This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2021.3062707, IEEE Transactions on Knowledge and Data Engineering

Gas-Theft Suspect Detection among Boiler Room Users: A Data-Driven Approach

Xiuwen Yi, Xiaodu Yang, Yanyong Huang, Songyu Ke, Junbo Zhang, Tianrui Li, Yu Zheng

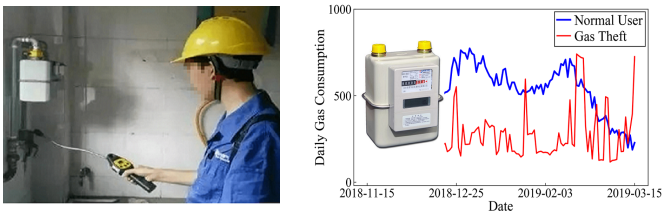
Abstract—The natural gas tightly correlates with our everyday life. However, driven by gray incomes, some users are prone to stealing gas by refitting the equipment without permission. Especially for the boiler room users in winter, this phenomenon appears more rampant. Traditional gas-theft detection methods highly rely on the on-site inspection, where exists ineffective and randomness. With the rapidly deployed IoT sensors, we can collect real-time gas consumption data to analyze users' behavior patterns, where the gas-theft suspects could be discovered early and accurately. In this paper, we propose a data-driven approach, named SVOC, to detect gas-theft suspects among boiler room users. Our approach consists of a scenario-based data quality detection algorithm, a deformation-based normality detection algorithm, and an One-Class Support Vector Machine (OCSVM) based anomaly detection algorithm. Specifically, considering the temporal proximity between the gas consumption and the outdoor temperature, the normality detection algorithm adopts a similarity-based deformation correlation to detect normal boiler room users out of abnormal ones. Then, we employ OCSVM as the anomaly detection algorithm to capture various features across multiple data sources, aiming to distinguish gas-theft suspects from the remaining irregular users. Here, the detected normal and abnormal users are fed into the OCSVM for training and prediction, respectively, which can overcome the label scarcity problem. We conduct extensive experiments on a real-world dataset during one heating season. The results demonstrate distinct advantages of our approach over various baselines. We have developed a real-time system on the cloud, providing daily gas-theft suspects for gas companies.

Index Terms—Gas-Theft Suspect; Normality Detection; Anomaly Detection; Urban Computing

1 Introduction

Natural gas is tightly relevant to millions of people's daily life [1]. However, driven by the gray income, i.e., to report less charged gas consumption than the used actual volume, some users are prone to stealing gas by refitting installations or pipelines without permission. It is illegal and dangerous, which is likely to harm the economic interests of gas companies and endanger public safety. Especially for the boiler room users for supplying heating for inhabitants in winter, the phenomenon of gas-theft appears more rampant [2]. Since in northern China cities, from November to March, many boiler rooms supply heating for inhabitants by consuming natural gas, where the cost of a boiler room is up to one million RMB over the entire heating period. Thus, gas companies need to fight against gas-theft behaviors timely and effectively for preventing illegal activities.

The traditional means of gas-theft detection are mainly carried out by internal inspections along with measurement and maintenance, which require the on-site examination by the staff of gas companies. As shown in Figure 1(a), a maintainer is manually checking the condition of meters and calculating the quantity difference between supply and marketing. To finish all these procedures, it costs considerable human resources while still exists the problem of inefficiency and delay caused by the random inspection without specific target suspects.



(a) Traditional method (b) Data-driven method

Fig. 1. Gas-theft suspects detection methods.

As shown in Figure 1(b), with the rapid development of IoT gas meter, we can collect the gas consumption data of boiler rooms by remote data transmission. Hence, it is likely to detect boiler room gas-theft suspects out of normal users using a data-driven approach, considering that different users' gas consumption patterns. Thus, the gas-theft suspects could be discovered early and accurately while reducing the cost of workforces. Nevertheless, to identify gas-theft suspects based on their daily gas consumption records, we encounter some challenges:

Firstly, the gas consumption of boiler rooms is diverse and complicated. As boiler rooms serve different end-users, they will present various gas consumption patterns. The boiler rooms of community residents will supply heating for 24 hours without interruption, while users of shopping malls only work during business hours. Besides, users of

- *Xiuwen Yi is with JD Intelligent Cities Research and Tsinghua University, China. E-mail: xiuwenyi@foxmail.com*
- *Xiaodu Yang and Tianrui Li are with School of Information Science and Technology, Southwest Jiaotong University, China. E-mail: xiaodu.yang@foxmail.com, trli@swjtu.edu.cn*
- *Yanyong Huang is with School of Statistics, Southwestern University of Finance and Economics, China. He is the corresponding author. E-mail: huangyy@swufe.edu.cn.*
- *Songyu Ke is with Shanghai Jiao Tong University. E-mail: songyuke@outlook.com.*
- *Junbo Zhang and Yu Zheng are with JD Intelligent Cities Research, China. E-mail: msjunbozhang,msyuzheng@outlook.com*

1041-4347 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: SOUTHWEST JIAOTONG UNIVERSITY. Downloaded on January 30,2022 at 13:55:15 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2021.3062707, IEEE Transactions on Knowledge and Data Engineering

office buildings are more complicated: some run during the weekend, yet others do not. Moreover, the gas consumption of a boiler room is not constant and may have some fluctuations. Sometimes, the boiler room may shut down for the equipment maintenance. Besides, it could be decreased deliberately by users to save the fee paid for heating. Due to such complex actual situations with diverse gas consumption patterns, a normal boiler room user's fluctuation is thus likely to be misjudged as anomalies.

Secondly, gas-thieves only account for a small fraction among all boiler room users, and the caught gas-thieves are scarce. Typically, a gas-stolen event usually appears with a low probability, and the gas company can only catch a few of them with the traditional on-site inspection method. There are merely 0.23% users who have been caught as gas-thieves during one heating season regarding the data we obtain. With labels in such limited quantity and imbalance problems, it requires an effective anomaly detection method. Furthermore, there is no specific definition of normal and irregular gas consumption patterns before. Consequently, it is hard to identify useful features for distinguishing gas-theft suspects from normal users, as well as to build a robust gas-theft suspects detection approach.

To address the challenges above, we propose a data-driven approach, named SVOC, to detect gas-theft suspects of boiler rooms. Our approach contains three components: 1) *scenario-based data quality detection*, which excludes zero-use & data-missing users and filters severe fluctuation and low usage users; 2) *deformation-based normality detection*, which detects the normal and abnormal boiler rooms users by calculating deformation similarity; 3) *OCSVM based anomaly detection*, which discovers gas-theft suspects among abnormal users considering various gas usage characteristics. Inspired by the domain knowledge of gas supply and usage, our method has more interpretations in the actual situation. The main contributions are as follows:

- To the best of our knowledge, this is the first data-driven approach to detect gas-theft suspects of boiler rooms. Thus, the detection will no longer depend on the human experience, but the rules learned from gas consumption data. Therefore, it can dramatically reduce the cost of workforces and increases the efficiency of gas companies.
- Considering the temporal proximity between gas consumption and outdoor temperature, we propose a deformation-based normality detection algorithm to detect normal and abnormal users, markedly decreasing the scope of suspects.
- Based on the separated normal and abnormal users, we propose an OCSVM based anomaly detection algorithm to capture multiple characteristic factors for identifying gas-theft suspects. It is seamlessly connected to the normality detection algorithm for overcoming the label scarcity problem.
- We conduct experiments on a real-world dataset over one heating season, where the results show the distinct advantages of our approach over baselines. Besides, we discuss the detected anomalies in realistic situations and state the reason why there are two anomaly labels undetected.
- We have developed a real-time system on the cloud, entitled GasShield, providing the daily user classification of boiler rooms, especially for the gas-theft suspects. Thus, the potential suspects could be discovered in the early stage with higher accuracy.

2 Overview

2.1 Problem Formulation

For a list of boiler rooms users R, given the boiler room attribute data $\{Att_R\}$, gas consumption data $\{Gas_R^t\}_{t=1}^T$, and the temperature data $\{Temp^t\}_{t=1}^T$, where T is the time length of day, we aim at detecting gas-theft suspects $R_{suspect}$ out of all boiler room users R.

