Activity Analysis Based on Low Sample Rate Smart Meters

Feng Chen[1] Jing Dai[2] Bingsheng Wang[3] Sambit Sahu[4] Milind Naphade[5] Chang-Tien Lu[6]

[1, 3, 6] Computer Science Department, Virginia Tech
7054 Haycock Road
Falls Church, VA
{[1]chenf, [3]claren89, [6]ctlu} @vt.edu

[2, 4, 5] Watson Research Center, IBM
19 Skyline Drive
Hawthorne, NY
{[2]jddai, [4]sambits, [5]naphade} @us.ibm.com

ABSTRACT

Activity analysis disaggregates utility consumption from smart meters into specific usage that associates with human activities. It can not only help residents better manage their consumption for sustainable lifestyle, but also allow utility managers to devise conservation programs. Existing research efforts on disaggregating consumption focus on analyzing consumption features with high sample rates (mainly between 1 Hz ~ 1MHz). However, many smart meter deployments support sample rates at most 1/900 Hz, which challenges activity analysis with occurrences of parallel activities, difficulty of aligning events, and lack of consumption features. We propose a novel statistical framework for disaggregation on coarse granular smart meter readings by modeling fixture characteristics, household behavior, and activity correlations. This framework has been implemented into two approaches for different application scenarios, and has been deployed to serve over 300 pilot households in Dubuque, IA. Interesting activity-level consumption patterns have been identified, and the evaluation on both real and synthetic datasets has shown high accuracy on discovering washer and shower.

Categories and Subject Descriptors

H.2.8 [Database Management]: Database Applications

General Terms

Algorithms, Design, Experimentation, Performance

Keywords

Smart Meter, Low Sample Rate, Disaggregation, Classification, Hidden Markov Model, Gaussian Mixture Model.

1. INTRODUCTION

Sustainability and design of sustainable technologies have become urgent and important priority for cities given the unprecedented level of resource demand - water, energy, transit, healthcare, public safety - to every imaginable service that makes a city attractive and desirable. At the same time, digital reification of cyber-physical world has been possible with widespread penetration of sensing and monitoring technologies. These two important catalysts have fuelled significant interest and cross organizational collaboration among researchers, industries, urban planners, and government. A lot of technology and research has recently focused on leveraging information from such digital reification of cyber-physical world to help manage various services more efficiently. Our paper takes a step in that direction - examines the feasibility and provides innovative approaches towards influencing people's consumption behavior. More precisely, we provide activity analysis based on smart water meter readings.

Given the real world constraints, we research the feasibility of activity analysis to identify activities from smart utility meter readings. Our study is based on the hypothesis that consumption activities disaggregated from meter readings will empower residents with appropriate insights to influence and shape their behavior. This has been rightly validated through a city-wide survey [1] followed by four-month-long experimentation with a real city [2]. In addition, from disaggregated consumption, utility managers can design and assess conservation programs, and prioritize energy-saving potential retrofits.

Research on disaggregating electricity or water load has been conducted on smart meter readings with fine granularity (mainly between 1 Hz ~ 1MHz). Existing approaches identify appliances/ fixtures based on analyzing steady state or transient state change in real-time consumption. However, they are not suitable for many existing smart meter infrastructure.

Real-world deployments of smart meters are designed for utility billing and some basic analysis requirement, but many of them are not suitable for consumption disaggregation. Smart meters transmit consumption readings using wireless protocols, which consume battery and have dependency on physical environments. Although the meters can sample at a rate even higher than 1MHz, many of existing deployments have chosen to accumulate to 15 min or even longer intervals to ensure reliable data transmission. However, physical environment may still affect the data transmission. This scenario brings the following challenges to consumption disaggregation: 1) Parallel usage activities, e.g., a toilet flush and shower in the same 15 minute interval. 2) Difficulty of aligning usage events temporally, e.g., a shower may appear in one or two intervals. 3) Lack of features, i.e., only aggregated consumption and start time of each interval can be used to identify usage activity. An example of such water meter data and expected disaggregated activities is illustrated in Figure 1.

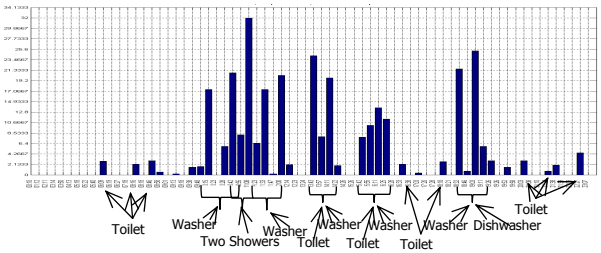


Figure 1. An Example of Data and Disaggregated Activities.

To handle these challenges, we have designed a novel statistical framework for activity analysis on coarse granular smart water

KDD'11, August 21–24, 2011, San Diego, California, USA.

meter readings, and deployed it as a component in Smarter Water Service for Dubuque, IA. In this framework, fixture characteristics, household behavior, and activity correlations are utilized to disaggregate consumption. To implement this framework, we propose two approaches to identify activities. The first approach applies hidden Markov model to capture the relationship among consumption events and hidden activities. The second approach utilizes classification techniques to learn from labeled activities, and a Gaussian mixture model is used for disaggregation. The proposed approaches have been validated using both real-world water consumption and synthetic datasets. The experiments have demonstrated the capability of the proposed disaggregation framework, illustrated the appropriate sample rate for disaggregation in various applications, and revealed interesting usage insights from 300+ pilot households. In summary, the major contributions of this work include:

- Providing activity-level consumption insights to residents and the city management team to support decision making.
- Designing a general disaggregation framework with two implementations for different scenarios.
- Exploring appropriate smart meter sample rate to enable consumption disaggregation.
- Revealing interesting consumption patterns from the disaggregation results.

This paper is organized as follows: Section 2 illustrates the application deployment for the proposed approach, and introduces the related challenges. A novel general statistical framework for disaggregation is proposed in Section 3. The detailed implementations for water consumption disaggregation are described in Section 4. Section 5 evaluates the performance of the proposed approaches under different scenarios with real-world and synthetic datasets and demonstrates some interesting findings from the pilot households. The related work is reviewed in Section 6. Finally, Section 7 concludes our work with future directions.

2. BACKGROUND & PROBLEM

2.1 Application Deployment

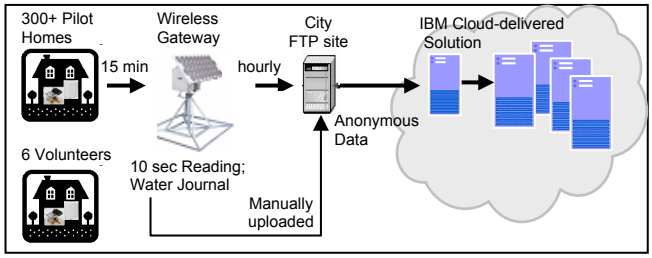


Figure 2. Data Acquisition.

The activity analysis is an important function provided in Smarter Water Service based on smart water meters. The deployed environment of our smart water meter infrastructure is shown in Figure 2. Since August 2010, over 300 pilot households have volunteered to install Neptune R900 smart water meters [3] with UFR (Unmeasured Flow Reducer), which transmit a new aggregated reading roughly every 15 minutes through 900MHz wireless connection. Each aggregated reading is broadcasted repeatedly within the entire interval to ensure the success of transmission. Wireless gateways have been deployed in the city to collect these readings, attach timestamps, and send to a data center through 3G network every hour. In addition, 6 volunteer households had applied data logger which records water consumption every 10 seconds, and had done water usage activity journaling accordingly for a week. All the meter readings have been anonymized and sent to IBM Computing Cloud for analytics.

