15TH IEA HEAT PUMP CONFERENCE
MAY 26–29, 2026 | HOFBURG
VIENNA 2026
DECARBONISATION THROUGH INNOVATION

Peer-benchmarking-based fault detection in a residential geothermal heat-pump fleet

P. Leroy Roger[a,b], E. de Rautlin de la Roy[b], S. Raji[a], T. Recht[b], L. Mora[b] *

[a]NOBATEK, National Institute for the Energy Transition in the Construction Sector, 64600 Anglet, France
[b]Institute of Mechanical Engineering (I2M), UMR CNRS 5295, Université de Bordeaux, 33400 Talence, France

Abstract

The growing adoption of heat pumps for heating and cooling in the residential sector raises the critical issue of ensuring their operational reliability over time. Fault Detection and Diagnosis (FDD) methods can identify soft faults at an early stage, thereby preventing their evolution into critical failures or long-term performance losses. Among the various FDD approaches, process history based techniques are becoming increasingly popular due to their minimal requirement for prior system knowledge. However, these methods typically require large volumes of labelled operational data, which may be unavailable — especially in cases of limited historical records (e.g., new machines or restricted data storage) or when no preliminary laboratory testing has been conducted.

This study proposes an unsupervised methodology based on peer-benchmarking to compare the performance of similar, though not strictly identical, heat pumps with the aim of identifying potential outliers. These outliers are then analysed to highlight true anomalies, indicative of soft faults. This methodology is tested on a real-world dataset of 199 residential geothermal heat pumps, with thermal capacities ranging from 8 to 14 kW, monitored over a one-year period.

This study assesses the suitability of peer benchmarking via outlier detection for fault diagnosis. In the case study, the methodology detected five outliers, three of which were associated with confirmed heat pump faults. These findings demonstrate the capability of this approach to identify malfunctioning units in a heterogeneous, label-free heat pump fleets.

Keywords: Heat pump fleets ; Fault Detection and Diagnosis ; Outlier Detection

1. Introduction

In France, the building sector accounts for 44% of final energy consumption [1]. This energy consumption represents 15.8% of the country's total CO_2 emissions [2]. In the residential sector, 66% of this energy is used for heating. To meet this need, 52% [3] of households use fossil fuels (fuel oil or gas) and 30% use resistive heating devices. The signatory states to the Paris Agreement have committed to keeping "the global average temperature increase well below 2°C above pre-industrial levels" [4], which implies a 43% reduction in greenhouse gas emissions by 2030. One way to achieve this goal is to decarbonise heating needs through increased use of low-carbon energy systems, such as heat pumps (HP), in both new buildings and renovations. This trend is already reflected in the increase in heat pump installations, with more than one million units sold in France in 2020 [1].

The effectiveness of this strategy depends on the installed systems achieving the performance levels anticipated during the design stage. However, the HP performance of heat pumps is influenced by various parameters, many of which lie beyond the designer's control, —such as user behaviour,on-site settings [5],

* Corresponding author.
E-mail address: laurent.mora@u-bordeaux.fr

15th IEA Heat Pump Conference 2026 - DIS2.1 - https://doi.org/10.23697/s4ga-6z12

and the occurrence of faults [6]. This often leads to a gap between expected and observed performances, a phenomenon repeatedly observed in the literature. For instance, Chesser et al. [5] found a drop of 16% between the in-situ Coefficient Of Performance (COP) and the laboratory-measured COP for an 8.5 kW air-source heat pump operating at 7 °C. Deng et al. [7] studied 28 different types of heat (e.g., aerothermal, small and medium-depth geothermal or surface water) and found that all machines, except two, showed a COP lower than that estimated at the design stage. Similarly, O'Hegarty et al. [8] observed an average difference of 40 % between the declared seasonal performance factor and the on-site performance factor for a set of 278 air-to-water heat pumps installed in the United Kingdom.

Therefore, the ability to detect underperforming machines as early as possible and to identify the causes of such underperformance is crucial. To address this issue, FDD approaches can be used. FDD refers to a systematic methodology aimed at identifying and isolating operational faults within a system [9]. Faults can be defined as "a deviation from an acceptable range of an observed variable or estimated parameter related to the process" [10]. The application of FDD approaches to heat pumps has been extensively studied by the scientific community [11], [12]. These studies generally focus on a limited number of machines and consider only system-level faults (e.g., inappropriate refrigerant charge) or component-level faults (e.g., fouling of heat exchangers) [13]. However, under real operating conditions, machine performance may also be influenced by installation-level fault, including improper sizing or incorrect adjustment (e.g., incorrect heating curve setting) [14]. In their study of 1 023 HPs, Brudermueller et al. [14] demonstrate that 7.5% of them are incorrectly sized. This result highlights the fact , when the analysis is extended from a small group of machines to a large set of HPs, the occurrence of installation-level faults becomes significant. It is therefore essential to extend FDD approaches to include installation-level faults.

According to Katipamula and Brambley [15], when applied to building systems, FDD approaches can be classified into three distinct categories:

- Quantitative model-based approaches ;
- Qualitative model-based approaches ;
- Process history-based approaches.

The first two approaches rely on an expert knowledge of the monitored system. The quantitative approach aims to create a physical models of the monitored system. By comparing the models' outputs with the physical measurements taken at the installation, it is therefore possible to detect a fault and identify its cause. The qualitative approach aims to create expert-knowledge rules or thresholds that must not be exceeded in order to detect and diagnose a fault. Conversely, process history based approaches do not require prior system knowledge and instead use statistical methods or machine-learning methods. These methods have been increasingly popular due to recent improvements in computer processing power and access to big data [16]. This last category is particularly suitable for a heterogeneous machine fleet, that includes a wide variety of use cases, complicates the development of universal detection rules. However, most of these rules require the use of labelled data, typically obtained from simulations [17], laboratory tests [18] or tests conducted on a limited number of on-site heat pumps [19]. For large machine fleets, the diversity of operating conditions makes manual data labelling impractical, and the performance status of each machines is unknown a priori. To address this issue, unsupervised FDD approaches (i.e. those that not requiring labelled data) can be adopted. Nevertheless, the literature contains only limited examples of such applications. In their review, Zhelun Chen et al. [20] identified only one study implementing these approaches for vapour compression systems. To date and to the authors' knowledge, no published work has applied them to large-scale datasets comprising several hundred heat pumps. Consequently, there is a need to develop an unsupervised FDD approach leveraging process history data that is suitable for application to a machine fleet.