2.2 Overview

Figure 2 shows the framework of our proposed data-driven approach SVOC for detecting gas-theft suspects of boiler rooms, considering the gas consumption data, boiler room attribute data, and outdoor temperature data. SVOC consists of three components: a scenario-based data quality detection algorithm, a deformation-based normality detection algorithm, and an OCSVM based anomaly detection algorithm. Specifically, in the scenario-based data quality detection algorithm, we analyze the gas consumption data to exclude data-deficient (data-missed and data-zero) users and data-abnormal (severe fluctuations and continuous low consumption) users. Among them, there may exist some potential gas-theft suspects. Considering the strong temporal proximity between daily gas consumption and daily outdoor temperature, in the deformation-based normality detection algorithm, we first analyze the consumption continuity and transform the data of gas consumption as well as temperature, then calculate the deformation correlation to detect out normal boiler rooms. In this way, all boiler rooms can be separated into normal ones and the abnormal ones. The abnormal boiler rooms can be further classified into the gas-theft suspects and users with irregular patterns. In the anomaly detection algorithm, we extract characteristic features across multiple data sources and then feed these extracted features into the OCSVM model. With the detected normal boiler rooms as positive samples to train the model, we can distinguish the gas-theft suspects from irregular users after feeding the detected abnormal users for prediction. Thus, normality and anomaly algorithms are seamlessly connected to overcome the label scarcity problem and achieve better suspects detection accuracy.

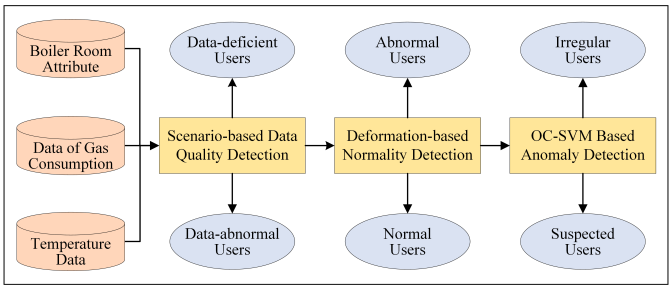


Fig. 2. The framework of our proposed approach SVOC.

1041-4347 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: SOUTHWEST JIAOTONG UNIVERSITY. Downloaded on January 30,2022 at 13:55:15 UTC from IEEE Xplore. Restrictions apply.

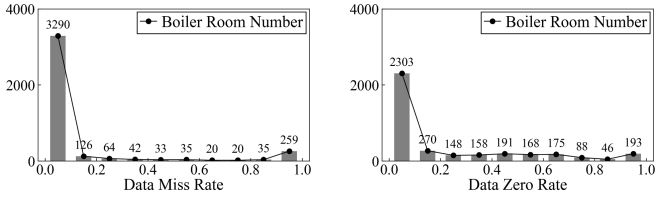
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2021.3062707, IEEE Transactions on Knowledge and Data Engineering

3 METHODOLOGY

In this section, we elaborate on our proposed approach SVOC: a scenario-based data quality detection algorithm, a deformation-based normality detection algorithm, and an OCSVM-based anomaly detection algorithm.

3.1 Scenario-based Data Quality Detection

The collected gas consumption data exist some data problems due to the realistic conditions, mainly consisting of 1) missing data; 2) zero consumption; 3) severe fluctuations; 4) continuous low consumption. Besides, some gas-thefts are hidden among these users with the data-quality problem, reporting the error readings. With such bad data quality, the detection algorithms can not help to detect gas-thefts. For achieving high-quality data, we detect and filter boiler rooms users having such one type of data quality issues.



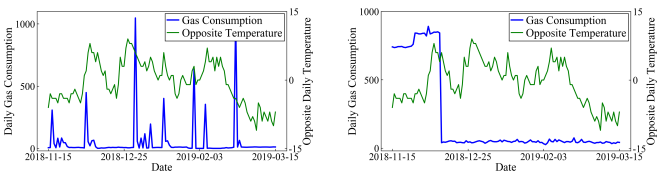
(a) Data miss rate distribution (b) Data zero rate distribution

Fig. 3. Statistics of gas consumption data quality.

First, we detect data-missing users. As shown in Figure 3(a), the overwhelming majority of boiler room users hold the miss rate of daily gas consumption smaller than 10%. While gas-theft users usually destroy gas equipment, which causes meter readings missed with higher frequency. So we exclude users whose data miss rate is higher than 10%.

Second, we detect zero-consumption users. As shown in Figure 3(b), the distribution of zero rate appears two obvious plunges, the first one after 10% and the second one after 70%. The high proportion of zero readings indicate that either longtime continuous or frequent irregular shutdown has occurred. It conflicts with normal operation patterns of boiler rooms and is highly suspicious of stealing gas. So we exclude users whose data zero rate is higher than 70%.

Third, we detect users whose gas consumption fluctuates severely. For the boiler room shown in Figure 4(a), spikes that exceed its usual gas consumption level appear in records, where extreme values are normally caused by meter failure. In this condition, records fluctuate regardless of actual gas consumption, which disturbs the analysis of overall patterns. So we exclude users whose maximum daily gas consumption is ten times greater than the median.



(a) Severe fluctuations (b) Continuous low consumption

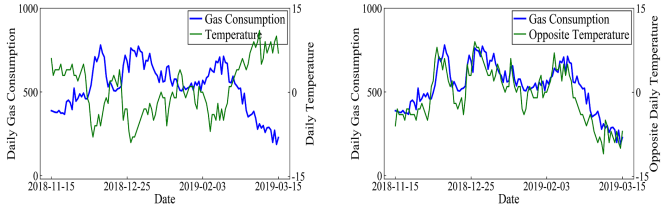
Fig. 4. Illustration of users having data quality problems.

Fourth, we detect users whose gas consumption is continuously low. For the boiler room shown in Figure 4(b), after some time point, its gas consumption remains comparatively lower than its former level. It indicates that this boiler room either operates at a low-temperature level or has fraudulent behaviors to report less gas consumption. Both situations are evidently abnormal and easy to identify in this component. So we exclude users whose daily gas consumption is lower than half of the maximum for more than 7 days (one week).

Here, all thresholds are defined based on the data distribution of realistic conditions. Thus, data-deficient (data-missed and data-zero) boiler room users and data-abnormal (severe fluctuations and continuous low consumption) boiler room users can be quickly detected and removed, providing higher-quality data for the following detection algorithms. Besides, we warn these filtered boiler rooms as one type of anomaly.

3.2 Deformation-based Normality Detection

As shown in Figure 5(a) and Figure 5(b), with the integrated analysis on gas consumption data and outdoor temperature data, we find that the daily gas consumption is strongly negatively related to the daily outdoor temperature. When it becomes colder, the gas consumption will increase in the upcoming days to offer the external heat supply, and vice versa [3]. Thus, it is important to take the opposite outdoor temperature as a reference. If the gas consumption curve fits the reference well, we can infer that the boiler room is normal. While for the remaining boiler rooms, we can judge them as abnormal users.



(a) Original (b) Opposite

Fig. 5. Daily gas consumption and daily temperature.

Considering the gas patterns of normal boiler rooms, we propose a deformation-based normality detection algorithm (TGSV) for detecting normal boiler rooms, as shown in Figure 6. It consists of three steps: continuity processing, data transformation, and normality detection. In detail, the continuity processing classifies boiler rooms into the weekday mode or the holiday mode and then processes the gas consumption of these two types of boiler rooms separately. Next, with the wavelet transformation, the phase calibration, and the min-max normalization, we transform the gas consumption data and temperature data into two cleaned time series by denoising and eliminating differences. Afterward, we calculate the deformation correlation with the defined temperature-gas shape variation to filter out normal boiler rooms, while the rest are abnormal. In this way, normal users can be excluded, which dramatically decrease suspects' scope and make the following anomaly detection algorithm more targeted.

1041-4347 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: SOUTHWEST JIAOTONG UNIVERSITY. Downloaded on January 30,2022 at 13:55:15 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2021.3062707, IEEE Transactions on Knowledge and Data Engineering

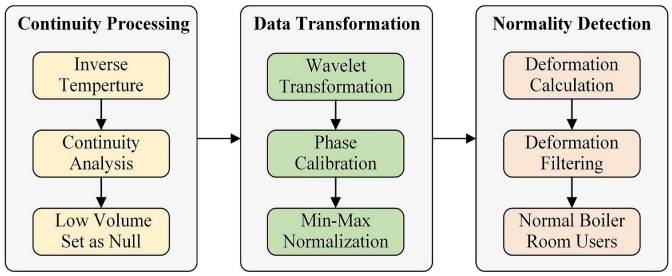
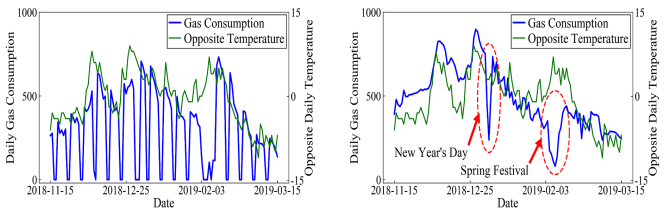


Fig. 6. Framework of normality detection algorithm.

