Taylor & Francis
Taylor & Francis Group

Science and Technology for the Built Environment

ISSN: 2374-4731 (Print) 2374-474X (Online) Journal homepage: http://www.tandfonline.com/loi/uhvc21

Data-quality detection and recovery for building energy management and control systems: Case study on submetering

Yangyang Fu, Zhengwei Li, Fan Feng & Peng Xu

To cite this article: Yangyang Fu, Zhengwei Li, Fan Feng & Peng Xu (2016) Data-quality detection and recovery for building energy management and control systems: Case study on submetering, Science and Technology for the Built Environment, 22:6, 798-809, DOI: 10.1080/23744731.2016.1195658

To link to this article: http://dx.doi.org/10.1080/23744731.2016.1195658

Accepted author version posted online: 31 May 2016.
Published online: 31 May 2016.

Science and Technology for the Built Environment (2016) **22**, 798–809

ISSN: 2374-4731 print / 2374-474X online
DOI: 10.1080/23744731.2016.1195658

Taylor & Francis
Taylor & Francis Group

Data-quality detection and recovery for building energy management and control systems: Case study on submetering

YANGYANG FU, ZHENGWEI LI, FAN FENG, and PENG XU*

School of Mechanical Engineering, Tongji University, No. 4800, Caoan Road, Jiading District, Shanghai, China

Most modern buildings are equipped with building energy management and control systems. These systems can store tremendous amounts of data on buildings' performance and energy usage. A significant amount of data on buildings' mechanical devices, particularly electricity-consumption data, is now available for analysis. However, the quality of the collected data is questionable. Some data are mislabeled, and others contain gaps and errors. In this article, a methodology based on a correlation coefficient and a wavelet-based support vector machine predictor is proposed to detect and recover the proportional deviation data faults and faults caused by network communication. After testing this methodology with electricity data collected from a large commercial building, it is found that a high accuracy of faulty data alerts and automated data recovery can be achieved. Considering the wide use of building energy management and control system data for performance monitoring, fault detection and diagnostics, and demand responsive control, this method is useful and practical in many engineering situations.

Introduction

The building sector consumes more than 30% of the total energy worldwide (IEA 2010). An efficient way to alleviate global warming and improve environmental sustainability is to enhance building energy efficiency. To better track building energy performance and power consumption, modern building energy management and control systems (EMCS) have increasingly paid attention to storing control data and recording energy usage. For example, all commercial buildings in California are required to have electrical meters that record data on 15-minute intervals. Building-control vendors are building new functions in their control system, such as demand response control, fault detection, and diagnostics. All of these functions need clear and reliable data to support them (Effinger et al. 2012).

Although modern EMCS have the ability and storage capacity to trend large amounts of data and perform preliminary analyses, engineers claim that these data are rarely or never used. One reason is that the quality of the trended data is poor. EMCS were originally designed for control, not record saving and performance monitoring. Some data are mislabeled, and others contain gaps and errors (Smothers 2002).

To better track buildings' electricity use, sometimes, more than a whole-building electric meter is needed. For example, the California Public Utilities Commission (CPUC) issued a decision on the sub-metering of electricity in multi-story commercial buildings. By identifying where power is used or wasted, submetering can help building owners and tenants not only solve the split incentive issue but also know which building equipment and system needs upgrades or better management and scheduling. Furthermore, to promote submetering, the U.S. Department of Energy's Buildings Technologies Program announced one of its latest challenges: an initiative to develop a $100 wireless submeter (DOE 2013).

Building EMCS data can be used in various ways, such as detecting abnormal electricity usage behavior in whole buildings (Dodier and Kreider 1999; Fontugne et al. 2013a, 2013b; Liu et al. 2010) and in components (Lee et al. 1996) and conducting building performance analyses using comparative methods (Guo et al. 2014). However, improper installation, faulty sensors (for example, those that have completely failed, are aging, or have calibration errors), environmental electromagnetic disturbance (O'Driscoll and O'Donnell 2013), and interrupted signal transmission processes account for the unreliability and inaccuracy of building EMCS data. These problematic data could be detrimental to various schemes that make decisions based on measurements. Therefore, the timely detection and diagnosis of the occurrence and reliable recovery of faulty data are of primary importance to performing the efficient operation and management of an EMCS.

A conventional engineering method to find and correct the faults is to follow the procedures that check and recalibrate the sensors periodically (Pike and Pennycook 1992). This approach does not satisfy the requirements of EMCS, which require reliable measurements for continuous online automated schemes. Some other simple recovery methods used in

Received November 18, 2015; accepted March 7, 2016
Yangyang Fu is a Master's Degree Student. **Zhengwei Li, PhD,** is an Assistant Professor. **Fan Feng** is a Master Student. **Peng Xu, PhD,** is a Professor.
*Corresponding author e-mail: xupeng@tongji.edu.cn
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/uhvc.

pre-processing problematic data, such as replacing them with new ones that are randomly regenerated according to the mean and the variance of the adjacent data, are efficient when handling discretely obvious faulty data, such as a small scale of zero and negative numbers, but are not applicable to continuous and more subtle faulty data. Therefore, effective methods for detecting, locating, and reconstructing poor sensor data is highly desirable.

Although many studies have focused on the detection and diagnosis of sensor faults, few have focused on data recovery, let alone the combination of both. Maquin et al. (2000) proposed an approach to finding an optimal solution of uncertain models to recover data. However, the fact that the mathematical model of the reconciliation process requires all available knowledge to prevent erroneous decisions makes it complicated and difficult. Lee et al. (1997) presented a regression model to recover estimates of supply air temperature after a successful diagnosis of a temperature sensor fault. Those recovered data were then used in a feedback control loop to bring the supply air temperature back to the setpoint value. However, this recovery method is highly limited by the accuracy of regression models. Many other data-driven methods have also been applied to address this problem. The principal component analysis (PCA) method has been presented for fault detection and identification and the data recovery of flow meters and temperature sensors in typical buildings' central air-conditioning systems (Chen and Lan 2010; Hao et al. 2005; Wang and Chen 2004). Artificial neural network models have also been proposed to detect, diagnose, and recover the faults of outdoor airflow rate sensors and supply airflow rate sensors, which accomplished the fault-tolerant control of outdoor airflow (Wang and Chen 2002). Yang et al. (2014) proposed the fractal correlation dimension (FCD) algorithm to detect sensor faults and the support vector regression (SVR) process model to provide references of sensor measurements.

To solve all EMCS data problems with one effort seems impossible. Flow sensors and temperature sensors are prone to mistake and errors. This study takes the first initial step in addressing the more important and less sophisticated data-quality problem for electric power meters. Although some data-quality methods in this study are designed specifically for power metering, the methodology and framework should also work with other types of EMCS data. To understand the common problems related to electricity meters currently being used, electricity sub-metering data collected from more than 100 buildings has been investigated in this study. It was found that installation faults and environmental disturbance are the two major causes of the significant deviation of the records of meters from the true value. Although explicit data-quality faults, such as zero values or extremely high or low values, are easily detected, some implicit data-quality faults with a small but erroneous deviation from true values are difficult to detect and recover.

To detect and recover faulty data, in either a continuous or discontinuous mode and in both an explicit and implicit mode, a methodology based on correlations and a wavelet-analysis-based support vector machine (WASVM) predictor is developed in this article. The content of this article is organized as follows: First, the common data-quality faults are investigated and categorized, followed by a complete data-quality detection and recovery methodology in section 3. Then, a case study on building sub-metering data is used to illustrate the use of this methodology, and the discussion of this method is provided in section 5. Finally, conclusions are given at the end of the article.