The software architecture of the deployment is visualized in Figure 4. The smart meter data are first cleaned and transformed by InfoSphere Information Server®(IIS), and then stored in a Smart Meter Database managed by DB2®. On top of this database, Cognos® is utilized to provide OLAP functions such as consumption metric and pattern monitoring; a java-based module is developed to perform advanced analytics functions such as disaggregation and prediction. IBM WebSphere Application Server®(WAS) hosts the service layer to allow users interact with the services. In addition, a community engagement component plays the role of motivating residents through competition and collaboration via multiple media channels. The whole system, as a $850K deployment engagement with Dubuque, IA, has been deployed on IBM Smarter Cities Sustainable Model Cloud, and provides services to residents (300+ pilot households) and the city management team (about 10 government employees)[2].

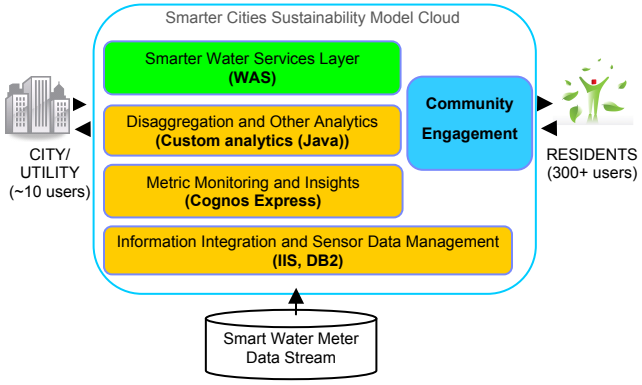


Figure 4. Smarter Water Service Architecture.

The main objective of this Smarter Water Service is to provide affective services that can help the volunteers modify their behavior to be more sustainable, in other words, let the residents know what they need to know to change their behavior. To achieve that goal, one important process is to reveal disaggregated water consumption, so that the users can know where in their houses they could conserve water, and sustainable operations or investment can be suggested. As a component of Smarter Water Service, activity analysis shared the computing resources with the other custom analytics. It works as a backend service that outputs activity-level consumption distribution reports every month from 15-minute aggregated consumption. This component will continuously provide consumption insights as part of the Smarter Water Service, and will be updated by enhancing learning ability and expanded to the expected 4000 households with hourly readings by 2013.

A preliminary summary has shown 6.6% normalized accumulative consumption reduction in 8 weeks after the Smarter Water Service was published in September 2010. In addition, a survey conducted in December 2010 showed that since September,

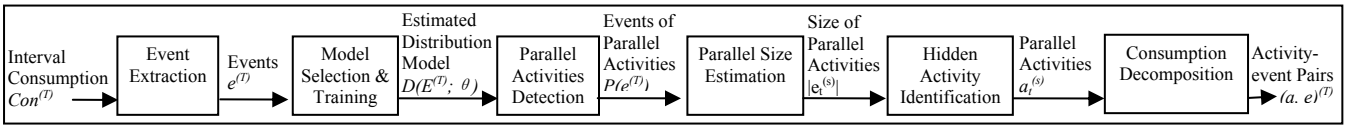


Figure 3. Disaggregation Framework.

out of 64 respondents, 15 households had fixed leaks, 13 respondents had shortened their showers, and 14 purchases on water-efficient toilet/appliances had been made.

2.2 Problem & Definitions

The problem of disaggregation from coarse granular smart water meter readings can be informally described as follows:

Problem: *Given a sequence of aggregated interval water consumption* $\boldsymbol{Con}^{(T)} = (Con_1, \dots, Con_T)$*, where* Con_i *refers to the aggregated water consumption at the i-th time interval, the proposed solution should return a set of activities* $\big((A_1, E_1), \dots, (A_k, E_k)\big)$ *that are most likely to cause the aggregated consumption* $\boldsymbol{Con}^{(T)}$*, where* A_i *refers to an activity state (e.g., washer, shower, or toilet uses), and* E_i *refers to an observation (event) of water consumption for this activity state and is represented by a vector of event features, including total water consumption and start/end time intervals.*

The related terms and their definitions are summarized in Table 1, and will be used in the rest of the paper. We use capital letters to denote random variables and small letters to denote observations.

Table 1. Terms & Definitions.

Term	Symbol	Definition
Consumption	Con	Amount of water used in terms of gallons.
Interval	Int	The time period between 2 consecutive meter readings.
Activity	A	Integer value that represents one of the following: sink, toilet, shower, and washer.
Event	E	A vector of features to represent an event. The event features include total consumption, start/end time, etc.
Event sequence	$(E_1, \dots, E_T)$	A sequence of events occurs in a time window (e.g., 24 hours), where T is the number of events.
Parallel activities	$(A_{t1}, \dots, A_{ts})$	s activities occur together in event E_t
Events of parallel activities	$P(E^{(T)})$	A set of events in $(E_1, \dots, E_T)$ generated by parallel activities.
Parallel sub-events	$(E_{t1}, \dots, E_{ts})$	A set of parallel sub-events whose aggregation generates the event E_t. Each sub-event E_{ti} is generated by a single activity A_{ti}.

2.3 Research Challenges

General challenges for usage disaggregation from single main meter include the following: 1) Appliances/fixtures with similar consumption patterns, e.g., certain sink usage and a toilet flush; 2) Appliances/fixtures with multiple settings, e.g., normal, dedicated, and permanent of a washer; 3) Load variation, e.g., low, medium, and full load of a washer, or length of showers; 4) Multiple cycles, e.g., washer and dishwasher; 5) Lack of real-world ground truth, i.e., hard to collect sufficient labeled data from consumers. Disaggregation with the above challenges can be treated as a real-world classification problem.

In addition, the specific application scenario introduced in the previous section brings more challenges because of the coarse granularity and unstable reading intervals caused by unreliable communication. These limitations cause: 1) Parallel usage activities, e.g., two toilet flushes and a shower in the same 15 minute interval. 2) Difficulty of aligning usage events temporally, e.g., a shower may appear in one or two intervals. 3) Lack of features, i.e., only aggregated consumption and start time of each interval can be used to identify usage activity. These specific challenges make the task of water usage disaggregation more than a classification problem and difficult to solve.

The existing disaggregation approaches focus on analyzing steady state or transient state changes. They cannot handle the specific challenges in this scenario, because no steady state or transient state can be detected with such a low sample rate.

2.4 Observations

Due to the challenges discussed, the aggregated consumption of each interval alone surely cannot provide confident disaggregation results. We need to investigate the available ground truth on what other factors may help improve the disaggregation accuracy. After a study over the activity journaling from the volunteers, we have found three useful characteristics of water usage activities: fixture-dependant, household-dependant, and time-dependant.

2.4.1 Fixture-dependant Pattern

Each fixture category has its own usage pattern in term of consumption and duration that can be used to distinguish it from the others. Specifically, the amount of water consumed in a toilet flush usually fell in several small ranges between 1.5 ~ 5 gallons, and was consistent for a specific toilet. A load of washer generally lasted between 30~60 minutes, and consisted of multiple cycles with similar water usage. Showers had consistent flow rate most of the time, and lasted from 5 minutes to 15 minutes in most cases. Sink usage was usually short in time and low in consumption. These patterns can help briefly categorize the usage events. For example, any interval with flow rate lower than 0.1 gallons per 15 minutes can be filtered out as sink usage. However, using a fixture specification library is not enough to identify parallel activities, or to deliver customized models for each household.

2.4.2 Household-dependant Pattern

Activity patterns heavily depend on the fixture models and occupants of a specific household. For example, households with kids generally spent more time on shower every day; households with open leaks showed continuous usage for a long time; some households have 3 toilets and each has a different specification. Therefore, each household needs to be modeled separately to ensure accurate disaggregation. These models can be learned from historical consumption records and household profiles if available.