3.2.1 Continuity Processing

Typically, most boiler rooms present a similar pattern regarding the opposite of outdoor temperature. However, as illustrated in Figure 7(a), it has regular sharp decreases on weekends compared with that on weekdays. The reason is that some boiler rooms only supply heating during weekdays. Moreover, as shown in Figure 7(b), during the Chinese official holidays, especially New Year's Day and the Spring Festival, the gas consumption drops significantly. That is also reasonable, as some enterprises will close during holidays. Without such knowledge, these fluctuations will be inclined detected as anomalies.



(a) On weekend (b) On holidays

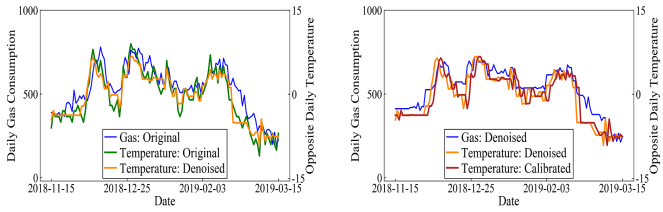
Fig. 7. Gas consumption decreased during troughs.

To avoid misjudging such reasonable fluctuations on weekends and holidays, we process the boiler rooms' gas consumption continuity. Firstly, we regard the opposite of daily outdoor temperature as a reference. Secondly, we calculate the ratio between average daily gas consumption on weekdays and that on weekends, separating users into the weekday-operating mode and everyday-operating mode. If the ratio is larger than a threshold (e.g., 1.3), we infer the boiler room is the weekday-operating mode. Otherwise, it is of the everyday-operating mode. After that, we set daily gas consumption on weekends as null for boiler rooms of weekday-operating mode. While during Chinese official holidays, we set the daily gas consumption of all boiler rooms as null. By this means, further analysis will be more accurate with misjudgments having been reduced.

3.2.2 Data Transformation

After continuity processing, some small short-term fluctuations still exist on the gas consumption data and outdoor temperature data. Besides, we notice some gas consumption delays compared with the temperature, as users usually need several days to adjust in response to the change of temperature. Moreover, the scale and the dimension of the two data sources are different. Hence, we transform the gas consumption data and the temperature data to denoise and eliminate these differences using three steps: wavelet transformation, phase calibration, and min-max normalization.

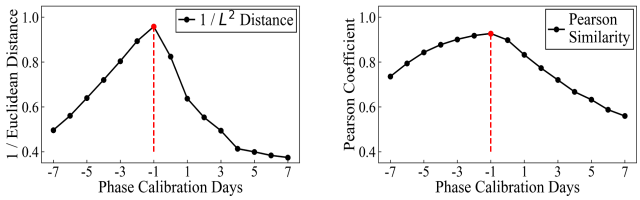
For denoising the small short-term fluctuations of gas consumption and outdoor temperature data, we choose a Mallat decomposition and reconstruction based wavelet transformation method [4], which is adaptive enough to represent localized signals in both the time and the frequency domain. Specifically, we conduct a multilevel 1-D discrete wavelet transformation for each time series firstly. The energy of dominating features will then be concentrated in a few large-magnitude wavelet coefficients, while noises will disperse on some small-magnitude coefficients. After that, we can remove noises while retaining useful information by thresholding coefficients. Here, we choose the Haar wavelet base [5] with a soft threshold function and set the decomposition level as 4 to approximate the optimal estimation. Finally, we obtain the denoised time series by reconstruction with the inverse wavelet transformation on the wavelet base and filtered coefficients. As shown in Figure 8(a), the processed time series of the opposite outdoor temperature is more smooth than the original one.



(a) Wavelet denoising (b) Phase correction

Fig. 8. Illustration on data transformation.

For rectifying the response delay, we implement the phase calibration on the opposite temperature rather than on the gas consumption for reducing computation complexity. As shown in Figure 9(a) and 9(b), the similarity between gas consumption and temperature varies with shifting the temperature forward and backward on different days. It can be seen that when the temperature is shifted forward one day, the similarity measured by both the Euclidean distance and the Pearson Correlation reaches the peak. Thus, we impose the phase calibration of one day forward on the reverse daily temperature. As shown in Figure 8(b), we obtain a rectified time series of temperature to offset the time delay of the gas consumption after phase calibration.



(a) Reciprocal of L^2 distance (b) Pearson correlation

Fig. 9. Similarity under different days shift.

Before measuring the similarity between gas consumption and temperature, it is worth noting that they are of different dimensions. Moreover, the scale of the gas consump-

1041-4347 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: SOUTHWEST JIAOTONG UNIVERSITY. Downloaded on January 30,2022 at 13:55:15 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2021.3062707, IEEE Transactions on Knowledge and Data Engineering

tion of each boiler room varies. Hence, it is supposed to normalize all the data into the same scale firstly, eliminating the dimension difference. We construct a pair of temperature series and gas consumption series for each boiler room, then perform the min-max normalization to scale them uniformly into $[0, 1]$. In this way, we obtain two time series with the same dimension and scale.

3.2.3 Normality Detection

After the data transformation, for each boiler room r, we get a pair of denoised and shifted time series of the daily gas consumption G_r and the opposite daily outdoor temperature T_r. If the curve of daily gas consumption fits the reference curve well, the boiler room can be inferred to be normal. For detecting such normal users, we define a temperature-gas shape variation $ShapeVar$ in the Equation 1. It measures the deformation correlation between the two time series. Its two components $\Phi(CORT)$ and $Diff$ are defined in Equation 2 and Equation 5, which characterize the trend consistency and the value deviation respectively.

$$ShapeVar_r^t = \Phi(CORT_r^t(G_r, T_r)) \times Diff_r^t(G_r, T_r) \quad (1)$$

$$\Phi(CORT_r^t(G_r, T_r)) = |1 - CORT_r^t(G_r, T_r)| \quad (2)$$

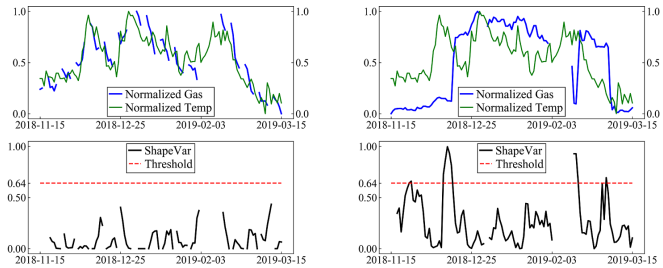
$$CORT_r^t(G_r, T_r) = \frac{\sum_{d=t-\omega}^{t} \Delta G_r^d \Delta T_r^d}{\sqrt{\sum_{d=t-\omega}^{t} (\Delta G_r^d)^2} \sqrt{\sum_{d=t-\omega}^{t} (\Delta T_r^d)^2}} \quad (3)$$

$$\Delta G_r^d = G_r^d - G_r^{d-1}; \Delta T_r^d = T_r^d - T_r^{d-1} \quad (4)$$

$$Diff_r^t(G_r, T_r) = \frac{\sum_{d=t-\omega}^{t} |G_r^d - T_r^d|}{\omega} \quad (5)$$

The component $\Phi(CORT)$ reflects how severely the gas consumption deviates from the reference, which can represent the trend consistency. Unlike the Pearson coefficient, $CORT$ represents the first-order temporal correlations, where the strength of monotonicity and the closeness of growth rates are both considered [6]. Moreover, the component $Diff$ portrays to what extent values of the gas consumption and the temperature diverge, namely their value deviation. Here, for one boiler room, to calculate its $ShapeVar_r^t$ at each time slot t, we consider recent influences inside pre-partitioned sliding windows, of which the size ω of both $\Phi(CORT)$ and $Diff$ is set as 3.

With the calculated $ShapeVar_r$, we can set a threshold to judge whether a boiler room is normal or not. Since the larger the $ShapeVar_r^t$ is, the more severely the gas consumption deviates from the normal reference level. As long as $ShapeVar_r$ on at least one timestamps surpasses the threshold, the boiler room is judged as an anomaly. Figure 10(a) shows a detected normal boiler room, of which the $ShapeVar_r$ is entirely below the threshold. Figure 10(b) shows an abnormal boiler room, which appears an obvious bias in comparison with the reference curve on surpassing several times. The threshold of $ShapeVar_r$ is set by considering both gas-theft labels and expert experience of the upper bound proportion of boiler room users who can be suspicious of stealing gas. The principle is that, with all gas-theft labels being detected, the least abnormal users will be reported. Based on it, normal boiler rooms can be excluded, which dramatically decreases the suspects' scope and makes the following detection more targeted.



(a) Normal user (b) abnormal user

Fig. 10. Results of normality detection.