Sub-metering platform and its common faults

A typical electric sub-metering system

Assuming there is a building where $N(N1 + N2)$ electric appliances or aggregated electric appliances need to be metered, the diagram of a typical electric sub-metering system is shown in Figure 1. As shown, it is a system with multiple sensors: In total, there are N electric sub-meters in the level 2 subsystem and two others (it can be more than two) in the level 1 system. Each meter measures the current I that goes through the appliance and sends it to the data-collection devices, from which the data are relayed to the central data-management platform. At the central data-management platform, current I is further processed to derive power P for each appliance using the following formula:

$$P = UI \tag{1}$$

where U is a parameter previously stored in the data-management platform, the value of which depends on the type of electric meter (three phase three wire, three phase four wire, or single phase).

Note that the utility meter provided by the utility company can also provide electric energy consumption, although in a much lower frequency. Denoting the total electricity energy consumption recorded by the utility meter as R_u, the electricity energy consumption by the ith level 2 electric sub-meter as R_i, and the total electricity energy recorded by the *level 1* electric sub-meter as R_s, then R_i, R_s, and R_u follow this relationship:

$$R_u = R_s = \sum_{i}^{N} R_i \tag{2}$$

Common data-quality problems related with electricity sub-metering system

To understand data-quality problems in sub-metering systems, more than 100 buildings with electricity sub-metering systems in Shanghai have been investigated. From their data-management platforms and maintenance history records of electricity meters, the authors have listed common data-quality problems in Table 1, grouping them into four types according to their consequences.

The first type is not a number (NAN), or zero fault, which includes transmission system failure (fault *a*), meter failure (fault *b*), and data-collection system failure (fault *c*). Faults in this category are primarily hardware failures, which generally lead to the problems of "data missing" or "zero value." The second type is the negative fault, where negative data are recorded, caused by an inverse current flow due to improper

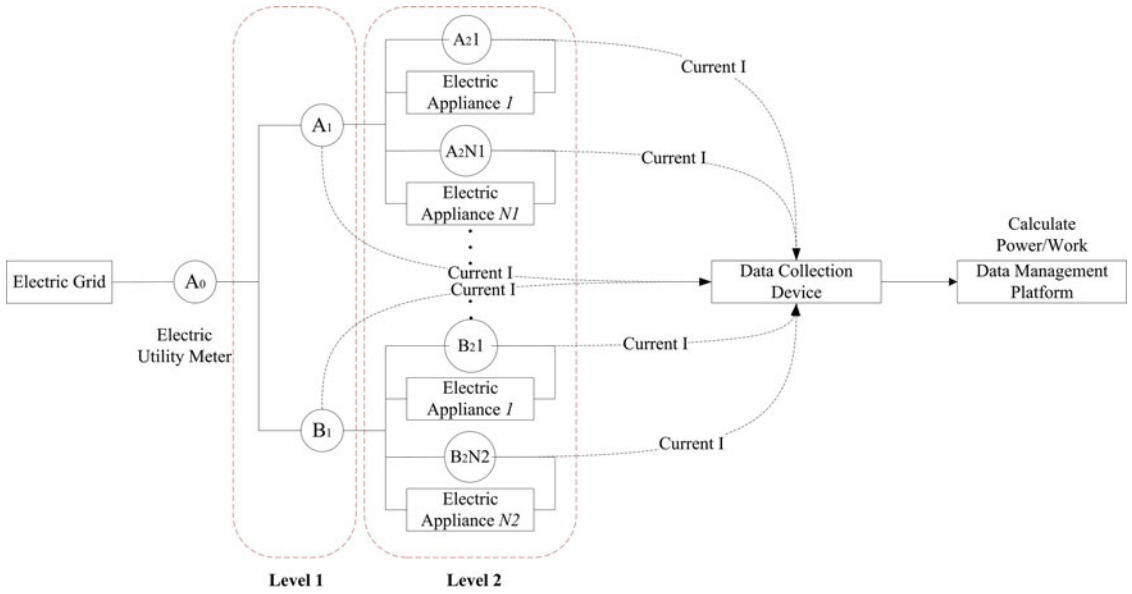

Fig. 1. A typical electric sub-metering system.

wiring. The third one is the proportional deviation fault, resulting in power-consumption readings that are proportionally smaller than the true value. This fault is commonly caused by the losing phase in the current or voltage and wrong post-processing parameters (faults *e*, *f*, and *g*). The fourth type is the disturbance fault (fault *h*), which is primarily caused by environmental disturbance around working meters. This fault has nothing to do with the sub-metering system itself, and the consequence is not as serious as the consequences of the other faults. It should be noted that the noise and calibration bias of the electric meter do not cause as much trouble as the faults mentioned above, so they are not included in the fault list.

A completer data-quality detection and recovery methodology

Data fault-detection and data-recovery procedure

Because other faults are easy to detect, such as missing data and zero or negative electricity consumption, only two types of faults are discussed here: Type 1 faults are proportional deviation faults, which cause collected data to deviate from their true values proportionally for a long time unless maintainers correct them (faults *e*, *f*, *g* in Table 1); type 2 faults are disturbance faults. Data with these types of faults change drastically within a very short time and then return to normal (fault *h* in Table 1).

A top-down fault-detection and data-recovery procedure is proposed for these two faults (Figure 2). The data required in this method include electric energy consumption recorded by electric meters of utility company W_u, level 1 electric meter W_s, and level 2 electric meters for all electric appliances W_i. Because the meter from the utility company is more accurate than the electric sub-meters, W_u is used as calibration benchmark at the very beginning. In the first step, three monthly electric energy-consumption data (R_u, R_s, and sum of R_i) are compared in a top-down approach; that is, the utility meter data was compared (as calibration) and level 1 sub-meter data (as calibrated) first, and then, level 1 sub-meter data (as calibration) and level 2 sub-meter data (as calibrated) were compared. If any deviations between the calibration meter and the calibrated meter are larger than a predefined threshold (Section 3.2), the authors continue to detect and recover the faulty calibrated meters (Section 3.3).

Otherwise, the second step is used. Because environmental disturbances and deviations last for a short time, high-resolution electricity-consumption data are required to detect

Table 1. Common data-quality problems with the electric sub-metering system.

Index	Fault type	Causes of the fault	Consequences of the fault
a	NAN or zero fault	Transmission system failure	No data transmitted to the central platform
b		Electric meter failure	No data recorded or zero
c		Data collection system failure	No data recorded or zero
d	Negative fault	Inverse current flow	Negative electricity consumption
e	Proportional deviation fault	Missing one phase in the current	Deviate from true value proportionally
f		Missing one phase in the voltage	Deviate from true value proportionally
g		Wrong post-processing parameter	Deviate from true value proportionally
h	Disturbance fault	Environmental disturbance	Deviate significantly from true value in a short time