This overview of the scientific literature highlights that, for a fleet of machines, there is a need, on the one hand, to adapt current FDD approaches so that they account for installation-level faults, and, on the other hand, to adopt approaches that can be applied in a fully unsupervised context. The main contribution of this study is to propose and assess an unsupervised fault-detection methodology, suitable for large and heterogeneous heat pump fleets. The methodology is built upon outlier detection algorithms that use macro-features aggregating daily operating indicators, enabling the identification of heat pumps whose behaviour diverges from the group. Its relevance is then tested on a fleet of 199 geothermal heat pumps in operation. The plausibility of the detected of HPs considered as outlined is subsequently assessed through detailed data analysis. This methodology is described in more detail in the following section.

2. Methods

2.1. Overview of the proposed approach

As discussed in the previous section, a fault can be defined as a significant deviation in one or more operational parameters of the heat pump. However, deviations of this nature may arise from factors other than faults. For example, the shutdown of a system in an unoccupied building does not indicate a malfunction. Assuming that the majority of systems in the fleet operate normally, both cases could be identified as outliers with respect to the fleet average. Consequently, it is crucial to distinguish between these two types of situations. If the discrepancy is actually due to a fault (e.g., refrigerant undercharge, incorrect sizing), it is further referred to as an anomaly. Otherwise, it is referred to as a true outlier, unrelated to equipment malfunction (e.g., a building that is underoccupied in winter). Proper selection of descriptive features is therefore essential to distinguish between these two situations and to guide maintainers towards the underlying cause of an abnormal behaviour. In this article, we therefore encompass the FDD approach within an outlier detection approach.

2.2. Description of the monitored fleet and data pre-processing

As part of our study, we had access to operating data from 207 HPs, all manufactured by the same French company. Although the methodology is designed to be applicable to a heterogeneous fleet, we preferred, as a proof of concept, to pre-filter the available machines in order to control variability within the case study. To do so, we established a list of criteria that a HP must meet in order to be included in the study. These criteria are summarised in Table 2. To build a consistent set of machines installation architecture between machines, we selected geothermal HP with a buffer tank. The refrigerant used by the manufacturer for this type of machine is R407C. We limited the power of the machines to 18 kW and the minimum power of the fleet to 4 kW. Additionally, as the selected features were obtained from the machines' operating cycles, only single-compressor technologies were considered to facilitate comparison. The HP in the fleet can be either single or dual-purpose (heating and domestic hot water), but this study focuses only on heating. The heat pumps are equipped with sensors that measure the operating status (e.g., compressor status, operating time since last start-up), cycle parameters (e.g., temperature and pressure at various points in the thermodynamic cycle) and inflows / outflows to the heating zones (setpoint and temperature). These variables are measured every 15 minutes one-year period,, representing 525 600 recording points per variable, assuming continuous data collection, without gaps.

Table 1. selection criterion used for the case study

Selection criteria	Value
Heat pump type	Water to water heat pump
Installation	With buffer tank
Refrigerant	R407C
Rated power	[4 kW ; 18 kW]
Number of compressors	1

To mitigate the potential influence of outliers in the data it is filtered using the inter-quantile method [21]. The features selected in the methodology are then extracted. A substantial discrepancy exists in the number of measurement available per days among the machines. To ensure that the features reflect the actual behaviour of the machine, we prefer to exclude machines with too few days of operation. To do this, we filter out machines that fall outside the interval centred on the mean and four standard deviations wide, equivalent to 95% of the values. In our case, this corresponds to machines with less than 8 days of operation. This brings the number of machines studied to 199. The features are then averaged annually per machine.

2.3. Outlier detection algorithms

Outlier detection has been the subject of in-depth study for decades, and numerous methods have been proposed by the scientific community. These methods can be classified according to the techniques used [21] or according to label availability [22]. In the context of unsupervised data from steam compression machines, Mtibaa [23]proposed the following classification:

- Statistics-based approaches: the statistical distribution of points allows to identify observations that deviate significantly from the group average. This category can itself be divided into two sub-groups. If the data distribution is known, parametric methods such as Elliptic Envelope (EE) [24]can be used. Otherwise, non-parametric methods such as Gaussian Mixture Model (GMM) [25] are suitable.
- Distance-based approaches: the distances between points allows to identify observations that deviate from the average group. This category can itself be divided into three sub-groups. The first, based on proximity, uses the distance between the observation and its closest neighbours as a sorting criterion. A common example is the K-Nearest Neighbours (KNN) [26]algorithm. The second is density-based and assumes that an outlier is an observation located in a low-density space, with density estimated from the k nearest neighbours. By comparing the local densities of each point, it is then possible to distinguish inliers from outliers. The Local Outlier Factor (LOF) [27] is a good example of a density-based approach. The third category adopts a clustering approach, the idea being that it is possible to form different clusters from the observations. Points that do not belong to any cluster or to small clusters are considered outliers. The Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [28]algorithm is most often cited for outlier detection.
- Angle-based: similar to distance-based approaches, it is possible to distinguish between outliers and inliers based on the angle formed between the data points. Angle-Based Outlier Detection (ABOD) [29] uses this approach.
- Classifier-based approaches: the use of labelled data allows to train a classifier model to recognise a normal observation, thereby enabling test data to be classified as inliers or outliers. These methods can be adapted to unsupervised cases by assuming that a small proportion of the dataset is outliers. An algorithm regularly cited as belonging to this category is Isolation Forest (IF) [30].

Without assuming which of these approaches is most suitable for the application, one algorithm from each family presented above has been selected based on their prevalence in the literature: EE, GMM, LOF, ABOD, and IF. This selection aims to compare methods founded on distinct theoretical principles.