3.3 OCSVM-based Anomaly Detection

With the deformation-based normality detection algorithm, we can exclude normal boiler rooms from abnormal ones. However, many normal users, consuming the natural gas irregularly at times, maybe misclassified as suspects. For distinguishing gas-theft suspects from irregular users, as shown in Figure 11, we propose an OCSVM based anomaly detection algorithm to capture multiple characteristic factors from different data sources. Specially, we employ the detected normal boiler rooms as the positive samples to train the OCSVM model. And then, we predict the detected abnormal users with the trained OCSVM to differentiate suspected users. Here, three categories of features are extracted to depict the characteristics of boiler rooms. They are boiler room attribute features, gas consumption features, and temperature-gas joint features. In this way, gas-theft suspects can be detected more accurately.

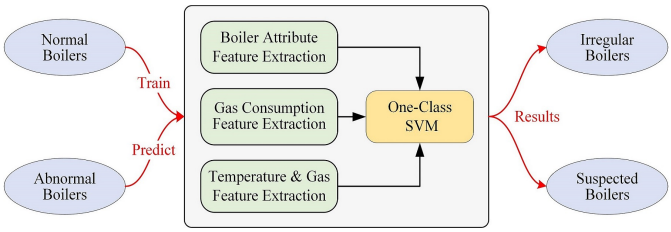


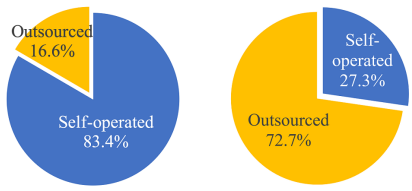
Fig. 11. Framework of anomaly detection algorithm.

3.3.1 Boiler Room Attribute Features

We extract six descriptive attribute features of boiler rooms listed in the first category of Table 1. In detail, the feature *Industry Types* describes types of heating entities, including offices, restaurants, business, and accommodation. The feature *Building Types* describes types of constructions, covering industry, public, civilian, and public-civilian shared. Besides, *Management Mode* indicates whether the boiler room is outsourced. As shown in Figure 12(a) and 12(b), the management mode appears a strong distinction between gas thefts and normal users. The majority of gas thefts are outsourced, while normal users are mainly self-operated. The reason behind it is that outsourcing operators are less restricted and more motivated to steal gas driven by gray profits. Moreover, a boiler room is usually characterized by the number of boilers, heating area, and heating level, which are denoted by features *Number of Boilers*, *Heating Area* and *Total Vapor Ton*, respectively.

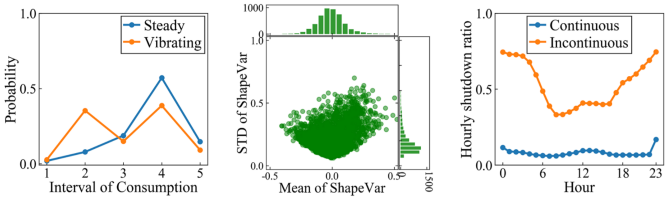
1041-4347 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: SOUTHWEST JIAOTONG UNIVERSITY. Downloaded on January 30,2022 at 13:55:15 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2021.3062707, IEEE Transactions on Knowledge and Data Engineering



(a) Normal user (b) Abnormal user

Fig. 12. Management Mode.



(a) Interval distribution (b) Shutdown distribution (c) Daily Continuity

Fig. 13. Gas Consumption Features.

3.3.2 Gas Consumption Features

We extract seven statistical gas consumption features listed in the second category of Table 1. The *Interval Distribution of* G_r feature presents the value distribution of daily gas consumption. For each boiler room, we map the daily consumption into 5 intervals: $(min, \mu_{neg} - \sigma_{neg})$, $(\mu_{neg} \pm \sigma_{neg})$, $(\mu_{neg} + \sigma_{neg}, \mu_{pos} - \sigma_{pos})$, $(\mu_{pos} \pm \sigma_{pos})$ and $(\mu_{pos} + \sigma_{pos}, max)$. Each dimension of the feature means the probability of corresponding interval. Thus, boiler room users can be clustered into two groups: the steady and the vibrating one, shown in Figure 13(a). For the steady group, their gas consumption mainly concentrates in the large-value interval while with minor probability in the small-value intervals. While for the vibrating group, they are turned down more often, so their consumption rises and falls.

Moreover, the *Hourly Shutdown Ratio* feature and *Daily Shutdown Ratio* feature describe the ratio of time slots when the gas consumption is turned down below a threshold. As illustrated in Figure 13(b), the overwhelming majority of boiler rooms have tiny shutdown ratios either in an hour or in a day. For each boiler room, the threshold is as 10% of its max daily gas consumption. Similarly, the feature *Daily Continuity* depicts the continuity of hourly gas consumption in the day. We calculate the proportion of days on each hour when the boiler room is shut down and then cluster all users into two groups: continuous and noncontinuous. As shown in Figure 13(c), the continuous group has few shutdowns in the day, while the noncontinuous one turns down their consumption more frequently, especially out of the working time. Besides, we extract the feature *Mean of* G_r and the features *Mean of* ΔG_r and *STD of* ΔG_r, where reveals the magnitude and sequential variation of gas consumption of each boiler room.

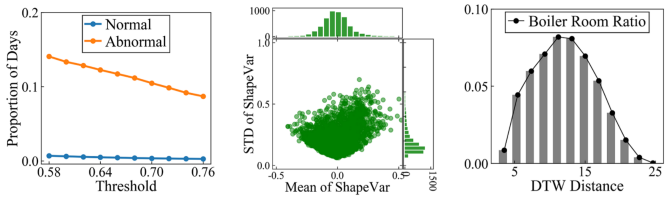
TABLE 1
Extracted features for OCSVM.

Category	Feature Description	Dimension
Boiler Room Attribute Features	Industry Types	4
	Building Types	4
	Management Mode	2
	Number of Boilers	1
	Heating Area	1
	Total Vapor Ton	1
Gas Consumption Features	Interval Distribution of G_r	5
	Daily Continuity	2
	Hourly Shutdown Ratio	1
	Daily Shutdown Ratio	1
	Mean of G_r	1
	Mean of ΔG_r	1
	STD of ΔG_r	1
Temperature - Gas Joint Features	Distribution of $ShapeVar$	5
	Mean of $ShapeVar$	1
	STD of $ShapeVar$	1
	DTW from G_r to T_r	1
	Mean of $\Delta G_r / \Delta T_r$	1
	STD of $\Delta G_r / \Delta T_r$	1

3.3.3 Temperature-Gas Joint Features

We extract six temperature-gas joint features listed in the third category of Table 1. The *Distribution of* $ShapeVar$ feature describes the ratio of abnormal days changing with the incremental threshold. As illustrated in Figure 14(a), for normal users, their $ShapeVar$ seldom exceeds thresholds. While for abnormal ones, their $ShapeVar$ deviates from the normal level more severely. The higher the threshold is, the less abnormal is detected. The co-distribution of features *Mean of* $ShapeVar$ and *STD of* $ShapeVar$ are displayed in Figure 14(b), which is scattered symmetrically. The more gas consumption deviates from the normal level, the larger *STD* and the absolute value of the mean of $ShapeVar$ are. Therefore, the *STD*, along with the larger positive *Mean*, is higher than that with the negative ones.

Besides, Dynamic Time Warping (DTW) distance measures the similarity between two time series, where the larger the DTW distance is, the less similar the two time series are. The feature *DTW from* G_r *to* T_r for the gas consumption G_r and the opposite outdoor temperature T_r, its distribution can be seen in Figure 14(c). Apart from the sole ΔG_r, we also calculate the normalized daily average temperature difference, denoted by ΔT_r. Considering the variation dependency between them, the features *Mean of* $\Delta G_r / \Delta T_r$ and *STD of* $\Delta G_r / \Delta T_r$ are extracted. Their co-distribution and that of ΔG_r are alike. It reveals that, for the majority of boiler rooms, the variation of gas consumption obeys consistent laws, which is tightly associated with that of temperature.



(a) ShapeVar distribution (b) Mean and STD (c) DTW distance

Fig. 14. Temperature-Gas Joint Features.

1041-4347 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: SOUTHWEST JIAOTONG UNIVERSITY. Downloaded on January 30,2022 at 13:55:15 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2021.3062707, IEEE Transactions on Knowledge and Data Engineering

3.3.4 One-Class Support Vector Machine

As we know, gas-thefts and normal users have different gas consumption patterns, which can be revealed in the feature space across different data sources. With the extracted features representing the multiple characteristics, we can use a machine learning method to reduce gas-theft suspects' scope further and avoid classifying normal users as gas-theft suspects. However, the percentage of confirmed gas-thefts is meager over the whole boiler rooms, resulting in the scarcity of labels and the data imbalance problem. Thus, it is difficult to design a supervised algorithm for classification, which may lead to significant false-positives. Luckily, we have detected normal and abnormal boiler rooms from the deformation-based normality detection algorithm, which can be viewed as the pseudo label for designing a self-supervised or one class classification model.