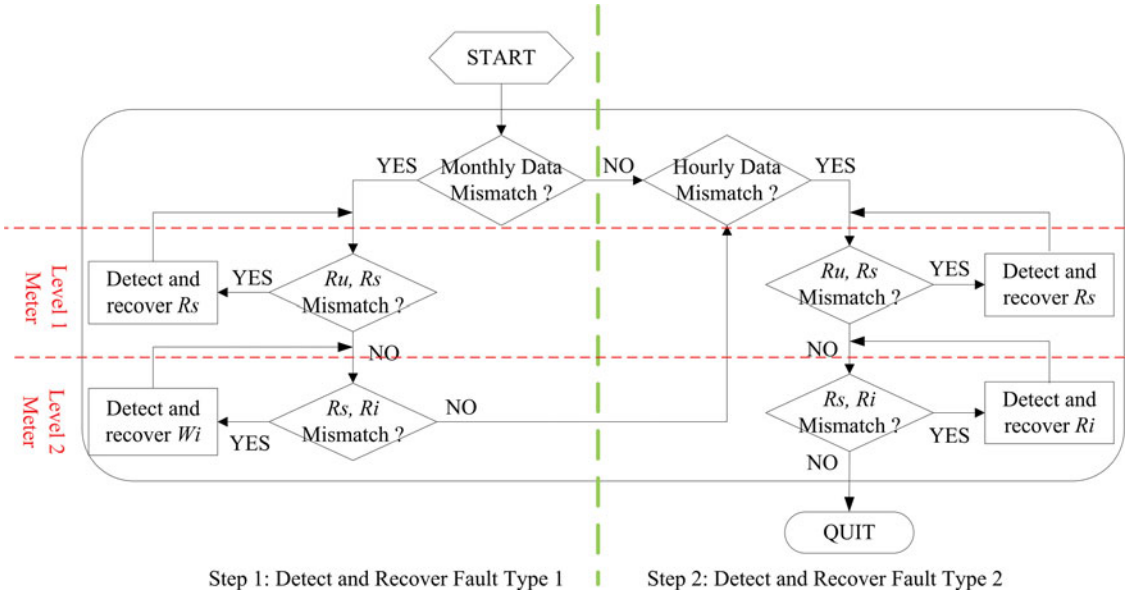

Fig. 2. A two-step top-down fault-detection and recovery procedure.

such faults. In this method, hourly data are used. Similarly, three hourly data (R_u, R_s, and sum of R_i) are compared in a top-down approach. If any deviations larger than a predefined threshold are found, then the authors go to type 2 faults for further detection and recovery (Section 3.4); otherwise, the detection processes were ended because no mismatch exists.

Determine the mismatch threshold

Denoting the calibration meter reading as R_a and the calibrated meter as R_c, the following index is used to evaluate the mismatch extent:

$$\delta = \frac{|R_a - R_c|}{R_a} \tag{3}$$

Because the electric energy consumption over a long period is relatively reliable, a value of 5% is tentatively set as the initial mismatch threshold. In other words, if $\delta > 5\%$, the calibrated meter reading R_c will be regarded as abnormal and needs to be calibrated.

Method to identify and recover type 1 fault

Once a monthly data mismatch is found among readings from the utility meter, level 1 meters and level 2 meters, we start the following process (Figure 3) to detect and recover type 1 faults, proportional deviation fault.

Denoting the difference between the electricity consumption recorded by calibration sub-meter R_a and the sum of the branch-calibrated electricity sub-meters $\sum_{i=1}^{N} R_{c,i}$ as R_d, the value of Pearson correlation coefficient F_i is used as a fault indicator of the ith sub-meter due to the linear relationship between the reading of the calibrated sub-meter and the value of R_d (ideally, if only this fault exists, R_d will be equal to $k \bullet R_{c,i}$, where k is the proportion, i is the index of faulty sub-meters). Thus, this fault indicator F_i of the ith branch electricity sub-meter is calculated as follows:

$$F = \frac{Cov\,(R_d,\, R_{c,i})}{\sigma\,(R_d)\,\sigma\,(R_{c,i})} \tag{4}$$

where σ denotes the standard deviation of the variable.

Based on the value of F_i, the fault counter for the ith sub-meter, CO_i is calculated following the procedure shown in Figure 3. First, for each month, a fault indictor for the ith sub-meter (F_i) is calculated. If F_i is larger than a threshold $F_{i,th}$ (e.g., 0.85), which has been trained and identified using a large amount of reliable testing data of the same building, the proportional deviation fault is identified, and the corresponding fault counter (CO_i) is added by 1. Otherwise, the authors think some electric appliances have not been metered in the lower branch, which accounts for the difference detected

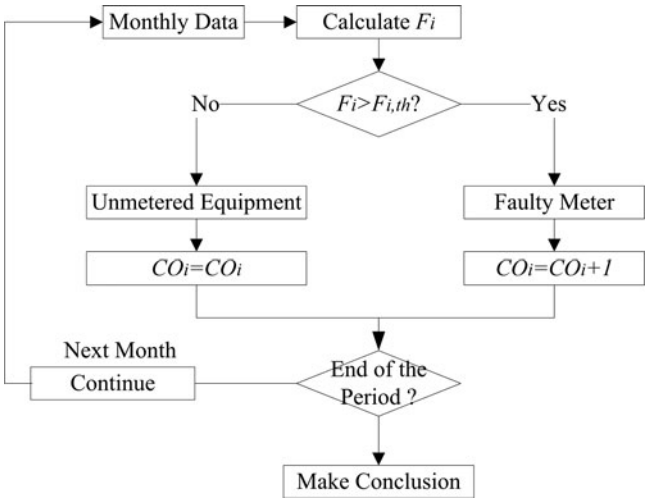

Fig. 3. Identification of proportional deviation fault.

in the very beginning. In this case, the authors propose the value of CO_i stay unchanged. When the iteration process finishes, if the value of CO_i exceeds fault counter threshold CO_{th} (e.g., 6 in the case of 1-year data), the ith sub-meter is declared as faulty.

Denoting the number of faulty branch electricity sub-meters as N_f, the reading of the kth faulty branch electricity sub-meter as R_k, and the difference between the sum of the calibrated sub-meters and the calibration meter as R_d, to find the true value of the kth faulty sub-meter $R_{t,k}$, the following algorithm is used:

$$r = \frac{R_k}{\sum_{k=1}^{N_f} R_k} \tag{5}$$

$$R_{t,k} = R_{f,k} + r_k R_d \tag{6}$$

Method to identify and recover the type 2 fault

Wavelet analysis

Wavelet analysis with the ability to reduce noise and analyze the local characteristics of the nonlinear and nonstationary signals is the localization analysis of the time (space) and frequency. Through the telescopic shifting operation, the signals will be refined multiple times to achieve time subdivision at a high frequency and frequency subdivision at a low frequency. In this article, the electricity-consumption series are treated as signals $R(t)$.

Discrete wavelet transform (DWT) is used as the pre-signal processor, which will obtain the approximation signal $a_i(t)$ and the detail signal $d_i(t)$ corresponding to resolution j ($1 \leq i \leq j$). Assuming that i gradually increased from 1 to j, the decomposition structure is shown in Figure 4. Therefore, the wavelet decomposition of signals $R(t)$ can be obtained with Equation 7.

$$R(t) = \sum_{i=1}^{j} d_i(t) + a_j(t) \tag{7}$$

Wavelet reconstruction can be achieved simply by summing up all detail signals and the last-level approximation signal based on Equation 7.

SVM

The SV algorithm is within the framework of statistical learning theory, which has been developed over the last four decades by Vapnik (Vapnik and Chervonenkis 1974). It is a nonlinear

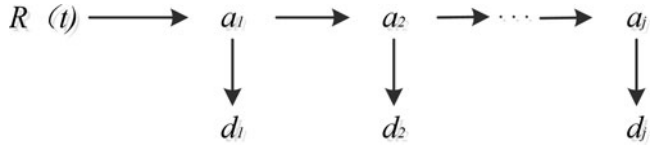

Fig. 4. Approximation signals and detail signals of signal $x(t)$ at resolution j.

generalization of the generalized portrait algorithm, based on the structural risk minimization (SRM) inductive principle. Due to the SV kernels, training SVM is equivalent to solving a linearly constrained quadratic programming problem; thus, a globally optimal solution can be found.