2.4.3 Time-dependant Pattern

According to human behavior, some activities may happen frequently during a specific time period, which can be used to distinguish ambitious water usage. One interesting example of such pattern is shower. Most of the labeled showers happened either close to the first event of usage in the morning or close to the first event after work. Although toilet flush occurred almost any time in a day, it was less frequent in working hours and midnight than the rest of a day. Not only time of day, but also day of week has been found drawing impacts on activity patterns. An example could be washer usage which happened mostly during weekends in some households. In addition, some activities are found temporally associated. For instance, a toilet flush in many cases was followed by a short sink usage for hand washing. According to the time-dependant activity patterns, timestamps of usage events should be able to improve disaggregation results significantly.

3. A NEW STATISTICAL DISAGGREGATION FRAMEWORK

Coarse granular smart meter readings cause a large portion of parallel activities, and disaggregation of parallel activities has become a critical and important challenge. This section introduces a new General Disaggregation Framework (GDF) to address the disaggregation problem. As illustrated in Figure 3, the GDF framework applies six phases to disaggregate water consumption. The work flow is described as follows:

Phase 1 Event extraction: Given a sequence of aggregated interval consumption $\boldsymbol{Con}^{(T)} = (Con_1, \dots, Con_T)$, the intervals with continuous consumption are grouped to generate events where

each represents one activity or parallel activities. The output of this phase is an event observation sequence of a given time window: $\boldsymbol{e}^{(T)} = (e_1, e_2, \dots, e_T)$. Hence, $\boldsymbol{e}^{(T)}$ is regarded as one observation of the event random variables $\boldsymbol{E}^{(T)} = (E_1, E_2, \dots, E_T)$. Each event E_i may be generated by a hidden activity (A_i) or several parallel hidden activities $(A_{i1}, \dots, A_{is})$.

Phase 2 Model selection and training: Select an appropriate stochastic model $\boldsymbol{D}(\boldsymbol{E}^{(T)}; \boldsymbol{\theta})$, such as *HMM* or *GMM*, and estimate parameters $\widehat{\boldsymbol{\theta}}$ based on historical labeled or unlabeled observations.

Phase 3 Parallel activity detection: Given the estimated stochastic model $\boldsymbol{D}(\boldsymbol{E}^{(T)}; \widehat{\boldsymbol{\theta}})$, the events with parallel activities $\boldsymbol{P}(\boldsymbol{e}^{(T)})$ can be identified from anomalous events $\boldsymbol{O}(\boldsymbol{e}^{(T)})$. Anomalous events can be obtained using leave-one-out test, i.e., $\boldsymbol{O}(\boldsymbol{e}^{(T)}) = \{e_t \mid e_t \in \boldsymbol{R}(\boldsymbol{E}^{(-t)} = \boldsymbol{e}^{(-t)}, \alpha)\}$, where $\boldsymbol{E}^{(-t)} = (E_1, \dots, E_{t-1}, E_{t+1}, \dots, E_T)$, $\boldsymbol{e}^{(-t)} = (e_1, \dots, e_{t-1}, e_{t+1}, \dots, e_T)$. $\boldsymbol{R}(\cdot)$ refers to the outlying region of normal event E_t that is defined based on the conditional distribution of $[E_t \mid \boldsymbol{E}^{(-t)} = \boldsymbol{e}^{(-t)}]$ and a confidence level α (e.g., 0.99). The calculation of outlying regions based on *HMM* and *GMM* models will be discussed in Section 4. This phase assumes all anomalous events are generated due to parallel activities. An anomalous event may also be generated by true abnormal activities such as a shower lasting more than an hour. However, it is difficult to differentiate these only based on coarse granular meter readings. Hence, we only consider parallel activities.

Phase 4 Parallel size estimation: For each anomalous event observation $e_t \in O(\boldsymbol{e}^{(T)})$, the number of parallel activities that generate e_t can be estimated by

$$s = \min\{s \mid e_t \in \boldsymbol{R}^{-}_{Agg}(\boldsymbol{E}^{(-t)} = \boldsymbol{e}^{(-t)}, Agg(E_{t1}, \dots, E_{ts}), \alpha)\} \quad (1)$$

where $\{E_{t1}, \dots, E_{ts}\}$ refers to the parallel activities (random variables) whose aggregation generates the event e_t, $Agg(\cdot)$ refers to the vector of aggregated features, and $\boldsymbol{R}^{-}_{Agg}(\cdot)$ refers to the normal region of the aggregated features $Agg(E_{t1}, \dots, E_{ts})$. $Agg(E_{t1}, \dots, E_{ts})$ returns aggregated features, such as the total water consumption, the earliest start time, and the latest end time of the sub-events $\{E_{t1}, \dots, E_{ts}\}$. The reason of selecting the minimal s is that heavy consumption (a washer load) can always be decomposed into a large number of small activities (e.g., toilet flushes), which is not reasonable.

Phase 5 Hidden activity identification: For each abnormal event $E_t \in O(E^T)$, given s, the estimated size of parallel activities, this phase estimates the disaggregated activities $\{a_{t1}, \dots, a_{ts}\}$:

$$(a_{t1}, \dots, a_{ts}) = \arg \max_{(a_{t1}, \dots, a_{ts}) \in \{1, \dots, m\}^s} \Pr(A_{t1} = a_{t1}, \dots, A_{ts} = a_{ts} \mid \boldsymbol{E}^{(-t)} = \boldsymbol{e}^{(-t)}, \mathrm{Agg}(E_{t1}, \dots, E_{ts}) = e_t), \quad (2)$$

where m is the total number of activity types (e.g., shower, washer).

Phase 6 Consumption decomposition: Given the hidden parallel activities $\{a_{t1}, \dots, a_{ts}\}$ estimated in Phase 5, the related water consumption of these hidden activities can be estimated as:

$$\begin{aligned} &(Con(e_{t1}), \dots, Con(e_{ts})) \\ &= \arg \max_{Con(e_{t1}), \dots, Con(e_{ts})} L(Con(E_{t1}) = Con(e_{t1}), \dots, Con(E_{ts}) \\ &\qquad = Con(e_{ts}) \mid \boldsymbol{E}^{(-t)} = \boldsymbol{e}^{(-t)}, A_{t1} = a_{t1}, \dots, A_{tm} \\ &\qquad = a_{ts}, Agg(E_{t1}, \dots, E_{ts}) = e_t), \end{aligned} \quad (3)$$

where L is the likelihood function, and $Con(e_{ti})$ is the consumption feature of the sub-event observation e_{ti}, i = 1,.., s.

To evaluate the correctness of *GDF*, we have the theorem as:

Theorem: *Given a sequence of aggregated consumption intervals* $\boldsymbol{Con}^{(T)} = (Con_1, \dots, Con_T)$*, GDF is able to identify true hidden activities* $((A_1, E_1), \dots, (A_k, E_k))$ *of* $\boldsymbol{Con}^{(T)}$*, if the following assumptions are satisfied: a) In Phase 1, The events can be correctly identified and the features extracted are sufficient; b) The distribution* $\boldsymbol{D}(\boldsymbol{E}^{(T)}; \boldsymbol{\theta})$ *is correctly selected and estimated; c) All anomalous events are due to parallel activities; d) The minimal s selected in Phase 4 is correct.*

Proof Sketchy: The four conditions stated above assure that the built statistical model by *GDF* is consistent with the true distribution of hidden activities of $\boldsymbol{Con}^{(T)}$. It follows that the activities identified by *GDF* are most probable results and should be consistent with true hidden activities.

4. DISAGGREGATION APPROACHES

This section presents two approaches based on *GDF* to handle different disaggregation scenarios. When there is no sufficient training data available, which is true in many real-world scenarios, we propose an approach to learn hidden relationship among consumption events and activities without user input based on hidden Markov model (*HMM*). When labeled activities are available for training, we design the second approach to construct statistical models using classification techniques and disaggregate parallel activities using Gaussian mixture model (*GMM*).

4.1 HMM-based Approach

This section presents an implementation of *GDF* based on *HMM*. It is trained based on unlabeled data and performs disaggregation without user input. For the purpose of simplicity, each event E_i is represented by a single feature, the total water consumption. Other features, such as start/end time intervals, and duration can be included to this approach in a straightforward manner.