Here, we select the OCSVM model due to its modeling flexibility, computing efficiency, and detection accuracy. OCSVM is widely used for anomaly detection, where is trained on the data that has only one "normal" class while do not have the label information [7], and then predict which examples are unlike the normal examples, called anomalies. With the detected normal and abnormal boiler rooms, we can treat the normal boiler rooms as positive samples to train the OCSVM model. Then, we predict the abnormal boiler rooms to differentiate gas-theft suspects from normal users with irregular patterns. From this perspective, the normal and anomaly detection algorithm are seamlessly integrated, overcoming the label scarcity problem and achieving better accuracy.

More specifically, with the extracted features, OCSVM adopt the *rbf* kernel function $exp(-\gamma(\| \ x - x^{'} \ \|)^2)$ to learn a decision boundary. It first maps the original features into a high dimensional space corresponding to the kernel function, and then separate them from the original one using a decision boundary, which maximizes the distance from this boundary to the origin [8]. For a new sample of abnormal boiler rooms, if it falls on the same side of the decision boundary where most training data fall, it will be classified as a normal sample, otherwise as an anomaly. The optimization of OCSVM is to solve the quadratic programming problem, where tuning the parameters ν and γ.

3.4 Algorithm Psudo-code

Algorithm 1 outlines the proposed gas-theft suspects detection approach. For the deformation-based normality detection algorithm, we first use a wavelet transform and phase calibration to pre-process the opposite of temperature data (Lines 1-3). For each boiler room, we use continuity processing and wavelet transform on the gas consumption data and then calculate the deformation correlation after min-max normalization (Lines 4-8). After that, we detect the normal and abnormal boiler room users with the calculated deformation correlation (Lines 9). For OCSVM based anomaly detection algorithm, we first extract boiler room features, gas consumption features and temperature-gas joint features for users (Line 10-13). Then, we train the OCSVM model using the detected normal users with the extracted features and predict the gas-theft suspects with the detected abnormal users (Lines 14-15).

Algorithm 1: Gas-Theft Suspects Detection (SVOC)

Input: List of all boiler rooms R_{all}; Gas consumption data $\{Gas_R^t\}_{t=1}^T$; Temperature data $\{Temp^t\}_{t=1}^T$; Boiler room attribute data $\{Att_R\}$;
Output: List of gas-theft suspects $R_{suspect}$;

1 T^i = Set_opposite($\{Temp^t\}_{t=1}^T$) ;
2 T^w = Wavelet_Transform(T^i) ;
3 T^p = Phase_Calibration(T^w) ;
4 **for** each boiler room r in R_{all} **do**
5 G_r^c = Continuity_Processing($\{Gas_r^t\}_{t=1}^T$) ;
6 G_r^w = Wavelet_Transform(G_r^c) ;
7 G_r, T_r = Min-max_Normalization(G_r^w, T^p) ;
8 D_r = Deformation_Calculation(G_r, T_r) ;
9 R_{nor}, R_{abn}= Normality_Detection($\{D_r\}_{r=1}^R$) ;
10 **for** each boiler room r in R_{all} **do**
11 F_r^b = Boiler_Feature_Extraction(Att_r);
12 F_r^g = GasConsumption_Feature_Extraction(G_r) ;
13 F_r^t = Temperature_Feature_Extraction(T_r, G_r) ;
14 $Model$ = OCSVM_Train($\{F_r^b, F_r^g, F_r^t\}_{R_{nor}}$) ;
15 $R_{suspect}$ = OCSVM_Predict($Model, \{F_r^b, F_r^g, F_r^t\}_{R_{abn}}$) ;

4 Experiments

4.1 Settings

4.1.1 Datasets

We conducted experiments on a real-world dataset in Beijing, which detailed in Table 2. The dataset is collected by three different gas companies, with 3,035 boiler rooms and 11 labeled gas-thefts. Each boiler room has the daily gas consumption data and the static attribute information. We also use daily outdoor temperature data for reference. For both gas consumption data and temperature data, the time-span lasts from November 15, 2018, to March 15, 2019, where cover one whole heating season.

Though the caught gas thefts are limited, we still can not generate some synthetic gas-theft labels, as the thefts only can be judged by on-site inspections. Besides, it is not easy to set criteria of gas-thefts regarding the degree and anomalies pattern. Thus, this gas-theft suspect detection task has limited labels as ground truths. For evaluation, we adopt cross-validation on three subsets, where we detect gas-theft suspects for each subset by tuning the hyper-parameters on the other two subsets.

4.1.2 Parameter Setting

- Scenario-based data quality detection algorithm. The threshold of missing rate is set to 0.1. The threshold of zero rate is set to 0.7. The threshold of reasonable maximum daily gas consumption is set to ten times

TABLE 2
Details of the datasets

Boiler Room Attributes	Company	#. boiler rooms / #. thefts
	A	584/4
	B	781/2
	C	1670/5
	Total	3035/11
	Time Slot	Time Span
Gas Consumption Data	Daily	2018/11/15 - 2019/03/15
Temperature Data	Daily	2018/11/15 - 2019/03/15

1041-4347 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: SOUTHWEST JIAOTONG UNIVERSITY. Downloaded on January 30,2022 at 13:55:15 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2021.3062707, IEEE Transactions on Knowledge and Data Engineering

of its own median. The threshold of low consumption is set to half of its own maximum, and "continuously" means that it appears more than 7 days.
- Deformation-based normality detection algorithm. For continuity processing, the ratio between the weekday and weekend consumption is set to 1.3. For wavelet transformation, the level of decomposition is set to 4. For the phase calibration, we shift the opposite temperature time series 1 day forward. For normality detection, the size of sliding window is set to 3. The threshold of $TGShapeVar$ is set to 0.64 uniformly for all the three datasets, based on both gas-theft labels and expert knowledge on the proportion of suspicious users.
- OCSVM based anomaly detection algorithm. The parameters ν and γ for each subset are set by grid search. For Company A, $\nu = 0.4$ and $\gamma = 1e-4$. For Company B, $\nu = 0.3$ and $\gamma = 1e-7$. For Company C, $\nu = 0.1$ and $\gamma = 1e-5$.

4.1.3 Baselines

- **LOF** [9]: *LOF* detect outliers by computing the local density deviation and considering the samples with a substantially lower density as outliers. The number of neighbors is set to 20 in our experiment.
- **iForest** [10]: *Isolated Forest* is a tree-ensemble based method for identifying anomalies instead of normal observations. The number of base estimators in the ensemble is set to 1000 in our experiment.
- **DBSCAN** [11]: *DBSCAN* is a density-based clustering method, where points lie in low-density regions are regarded as outliers. Here, MinPts is set to 4 and eps is set to 0.75.
- **DONUT** [12]: *Donut* is an unsupervised anomaly detection algorithm based on VAE, targeting for seasonal KPIs (time series for monitoring machine services). Parameters are set as [12] suggests.
- **DAGMM** [13]: *DAGMM* combines the deep auto-encoder and the Gaussian mixture model for unsupervised anomaly detection. Here, the number of training epochs is set to 200, the size of mini-batches 256 and other parameters are set as [13] suggests.
- **SRCNN** [14]: *SRCNN* is a state-of-the-art method for time series anomaly detection by combining the SR and CNN. It adopts the Spectral Residual in the domain of computer vision to strengthen anomalies. Parameters are set as [14] suggests. Positive samples are gas-theft labels and negative samples are selected randomly from all other users.

Apart from the above baselines compared with our **SVOC** (TGSV&OCSVM), we also compare TGSV with popular similarity measurements and compare OCSVM with common models adopted in utility fraud detection.

4.1.4 Evaluation Metrics

We use precision (PR) and recall (RC) for evaluation. Due to the scarcity of labels, the detected anomalies should cover labels as many as possible and avoid false alarms. Therefore, with the same RC nearly equal to 1, the higher the PR is, the better the model performs. Also, we use #. detected suspects / #. hit thefts for demonstration.

4.2 Performance Comparison

4.2.1 Comparison with Baselines

As Table 3 illustrates, our approach achieves the best performance on all subsets compared with various baselines. LOF seldom hits labels across the three subsets; iForest may overfit on the subset B while has bad results on others; DBSCAN performs better than iForest but yet not well enough. The reason behind it is that outliers usually take a tiny proportion and show distinct patterns with normal samples, where suspicious users may gather into groups in the form of small clusters instead of scattering away from normal ones in the feature space. For Donut, DAGMM, and SRCNN, they perform not ideally enough in our scenario. The training data of Donut and DAGMM should better be clean normal samples, and that of SRCNN should be confirmed gas-theft labels together with clean normal samples. However, due to the label scarcity of the realistic condition, boiler room users' data is mixed by normal and unlabeled abnormal samples. Such dirty data would degrade the performance of these detection methods.

