree of accuracy attained in recognising extracted single events of each category. For the individual combined event examination, the method effectiveness was considered in terms of

number of events, which aimed to find the ratio between the number of correctly classified events over the total number of single events within the combined one; and the accuracy in terms of volume, which aimed to determine the ratio between the correctly classified volume over the actual volume of each single event within the combined one. These two validation indices were clearly presented in Table 2 for the three different types of event combination discussed prior.

Table 2 Average separation accuracy

Type	Average accuracy in terms of number of event	Average accuracy in terms of volume
1	100	98.9
2	81.4	82.6
3	82.7	82.4
Overall	88.0	88.0

From Table 2, it can be seen that all type 1 combined events have been accurately separated and identified. The accuracy of 98.9% in terms of volume is due to the fact that during the separation process, the starting and ending points of each single event component were not perfectly determined, which led to the slight difference in volume of the separated event compared to that of the actual single one. The accumulation of all volume differences has resulted in this minor reduction in the average accuracy for this type of combined event.

As agreed previously, type 2 combined events are the most commonly occurring in residential households. For type 2 combined event analysis, 81.4% of the total events extracted from the 10 samples were accurately classified, while 82.6% of the total volume has been properly recognised.

The final type 3 combined event to be tested includes two overlapping layer as discussed earlier. It was surprising that the accomplished accuracy in terms of number of event for this very complex type was higher than that of type 2 (82.7% compared to 81.4%), and the accuracy for volume is almost similar (82.4% compared to 82.6%). The results demonstrated that the proposed method was effective in analysing most types of combined events, from two simple simultaneous events to the most complicated sample with two overlapping layers of many simultaneous events.

However, for a more comprehensive understanding on how the technique performed on each end use category, another accuracy index was employed, which was actually the ratio between the number of correctly classified events over the total events of that particular end use category participating in the verification process. Table 3 was established to indicate the accuracy achieved for all end use categories for the three tested types.

Table 1 Disaggregation accuracy for each end use category

Event category	Number of event participating in the verification	Number of correctly disaggregated events	Accuracy
Shower	13	11	84.6
Faucet	79	69	87.3
Clothes washer	16	15	87.5
Dishwasher	14	12	85.8
Toilet	36	30	83.3
Bathtub	8	6	75.0
Irrigation	18	14	77.7

Table 3 shows that the disaggregation accuracy was at least 75% for all water end use categories, with the highest of 87.5% for clothes washer and lowest of 75.0% for bathtub. Bathtub and irrigation had similar patterns to other end use categories and were often misclassified as the others, which reduced their end use accuracies to 75% for bathtub and 77.7% for irrigation. Higher classification accuracies were evident for clothes washer (87.5%), dishwasher (85.8%) and faucet 87.1%. In terms of toilet event classification, the accuracy is a bit lower at 83.3%. It was found that most misclassified toilet events were placed into the faucet group since their patterns have been distorted considerably due to the pressure loss that happens when many events take place concurrently.

4 CONCLUSION

The establishment of an integrated water management system, which employs smart water metering in conjunction with an intelligent algorithm to automate flow trace analysis process, is becoming increasingly feasible due to the work of the research team. The first fundamental step to extract single events from the flow rate series and assign them to appropriate classification was completed in a preceding study using a hybrid combination of HMM and DTW algorithms (Nguyen et al. 2012). While this existing research enabled single events to be classified into their respective end use categories, the problem of dissecting simultaneously occurring events still remained. Therefore, the completion of a comprehensive flow disaggregating method would not be complete until combined event classification has been tackled. The present study has proposed a novel approach for such a complex issue, which incorporates the HMM method in conjunction with a developed filtering technique. Through method testing with 20 selected combined events of increasing difficulty types, the technique has been proved to be effective for this complex disaggregation problem, as the accomplished average recognition accuracies both in terms of number of event and volume were 88%.

The developed analytical method has transferability to other the blind source separation issues in other fields which have to deal with concurrent recognition signal. The key to solving this issue was the integration of appropriate physical characteristics into the HMM model to achieve reliable outcomes**.**

5 REFERENCE

Chien, J.-T., Wang, H.-C. 1997. Telephone speech recognition based on Bayesian adaptation of hidden Markov models. Speech Communication, 22**,** 369-384.

Cho, W., Lee, S.-W., Kim, J.H., 1995. Modeling and recognition of cursive words with hiddenMarkov models. Pattern Recognition 28(12): 1941-1953. doi: 10.1016/0031-3203(95)00041-0

CSIRO (2010) State of the Climate. Bureau of Meteorology, Commonwealth of Australia, Melbourne.

Nguyen,T. H. L, Jutten., C.,1995. Blind source separation for conclusive mixture. Elsevier, 45, 209-229

Nguyen, A. K., Zhang, H., Stewart, R. A.,2012. Development of an intelligent model to categorise different water end use categories: I. Single event. Journal of Hydro-Enviroment. [under review]

Stewart, R. A., Willis, R., Giurco, D., Panuwatwanich, K., Capati, G., 2010. Web-based knowledge management system: linking smart metering to the future of urban water planning. Australian Planner, 47**,** 66 - 74.