4.1.1 Event Extraction (GDF Phase 1)

The key challenge of event extraction is the segmentation process. Without labeled historical data, it is necessary to define a set of heuristic rules to generate meaningful events based on domain knowledge. The basic criterion is to keep adjacent interval consumption in a single event if they possibly relate to one activity or parallel activities. This is to avoid the situation where one activity is divided to two separate events, which is not recoverable in our approach. If two nonparallel activities are mistakenly grouped to one event, they can still be identified in the consequent disaggregation process.

Similar to the idea of hierarchical clustering, a bottom-up based segmentation algorithm is proposed as follows:

Step 1: Preprocessing. Remove leaking effects, and filter out all zero-consumption intervals.

Step 2: Initialization. Regard each left interval as one event. Then we have the sequence of initial events $(e_1, \dots, e_k)$, where k is the number of nonzero consumption intervals.

Step 3: Merging heavy events. Define a water consumption threshold ϑ (e.g., 5.5 gallons for 15-minute-size intervals). For each continuous event pair (e_i, e_{i+1}), if $Con(e_i) > \vartheta$ and $Con(e_i) > \vartheta$, merge e_i and e_{i+1}. Repeat until no such pair exists.

Step 4: Merging light events. For each event e_i with $Con(e_i) > \vartheta$, if $0 < Con(e_{i-1})$, then merge e_i and e_{i-1}. Similarly, if $0 < Con(e_{i+1})$, then merge e_i and e_{i+1}. If there is an event e_i with $0 < Con(e_i)$, and both $Con(e_{i-1})$ and $Con(e_{i+1})$ greater than ϑ, then e_i is merged to the segment with the smallest consumption.

Step 5: Merging peak events. Merge two peak events $(Con(e_i), Con(e_j))$ if $dist(e_i, e_j) \le \tau$, where $dist(e_i, e_j) = t_{start}(e_j) - t_{end}(e_i)$, and $t_{start}(\cdot)$ and $t_{end}(\cdot)$ refer to the start and end time of an event respectively. We define an event as a peak if its total water consumption is greater than a threshold γ (e.g., 20 gallons). This step is specifically designed for fixtures like washers,

which consists of multiple peaks with more than 15 minutes empty cycle (no water consumption) between peaks.

4.1.2 HMM Parameter Estimation (GDF Phase 2)

A hidden Markov model is usually trained based on EM algorithm, which can only guarantee local optimum. Given a large number of parameters to be estimated in a HMM model, including the number of hidden states, the initial probabilities, the emission distribution of each state, and the transition matrix, it is critical to find appropriate initial settings for these parameters. By empirical evaluation, we decided a mixture model of three Gaussians for sink events, and Gaussian models for other activity events. This section presents a heuristic based approach to seek initial settings for each household based on generic domain knowledge:

Step 1: Toilet identification. Hierarchical clustering is applied on events to identify toilet clusters. By domain knowledge, toilet clusters could be identified by requiring the cluster size to be greater than 3 times the total number of days in the training data, and the consumption standard deviation smaller than 0.5 gallons.

Step 2: Sink identification. Sink events can be identified as the events with consumption lower than $(\mu_i - 2 * \sigma_i)$, where μ_i and σ_i are the mean and standard deviation of the toilet cluster with the smallest mean consumption in all toilet clusters.

Step 3: Frequent pattern identification. After removing sink events and toilet clusters, hierarchical clustering is applied on the remaining events to identify other qualified clusters. In order to control the HMM complexity, we only keep the 12 clusters with the smallest standard deviation.

Step 4: Cluster labeling. This step gives labels to the qualified clusters based on predefined rules such as a shower usage should be within 5~25 gallons. If some clusters are still not labeled, we label these clusters as "others", which may relate to some unknown activity state or frequent combination of parallel activities.

Step 5: Anomaly removal. The anomalous events are identified based on a Gaussian mixture distribution estimated from qualified clusters. These outliers will impact the training of HMM, therefore they are removed from training data.

Step 6: Probability estimation. Regarding each qualified cluster as a hidden state, we can get the number of hidden states, the mean and standard deviation of each hidden state. The transition matrix and initial probabilities can be estimated based on labeled events.

4.1.3 Disaggregation and Labeling (GDF Phase 3-6)

First, several notations are defined as follows. The set of activity states is $\{1, \dots, m\}$, D is an m by m transition matrix, π is the initial probability of the m states, $p_i(e_t) = Pr(E_t = e_t \mid A_t = i)$, and $u_i(t) = Pr(A_t = i)$. For the purpose of simplicity, we assume that each event E_t conditioned on activity state A_t follows a Gaussian distribution $[E_t|A_t = i] \sim \mathcal{N}(\mu_i, \sigma_i^2)$. Note that the following derivations can also be straightforwardly extended to Gaussian mixture distributions.

Let $P(e) = \begin{bmatrix} p_1(e) & 0 & 0 \\ 0 & \dots & 0 \\ 0 & 0 & p_s(e) \end{bmatrix} \in \mathbb{R}^{s\times s}, \alpha_t = \Pr(e_1, \dots, e_t, A_t) \in \mathbb{R}^s,$
$\alpha_t(a_t) = \alpha_t = \Pr(e_1, \dots, e_t, A_t = a_t) \in \mathbb{R}, \beta_t = \Pr(e_{t+1}, \dots, e_T|A_t) \in \mathbb{R}^s,$
$\beta_t(a_t) = \Pr(e_{t+1}, \dots, e_T|A_t = a_t) \in \mathbb{R}$, and $B_t = DP(e_t)$.

The *HMM* implementations of *GDF* Phase 3 to 6 are as follows:

***GDF* Phase 3: Parallel activity detection**

The probability density function

$$P\left(E_t = e \mid E^{(-t)} = e^{(-t)}\right) = \frac{\alpha_{t-1}^T DP(e)\beta_t}{\alpha_{t-1}^T D\beta_t} = \sum_i w_i(t) p_i(e),$$

where $w_i(t) = \frac{d_i(t)}{\sum_{j=1}^m d_j(t)}, d_i(t) = [\alpha_{t-1}^T D]_i[\beta_t]_i$. It indicates that $\left[E_t = e \mid E^{(t)} = e^{(-t)}\right]$ follows a *GMM*:

$$\left[E_t = e \mid E^{(-t)} = e^{(-t)}\right] \sim \sum_i w_i(t)\mathcal{N}(x|\mu_i, \sigma_i^2)$$

The outlying region of the *GMM* model can be calculated as

$$\boldsymbol{R}(\boldsymbol{e}^{(-t)}, \alpha) = \left\{e \middle| \, |e - \mu_{k^*}| > \sigma_{k^*}\Phi^{-1}\left(\frac{1-\alpha}{2}\right)\right\},$$

where k^* is the Gaussian component closest to e, and $\Phi(\cdot)$ is the cumulative density function (CDF) of a standard Gaussian distribution. Here, we assume that the statistics of outlying events are dominated by the component closest to the observation. This outlying region estimation has been justified in [4] using extreme value statistics.

***GDF* Phase 4: Parallel size estimation**

The probability density function

$$P\left(E_{t1} = e_{t1}, \dots, E_{ts} = e_{ts} \mid \boldsymbol{e}^{(-t)}\right) = \frac{\alpha_{t-1}^T \prod_{i=1}^s \{DP(e_{ti})\}\beta_t}{\alpha_{t-1}^T D^s \beta_t} = \sum_{(l_1,\dots,l_s)\in\{1,\dots,m\}^s} w_{l_1,\dots,l_s} P_{l_1}(e_{t1}) \cdots P_{l_m}(e_{ts}),$$

where $w_{l_1,\dots,l_m}$ is the weight that can be calculated based on the form $\{\alpha_{t-1}^T \cdot \prod_{i=1}^s \{DP(e_{ti})\} \cdot \beta_t\}/\alpha_{t-1}^T D^s \beta_t$.