2.4. Outlier detection algorithms tuning

As detection is performed in an unsupervised context, it is difficult to know in advance which of the methods' hyperparameters are most effective for detection. We have chosen to keep the default parameters of the library algorithms (i.e. Scikit learn for EE, GMM, LOF, IF and PyOD for ABOD), with the exception of the contamination rate. This choice aims to limit biases linked to empirical hyperparameter tuning, which is difficult to compare in an unsupervised context [31]. The comparison between methods is therefore based on their standard behaviour.

The contamination parameter represents the expected proportion of outliers in a dataset. In the case of unsupervised detection It is not possible to access a training dataset that would allow this to be estimated. To attain a realistic value of contamination parameter, we propose to optimise it based on the percentage of agreement between the different algorithms, reflecting the overlap between methods. This involves adjusting the contamination to maximise the consistency of predictions, according to an ensemble agreement principle [32]. To do this, we simultaneously vary the contamination scores of the different algorithms. As the contamination score is unknown, a conservative approach is preferred. The range of variation chosen is [0.01; 0.49], with 0.49 being the limit above which the assumption that a majority of HPs exhibit normal behaviour is no longer satisfied. At each iteration, an Average inter-method Concordance Score (ACS) is calculated, representing the number of times an outlier has been detected by several approaches. This score is calculated using the equation Eq. (1).:

$$ACS = \frac{1}{n}\sum_{j=1}^{n} A_j \tag{1}$$

where A_j is percentage of agreement of one algorithm, is calculated using the equation Eq. (2).:

$$A_j = \frac{1}{n-1}\sum_{\substack{i=1 \\ i\neq k}}^{n} M_{ij} \tag{2}$$

where M_{ij} is the percentage of agreement between two algorithms, calculated using the following formula:

$$M_{ij} = \frac{O_i \cap O_j}{\max(O_i, O_j)} * 100 \quad (3)$$

where O_j being the outliers detected by algorithm i, and O_j by algorithm j.

Subsequently, the contamination score that maximises this indicator is retained as a hyperparameter. All of the methods selected, along with their associated hyperparameters, are summarised in Table 2.

Table 2. outlier detection method selected and associated hyperparameters

Category	Method	Hyper parameter
Parametric Statistics-based approaches	EE	Assume centered : False, Support Fraction : None, Random State=None
Non parametric Statistics-based approaches	GMM	Number of mixture components : 1, Covariance type : full, Convergence threshold : 10^{-3}, Non-negative regularization : 10^{-6}, Max number of iteration : 100, Number of initializations to perform : 1, Method used to initialize the weights : 'kmeans'
Distance -based approaches	LOF	Number neighbours : 20, Algorithm : 'auto', Leaf Size : 30, Metric : 'minkowski', Parameter for the Minkowsk : 2, Novelty : False
Angle-based approaches	ABOD	Number neighbours : 5
Classifier-based approaches	IF	Number of base estimators : 100, Number of samples to draw : 'auto', , Number of features to draw =1.0, Bootstrap=False

2.5. Feature selection

Generally, the first step in an FDD approach is to select the appropriate features for detection and diagnosis [33] or to construct them from other physical quantities. Although it is possible to automate feature selection using wrapper methods [23], we prefer to select and construct them manually in our approach in order to maintain greater interpretability. The context of a heterogeneous group of machines encourages the use of features derived from easily measurable physical quantities that are common to all machines in the group. Brudermueller et al [34] adopt such an approach and propose indicators derived from energy metering in relation to machine cycle behaviour, assuming that an atypical cycle is an indicator of an underlying problem. They therefore select four features calculated per day: the *number of operating hours*, the *number of cycles*, the *average cycle time* calculated using the ratio between the number of operating hours and the number of cycles, and the *inverse of this ratio*. In another publications [14], authors evaluate the sizing of machines using, in particular, the utilisation ratio calculated from the compressor rotation speed. As we wanted to include both fixed and variable speed machines in the study, we use the approximation proposed in [34] Eq. (4).:

$$u_{day} = \frac{E_{day} * 100\%}{P_{hp} * 24h} \quad (4)$$

E_{day} is the energy produced by the heat pump in one day, expressed in W.h and P_{hp} is the nominal power of the heat pump in W. In the rest of this article, we adopt the viewpoint of Brudermueller et al. [34] by using the aforementioned indicators, listed in Table 3 of Section 3, as features for outlier detection. Although relevant to this approach, we did not use the COP, which is nevertheless the most frequently cited indicator when discussing heat pump performance monitoring [6]. In this case study, approximately 39% of the machines lack the necessary equipment to perform this measurement and are therefore excluded from the selection. The behaviour of a heat pump is greatly influenced by its environment, particularly the outside temperature, which affects both the temperature of the source (aerothermal or geothermal) and the building's heating requirements. A common solution is to divide the selected indicators by the average daily temperature. In our case, we prefer to retain them as they are, in order to avoid compensatory effects in which an indicator would be artificially amplified because it is divided by an excessively low temperature. Although the performance of a heat pump may evolve over time, whether due to a calendar effect or gradual refrigerant leakage, the indicators in this study were averaged over the entire observation period. It allows to obtain an overview of the machine fleet and limit the amount of data input into the outlier detection algorithms.

Finally, one of the recurring problems with outlier detection is the ‘curse of dimensionality’ [21]: increasing the number of features tends to spread the points further apart, making the task of detecting outliers more complex. To guard against this, we first perform a correlation analysis using Pearson's coefficient [23]. Subsequently, if two or more features have a coefficient greater than 0.8 [35], we will only retain one.

2.6. Evaluation procedure

Reducing the number of false positives is one of the key to operators' confidence in the FDD process [36]. This choice is made at the expense of the possibility of detecting all true positives, however, since the work is still at the proof-of-concept stage, a conservative approach is preferred. To reduce the number of false positives, only outliers detected by all detectors are considered true positives.

The next step is to classify these true positives as anomalies or true outliers, if possible. To do this, we rely on the analysis of Z-scores per feature and per detection. This score allows us to approximate the weight of the features in the detection. Z-score is expressed as:

$$Z = \frac{x - \mu}{\sigma} \tag{5}$$

where x is the observation, μ is the mean of the observations and σ is the standard deviation of the observations. The z-score reflects the distance of the observation from the group mean. The greater the absolute value of the z-score, the further the observation is from the mean. Thus, an analysis of the classification of z-scores allows us to understand which feature makes the machine appear abnormal compared to the fleet. In an unsupervised context, we do not have access to machine labels. The usual techniques for evaluating outlier detectors such as AUC-ROC [32] cannot therefore be applied. Instead, we propose an exploratory evaluation in the next section.