In this article, ε-SV regression is considered. Suppose training data was given $\{(x_1,y_1)\ldots(x_i,y_i)\subset \mathrm{X}\times\mathbb{R}\}$, where X denotes the space of the input patterns (e.g., $X = \mathbb{R}^d$), and l is the number of training samples. The goal is to find a function $f(x)$ that has at most ε deviations from the actually obtained targets y_i for all training data and is simultaneously as flat as possible. In other words, the authors do not care about errors as long as they are less than ε, but will not accept any deviation larger than this. SVM approximates the functions f, taking the following form:

$$f(x) = \langle w, \phi(x)\rangle + b \tag{8}$$

where $\langle,\rangle$represents the dot product in X, and $\varphi(x)$ represents the high-dimensional feature spaces that are nonlinearly mapped from input space x. Flatness in the case of Equation 8 means that one seeks a small w; one way to find a small w is to minimize the norm. Coefficients w and b are estimated by minimizing the following regularized risk function:

$$\frac{1}{2}\left\|w^2\right\| + C\sum_{i=1}^{l} \zeta_\varepsilon(y_i, f(x_i)) \tag{9}$$

where the first term $\frac{1}{2}\|w^2\|$ is called regularized term, and the second term is related to empirical risk function. The standard setting in the SV case is the ε-insensitive loss function, taking the following form:

$$\zeta_\varepsilon(y_i, f(x_i)) = \begin{cases} |y_i - f(x_i)| - \varepsilon, |y_i - f(x_i)| \geq \varepsilon \\ 0, otherwise \end{cases} \tag{10}$$

Figure 5 depicts the situation graphically. If the predicted value is within the ε tube (the shaded region), the loss is zero, whereas if the predicted point is outside the tube, the loss is penalized in a linear fashion. As the regularization constant, C determines the trade-off between the flatness of f and the amount up to which deviations larger than ε are tolerated.

So in standard SVM, Equation 9 can be formulated as Equation 11:

$$\min_{w,b} \frac{1}{2}\left\|w^2\right\| \quad s.t. \begin{cases} y_i - \langle w, \phi(x_i)\rangle - b \leq \varepsilon \\ \langle w, \phi(x_i)\rangle + b - y_i \leq \varepsilon \end{cases} \tag{11}$$

The tacit assumption in Equation 11 is that such a function f actually exists and approximates all pairs (x_i, y_i) with ε precision, or in other words, the convex optimization problem is *feasible*. However, this may not be the case, or there might be an allowance for some errors. Hence, slack variables ξ_i, ξ_i^* are introduced to cope with the otherwise infeasible constraints

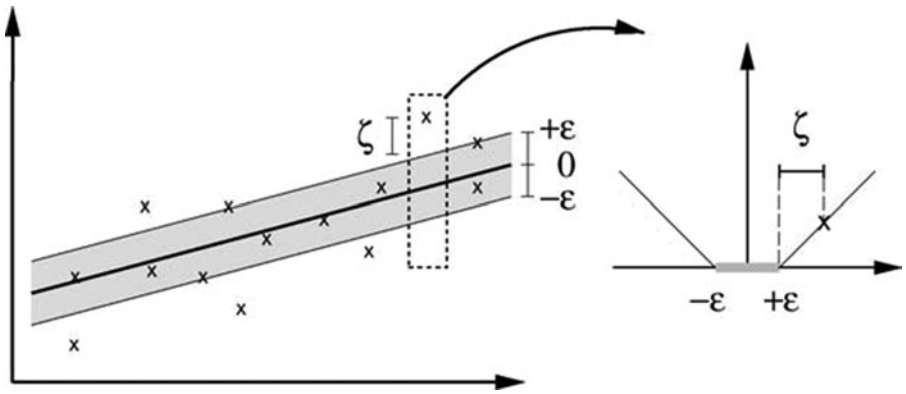

Fig. 5. The soft margin loss setting for a linear SVM (Smola and Schölkopf 2004).

of the optimization problem (Equation 9) and then a new optimization problem is formulated in Equation 12:

$$\min_{w,b,\xi_i,\xi_i^*} \frac{1}{2}\left\|w^2\right\| + C\sum_{i=1}^{l}\left(\xi_i + \xi_i^*\right) \quad s.t. \begin{cases} y_i - \langle w, \phi(x_i)\rangle - b \le \varepsilon + \xi_i \\ \langle w, \phi(x_i)\rangle + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \end{cases} \tag{12}$$

In most cases, the optimization problem (Equation 12) can be solved more easily in its dual formulation:

$$\begin{aligned} L = \frac{1}{2}\left\|w^2\right\| + C\sum_{i=1}^{l}\left(\xi_i + \xi_i^*\right) - \sum_{i=1}^{l}\left(\eta_i\xi_i + \eta_i^*\xi_i^*\right) \\ - \sum_{i=1}^{l} a_i\left(\varepsilon + \xi_i - y_i + \langle w, \phi(x_i)\rangle + b\right) \\ - \sum_{i=1}^{l} a_i^*\left(\varepsilon + \xi_i^* + y_i - \langle w, \phi(x_i)\rangle - b\right) \end{aligned} \tag{13}$$

Here, L is the Lagrangian, and η_i, η_i^*, α_i, α_i^* are Lagrange multipliers. Hence, the dual variables in Equation 13 have to satisfy positivity constraints, such as:

$$\eta_i, \eta_i^*, a_i, a_i^* \ge 0 \tag{14}$$

It follows from the saddle point condition that the partial derivatives of L with respect to the primal variables (w, b, ξ_i, ξ_i^*) have to vanish for optimality.

$$\begin{cases} \partial_b L = \sum_{i=1}^{l}(a_i^* - a_i) = 0 \\ \partial_w L = w - \sum_{i=1}^{l}(a_i - a_i^*)\phi(x_i) = 0 \\ \partial_{\xi_i} L = C - a_i - \eta_i = 0 \\ \partial_{\xi_i^*} L = C - a_i^* - \eta_i^* = 0 \end{cases} \tag{15}$$

Substituting Equation 15 into Equation 13 and eliminating η_i, η_i^*, one can yield the dual optimization problem:

$$\max_{a_i, a_i^*} \begin{cases} -\frac{1}{2}\sum_{i,j=1}^{l}\left(a_i - a_i^*\right)\left(a_j - a_j^*\right)\left(\phi(x_i), \phi\left(x_j\right)\right) \\ -\varepsilon\sum_{i=1}^{l}\left(a_i + a_i^*\right) + \sum_{i=1}^{l} y_i\left(a_i - a_i^*\right) \end{cases}$$
$$s.t. \sum_{i=1}^{l}\left(a_i - a_i^*\right) = 0, a_i, a_i^* \in [0, C] \tag{16}$$

By solving Equation 16, we can get Lagrange multipliers α_i, α_i^*. Replacing w with the results from equations in Equation 15 can obtain the regression function:

$$f(x) = \sum\nolimits_{i=1}^{l}\left(a_i - a_i^*\right)\phi(x_i)\cdot\phi(x) + b \tag{17}$$

By introducing kernel function $K(x,y)$, Equation 17 can be rewritten as follows:

$$.f(x) = \sum\nolimits_{i=1}^{l}\left(a_i - a_i^*\right)K(x_i, x) + b \tag{18}$$

Kernel function $K(x_i, x)$ is also called an SV kernel if it satisfies Mercer's condition (Campbell et al. 2006). Typical kernel functions include, for example, a linear function, a polynomial function, and the Gaussian function. Among these functions, the Gaussian function is well-suited for representing the complex nonlinear relationship between the input and output. Furthermore, using the Gaussian kernel functions, the computations can be directly performed in the input space, rather than in the feature space; thus, computational cost can be reduced. The Gaussian function is shown as follows:

$$K\left(x_i, x_j\right) = e^{-\gamma\cdot\left\|x_i - x_j\right\|^2} \tag{19}$$

where the γ is the kernel parameter.