TABLE 3
Performance comparison with baselines

Method	Company A		Company B		Company C	
	PR	RC	PR	RC	PR	RC
LOF	0	0	0	0	0.006	0.2
iForest	0	0	0.026	1	0.012	0.4
DBSCAN	0.014	1	0.006	1	0.006	0.6
DONUT	0.015	1	0.003	0.5	0.003	0.4
DAGMM	0.010	1	0.009	1	0.003	0.6
SRCNN	0.030	1	0.004	1	0.006	0.6
SVOC	**0.069**	**1**	**0.016**	**1**	**0.007**	**0.6**

4.2.2 Comparison with TGSV Variants

As illustrated in Table 4, we compare $TGSV$ algorithm with its variants of different combinations of data transformation modules before calculating the $ShapeVar$. Here, CP stands for the continuity processing, WT for the wavelet transformation, and PC for the phase calibration. Results tell that the CP decreases misjudgments during holidays and weekends significantly; based on that, the WT and the PC diminish the false-positives caused by short-term vibrations and response delay to the temperature. Therefore, we select CP & WT & PC as the process of data transformation for calculating the deformation $ShapeVar$, which can further reduce 40.6% misjudgments compared to that without these data transformation modules. Overall, TGSV algorithm can exclude 61.9% of boiler rooms as normal users.

TABLE 4
Comparison with different data transformation

#. suspects / #. hit thefts	A	B	C	Overall
	584/4	781/2	1670/5	3035/11
-	396/4	503/2	1046/3	1945/9
CP	297/4	322/2	700/3	1319/9
CP & WT	252/4	308/2	640/3	1200/9
CP & PC	228/4	328/2	656/3	1212/9
CP & WT & PC (TGSV)	**218/4**	**310/2**	**628/3**	**1156/9**

1041-4347 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: SOUTHWEST JIAOTONG UNIVERSITY. Downloaded on January 30,2022 at 13:55:15 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2021.3062707, IEEE Transactions on Knowledge and Data Engineering

As shown in Table 5, we compare *ShapeVar* with three commonly-used similarity measurements, Pearson after same data processing and transformation. For these similarity measurements, none of them performs well. Pearson correlation and Euclidean distance do not consider the time series's internal temporal dependency, while DTW distance cannot reflect the information we want by the single value. Our TGSV algorithm defines a temperature-gas shape variation, named *ShapeVar*, characterizing both trend consistency and value deviation to model the deformation correlation.

TABLE 5
Comparison with different similarity measurements

Method	Company A		Company B		Company C	
	PR	RC	PR	RC	PR	RC
Pearson	0.01	1	0.004	1	0.003	0.6
Euclidean	0.009	1	0.003	1	0.003	0.6
DTW	0.01	1	0.004	1	0.003	0.6
ShapeVar	**0.018**	**1**	**0.006**	**1**	**0.005**	**0.6**

4.2.3 Comparison with OCSVM variants

As illustrated in Table 6, we compare OCSCM model with different combinations of three extracted features: gas consumption features (GC), temperature-gas joint features (TG), and boiler attribute feature (BA). The TG features can still improve detection efficiency as it contains more information than the TGSV algorithm. The GC features perform much better than TG, very close to our best results, which indicates the importance of this feature. While with only the GC features, several labels are missed. By combining the GC and TG features, we achieve a higher recall. Furthermore, when incorporating the BA with GC and TG, it will further improve the precision. Overall, our method can further detect 21% users as gas-theft suspects with high recall.

TABLE 6
Comparison with different feature combinations

#. suspects / #. hit thefts	A	B	C	Overall
	218/4	310/2	628/3	1156/9
TG	145/2	168/2	551/3	864/7
GC	55/3	127/2	401/1	583/6
TG & GC	76/4	127/2	556/3	759/9
GC & TG & BA	**57/4**	**127/2**	**457/3**	**641/9**

As for two-step methods presented in Table 7, the TGSV algorithm can first tell normal users apart from abnormal ones, then RF, GBDT, MLP, and VAE can be trained with normal samples and predict on abnormal ones. In this way, similarity-based algorithms and model-based algorithms can integrate seamlessly. Since this task can be regarded as a classification task, we compare OCSVM with several typical classifiers. RF, GBDT, and MLP will distinguish the normal and abnormal ones, while abnormal ones contain many users with irregular patterns. In comparison with these classifiers, OCSVM focuses on capturing normal patterns and thus generally outperforms them. We also compare OCSVM with VAE, since they both capture normal patterns. Although VAE also models normal patterns, it performs not well as it is confused with abnormal fluctuations and gas-theft behaviors.

TABLE 7
Comparison with substitutes for OCSVM

Method	Company A		Company B		Company C	
	PR	RC	PR	RC	PR	RC
TGSV & RF	0.014	0.75	0.006	1	0.005	0.6
TGSV & GBDT	0.017	0.75	0.006	1	0.005	0.6
TGSV & MLP	0.017	1	0.006	1	0.004	0.6
TGSV & VAE	0.020	0.5	0	0	0.007	0.4
TGSV & OCSVM	**0.069**	**1**	**0.016**	**1**	**0.007**	**0.6**

4.3 Case Study

4.3.1 Detectable Anomalies

The detected anomalies can be categorized into four types of cases as follows.

Against the common sense. It is common that the higher the temperature is, the less the natural gas consumed to supply heating, and vice versa. As Figure 15(a) illustrated, in the red circle, the monotonicity of gas consumption is following that of temperature. It means that the warmer it is, the more heating supplied, which is contradictory to the objective law.

Consumption lower than that under similar temperatures. For each boiler room with fixed facilities, the gas consumption under similar temperatures should be on the same level. As Figure 15(b) shows, within the two periods inside red circles, the trend of gas consumption is roughly consistent with that of temperature, whereas the volume of gas consumption is lower than that under a similar level of temperature before.

Varying seldom with the changing temperature. Since it exists a strong negative correlation between gas consumption and temperature. If the gas consumption of a boiler room keeps on the same level regardless of the changing temperature, it is impossible to satisfy the heating demand. As Figure 15(c) demonstrated, except for several transient increases, the gas consumption is approximately the same as that at the beginning of the heating period.

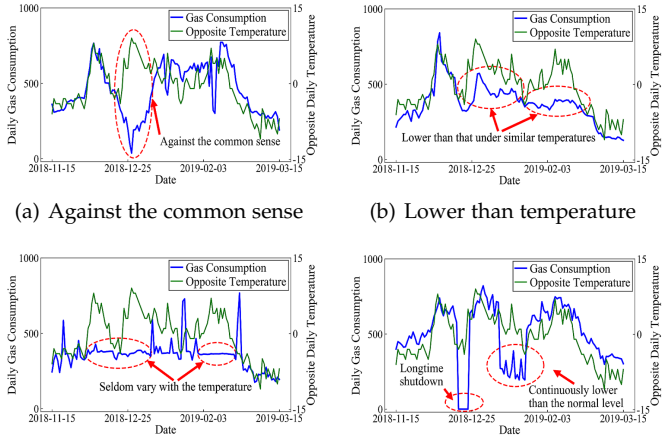
Being turned down or shut down continuously. It is tolerable that some boiler rooms are shut down occasionally due to something exceptional or the need for maintenance. However, it will turn into potential anomalies if this situation lasts for some time. As shown in Figure 15(d), around Christmas, it appears a shutdown lasting for several days. Then for about one week, the gas consumption keeps on a visibly lower level than the normal condition. Both periods do not belong to the Chinese official holidays.

4.3.2 Undetectable Anomalies

As mentioned in Section 4.2, there exist two labels that neither our method nor baselines can hit. Their gas consumption data fits the opposite temperature exactly, which is similar to normal users. We contact the gas company to investigate the reasons behind it. According to the feedback, the two boiler rooms steal gas by the same means. They connect pipes before meters without permission, which will not affect their gas consumption data. Naturally, our data-driven approach cannot help to detect such gas-theft behaviors. Thus, if a user only behaves abnormally in the magnitude of usage while behaves normally in cyclic behavior, the data-driven approach still can not help.

1041-4347 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: SOUTHWEST JIAOTONG UNIVERSITY. Downloaded on January 30,2022 at 13:55:15 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2021.3062707, IEEE Transactions on Knowledge and Data Engineering



(a) Against the common sense (b) Lower than temperature

(c) Seldom varying along time (d) Continuously turned down

Fig. 15. Categories of detectable anomalies

4.3.3 *Detected Suspicious Users*

As concluded in Table 6, our approach detect about 21% gas-theft suspects among all boiler rooms, which is reasonable considering the actual situation. Firstly, our dataset covers one whole heating season lasting for four months, where experiments are conducted in an offline scenario with limited gas-theft labels. Secondly, the data collected by sensors are incomplete due to the realistic physical condition, where much work can be done by the gas company staff to maintain the data quality. Last but not least, gas consumption patterns are diverse for all the boiler room users, for which we need to trade off between individuality and generality.