3. Results and discussion

3.1. Results

In accordance with the methodology described in section 2.5, we calculated Pearson's correlation coefficient for the various features identified in the literature. The correlation matrix is shown in Table 3. The only two features exceeding the threshold of 0.8 are the utilisation ratio and the total daily operating time. To reduce the dimensionality of the problem, we decided to keep only the utilisation ratio feature. This feature has an immediate significance that facilitates the understanding of the detection.

Table 3. Correlation matrix of the features based on Pearson's coefficient

	Total daily operating time [min]	Number of daily starts-up [/]	Utilisation [%]	Average cycle operating time [min]	Number of starts per operating time [1/min]
Total daily operating time [min]	/	-0.128660	0.861106	0.707890	-0.589525
Number of daily starts-up [/]	/	/	0.073183	- 0.574844	0.538315
Utilisation [%]	/	/	/	0.565671	- 0.354455
Average cycle operating time [min]	/	/	/	/	-0.570471
Number of starts per operating time [1/min]	/	/	/	/	/

We then optimize the contamination parameter of the various outlier detection algorithms using the average concordance percentage indicator, see Figure 1.a. For our application, we have a local maximum of 4 % and a global maximum of 15 % .The respective average concordance percentages are 82.8% and 85.7%,. We retain the former as the input hyperparameter, assuming that outliers detected for a lower percentage are further from the average group and therefore more abnormal. The correspondence matrix associated with a contamination percentage of 4 % is shown in Figure 1.b. Among the different outlier detection methods, GMM has the best percentage of agreement (84.4 %), followed by ABOD and IF (81.3 %), then LOF (78.13 %). EE has the worst concordance percentage (68.75 %).

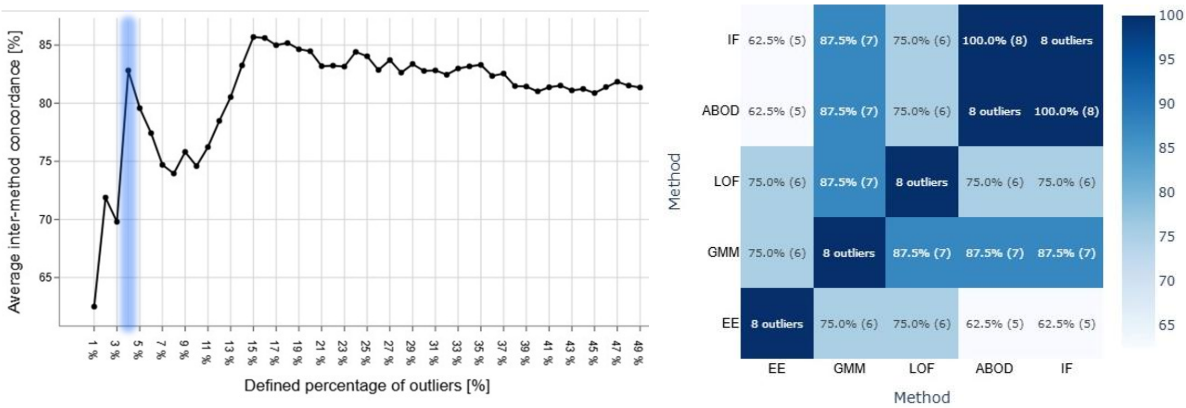

Figure 1.a. Evolution in inter-method concordance score based on the contamination rate used. Figure 1.b. Inter-method concordance matrix for the optimum of 6% selected.

Figure 2 shows, on the left-hand side, all the machines in the fleet that were detected as outliers, and which algorithms classified them as such. The right-hand side of the graph shows the Z-score by outlier and by feature. Identification numbers have been assigned to the detected heat pumps to improve readability.

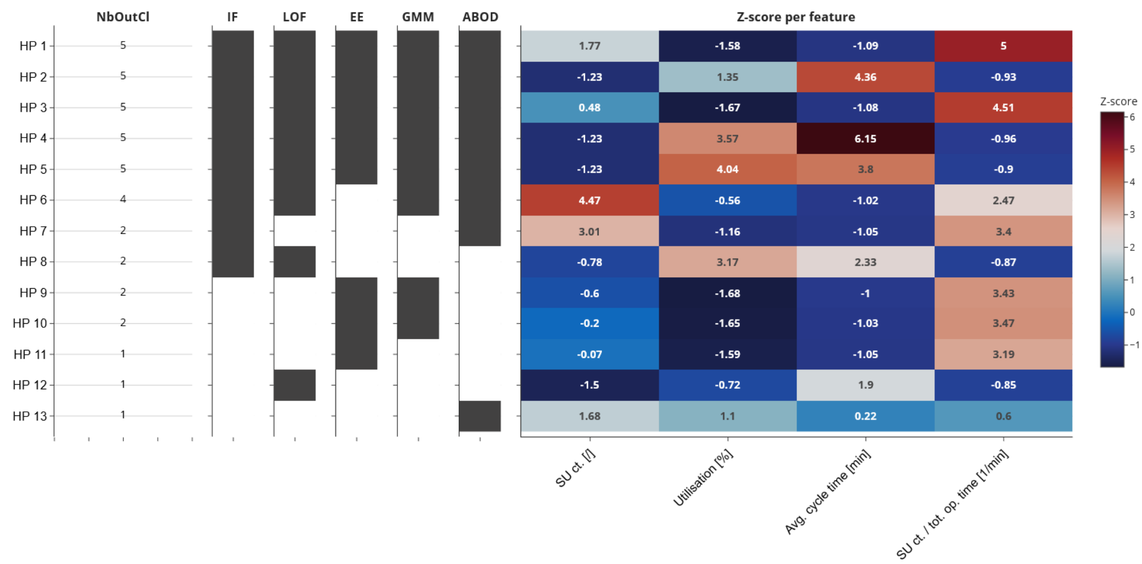

Figure 2.: count of detections by different algorithms, and algorithm that detected the machine. right: Z-score by feature and by machine.