It implies that

$$\left[E_{t1}, \dots, E_{ts} \mid \boldsymbol{E}^{(-t)} = \boldsymbol{e}^{(-t)}\right] \sim \sum_{(l_1,\dots,l_s)\in\{1,\dots,m\}^s} w_{l_1,\dots,l_s} \mathcal{N}\left(\left[\mu_{l_1}, \dots, \mu_{l_s}\right]^T, \mathrm{diag}\left(\sigma_{l_1}^2, \dots, \sigma_{l_s}^2\right)\right)$$

By linear transformation, we have that

$$\left[E_{t1} + \cdots + E_{ts} \mid \boldsymbol{E}^{(-t)} = \boldsymbol{e}^{(-t)}\right] \sim \sum_{(l_1,\dots,l_s)\in\{1,\dots,m\}^s} w_{l_1,\dots,l_s} \mathcal{N}\left(\sum_{k=1}^s \mu_{l_i}, \sum_{k=1}^s \sigma_{l_i}^2\right).$$

Note that here $Agg(E_{t1}, \dots, E_{tm}) = E_{t1} + \cdots + E_{ts}$. Since $\left[Agg(E_{t1}, \dots, E_{tm}) | \, \boldsymbol{E}^{(-t)} = \boldsymbol{e}^{(-t)}\right]$ follows a Gaussian mixture distribution, the normal region $\boldsymbol{R}_{Agg}^{-}(\cdot)$ can be estimated similarly as in the above *GDF* Phase 3.

***GDF* Phase 5: Hidden activity identification**

The probability density function

$$\Pr\left(A_{t1} = a_{t1}, \dots, A_{ts} = a_{ts} \mid \boldsymbol{E}^{(-t)} = \boldsymbol{e}^{(-t)}, E_{t1} + \cdots + E_{ts} = e_t\right) = \frac{\alpha_{t1}(a_{t1}) \prod_{i=1}^{s-1} \Pr\left(a_{t(i+1)}|a_{ti}\right) \Pr\left(\sum_k E_{tk} = e_t|a_{t1}, \dots, a_{ts}\right) \beta_{ts}(a_{ts})}{L_T},$$

where L_T is the likelihood of the whole sequence and can be neglected when solving the problem (2). Note that the random variables $E_{t1}, \dots, E_{ts}$ are independent to each other given their hidden activity states $A_{t1}, \dots, A_{ts}$. The probability density function $\Pr(\sum_k E_{tk} = e_t|a_{t1}, \dots, a_{ts})$ can be calculated by simple linear transformation of independent Gaussian random variables.

***GDF* Phase 6: Consumption decomposition**

Given the hidden activity states $\{a_{t1}, \dots, a_{ts}\}$, we have that

$$[E_{t1}, \dots, E_{ts} \mid a_{t1}, \dots, a_{ts}] \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\mu} = \left[\mu_{a_{t1}}, \dots, \mu_{a_{ts}}\right]^T, \boldsymbol{\Sigma} = \mathrm{diag}(\sigma_{a_{t1}}^2, \dots, \sigma_{a_{ts}}^2)$. The optimal solution of the problem (3) can be obtained as [5]

$$[e_{t1}, \dots, e_{ts}]^{\mathrm{T}} = \boldsymbol{\mu} - \boldsymbol{\Sigma}^{-1}\mathbf{1}^T(\mathbf{1}^T\boldsymbol{\Sigma}\mathbf{1})^{-1}(\mathbf{1}^T\boldsymbol{\mu} - e_t).$$

4.2 Classification-GMM-based Approach

Different from the *HMM*-based approach, this section presents a mixed model approach to the disaggregation problem that requires labeled data for training. It first applies a classification model (e.g., support vector machine, neural network, and *k*-nearest neighbor classifier) to classify each event as a single activity, or a known

frequent combination of parallel activities, or an unknown infrequent combination of parallel activities. For the events classified to the last category (unknown infrequent combinations), it applies an implementation of the *GDF* framework based on *GMM* to disaggregate parallel activities.

Assume that we are given a sequence of aggregated interval consumption $\boldsymbol{Con}^{(T_1)} = \left(Con_1^*, \dots, Con_{T_1}^*\right)$ and the related hidden activities $\left((a_1^*, e_1^*), \dots, (a_k^*, e_k^*)\right)$ as the labeled training data. The objective is to build a model on $\boldsymbol{Con}^{(T_1)}$ that can identify unknown hidden activities $\left((a_1, e_1), \dots, (a_k, e_k)\right)$ of a new aggregated intervals consumption sequence $\boldsymbol{Con}^{(T)} = (Con_1, \dots, Con_T)$.

4.2.1 Event Extraction (GDF Phase 1)

This phase first applies the same procedure as in Section 3.2.1 to identify a sequence of events. Here each e_i has six features, which include the start time, duration, total consumption, minimal interval consumption, maximal interval consumption, and number of peaks.

4.2.2 Classification (GDF Phase 2)

The event extraction phase returns an event sequence $(e_1, \dots, e_k)$, where each e_i is represented by a vector of six features ($e_i \in \mathbb{R}^6$). Note that all the features are mapped to real type values, in order to apply classification models such as SVM and neural network.

Here, we neglect the dependencies between events and treat $(e_1, \dots, e_k)$ as a set of independent training instances: $\{e_1, \dots, e_k\}$. Based on the labels $(a_1^*, e_1^*), \dots, (a_k^*, e_k^*)$, it is able to identify hidden activities of each event e_i. To decide class labels, not only single activities (e.g., toilet, shower, and washer) are treated as distinct classes, but also frequent combinations of parallel activities are regarded as distinct classes. The current setting is that frequent parallel activities should occur at least once per week.

4.2.3 GMM-based Disaggregation (GDF Phase 3-6)

After the classification process, each event has been labeled as a single activity, or known/unknown combination of parallel activities. For parallel activities, a *GMM*-based implementation of the *GDF* framework is proposed to disaggregate parallel activities. The basic procedures are as follows:

Based on the labels of training events $\{e_1, \dots, e_k\}$, it is able to collect training instances for each activity state, such as toilet, shower, and washer. For simplicity, in this disaggregation step, we only consider a single feature (the total water consumption), for each event e_i. Each single-activity related event (E_t) can modeled by a Gaussian mixture distribution as $E_t \sim \sum_{i=1}^{m} \pi_i \mathcal{N}\left(\mu_i, \sigma_i^2\right)$, where π_i is the prior probability of the activity state i, and $\mathcal{N}\left(\mu_i, \sigma_i^2\right)$ is the event distribution of activity i.

Given an event e_t that is classified as parallel activities, the objective is to identify the most probable hidden activities $\left((a_{t1}, e_{t1}), \dots, (a_{ts}, e_{ts})\right)$ with $Agg(e_{t1}, \dots, e_{ts}) = e_t$. Here the aggregation function Agg is the summation function $\sum(\cdot)$. The GDF disaggregation framework can be employed here, which can be regarded a simplified case of HMM based approach. Readers are referred to [15] for detailed specifications.

5. EVALUATION & FINDINGS

The framework has been implemented using JDK 1.5 and deployed in the Custom Analytics Layer of the Smarter Water Service (Figure 4). Pie charts of activity consumption distribution are generated to illustrate how each fixture has been used on monthly basis. From the Smarter Water Service layer interface, the residents can browse their own consumption distribution; meanwhile, the government agency and utility manager can explore how water has been consumed by each activity at regional level.

Both HMM-based and GMM-based approaches have been implemented and evaluated. Specifically, for the GMM-based approach, we have assessed three classification methods, k-Nearest Neighbor classification (kNN-GMM), Artificial Neural Network (ANN-GMM), and Support Vector Machine (SVM-GMM) accordingly. Given the available labeled activities, the evaluation focused on identifying toilet flushes, showers, and washer loads.