When training SVM models, two free parameters need to be identified: kernel parameter γ and regularization constant C. Many articles have discussed the influence of SVM parameters on the SVM training models (Chang and Lin 2011; Dong

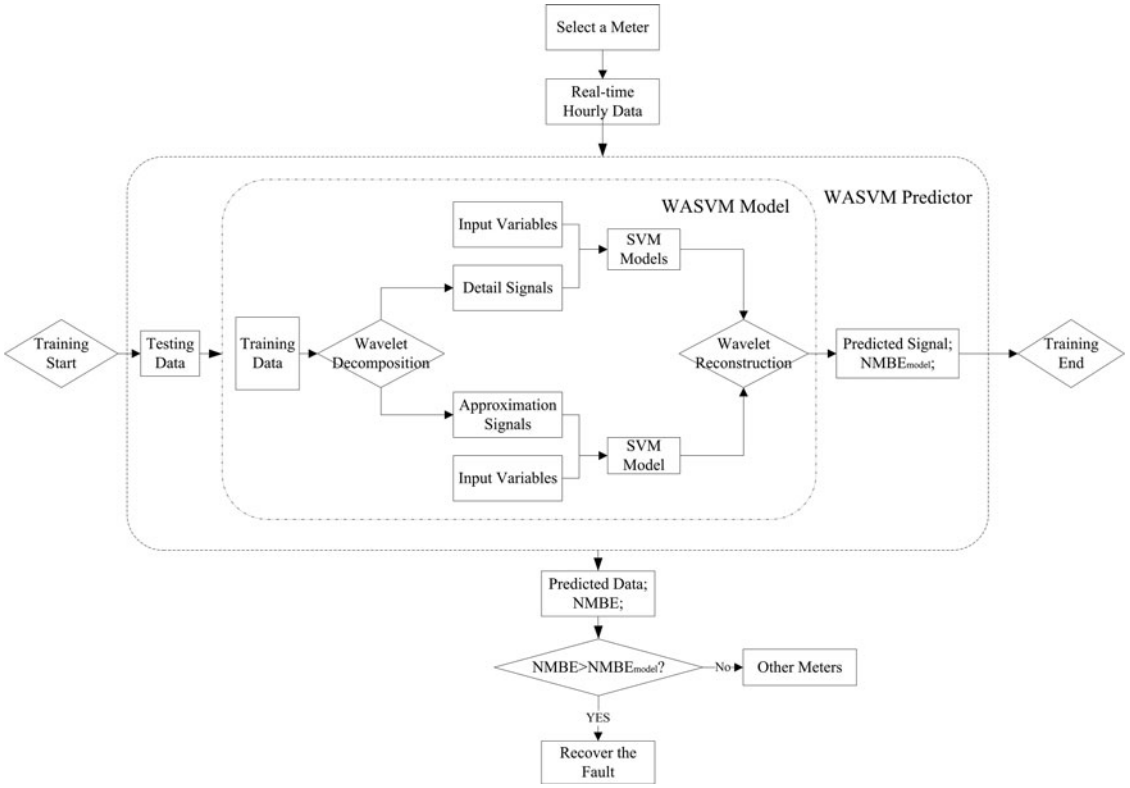

Fig. 6. WASVM predictor-based data problem detection and recovery for environmental disturbance.

et al. 2005). Theoretically, Parameter C determines the trade-off between the flatness of f and the amount up to which deviations larger than ε are tolerated. For example, a small value of C will under-fit the training data because the weight placed on the training data is too small thus resulting large training error. And when C is too large, SVM will over-fit the training data, which means $\frac{1}{2}\|w^2\|$ will lose its meaning and the objective goes back to minimize the empirical risk only. Here, to avoid mistakes resulting from user-supplied parameters, the authors use the genetic algorithm to minimize the training error by tuning these two parameters.

Detection and recovery for type 2 fault

A WASVM predictor is introduced to detect the data-quality problem caused by environmental disturbance, which is illustrated in Figure 6.

Before the start of scheme, the WASVM predictors for each branch are trained offline. The training process is described in Figure 6. First, because it was known exactly in which hours fault 2 happens via the "two-step top-down procedure" in Figure 2, the historical electricity consumption data before that hour are chosen as training and testing data. Then, the training data are decomposed into approximation signals and detail signals by DWT according to resolution j. These decomposed signals as separate target values, along with input variables X, are then used to train SVM models. The prediction results are summed up to reconstruct the final electricity consumption prediction value. Testing data are used to evaluate the performance after the WASVM models are trained.

To evaluate the prediction model's performance, two indexes are introduced: normalized root mean square error (CVRMSE) and normalized mean bias error (NMBE). Denoting the number of prediction points as l, the predicted electricity usage as $R_{p,i}$ $(0 \le i \le l)$, and the true electricity usage as $R_{t,i}$ $(0 \le i \le l)$, the values of CVRMSE and NMBE are calculated as follows:

$$CVRMSE = \frac{\sqrt{l\sum_{i=1}^{l}\left(R_{p,i} - R_{t,i}\right)^2}}{\sum_{i=1}^{l} R_{t,i}} \tag{20}$$

$$NMBE = \frac{\sum_{i=1}^{l}\left|R_{p,i} - R_{t,i}\right|}{\sum_{i=1}^{l} R_{t,i}} \tag{21}$$

Finally, the authors use the reconstruction value to detect the environmental disturbance fault, which lasts a short time. If the difference between the reconstruction value and the reading of the detected meter is larger than the WASVM predictor error $NMBE_{model}$, then this type of fault is detected and consequently recovered corresponding to that predicted value. Otherwise, the next meter data was checked with the same procedure.

Case study

In this case study, electricity submetering data collected from a large office building archive in Shanghai during 2013 are

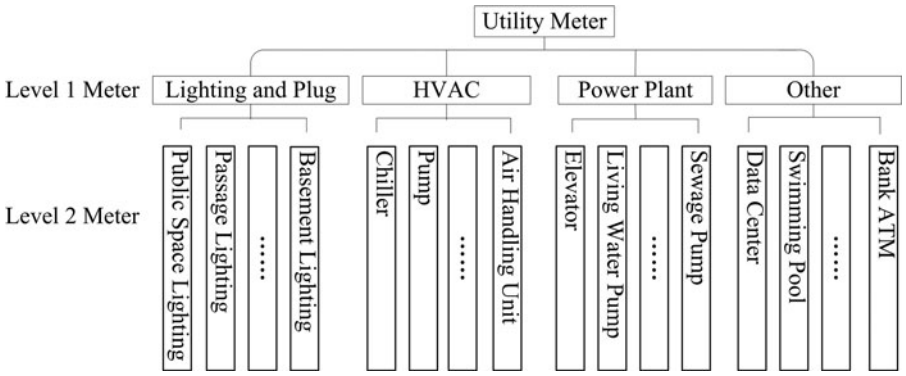

Fig. 7. A sub-metering system in the case-study building.

used. The electricity sub-metering system in this building is illustrated in Figure 7. This system includes 48 sub-meters in total: 4 level 1 meters and 44 level 2 meters. Major electricity consumers in this building are listed in Table 2; they include chillers (No. 1-No. 3), boilers (No. 1-No. 3), and a lighting system for, e.g., office buildings and storage space.