Of the 13 HPs outliers detected in total, 5 were detected by all five detectors and can therefore be considered true positives. The remaining 8 machines can be kept under surveillance without undergoing a more detailed diagnostic process. The features *average cycle duration* and *number of starts per operating time* allowed two outliers to be detected each. *Utilisation* allowed one outlier to be detected. Conversely, the *number of starts* did not allow outliers to be detected in our case study. The colour gradient indicates how much a feature stands out from the others statistically speaking. For example, HP_2 has the highest z-score for the average cycle time feature (4.36), clearly stands out from the features of number of starts (-1.23), utilisation (1.13) and number of starts per operating time (-0.91). This suggests that the average cycle time feature plays a key role in detection and, as a result, the machine stands out from the fleet according to this criterion. Although machine HP 5 was probably also detected based on the average cycle time feature (6.15), the z-score associated with usage is also clearly evident (3.45). This nuance in the detection and opens up avenues for diagnosis. In the second case, the cause of the long operating cycles is probably the same as the one causing intensive use of the machine. This could be due to difficult environmental conditions or undersizing. In the first case, however, the cause of abnormally long cycles is harder to identify. Possible diagnostic explanations will be discussed in the next section.

3.2. Discussion

As explained in section 2.6, the unsupervised context of the study does not allow for the use of standard assessment tools. Instead, we propose an exploratory evaluation. To do this, we analyse the time series of the five true outliers in order to manually evaluate the detectors and propose diagnostic avenues. Figure 3 shows a day of operation for HP 1 that reflects the abnormal nature of the machine. In 24 hours, the heat pump started 43 times, with each cycle lasting less than 5 minutes. This results in a high number of starts per operating time, which is consistent with the z-score of 5 for this feature observed in Figure 2. The ambient temperature in the heated area shows minimal fluctuations,, even though the heating circulation control is always set to maximum. Several factors could explain this, such as a problem with the heat emitters or excessive heat loss from the heated area. In addition, the buffer tank only fluctuates within a range of 4°C around an average temperature of 54°C. One possible cause of this behaviour could be that the tank control hysteresis setting is too tight. Without determining the precise cause, this heat pump is clearly operating abnormally. The HP 3 shown in Figure 4 behaves in a similar manner. In 24 hours, it restarts 36 times and does not have any cycles longer than 5 minutes. This result is consistent with the Z-scores obtained: in both cases, the feature of the number of starts per operating time is the one that enabled detection. The probable causes of malfunction are similar to those described above. This heat pump can therefore be considered an anomaly.

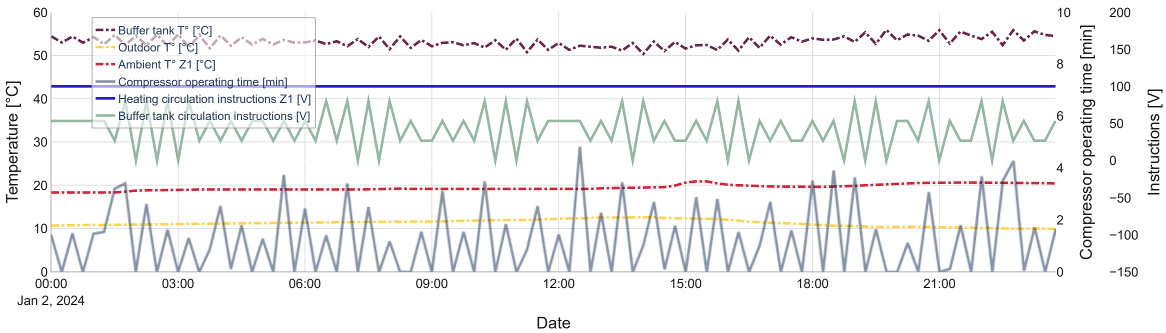

Figure 3. Evolution of the main operating indicators for heat pump 1. The dotted lines show the outside temperature, the ambient temperature in the heated area and the buffer tank temperature. The solid lines show the circulation setpoints to the tank and to the heated area, as well as the compressor operating time.

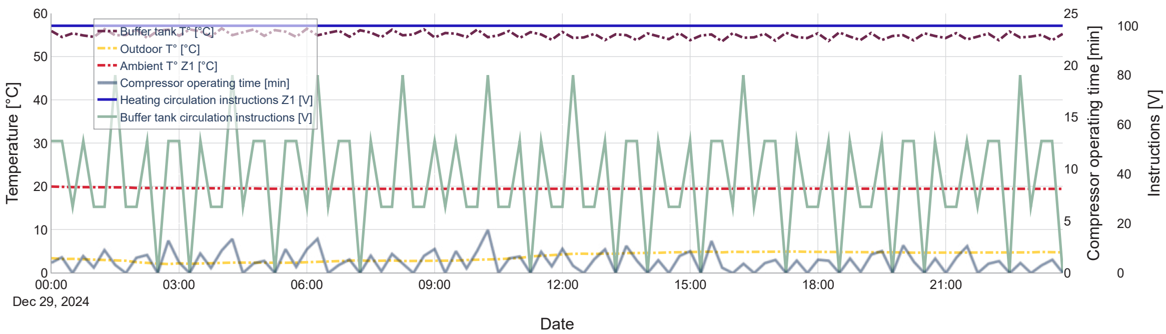

Figure 4. Evolution of the main operating indicators for heat pump 3. The dotted lines show the outside temperature, the ambient temperature in the heated area and the buffer tank temperature. The solid lines show the circulation setpoints to the tank and to the heated area, as well as the compressor operating time.

Hp 5 has the highest utilisation z-score (3.94 – see Figure 2). This heat pump differs from the previous two in that it supplies heating to two zones. Figure 5 shows a day during which both zones were supplied simultaneously. The moment of simultaneity is highlighted by an orange rectangle. Unlike the two previous cases, the heat pump's operating cycles are few and long. Despite this, the temperatures in the two zones vary slowly, by 0.8°C in 6.5 hours for zone 1 (Z1), and by 1.5°C in approximately three hours for zone 2 (Z2). We note that during the simultaneous phase, the temperature of the buffer tank stabilises, even though the outside temperature is at 5 °C. These two signals could therefore indicate that this machine is undersized, confirming that it is not functioning normally. This diagnosis could be confirmed by more contextual information such as the energy injected into the building, but this would take us away from the machine fleet approach presented here.