To evaluate the effectiveness of consumption disaggregation on identifying these activities, we adopted three metrics, *precision*, *recall*, and *F-measure*. The major reason of using these metrics is that the disaggregation evaluation is similar to an information retrieval process, where subsets of intervals represent certain true activities and the testing results are also subsets of intervals labeled as activities. The metrics need to capture not only how many labels are matched, but also how many true activities are missed and how many false labels are placed. These metrics are defined as follows: *Precision* refers to the portion of matched activities within the corresponding disaggregation results; *Recall* refers to the portion of matched activities within the corresponding true activities; *F-measure* is the harmonic mean of precision and recall.

To evaluate the proposed disaggregation solution, we have applied both HMM-based and GMM-based approaches on the consumption of 6 volunteer households, as well as 50 simulation datasets that were generated based on their labeled consumption. In addition, we varied the sample rate in these datasets to investigate its impact on disaggregation results. The correlation between sample rate and effectiveness can provide guidance to future planning and deployment of human activity analysis applications.

Due to the lack of labeled activities from most of the pilot households, we only applied the HMM-based model to analyze activities of the 300+ pilot households. Some interesting patterns discovered can illustrate common human behavior characteristics.

5.1 Datasets

A real-world dataset was collected from 6 volunteer households. It consists of 1/10 Hz water reading and the corresponding usage journaling records for 7 days. The usage journaling was input manually by these volunteers, so it always has approximated timestamps and missing activities, which introduce inaccuracy which needs to be handled carefully. Note that these households came from various demographic categories and showed significantly different consumption patterns. A summary of labeled activities from one volunteer is listed in Table 2 as an example.

Table 2. Water Journaling of One Household.

Fixture	Occurrences	Total Amount	Percentage
Shower 1	5	71	7%
Shower 2	5	57	6%
Washer	9	366	38%
Toilet 1	43	217	24%
Toilet 2	33	68	7%
Others (sink & unlabeled)	N/A	186	19%

50 simulation datasets were generated by simulating occurrences and corresponding consumption of activities according to their distributions in the labeled dataset from the 6 volunteer households. Firstly, from the labeled activities, the number of instances of each activity in a week was estimated using Poisson distribution. Each instance was randomly assigned to a day and time according to the distributions of labeled activities in day-of-week and time-of-day domains. These distributions were captured by activity occurrence histograms generated from labeled activities and smoothed by kernel density. Once date and start time of an instance was determined, its consumption and duration was randomly picked from a dictionary of the corresponding labeled activities. Finally, consumption noise of each day was randomly picked from 42 (6

households * 7 days) samples, of which each contains unlabeled consumption (<2 gallons) of a whole day. In this way, simulated consumption data for 6 months were generated in each dataset.

A live dataset was constructed from the 15-min consumption of all the pilot households since August 2010. This dataset has inconsistent reading intervals all the time, missing readings due to communication failure, and even water leaks that can impair the disaggregation results.

5.2 Parameter Settings & Baseline Methods

For HMM-based approach, the major settings are as follows: 1) in GDF Phase 1 (event extraction) Step 3 (merging heavy events), the threshold ϑ was set to 5.5 gallons; 2) in GDF Phase 1 (event extraction) Step 5 (merging peak events), the thresholds τ and γ were set to 15 minutes and 20 gallons, respectively; 3) in GDF Phase 2 (HMM parameter estimation) Step 4 (cluster labeling), the clusters with mean consumption between 1.2 gallon and 6 and frequency greater than two times per day were labeled as toilets; the clusters with mean consumption between 8 and 30 were labeled as showers; the clusters with mean consumptions between 30 and 55 gallons were labeled as washers; the clusters with frequency smaller than 1 times per day were disregarded; and the left clusters were labeled as "others"; 4) the number of states in *HMM* was decided automatically (See GDF Phase 2 step 3). Note that all the preceding parameters were decided based on domain experiences.

For kNN-GMM-based approach, the event extraction phase was the same as that in HMM-based approach. Note that the same event extraction process was also used in all other compared approaches. The kNN classifier used in the experiments was provided by MATLAB-2008a Bioinformatics Toolbox. One major parameter is the number of nearest neighbors used in the classification. We applied 10-folder cross validation to select the best k from the candidate values from 5 to 15.

For ANN-GMM-based approach, the neural network classifier was provided by MATLAB 2008a Neural Network Toolbox. We used one-per-class cording for multiclass classification. In one-per-class coding, each output neuron is designated the task of identifying a given class. The output code for that should be 1 at this neuron and 0 for others. We used Levenberg-Marquardt backpropagation, which is the default training algorithm in MATLAB. 10-folder cross validation was used to select the best parameter "the number of hidden layers" in the range from 2 layers to 8 layers. Other parameters were the default settings. Note that, another popular training algorithm is "Gradient descent back propagation" with two major parameters "learning rate" and "the number of hidden layers". We have also tried this training algorithm in experiments. But results indicate that the Levenberg-Marquardt backpropagation method is more accurate and efficient.

For SVM-GMM-based approach, the SVM classifier was provided by LIBSVM [6]. We used the popular radial basis function as the kernel function. There are two parameters including cost (c) and gamma (g). These two parameters were tuned by 10-folder cross validations, and the best parameters was selected from different combinations of the cost parameter (c) range: $log_2(c) = 1{:}\,0.25{:}\,5$, and the gamma parameter (g) range: $log_2(g) = -7{:}\,0.25{:}-1$. We used the "one-against-one" method for multiclass classification.

Two baseline approaches, named random-pick and knapsack based, were applied to evaluate the effectiveness of the above four proposed methods. The random-pick method is described as follows: First, conduct the same event extraction as in HMM-based method; second, the events with consumption smaller than 2 gallons are labeled as sink uses; third, the left events are randomly labeled to toilet, shower, and washer uses.

The knapsack based method is described as follows: First, conduct the same event extraction as in HMM-based method; second, knapsack each segment to the best combination of the following activities: "Toilet-old (1.6 gallons)", "Toilet-new (4 gallons)", "Shower-Low-flow (15 gallons)", "Shower-Standard (30 gallons)", "Laundry (50 gallons)", and "Sink (<=1.6)".

5.3 Effectiveness Comparison

To demonstrate the effectiveness of proposed approaches, we used the labeled activities from water journaling and the simulation datasets as ground truth, and compared the proposed approaches. The comparison was conducted among 4 versions of disaggregation approaches, HMM, kNN-GMM, ANN-GMM, and SVM-GMM; and the two baseline solutions, random pick and knapsack. Cross validation was applied to find the best parameters for the corresponding classification methods.

As shown in Table 3, all the proposed approaches achieved about 95% precision on shower identification, while the recall was relatively low (77~81%). It was because the deviation of shower consumption is very high in real life. In many cases, consumption of a shower may be similar to that of two toilet flushes, or a front-load washer. Therefore, some true showers could not be correctly identified. But once an activity is labeled as a shower, it's very likely to be true. Although these four methods performed similarly on labeling showers, SVM-GMM achieved the highest scores.

Table 3. Precision, Recall, and F-measure on Simulation Data.