In the original data, electric utility bills, the readings of level 1 electricity meters and the sum of level 2 sub-meters match very well (both within 2%). To illustrate the methodology proposed in this paper, faults A, B, and C are introduced into the data, as shown in Table 3.

Fault A: Proportional deviation (multiple faulty meters calibrating similar equipment)

In this testing case, the readings of two branch sub-meters (W1 and W2) deviate proportionally (one-third smaller) from the true data. As a result, the mismatch extent δetween level 1 submeter and the sum of the branch submeters is 7.2%.

Table 2. Major electricity consumers in Shanghai archive.

Index	Type of system	Component name	Sub-meter name	Energy consumption percentage
1	Air-conditioning	No. 1 chiller	W1	11.2%
2	Air-conditioning	No. 2 chiller	W2	10.5%
3	Air-conditioning	No. 3 chiller	W3	9.4%
4	Air-conditioning	No. 1 boiler	W4	15.5%
5	Air-conditioning	No. 2 boiler	W5	10.8%
6	Air-conditioning	No. 3 boiler	W6	16.4%
7	Lighting	Office building lighting	W7	9.4%

Based on the detection procedure in Figure 3, the fault counters for W1 (CO_{W1}) and W2 (CO_{W2}) are, respectively, 11 and 10, whereas the fault counters for other sub-meters are all below 6. Thus, the proposed algorithm successfully detects the faulty sub-meters (Table 4). The faulty sub-meters are then recovered according to Equations 5 and 6. To quantify the performance of the recovery algorithm, the Euclidean norm of the sub-meter reading error (EN) before and after the data recovery is used. The results are shown in Table 5. It can be seen that the proposed algorithm is able to fully recover the faulty meter, with EN reduced to 0 for both W1 and W2.

Table 3. Data faults in the sub-metering system.

Index	Fault	Consequence	Faulty sub-meter name
A	Missing one phase in the current	Proportional deviation (deviate 33% from true data)	W1 and W2
B	Missing one phase in the current	Proportional deviation (deviate 33% from true data)	W1 and W4
C	Environmental disturbance	Short-term dramatic change (deviate 40∼60% from true data)	W1

Table 4. Faulty meter detection results.

Testing case	Results
Fault A	Faulty meters (W1 and W2) are successfully detected, no false detection
Fault B	Faulty meters (W1 and W4) are successfully detected, normal meters W5 and W6 are wrongly detected
Fault C	Faulty meter (W1) is successfully detected, no false detection

Table 5. Faulty meter recovery results.

Testing case	EN_{W1} (before recover)	EN_{W1} (after recover)	EN_{W2} (before recover)	EN_{W2} (after recover)	EN_{W4} (before recover)	EN_{W4} (after recover)
Fault A	482.2	0	413.9	0	/	/
Fault B	482.2	0	/	/	2521.0	545.9
Fault C	219.4	20.3	/	/	/	/

Fault B: Proportional deviation (multiple faulty meters calibrating different equipment)

In this testing case, the readings of two branch sub-meters (W1 and W4) deviate proportionally (one third smaller) from the true data. As a result, the mismatch extent δ between level 1 sub-meter and the sum of the branch submeters is 8.9%.

Applying the fault-detection procedure shown in Figure 3, the fault counters for W1 (CO_{W1}) and W4 (CO_{W4}) are 6 and 7, respectively. In addition to W4, the fault counters for W5 and W6 exceed 6. Based on the threshold setting ($CO_{th} = 6$), faulty meters W1 and W4 are successfully detected, but false alarms are issued for normal meters W5 and W6. The faulty sub-meters are then recovered according to Equations 5 and 6. It can be seen that although W1 is fully recovered, W4 is only partially recovered. The successful recovery of W1 should be attributed to different operation periods of chillers and boilers, which are used in summer and winter separately. Because W4, W5, and W6 have similar operation schedules, recovering W4 is interfered with by W5 and W6. As a result, a reading mismatch (R_d) between level 1 meters and branch meters is distributed among these three meters. However, this kind of false alarm influences little in a real building, because maintainers will check the faulty meters once they have received the detection signal. If they find that W5 and W6 have nothing wrong with their current or voltage phase, then the reading mismatch will not be distributed to the incorrectly detected meters.

Fault C: Environmental disturbance

In this testing case, the external environment causes a sudden change (60, 40, and 50%) in the reading of W1 from 12:00 pm to 2:00 pm on September 21, 2013. The consequence of this change for the level 1 meter reading is shown in Figure 8. Due to the large difference between the level 1 meter and the sum of the branch sub-meters from 12:00 pm to 2:00 pm (14.2, 12.1, and 12.6%), the fault is detected in the hourly data mismatch.

Then, the authors need to train the proposed WASVM model to predict each branch to locate and recover the faulty one. The training data are selected from sub-metering data from June 1 to September 17. The sub-metering data from September 18 to September 20 are selected as testing data. Training inputs in t moment include dry bulb temperature of current and previous s time period ($T_{db}[t], T_{db}[t-1], \ldots, T_{db}[t-s]$), radiation of current and previous time period ($GSR[t]$, $GSR(t-1], \ldots, GSR[t-s]$), dew point temperature($T_{dp}[t]$) and nonworkday/workday information encoding with 0 and 1. In this case, s equals 5, so there are, in total, 14 inputs in t moment.

To detect fault C, WASVM models are trained in the level 2 branch. Before the training, the authors need to decide the value of resolution j in wavelet decomposition. The larger the resolution j is, the more SVM models are trained. In branch W1, the WASVM predictors performance NMBE varies little as the resolution j increases from 4 to 5 (Figure 8). So along with considerations about time cost in training SVM models, the decision was made to choose $j = 4$ in this case study. The wavelet decomposition of branch W1 with resolution $j = 4$ is plotted as an example in Figure 9. The original signal S is decomposed into high-frequency detail signals ($d_1 - d_4$) and low-frequency approximation signals (a_1), which are then used to train the SVM models as output variables. Figure 10 shows the reconstructed testing value and true testing value of W1. The performances of some other WASVM models are listed in the Table 6.

After the predictor is established, the authors have the prediction error interval with the $\pm NMBE_{model}$. From Figure 11, it can be seen that the faulty readings located beyond the error interval can be easily detected and sequentially recovered with the predicted value.