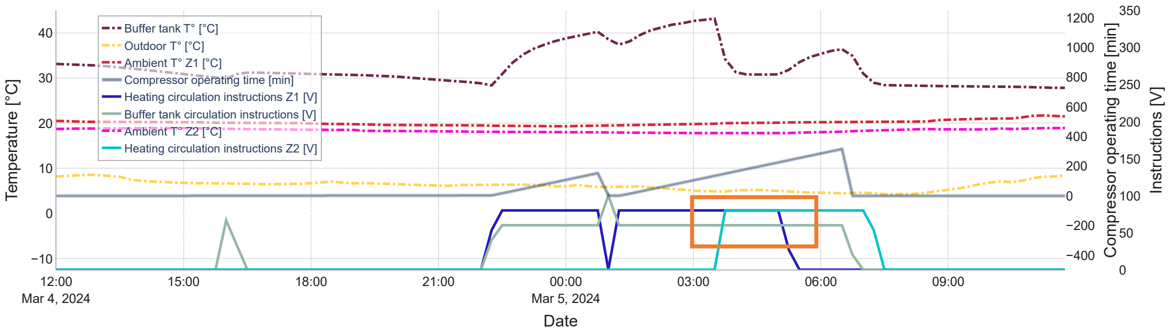

Figure 5. Evolution of the main operating indicators for heat pump 5. The dotted lines show the outside temperature, the ambient temperature in the heated area and the buffer tank temperature. The solid lines show the circulation setpoints to the tank and to the heated area, as well as the compressor operating time.

The HP 2 is also dual-zone. Unlike the previous one, the simultaneous periods highlighted in orange do not prevent the buffer tank from heating, even though the outside temperature is similar, see Figure 6. The room temperature observed in Z1 warranted further investigation. Although the heating circulation for this zone is activated from 9:45 p.m. to 9:15 a.m., the room temperature drops from 18.5 °C to 15.7 °C. As a result, the heat pump runs for 5 hours during the night. After that, the room temperature rises rapidly to reach 20.5 °C at 11 a.m., without any change in the machine's behaviour. This change in room temperature could be attributed to the user adopting a particular behaviour upon waking up, but this cannot be verified. As things stand, we are unable to attribute the machine's operating deviation from the fleet average to the machine itself. This presents a first limitation of the proposed methodology, which is that it is not always possible to classify the true outlier without external knowledge.

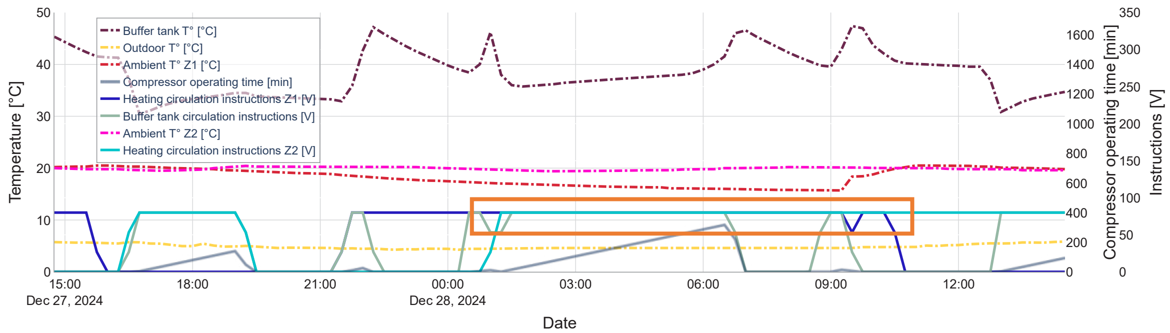

Figure 6. Evolution of the main operating indicators for heat pump 2. The dotted lines show the outside temperature, the ambient temperature in the heated area and the buffer tank temperature. The solid lines show the circulation setpoints to the tank and to the heated area, as well as the compressor operating time.

Finally, when attempting to analyse HP 4, we realised that the operating history file we had access to contained numerous gaps, even though this heat pump had passed the first filter described in section 3.1. Most of the data we had access to covers a period of twenty days, between 11 and 31 December 2024. During this period, the average outdoor temperature was 0.3 °C, which implies significant use of the heat pump. However, the prominent z-scores for this machine are those associated with the cycle duration (6.15) and usage (3.45) features. This is consistent with the expected behaviour of a heat pump in winter, which will tend to run longer and at a higher load, but it does not indicate abnormal machine behaviour. Here, the HP is therefore a true outlier. It brings to light a limitation within the proposed methodology. By not including the environment in the initial features, a heat pump could potentially be categorised as true positive when in fact it is simply reacting to its environment. In our case, this is due to a bias related to the recording period, but it could also be true for a machine located in a harsher climate than the average for the fleet.

3.3. Limitations and future work

The methodology presented in this article has demonstrated its effectiveness in detecting abnormal machines within a fleet in an unsupervised manner. Nevertheless, several limitations have been identified, which are discussed below.

As shown in section 3.2, certain variables such as the number of daily starts, were not decisive in identifying outliers. In this work, we chose to select features based on their physical significance and availability, attempting to keep the number as low as possible to avoid overfitting. This choice favours the interpretability of the results but does not ensure that the features chosen are the most relevant for separating machines between normal and abnormal behaviour. Future research could include a broader set of features, with particular attention to variables reflecting temperature conditions and seasonal dynamics. Furthermore, the decision to retain only the default hyperparameters of the detection methods is also questionable, as tuning these parameters can strongly affect detection performance [37]. A first direction for improvement would be to introduce unsupervised internal evaluation metrics, such as the IREOS proposed by Marques et al [31]. Subsequently, a comparison with labelled data would allow for a more accurate assessment of the methodology's performance.