Precision, Recall, F-measure	**Toilet** Mean (Standard Deviation)	**Shower** Mean (Standard Deviation)	**Washer** Mean (Standard Deviation)
HMM	0.7704 (0.08), 0. 6651 (0.04), 0.7110 (0.04)	0.9471 (0.04), 0.7883 (0.04), 0.8594 (0.03)	0.7839 (0.06), 0.9610 (0.04), 0.8620 (0.04)
kNN-GMM	0.7291 (0.07), 0.8552 (0.03), **0.7850 (0.04)**	0.9552 (0.02), 0.7723 (0.05), 0.8530 (0.03)	0.8536 (0.06), 0.8937 (0.09), 0.8702 (0.06)
ANN-GMM	0.5982 (0.05), 0.8709 (0.03), 0.7075 (0.04)	0.9584 (0.03), 0.7670 (0.06), 0.8505 (0.04)	0.8554 (0.08), 0.8994 (0.12), 0.8710 (0.09)
SVM-GMM	0.4669 (0.07), 0.8873 (0.02), 0.6086 (0.06)	0.9622 (0.02), 0.8057 (0.05), **0.8761 (0.03)**	0.8613 (0.06), 0.9329 (0.06), **0.8940 (0.04)**
Random Pick	0.1022 (0.03), 0.0531 (0.01) 0.0699 (0.02)	0.1514 (0.03), 0.1608 (0.04) 0.1560 (0.03))	0.0737 (0.02), 0.3237 (0.10) 0.1201 (0.07)
Knapsack	0.0655 (0.01), 0.1534 (0.02) 0.0918 (0.02)	0.4570 (0.05), 0.3294 (0.05), 0.3828 (0,05)	0.8619 (0.16), 0.3516 (0.13) 0.4995 (0.19)

Different to shower, washer loads were disaggregated with very high recall (89~96%), and relatively low precision (78~86%). Generally, cloth washer is the heaviest and meanwhile the least frequent activity on water consumption in a household. Based on the specifications and settings of a washer, its water consumption is usually consistent. That's the reason why almost all of the washer instances can be learned and identified. On the other hand, a washer usage usually crosses multiple intervals. This usage pattern may be similar to certain combinations of other consumption. Therefore, some other consumption was classified as washer by the disaggregation approaches. In overall, SVM-GMM achieved the best overall performance, and HMM got the highest recall.

Detecting toilet flushes is the most difficult task comparing to shower and washer. Because toilet usage typically happens very frequently and costs a small amount of water, it is hard to be distinguished from sink usage in 15-minute interval, or be identified when combined with heavy activities such as a shower or a washer load. All the four approaches had F-measure between 61% and 78%. HMM was the only approach with precision higher than recall. KNN-GMM performed the best in terms of F-measure.

Due to the small number of training data (<= 4 days per house), GMM-based approaches failed to disaggregate consumption on the

volunteer households. As shown in Table 4, HMM perfectly identified the washer usage, and disaggregated showers with high scores. The F-measure for toilet disaggregation with HMM only achieved 55%, although still much better than the baselines.

Table 4. Precision, Recall, and F-measure on Volunteers.

Precision, Recall, F-measure	**Toilet** Mean (Standard Deviation)	**Shower** Mean (Standard Deviation)	**Washer** Mean (Standard Deviation)
HMM	0.516 (0.27), 0.597 (0.17) **0.5536 (0.22)**	0.831(0.138), 0.818 (0.144), **0.8244 (0.14)**	1 (0), 1 (0) **1 (0)**
Random Pick	0.20 (0.18), 0.19 (0.08) 0.1949 (0.13)	0.08 (0.09), 0.19 (0.16) 0.1126 (0.17)	0.07 (0.09), 0.29 (0.34) 0.1128 (0.27)
Knapsack	0.20 (0.10), 0.904 (0.01) 0.3275 (0.05)	0.52 (0.34), 0.47 (0.16) 0.4937 (0.25)	0.44 (0.52), 0.23 (0.27) 0.3021 (0.39)

5.4 Impact of Sample Rate

Choosing an appropriate sample rate for smart meter deployment is a very important decision that may affect hardware and maintenance cost. This set of experiments can provide practical suggestions from the requirement of activity analysis. Reading intervals of the simulation datasets were varied from 15 min to 3 hours in this set of experiments to evaluate its impact on the accuracy of disaggregation results. Both HMM and GMM methods were evaluated in this set of experiments. SVM-GMM was selected to represent GMM, because it had shown practically good accuracy and efficiency in previous experiments. As suggested in Figure 5, both 15 and 30 min intervals provide acceptable results. 1 hour interval supports fair disaggregation of washer and shower, but cannot identify more than half of toilet flushes.

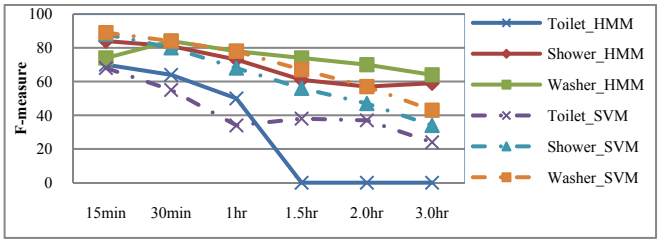
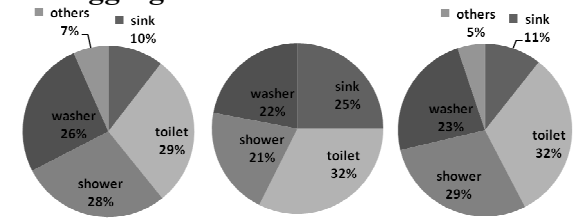
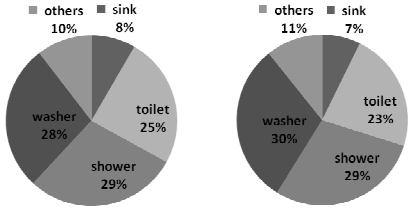


Figure 5. Impact of Interval Length.

5.5 Disaggregation for Pilot Households



a) Pilot Households b) Single Adult c) Two Adults



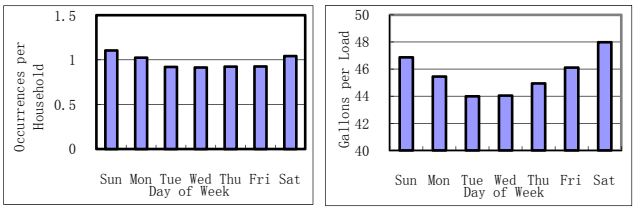
d) Households with Kids e) Households with Toddlers

Figure 6. Distribution vs. Demographic Info.

The proposed HMM-based approach has been applied on 300+ pilot households with 15 minute meter readings. Hidden Markov models were constructed for each household, and water consumption since August 2010 was disaggregated into activities to provide insights to residents and the city management team. Some interesting usage patterns discovered from the disaggregation results are illustrated in the following paragraphs.

By combining with demographic survey results, we first summarize the consumption distribution of different types of households in pie charts as shown in Figure 6. Each pie chart shows the portion of water each activity used by a given group of households. The consumption that cannot be disaggregated is included in category 'others'. The consumption distribution of all the pilot households is illustrated in Figure 6 a), where toilet and shower used about 30% each, and washer used about 25%. Households with single occupant (Figure 6 b)) showed different usage pattern, where shower only consumed 21% of the overall usage and washer reduced to 22%. Figure 6 c) shows the pie chart for households with two adults only. Compared to the single adult households, households of two adults consumed significantly higher in shower. On the other hands, kids in general caused more washer usage. As shown in Figure 6 d) and e), households with kids brought washer usage to 28%, and more specifically, households with toddlers had increased washer usage further to 30%. By comparison, a resident can easily figure out on which activity his or her household needs more efforts to conserve water.

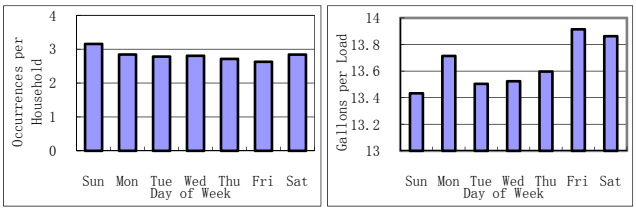
Temporal patterns of washer and shower usage have been identified from the disaggregation results. As shown in Figure 7, the pilot households preferred to use washer in weekends, and each weekday there was about 0.9 load per household in average. Not only the number of loads, but also the size of each load increased in weekends. Figure 7 b) illustrates that each load on Saturday used 9% more water than a load on Tuesday or Wednesday. This is reasonable because usually heavy laundry is saved to weekends.



a) Daily Occurrences b) Gallons per Load

Figure 7. Washer Usage vs. Day of Week.