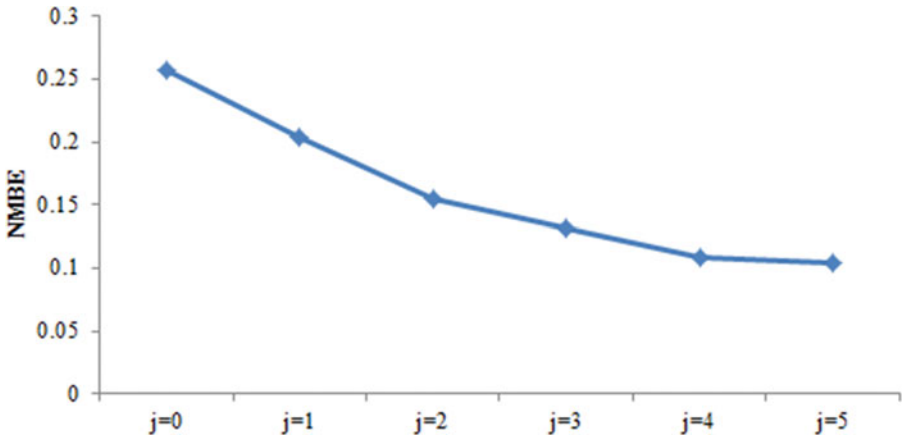

Fig. 8. WASVM predictors performance NMBE in branch W1 with different resolution j.M

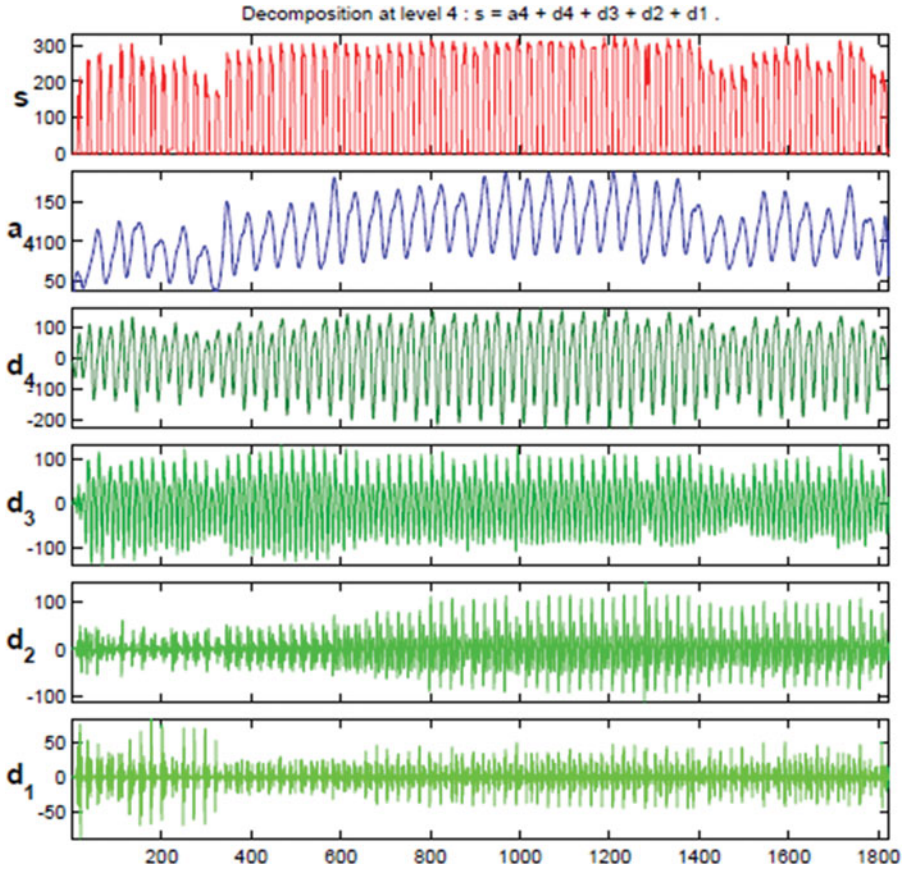

Fig. 9. Wavelet decomposition of branch W1.

Discussions and limitations

In previous testing, the number of faulty sub-meters is limited to two in the case of the proportional deviation fault and limited to one in the case of the environmental disturbance fault. To understand the limitation of the new method, testing is further conducted when the number of faulty sub-meters increases. The following two observations have been made:

For proportional deviation faults, when the number of faulty meters increases, the correlation between the mismatch quantity and the faulty sub-meter readings decreases. For example, when the proportional deviation fault occurs to three sub-meters (W1, W2, and W4) simultaneously, the fault counters for W1 (CO_{W1}), W2 (CO_{W2}), and W4 (CO_{W4}) drop to 5, 4, and 6, respectively. Considering the existence of other sub-meters with similar operation schedules, lowering the fault diagnostics threshold CO_{th} (six times a year) will increase the

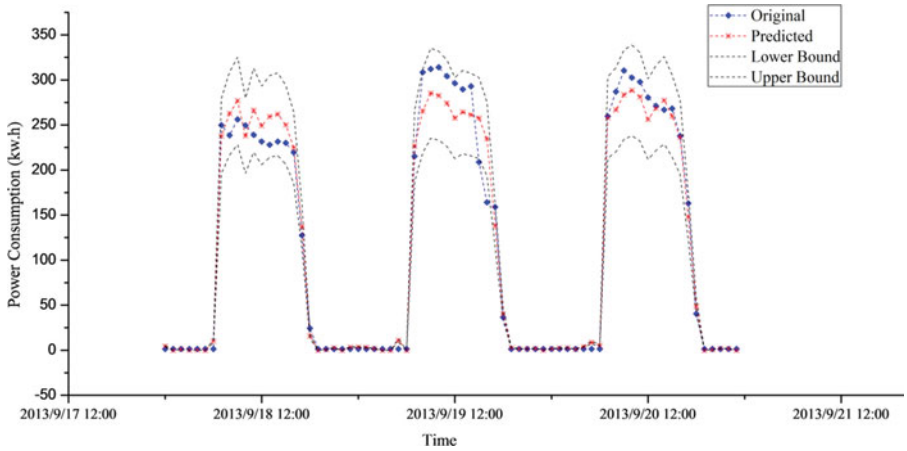

Fig. 10. Comparison of reconstructed signals and original signals.

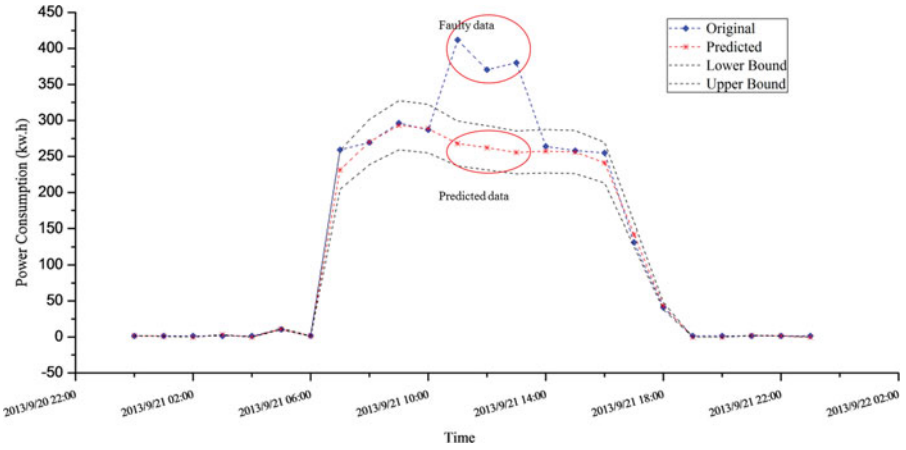

Fig. 11. Detected faulty data in fault C.

Table 6. WASVM predictor performance of some branches.

Branch	NMBE	CVRMSE
W1	0.11	0.20
W2	0.13	0.25
W3	0.13	0.27
W4	0.09	0.21
W5	0.13	0.30
W6	0.12	0.23
W7	0.05	0.09

false alarm rate. Thus, in this particular case, the performance of the proposed method is acceptable only when the number of faulty sub-meters is less than three.