5 GasShield System

We have developed a real-time system on the cloud, entitled *GasShield*, providing the daily user classification of boiler rooms. Figure 16 illustrates the system interface, consisting of two panels: 1) *Overview* , which display the user statistics and gives the anomaly type distribution based on daily predicted results; 2) *Suspicious Users List*, which lists the detected suspicious users so that operators can conduct more targeted on-site inspections. The user can click the view button to visualize the curve of gas consumption data and temperature data and click the download button to download the suspects' list.

Here, we extend our proposed approach to an online system running every day, where we use the sliding window to extract data of the past 15 days for each detection. Considering newly inspected gas-theft labels, we re-trained the OCSVM model and adjusted the threshold for TGSV model every month. After transferring to the online scenario, the proportion of daily suspects decreases to ∼1.5%.

After deploying the system in Beijing Gas Group Co., Ltd., the gas company's staff conduct an on-site inspection based on the detection results during the 2019-2020 heating season. After inspecting 52 suspicious boiler rooms, they found that all users have some problems. As a result, the 44% users belong to the data-abnormal users, 48% users are irregular users, and 8% users are confirmed gas-thefts. With this system, the potential suspects could be discovered in the early stage with higher accuracy.

Fig. 16. User interface of *GasShield* system.

6 Related Work

6.1 Gas-Theft Suspect Detection

Fighting against gas stolen behaviors is a vital task for gas companies. The way to find out gas stolen behavior requires on-site field inspection, where the staff of gas companies mainly carries out the action during meter charging, measurement operation, and maintenance inspections. Hence, it costs lots of human resources and exists randomness and hysteresis. Different from that, we propose a data-driven approach to detect gas-theft suspects of boiler rooms. Thus, the potential gas theft suspects could be known early-stage with more accuracy while significantly reducing the workforce cost and increasing efficiency. Beside boiler room users, we also proposed another data-driven approach to detect gas-theft suspects among restaurant users [15]. To the best of our knowledge, our proposed methods are the first data-driven approaches instead of manual inspections.

6.2 Utility Fraud Detection

Utility fraud is a common issue for the energy industry (e.g., electricity, water). Many detection techniques have been proposed, which can be divided into hardware solutions and non-hardware ones [16]. Hardware solutions focus on preventing users' fraudulent behaviors by protecting meters [17], [18], while they are rather costly in equipment. Non-hardware solutions are mostly data-driven with energy consumption data, where classification-based techniques [19], [20], [21] and clustering-based methods [22], [23] are commonly adopted. However, it is hard to collect fully labeled data in real-life datasets for classification-based methods. And it is hard to tell whether minor clusters or outliers out of formed clusters are indeed fraudulent for clustering-based methods. Unlike these methods, we first calculate the temperature-gas shape variation to detect normal boiler rooms. Then, we adopt OCSVM to capture multiple characteristic factors for detecting gas-theft suspects.

6.3 Anomaly Detection Methodology

Similarity-based approaches and model-based approaches are widely used in the field of time series anomaly detection [24]. For similarity-based approaches, it mainly chooses a proper similarity measurement and identifies anomalies based on the similarity or dissimilarities between data samples, e.g., as measured by Euclidean distances or Pearson

1041-4347 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: SOUTHWEST JIAOTONG UNIVERSITY. Downloaded on January 30,2022 at 13:55:15 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2021.3062707, IEEE Transactions on Knowledge and Data Engineering

correlation [25]. Model-based approaches [26] mainly depend on the features extracted from original data to learn a hyper-plane for splitting the anomalies (e.g., OCSVM), minimize the reconstruction error of normal samples (e.g., VAE [27], Donut [12], DAGMM [13]) or treat it as a classification problem (e.g., SRCNN [14]). Different from that, we combine the similarity-based and model-based algorithms, which are seamlessly integrated for overcoming the label sparsity problem and achieve better accuracy.

6.4 Urban Anomaly Detection Application

Several previous works focus on detecting anomalies in the urban computing scenario [28] with the cross-domain data fusion methods [29]. Chawla et al. infer the root cause of road traffic anomalies with Principal Component Analysis [30]. Borges et al. monitor the urban infra-structure considering the heterogeneous attributes and relationships in the data [31]. Zhang et al. detected urban anomalies with multiple spatio-temporal data sources [32]. Du et al. developed an anomaly detection system for identifying pickpocket suspects with transit records [33]. Furthermore, Zhao et al. detected pickpocketing gangs on buses with a graph-based community detection [34]. He et al. detected vehicle illegal parking events using sharing bikes' trajectories [35]. Most of these methods are designed to detect anomalies for the traffic flow and crowd using the human movement data. Unlike these scenarios, we detect the gas-theft suspects in the urban infrastructures of gas supply.

7 Conclusion And Future Work

In this paper, we propose a data-driven approach SVOC to detect gas-theft suspects of boiler rooms. In this way, gas-theft suspects can be discovered in the early stage with higher accuracy. Considering the temporal proximity between gas consumption and temperature, we first calculate the temperature-gas deformation variation to detect the normal and abnormal boiler rooms. Based on the detection results, the OCSVM based algorithm captures the different characteristic factors across multiple data sources for detecting gas-theft suspects. We conduct experiments on a real-world dataset covering one heating season, where the results demonstrate advantages of our approach. With the normality detection algorithm TGSV, we can exclude 62% boiler rooms as normal users; with the anomaly detection algorithm, we can further detect 21% boiler rooms as gas-theft suspects with high recall. We have developed a real-time system on the cloud, providing daily gas-theft suspects for gas companies.

In the future, we will upgrade our system to improve detection accuracy, collaborating with gas companies' staff on the feedback of inspections and maintenance records. Besides, we want to generalize our method to more types of gas users and other utility fraud detection tasks.

Acknowledgment

This work was supported by the National Key R&D Program of China (2019YFB2101800), National Natural Science Foundation of China Grant (61773324), and China Postdoctoral Science Foundation.

References

[1] T. Wang and B. Lin, "China's natural gas consumption and subsidies: From a sector perspective," *Energy Policy*, vol. 65, pp. 541–551, 2014.

[2] S. Paltsev and D. Zhang, "Natural gas pricing reform in china: Getting closer to a market system?" *Energy Policy*, vol. 86, pp. 43–56, 2015.

[3] R. P. Timmer and P. J. Lamb, "Relations between temperature and residential natural gas consumption in the central and eastern united states," *Journal of Applied Meteorology and Climatology*, vol. 46, no. 11, pp. 1993–2013, 2007.

[4] I. Daubechies, "The wavelet transform, time-frequency localization and signal analysis," *IEEE transactions on information theory*, vol. 36, no. 5, pp. 961–1005, 1990.

[5] C. Chen and C. Hsiao, "Haar wavelet method for solving lumped and distributed-parameter systems," *IEE Proceedings-Control Theory and Applications*, vol. 144, no. 1, pp. 87–94, 1997.

[6] A. D. Chouakria and P. N. Nagabhushan, "Adaptive dissimilarity index for measuring time series proximity," *Advances in Data Analysis and Classification*, vol. 1, no. 1, pp. 5–21, 2007.

[7] B. Schölkopf, R. C. Williamson, A. J. Smola, J. Shawe-Taylor, and J. C. Platt, "Support vector method for novelty detection," in *Advances in neural information processing systems*, 2000, pp. 582–588.

[8] K.-L. Li, H.-K. Huang, S.-F. Tian, and W. Xu, "Improving one-class svm for anomaly detection," in *Proceedings of the 2003 International Conference on Machine Learning and Cybernetics*, vol. 5. IEEE, 2003, pp. 3077–3081.

[9] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: identifying density-based local outliers," in *ACM sigmod record*, vol. 29, no. 2. ACM, 2000, pp. 93–104.

[10] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation-based anomaly detection," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 6, no. 1, p. 3, 2012.

[11] M. Ester, H. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," pp. 226–231, 1996.

[12] H. Xu, W. Chen, N. Zhao, Z. Li, J. Bu, Z. Li, Y. Liu, Y. Zhao, D. Pei, Y. Feng *et al.*, "Unsupervised anomaly detection via variational auto-encoder for seasonal kpis in web applications," in *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, 2018, pp. 187–196.

[13] B. Zong, Q. Song, M. R. Min, W. Cheng, C. Lumezanu, D. Cho, and H. Chen, "Deep autoencoding gaussian mixture model for unsupervised anomaly detection," 2018.