Although the case study includes 199 geothermal heat pumps, it remains limited in terms of the heterogeneity. This limits the diversity of anomalies that can be encountered across the fleet. The variability within the case study was intentionally restricted, since the present work serves as an initial proof of concept., such as air source heat pumps. Future work will therefore aim to broaden the scope of the study to a larger and more heterogeneous fleet of heat pumps in order to assess the robustness and generalisability of the methodology proposed in this article.

4. Conclusion

In this article, we have proposed a methodology for the unsupervised detection of faulty heat pumps within a fleet of machines, based on inter-comparison. This methodology relies on the assumption that the average behaviour of the fleet reflects normal operation, and that a faulty machine deviates from this reference. In this respect, detecting defective machines is equivalent to identifying outliers using macro-features derived from cycling patterns and produced energy. Five detection methods were compared on a dataset of 199 geothermal heat pumps. Only machines detected by all approaches were considered true outliers. In our case study, five heat pumps were classified as true outliers, three of which exhibited abnormal behaviour after manual analysis of their operating histories. Among the probable causes of malfunction, the hysteresis setting of the buffer tank and the undersizing of the machine were suggested, highlighting once again the crucial role of proper sizing and configuration in ensuring on-site performance. The results suggest that, once integrated into a maintenance programme, outlier detection could become a valuable tool for maintaining the performance of installations.

Acknowledgements

This research was conducted within the framework of the *PacPredict* project within the joint research team GP2E (as in French, Guarantee of Buildings Energy and Environmental Performances), between the I2M laboratory and the French Institute for Energy and Environmental Transition NOBATEK/INEF4, with funding from the Regional Council of Nouvelle Aquitaine (CRNA).

References

[1] Ministère de la Transition écologique, “Énergie dans les bâtiments.” Accessed: Nov. 12, 2025. [Online]. Available: https://www.ecologie.gouv.fr/politiques-publiques/energie-batiments?

[2] Syndicat des énergies renouvelables, “Energies renouvelables et bâtiment, état des lieux”, Accessed: Jan. 18, 2024. [Online]. Available: https://www.syndicat-energies-renouvelables.fr/wp-content/uploads/basedoc/brochure-enr-et-batiments-pap.pdf

[3] INSEE, “La facture énergétique du logement a baissé depuis 2013 malgré la hausse de la fiscalité,” 2017. Accessed: May 12, 2025. [Online]. Available: https://www.insee.fr/fr/statistiques/3280930

[4] UNFCCC, “Paris Agreement French,” 2015. Accessed: May 12, 2025. [Online]. Available: https://unfccc.int/sites/default/files/french_paris_agreement.pdf

[5] M. Chesser, P. Lyons, P. O'Reilly, and P. Carroll, "Air source heat pump in-situ performance," *Energy Build*, vol. 251, Nov. 2021, doi: 10.1016/j.enbuild.2021.111365.

[6] I. Bellanco, E. Fuentes, M. Vallès, and J. Salom, "A review of the fault behavior of heat pumps and measurements, detection and diagnosis methods including virtual sensors," *Journal of Building Engineering*, vol. 39, Jul. 2021, doi: https://doi.org/10.1016/j.jobe.2021.102254.

[7] J. Deng, Q. Wei, M. Liang, S. He, and H. Zhang, "Does heat pumps perform energy efficiently as we expected: Field tests and evaluations on various kinds of heat pump systems for space heating," *Energy Build*, vol. 182, pp. 172–186, Jan. 2019, doi: 10.1016/j.enbuild.2018.10.014.

[8] R. O'Hegarty, O. Kinnane, D. Lennon, and S. Colclough, "Air-to-water heat pumps: Review and analysis of the performance gap between in-use and product rated performance," *Renewable and Sustainable Energy Reviews*, vol. 155, Mar. 2022, doi: 10.1016/j.rser.2021.111887.

[9] V. Moudgil, K. Hewage, S. A. Hussain, and R. Sadiq, "Integration of IoT in building energy infrastructure: A critical review on challenges and solutions," *Renewable and Sustainable Energy Reviews*, vol. 174, Mar. 2023, doi: 10.1016/j.rser.2022.113121.

[10] V. Singh, J. Mathur, and A. Bhatia, "A comprehensive review: Fault detection, diagnostics, prognostics, and fault modeling in HVAC systems," *International Journal of Refrigeration*, vol. 144, pp. 283–295, Dec. 2022, doi: 10.1016/j.ijrefrig.2022.08.017.

[11] J. Bi *et al.*, "AI in HVAC fault detection and diagnosis: A systematic review," *Energy Reviews*, vol. 3, no. 2, Jun. 2024, doi: 10.1016/j.enrev.2024.100071.

[12] S. P. Melgaard, K. H. Andersen, A. Marszal-Pomianowska, R. L. Jensen, and P. K. Heiselberg, "Fault Detection and Diagnosis Encyclopedia for Building Systems: A Systematic Review," *Energies (Basel)*, vol. 15, no. 12, Jun. 2022, doi: 10.3390/en15124366.

[13] P. Barandier and A. J. Marques Cardoso, "A review of fault diagnostics in heat pumps systems," *Appl Therm Eng*, vol. 228, Jun. 2023, doi: 10.1016/j.applthermaleng.2023.120454.

[14] T. Brudermueller, U. Potthoff, E. Fleisch, F. Wortmann, and T. Staake, "Estimation of energy efficiency of heat pumps in residential buildings using real operation data," *Nature Communications* , vol. 16, no. 1, Dec. 2025, doi: 10.1038/s41467-025-58014-y.

[15] S. Katipamula and M. R. Brambley, "Review article: Methods for fault detection, diagnostics, and prognostics for building systems—A review, part I," *HVAC and R Research*, vol. 11, no. 1, pp. 3–25, 2005, doi: 10.1080/10789669.2005.10391123.