For environmental disturbance faults, the following should be noted: This method depends heavily on the prediction accuracy of WASVM models, which primarily rely on the quality of training and testing data. Therefore, before the first WASVM models are trained, the data quality manually should be checked to get rid of obvious faults. Once those data are checked one time, there is no need to do so again in future because the dependable WASVM model can check the coming data. Furthermore, it is time-consuming to train these WASVM models when the size of the training data is large, so the appropriate size of the training data should be considered.

The accuracy of the method strongly depends on the thresholds. In the method, when judging the mismatch between the level 1 meter and the branch sub-meter, a difference of 5% is used as the threshold, and the results seem to be reasonable. When identifying the level 2 sub-meter with fault 1, a correlation coefficient of 0.85 is used as the threshold. However, this choice is dependent on the particular case. For example, if the building type is no longer the office building used in our case and has a different operation schedule (e.g., electric appliances in hotel buildings), 0.85 might be too high to exclude normal meters. Thus, user needs to retrain and readjust the thresholds when the situation changes.

Conclusions

The EMCS data-quality problem has been a long-standing issue in the building industry. This article proposes a systematic methodology to detect and diagnose data faults in the EMCS system and use electricity sub-metering as the first step of the trial. Data faults that are hard to detect by simple rules are grouped into type 1 and type 2 faults. They are proportional deviation faults and environmental disturbance faults. A correlation coefficient method is presented to detect and recover proportional deviation faults, and a WASVM predictor is introduced for environmental disturbance faults. Testing this methodology using data collected from an office building was successful. It was found that a high accuracy can be achieved given a small number of faulty sub-meters (less than three for type 1 faults) during the period. Considering the low possibility of simultaneous multiple faults, this method is practical in real situations.

Nomenclature

Roman symbols

a = approximation signals in the wavelet decomposition
b = coefficient of the hyperplane f
C = regularized constant in regularized risk function
CO = fault counter in proportional deviation fault
Cov = covariance
d = detailed signals in the wavelet decomposition
EN = euclidean norm of the sub-meter reading error
f = the hyperplane SVM model approximates
F = Pearson correlation coefficient between the reading of the calibrated sub-meter and the mismatch difference, used as fault indicator in Proportional Deviation Fault.
I = current in a submeter
K = kernel function
l = number of training samples
N = total number of specific submeters
P = power measured by a submeter

R = reading of an electric meter
U = voltage in a submeter
w = coefficient of the hyperplane f
W = the branch submeter
x = input space formed by SVM model independent variables
y = SVM model dependent variables

Greek symbols

ζ = risk function
ε = tolerance in ε-insensitive loss function
δ = mismatch extent between the calibration meter and the calibrated meter
ξ, ξ^* = slack variables
ϕ = the high-dimensional feature spaces that are nonlinearly mapped from input space X
$\eta_i, \eta_i^*, \alpha_i, \alpha_i^*$ = Lagrange multipliers, also dual variables in the dual optimization problem
γ = kernel parameter in the Gaussian kernel function
σ = standard deviation

Subscripts

a = calibration meter
c = calibrated meter
i,k = number index
j = resolution in the wavelet decomposition
s = the summed electricity consumption recorded by level 1 or level 2 electric sub-meters
th = threshold
t = true value
u = utility meter

Funding

The authors would like to thank Tengtian Energy Conservation Inc. for the financial and data support under project number 12dz1202000.

References

Campbell, W. M., D. E. Sturim, D. A. Reynolds, and A. Solomonoff. 2006. SVM based speaker verification using a GMM supervector kernel and NAP variability compensation. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on* (Vol. 1, pp. I-I), Toulouse, France, May 14–19. IEEE.

Chang, C.-C., and C.-J. Lin. 2011. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)* 2(3):27.

Chen, Y., and L. Lan. 2010. Fault detection, diagnosis and data recovery for a real building heating/cooling billing system. *Energy Conversion and Management* 51(5):1015–24.

Dodier, R. H., and J. F. Kreider. 1999. Detecting whole building energy problems. *ASHRAE Transactions* 105(1):579.

DOE. 2013. Federal and industry partners issue challenge to manufacturers. http://energy.gov/articles/federal-and-industry-partners-issue-challenge-manufacturers.

Dong, B., C. Cao, and S. E. Lee. 2005. Applying support vector machines to predict building energy consumption in tropical region. *Energy and Buildings* 37(5):545–53.

Effinger, J., M. Effinger, and H. Friedman. 2012. Overcoming barriers to whole building M&V in commercial buildings. In *ACEEE Conference, Pacific Grove, CA, August 12–17*.

Fontugne, R., J. Ortiz, N. Tremblay, P. Borgnat, P. Flandrin, K. Fukuda, and H. Esaki. 2013a. Strip, bind, and search: a method for identifying abnormal energy consumption in buildings. *Proceedings of the 12th international conference on Information processing in sensor networks ACM Philadelphia, PA, April 8–11*, pp. 129–40.

Fontugne, R., N. Tremblay, P. Borgnat, P. Flandrin, and H. Esaki. 2013b. Mining anomalous electricity consumption using ensemble empirical mode decomposition. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference, Vancouver, BC, Canada, May 26–31,* pp. 5238–42).

Guo, M., J. Xia, Q. Shen, and L. Yang. 2014. Comparative analysis on energy consumption of commercial buildings based on sub-metered data. *International High Performance Buildings Conference, West Lafayette, IN, July 14–17*.

Hao, X., G. Zhang, and Y. Chen. 2005. Fault-tolerant control and data recovery in HVAC monitoring system. *Energy and Buildings* 37(2):175–80.

IEA. 2010. *Key World Energy Statistics*. Paris, France: International Energy Agency.

Lee, W.-Y., J. M. House, C. Park, and G.E. Kelly. 1996. *Fault diagnosis of an air-handling unit using artificial neural networks*. *ASHARE Transactions* 102:540–9.

Lee, W. Y., J. M. House, and D. R. Shin. 1997. Fault diagnosis and temperature sensor recovery for an air-handling unit. *ASHARE Transactions* 103:621–33.

Leene, J. A. 1981. Natural ventilation of parking garages. *CSTB Seminar on Designing with the Wind, Nantes, France*, June 15–19.

Liu, D., Q. Chen, K. Mori, and Y. Kida. 2010. A method for detecting abnormal electricity energy consumption in buildings. *Journal of Computational Information Systems* 6(14):4887–95.

Maquin, D., O. Adrot, and J. Ragot. 2000. Data reconciliation with uncertain models. *ISA Transactions* 39(1):35–45.

O'Driscoll, E., and G. E. O'Donnell. 2013. Industrial power and energy metering—a state-of-the-art review. *Journal of Cleaner Production* 41:53–64.

Pike, P., and K. Pennycook. 1992. *Commissioning of BMS: Code of Practice*. Chicago, IL: Building Services Research and Information Association.

Smothers, F.J. 2002. Benefits of Enhanced Data Quality and Visualization in a Control System Retrofit. *Teaming for Efficiency: Information and Electronic Technologies: Promises and Pitfalls* 7:221.

Wang, S., and Y. Chen. 2004. Sensor validation and reconstruction for building central chilling systems based on principal component analysis. *Energy Conversion and Management* 45(5): 673–95.

Wang, S.W., and Y.M. Chen. 2002. Fault-tolerant control for outdoor ventilation air flow rate in buildings based on neural network. *Building and Environment* 37(7):691–704.

Yang, X.-B., X.-Q. Jin, Z.-M. Du, B. Fan, and Y.-H. Zhu. 2014. Optimum operating performance based online fault-tolerant control strategy for sensor faults in air conditioning systems. *Automation in Construction* 37:145–54.