[14] H. Ren, B. Xu, Y. Wang, C. Yi, C. Huang, X. Kou, T. Xing, M. Yang, J. Tong, and Q. Zhang, "Time-series anomaly detection service at microsoft," in *KDD*, 2019, pp. 3009–3017.

[15] X. Yang, X. Yi, S. Chen, S. Ruan, J. Zhang, Y. Zheng, and T. Li, "You are how you use: Catching gas theft suspects among diverse restaurant users," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020, pp. 2885–2892.

[16] G. M. Messinis and N. D. Hatziargyriou, "Review of non-technical loss detection methods," *Electric Power Systems Research*, vol. 158, pp. 250–266, 2018.

[17] K. Dineshkumar, P. Ramanathan, and S. Ramasamy, "Development of arm processor based electricity theft control system using gsm network," in *2015 International Conference on Circuits, Power and Computing Technologies [ICCPCT-2015]*. IEEE, 2015, pp. 1–6.

[18] B. Khoo and Y. Cheng, "Using rfid for anti-theft in a chinese electrical supply company: A cost-benefit analysis," in *2011 Wireless Telecommunications Symposium (WTS)*. IEEE, 2011, pp. 1–6.

[19] B. Coma-Puig, J. Carmona, R. Gavalda, S. Alcoverro, and V. Martin, "Fraud detection in energy consumption: A supervised approach," in *2016 IEEE international conference on data science and advanced analytics (DSAA)*. IEEE, 2016, pp. 120–129.

[20] D. R. Pereira, M. A. Pazoti, L. A. Pereira, D. Rodrigues, C. O. Ramos, A. N. Souza, and J. P. Papa, "Social-spider optimization-based support vector machines applied for energy theft detection," *Computers & Electrical Engineering*, vol. 49, pp. 25–38, 2016.

[21] M. Di Martino, F. Decia, J. Molinelli, and A. Fernández, "A novel framework for nontechnical losses detection in electricity companies," in *Pattern Recognition-Applications and Methods*. Springer, 2013, pp. 109–120.

[22] L. A. P. Júnior, C. C. O. Ramos, D. Rodrigues, D. R. Pereira, A. N. de Souza, K. A. P. da Costa, and J. P. Papa, "Unsupervised

1041-4347 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: SOUTHWEST JIAOTONG UNIVERSITY. Downloaded on January 30,2022 at 13:55:15 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2021.3062707, IEEE Transactions on Knowledge and Data Engineering

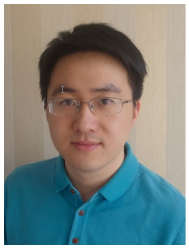
non-technical losses identification through optimum-path forest," *Electric Power Systems Research*, vol. 140, pp. 413–423, 2016.
[23] E. W. S. Angelos, O. R. Saavedra, O. A. C. Cortés, and A. N. de Souza, "Detection and identification of abnormalities in customer consumptions in power distribution systems," *IEEE Transactions on Power Delivery*, vol. 26, no. 4, pp. 2436–2442, 2011.
[24] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 3, p. 15, 2009.
[25] Z. Fu, W. Hu, and T. Tan, "Similarity based vehicle trajectory clustering and anomaly detection," in *IEEE International Conference on Image Processing 2005*, vol. 2. IEEE, 2005, pp. II–602.
[26] S. Agrawal and J. Agrawal, "Survey on anomaly detection using data mining techniques," *Procedia Computer Science*, vol. 60, pp. 708–713, 2015.
[27] Y. Pu, Z. Gan, R. Henao, X. Yuan, C. Li, A. Stevens, and L. Carin, "Variational autoencoder for deep learning of images, labels and captions," in *Advances in neural information processing systems*, 2016, pp. 2352–2360.
[28] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, "Urban computing: concepts, methodologies, and applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 5, no. 3, p. 38, 2014.
[29] Y. Zheng, "Methodologies for cross-domain data fusion: An overview," *IEEE transactions on big data*, vol. 1, no. 1, pp. 16–34, 2015.
[30] S. Chawla, Y. Zheng, and J. Hu, "Inferring the root cause in road traffic anomalies," in *2012 IEEE 12th International Conference on Data Mining*. IEEE, 2012, pp. 141–150.
[31] J. De Melo Borges, T. Riedel, and M. Beigl, "Urban anomaly detection: A use-case for participatory infra-structure monitoring," in *Proceedings of the Second International Conference on IoT in Urban Space*. ACM, 2016, pp. 36–38.
[32] H. Zhang, Y. Zheng, and Y. Yu, "Detecting urban anomalies using multiple spatio-temporal data sources," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, pp. 1–18, 03 2018.
[33] B. Du, C. Liu, Z. Hou, and H. Xiong, "Catch me if you can: Detecting pickpocket suspects from large-scale transit records," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 87–96.
[34] X. Zhao, Y. Zhang, H. Liu, S. Wang, Z. Qian, Y. Hu, and B. Yin, "Detecting pickpocketing gangs on buses with smart card data," *IEEE Intelligent Transportation Systems Magazine*, vol. 11, no. 3, pp. 181–199, 2019.
[35] T. He, J. Bao, R. Li, S. Ruan, Y. Li, C. Tian, and Y. Zheng, "Detecting vehicle illegal parking events using sharing bikes' trajectories." in *KDD*, 2018, pp. 340–349.

Xiuwen Yi is currently a Data Scientist of JD Intelligent Cities Research and Postdoctoral Researcher at Tsinghua University. He got his Ph.D. degree in Computer Science and Technology from Southwest Jiaotong University in 2018. He was an intern in Urban Computing Group at MSR Asia from 2014 to 2017. His research interests include: Spatiotemporal Data Mining, Deep Learning, and Urban Computing. He serves as associate editor of IET Smart Cities journal. He has published over 20+ research papers in refereed conferences and journals.

Xiaodu Yang is a Master student of both Southwest Jiaotong University and CentraleSupélec, Université Paris-Saclay. She majors in computer science and technology. Her research interests mainly include spatio-temporal data mining with deep learning, urban computing. She is also an intern student in JD Intelligent Cities Business Unit.

Yanyong Huang received his double Ph.D. degrees from the Southwest Jiaotong University and FernUniversität in Hagen in 2018 and 2020, respectively. He is currently an Associated Professor in the School of Statistics, Southwestern University of Finance and Economics. His research interests include big data, data mining, granular computing and rough sets.

Songyu Ke is a PhD student in Shanghai Jiao Tong University, majoring in computer science and technology. His research interests mainly include spatio-temporal data mining with deep learning, urban computing. He is also an intern student in JD Intelligent Cities Business Unit.



Junbo Zhang is a Senior Researcher of JD Intelligent Cities Research and the head of AI Platform Division of Intelligent Cities Business Unit, JD Digits. Prior to that, he was a researcher at MSRA from 2015 - 2018. His research interests include urban computing, machine learning, and data mining. He currently serves as Associate Editor of ACM Transactions on Intelligent Systems and Technology. He has published over 30 research papers in refereed journals and conferences, among which one paper was selected as the ESI Hot Paper, three as the ESI Highly Cited Paper. He is a member of IEEE, ACM, CAAI and China Computer Federation.

Tianrui Li received the Ph.D. degree from Southwest Jiaotong University in 2002. He was a postdoctoral researcher with SCK·CEN from 2005 to 2006, and a visiting professor with Hasselt University in 2008, the University of Technology in 2009, and the University of Regina in 2014. He is currently a professor and the director of the Key Laboratory of Cloud Computing and Intelligent Techniques, Southwest Jiaotong University. He has authored or coauthored more than 300 research papers in refereed journals and conferences. His research interests include big data, cloud computing, data mining, granular computing and rough sets. He is a fellow of IRSS and senior member of ACM and IEEE.

Yu Zheng is a Vice President and Chief Data Scientist at JD Digits, passionate about using big data and AI technology to tackle urban challenges. He is the general manager of the JD Urban Computing Business Unit and serves as the director of the JD Intelligent City Research. Before that, he was a senior research manager at Microsoft Research. Zheng currently serves as the Editor-in-Chief of ACM Transactions on Intelligent Systems and Technology currently serves as the Editor-in-Chief of ACM Transactions on Intelligent Systems and Technology. He has served as chair on over 10 prestigious international conferences, e.g. as the program co-chair of CIKM 2017 (Industrial Track). In 2013, he was named one of the Top Innovators under 35 by MIT Technology Review (TR35) and featured by Time Magazine for his research on urban computing. In 2014, he was named one of the Top 40 Business Elites under 40 in China by Fortune Magazine. In 2017, Zheng was named an ACM Distinguished Scientist.

1041-4347 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: SOUTHWEST JIAOTONG UNIVERSITY. Downloaded on January 30,2022 at 13:55:15 UTC from IEEE Xplore. Restrictions apply.