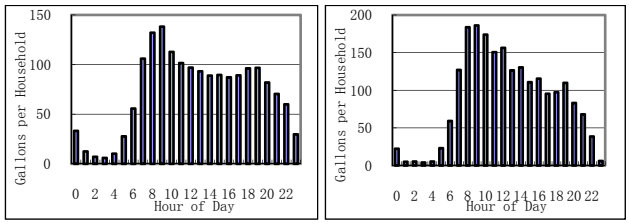
Similar to washer, as can be seen in Figure 8 a), more showers happened during the days in weekends. However, interestingly, an average shower on Sunday used the least water in a week, which was 10% less than one on Saturday. Furthermore, a shower on Friday consumed the highest amount of water in a week. It seemed that people wanted to relax and enjoyed longer showers on Friday, while the stress from work arrived early on Sunday.



a) Daily Occurrences b) Gallons per Load

Figure 8. Shower vs. Day of Week.

Figure 9 demonstrates the time of day distributions of shower and washer across the pilot households. As expected, the peaks of showers happened during 8~9 am and 6~7 pm in a day, which are before and after work. Washer usage showed a similar distribution in b), although the pm peak was not significant. That consistency could be explained as that many washer loads occurred right after a shower to handle the changed clothes.



a) Shower b) Washer Usage

Figure 9. Shower/Washer vs. Time of Day.

6. RELATED WORK

Non-intrusive load monitoring has been proposed based on analyzing steady state change and transient state change. So far most of the research effort has been focused on electricity load disaggregation with high sample rate [7-13]. A power meter with high sample rate (>= 1Hz) can identify most of the state changes of multiple metrics (e.g., power, reactive power, voltage, and harmonics) caused by individual appliances in a real-world home. Based on state change of current and voltage, a non-intrusive load monitoring approach [7] was proposed to determine power consumption of individual appliances. An electrical noise sensor has been used to disaggregate consumption by running SVM on transient noise of turning on and off appliances [8]. By measuring voltage of each outlet in a house, one approach [12] applied kNN and SVM to classify appliances. This approach collected peak, average, and RMS of voltage of a single target with 4kHz sample rate, and achieved best results using an NN classifier. An NN-based disaggregation approach has been proposed to identify appliances with 90% accuracy using only the main power meter [9, 13]. The features it used consist of power, reactive power, voltage RMS, and harmonics for state transition. RECAP has recently been proposed using artificial neural network (ANN) to disaggregate electricity usage [11]. Features including power factor, peak and RMS of voltage and current were aggregated every minute and analyzed in a 3-layer ANN. To extract better features, Matrix Pencil [10] has been proposed to model each signal as complex plan, and use residues and poles as features for disaggregation. Improved disaggregation results have been demonstrated.

Compared with electricity disaggregation, residential water disaggregation has attracted much less research effort. To the best of our knowledge, there has not been any design that can disaggregate water consumption either using a single water meter or from a sample rate lower than 500Hz. Microphone-based sensors were applied on major water pipes (cold inlet, hot inlet, and sewing) to recognize usage activities [14]. Combining the timestamps that these microphones detect noise, the authors identified most of the water usages. However, this approach has difficulties to disaggregate concurrent activities and cannot determine water volume. Integration of a water meter and a network of accelerometers [15] has been proposed to estimate the flow rates based on pipe vibration. This approach has been applied in laboratory environments to disaggregate water usage. To avoid accessing water pipes, an approach using pressure sensor on main source [16] was proposed to identify fixtures. This approach applies hierarchical classifiers to first detect valve open and close events, and then label fixtures. Due to the 1 kHz sample rate, it can clearly capture *on* and *off* signals of fixtures from water pressure.

7. CONCLUSION

This paper describes a design and deployment of activity disaggregation using low sample rate smart water meters in Dubuque, IA. In the proposed general disaggregation framework, fixture characteristics, household behavior, and activity correlations are modeled to disaggregate water consumption. Implementations based on Hidden Markov Model and Gaussian Mixture Model have been developed accordingly to provide insights for helping residents improve their behavior and supporting utility manager's decision making. Evaluation on both real and simulation datasets have demonstrated the effectiveness of the disaggregation approaches, and revealed some interesting patterns from pilot households. Future efforts may include providing user annotation interface to support learning from feedback; and expanding the disaggregation service to electricity smart meters.

8. ACKNOWLEDGEMENTS

The authors would like to gratefully acknowledge the collaboration from the Information Services Department of City of Dubuque on smart meter data transmission; and the support from Dubuque 2.0, an NGO, for engaging the volunteer households.

9. REFERENCES

[1] Dubuque2.0, "Inspiring Sustainability," 2010.
[2] T. Woody, "Smart Water Meters Catch On in Iowa," in *The New York Times* New York City, 2010.
[3] NeptuneTechnologyGroup, "R900 RF Wall or Pit MIU Product Sheet," 2009.
[4] S. J. Roberts, "Novelty Detection using Extreme Value Statistics," *IEE-VISP,* vol. 146, pp. 124-129, Jun. 1999.
[5] H. Rue, "Fast Sampling of Gaussian Markov Random Fields," *JRSS: Series B,* vol. 63, pp. 325-338, 2001.
[6] C. C. Chang and C. J. Lin, "LIBSVM: a library for support vector machines," 2001.
[7] G. W. Hart, "Nonintrusive Appliance Load Monitoring," *Proceedings of the IEEE,* vol. 80, pp. 1870 - 1891, Dec. 1992.
[8] S. N. Patel, T. Robertson, J. A. Kientz, M. S. Reynolds, and G. D. Abowd, "At the Flick of a Switch: Detecting and Classifying Unique Electrical Events on the Residential Power Line," in *ACM-ICUC*, Innsbruck, Austria, 2007, pp. 271-288.
[9] M. Berges, E. Goldman, H. S. Matthews, and L. Soibelman, "Learning Systems for Electric Consumption of Buildings," in *ASCE International Workshop on Computing in Civil Engineering*, Austin, TX, 2009, pp. 1-10.
[10] H. Najmeddine, K. E. K. Drissi, C. Pasquier, C. Faure, K. Kerroum, T. Jouannet, M. Michou, and A. Diop, "Smart metering by using "Matrix Pencil"," in *IEEE EEEIC*, Prague, Czech Republic, 2010, pp. 238-241.
[11] A. G. Ruzzelli, C. Nicolas, A. Schoofs, and G. M. P. O'Hare, "Real-Time Recognition and Profiling of Appliances through a Single Electricity Sensor," in *IEEE SECON*, Boston, MA, 2010, pp. 1-9.
[12] T. Saitoh, T. Osaki, R. Konishi, and K. Sugahara, "Current Sensor Based Home Appliance and State of Appliance Recognition," *SICE JCMSI,* vol. 3, pp. 86-93, Mar. 2010.
[13] M. Berges, E. Goldman, H. S. Matthews, and L. Soibelman, "Enhancing Electricity Audits in Residential Buildings with Nonintrusive Load Monitoring," *Journal of Industrial Ecology,* vol. 14, pp. 844-858, Oct. 2010.
[14] J. Fogarty, C. Au, and S. E. Hudson, "Sensing from the Basement: A Feasibility Study of Unobtrusive and Low-Cost Home Activity Recognition," in *ACM UIST*, Montreux, Switzerland, 2006, pp. 91-100.
[15] Y. Kim, T. Schmid, Z. M. Charbiwala, J. Friedman, and M. B. Srivastava, "NAWMS: Nonintrusive Autonomous Water Monitoring System," in *ACM SenSys*, Raleigh, NC, 2008, pp. 309-322.
[16] J. Froehlich, E. Larson, T. Campbell, C. Haggerty, J. Fogarty, and S. N. Patel, "HydroSense: Infrastructure-Mediated Single-Point Sensing of Whole-Home Water Activity," in *ACM ICUC*, Orlando, FL, 2009, pp. 235-244.