[16] Y. Hwan Eom, S. Bin Hong, J. Woo Yoo, and M. Soo Kim, "Deep learning-based refrigerant charge fault detection method of air-source heat pump system," in *13th IEA Heat Pump Conference 2021: Heat Pumps*, Jeju, Republic of Korea, 2021, pp. 552–558. Accessed: Nov. 17, 2025. [Online]. Available: https://iifiir.org/en/fridoc/13-lt-sup-gt-th-lt-sup-gt-iea-heat-pump-conference-2021-amp-nbsp-heat-pumps-145649

[17] S. Borges, L. Jöhnk, T. Klebig, C. Vering, and D. Müller, "Fault Detection and Diagnosis by Machine Learning Methods in Air-to-Water Heat Pumps: Evaluation of Evaporator Fouling," in *36th International Conference on Efficiency, Cost, Optimization, Simulation and Environmental Impact of Energy Systems, ECOS 2023*, Las Palmas de Gran Canaria, Spain, 2023, pp. 815–826. doi: 10.52202/069564-0074.

[18] G. Li *et al.*, "An improved decision tree-based fault diagnosis method for practical variable refrigerant flow system using virtual sensor-based fault indicators," *Appl Therm Eng*, vol. 129, pp. 1292–1303, Jan. 2018, doi: 10.1016/j.applthermaleng.2017.10.013.

[19] W. V. Payne, J. Heo, and P. A. Domanski, "A Data-Clustering Technique for Fault Detection and Diagnostics in Field-Assembled Air Conditioners," *International Journal of Air-Conditioning and Refrigeration*, vol. 26, no. 2, Jun. 2018, doi: 10.1142/S2010132518500153.

[20] Z. Chen *et al.*, "A review of data-driven fault detection and diagnostics for building HVAC systems," *Appl Energy*, vol. 339, Jun. 2023, doi: 10.1016/j.apenergy.2023.121030.

[21] A. Smiti, "A critical overview of outlier detection methods," *Comput Sci Rev*, vol. 38, Nov. 2020, doi: 10.1016/j.cosrev.2020.100306.

[22] A. Boukerche, L. Zheng, and O. Alfandi, "Outlier Detection: Methods, Models, and Classification," *ACM Comput Surv*, vol. 53, no. 3, May 2021, doi: 10.1145/3381028.

[23] A. Mtibaa, "Refrigerant Leak Detection in Industrial Vapor Compression Refrigeration Systems," Université PSL, Paris, 2025.

[24] P. J. Rousseeuw and K. Van Driessen, "A Fast Algorithm for the Minimum Covariance Determinant Estimator," *Technometrics*, vol. 41, no. 3, pp. 212–223, 1999, doi: 10.2307/1270566.

[25] W. Liu, D. Cui, Z. Peng, and J. Zhong, "Outlier Detection Algorithm Based on Gaussian Mixture Model," in *2019 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS)*, Shenyang, China: IEEE, Jul. 2019, pp. 488–492. doi: 10.1109/ICPICS47731.2019.8942474.

[26] S. Ramaswamy, R. Rastogi, and K. Shim KAIST, “Efficient Algorithms for Mining Outliers from Large Data Sets,” in *SIGMOD/PODS00: ACM International Conference on Management of Data and Symposium on Principles of Database Systems*, Dallas Texas USA, 2000, pp. 427–438. doi: 10.1145/342009.335437.

[27] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, “LOF: Identifying Density-Based Local Outliers,” in *SIGMOD ’00: Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, Dallas Texas USA, 2000, pp. 93–104. doi: 10.1145/342009.335388.

[28] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise,” in *KDD’96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, Portland Oregon USA, 1996, pp. 226–231. [Online]. Available: www.aaai.org

[29] H.-P. Kriegel, M. Schubert, and A. Zimek, “Angle-based outlier detection in high-dimensional data,” in *KDD ’08: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, Las Vegas Nevada USA: ACM, 2008, pp. 444–452. doi: 10.1145/1401890.140194.

[30] F. Tony Liu, K. Ming Ting, and Z.-H. Zhou, “Isolation Forest,” in *2008 Eighth IEEE International Conference on Data Mining*, Pisa, Italy: IEEE, 2008, pp. 413–422. doi: 10.1109/ICDM.2008.17.

[31] H. O. Marques, R. J. G. B. Campello, A. Zimek, and J. Sander, “On the internal evaluation of unsupervised outlier detection,” in *SSDBM ’15: Proceedings of the 27th International Conference on Scientific and Statistical Database Management*, La Jolla California USA, Jun. 2015, pp. 1–12. doi: 10.1145/2791347.2791352.

[32] E. Schubert, R. Wojdanowski, A. Zimek, and H.-P. Kriegel, “On Evaluation of Outlier Rankings and Outlier Scores,” in *Proceedings of the 2012 SIAM International Conference on Data Mining*, Anaheim, California USA, 2012, pp. 1047–1058. doi: 10.1137/1.9781611972825.90.

[33] F. Guo, Z. Chen, and F. Xiao, “Fault detection and diagnosis of electric bus air conditioning systems incorporating domain knowledge and probabilistic artificial intelligence,” *Energy and AI*, vol. 16, May 2024, doi: 10.1016/j.egyai.2024.100364.

[34] T. Brudermueller, M. Kreft, E. Fleisch, and T. Staake, “Large-scale monitoring of residential heat pump cycling using smart meter data,” *Appl Energy*, vol. 350, Nov. 2023, doi: 10.1016/j.apenergy.2023.121734.

[35] Y. Zeng, H. Chen, C. Xu, Y. Cheng, and Q. Gong, “A hybrid deep forest approach for outlier detection and fault diagnosis of variable refrigerant flow system,” *International Journal of Refrigeration*, vol. 120, pp. 104–118, Dec. 2020, doi: 10.1016/j.ijrefrig.2020.08.014.

[36] S. Frank *et al.*, “Hybrid Model-Based and Data-Driven Fault Detection and Diagnostics for Commercial Buildings: Preprint,” in *2016 ACEEE Summer Study On Energy Efficiency in Buildings*, Pacific Grove, California USA, 2016. Accessed: Nov. 17, 2025. [Online]. Available: https://www.osti.gov/biblio/1290794

[37] Y. Zhao and L. Akoglu, “HPOD: Hyperparameter Optimization for Unsupervised Outlier Detection,” in *Proceedings of the Third International Conference on Automated Machine Learning*, Paris, France, 2024, pp. 1–24. Accessed: Nov. 17, 2025. [Online]. Available: https://proceedings.mlr.press/v256/zhao24a.html