LARGE SCALE DATA ANALYTICS FOR FAULT DETECTION AND DIAGNOSIS OF RESIDENTIAL HVAC SYSTEMS

A Dissertation

by

FANGZHOU GUO

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Bryan Rasmussen
Committee Members,	Xia Ben Hu
	David Claridge
	Richard Malak
Head of Department,	Andreas Polycarpou

August 2021

Major Subject: Mechanical Engineering

ABSTRACT

Residential heating, ventilation, and air conditioning (HVAC) equipment maintains the indoor environment with appropriate temperature and humidity levels. Meanwhile, it accounts for 51.3% of annual energy use and 40.1% of annual energy expenditures in the residential buildings in the U.S. However, residential HVAC systems often suffer from installation faults and operational faults leading to degradation in system capacity or even complete breakdowns, causing extra energy consumption and occupant discomfort. Fault detection and diagnosis (FDD) methods assist in identifying specific system faults, predicting gradual degradation and prompting necessary maintenance. Though plenty of researches have been conducted to develop FDD methods for commercial HVAC systems, relatively few researches focus on residential systems, mainly because FDD requires installation of additional sensors on each HVAC equipment, which is not cost-effective for the mass-produced residential systems.

This research fills this gap by developing statistics-based FDD methods to identify faults and monitor behavior changes simultaneously from a large number of residential HVAC systems using smart thermostat data. Two main approaches for fault detection and preliminary diagnosis are proposed in this research, namely: comparing operational features between multiple systems and monitoring the changes of operational features within each system. Following the idea of each approach, a few useful FDD algorithms are developed, including the setpoint tracking failure detector, inadequate capacity detector, control problem detector, and degradation trend detector. Additionally, the research provides general preprocessing procedures for the smart thermostat data, which could be applied to all fault detectors. The preprocessing procedures ensure the data is clean and critical features are extracted that representative of the operational conditions of each system.

The main body of this thesis presents each of the proposed detector. The setpoint tracking failure detector identifies degraded systems that cannot effectively regulate the indoor temperature around the desired setpoints. The inadequate capacity detector identifies systems with much lower cooling/heating capacity compared to other systems in the similar climate region, and in majority

of the time the degradation of system capacity is imperceptible for home occupants. The control problem detector identifies systems with abnormally high cycle frequency and setpoint error in a large population, which is usually caused by control faults. Lastly, the degradation trend detector is able to detect slow system capacity degradation over time and quantify the magnitude of the degradation.

Finally, the author proposes a few future research directions of FDD for residential HVAC systems. Possible research directions include (1) improving the performance of fault detectors through verified *in situ* faulty systems, (2) developing deep learning models such as the recurrent neural network and the Siamese network, and (3) incorporating additional features from limited numbers of low-cost sensors.

DEDICATION

To my mother, my father, and my grandparents.

ACKNOWLEDGMENTS

First of all, I would like to thank Dr. Bryan Rasmussen for his excellent mentorship. Initially when I joined the thermo-fluids control research lab (TFCL), I had very little research experience and basically knew nothing about Python and data science. You gave me tremendous support and helped me accumulate research experience gradually. Every time I had difficulties in my research, you brought up good ideas or possible solutions of problems. Besides, you also work on improving our writing skills, training our leadership, and assisting in our career development. With your mentorship, I made great progress in my research and learned a lot of useful skills. I will not forget your help in these four years. I also want to thank my committee members, Dr. David Claridge, Dr. Xia Hu and Dr. Richard Malak for their time and service on my committee.

Thank you to my friends in the lab: Austin, Chris, Deokgeun, Marcus, Arvind, Rafael, Taek-Soo, Saurabh, and Hussein. I remember the time when we discuss new research ideas and share our culture difference. Also, Austin and Chris helped me a lot and answered me many research questions at the beginning. We had a great time together.

Thank you to my parents and grandparents. Without your support, I do not have the courage to go abroad and pursue a Ph.D. degree. I know you care for me every day and hope me to eventually achieve my goal. As a young adult who never studied outside of my hometown, I felt I grew up to be independent and learned useful social skills.

Finally, thank you to my roommates and friends that I met in College Station. I remember the time we chatted and traveled together. You made me not feel lonely in a foreign country.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a dissertation committee consisting of Dr. Bryan Rasmussen (advisor), Dr. David Claridge, and Dr. Richard Malak of the Department of Mechanical Engineering, and Dr. Xia Ben Hu of the Department of Computer Science & Engineering.

Funding Sources

Graduate study was supported by the J. Mike Walker ‘66 Graduate Fellowship from the Department of Mechanical Engineering at Texas A&M University, and by Trane Technologies Inc. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Trane Technologies.

NOMENCLATURE

List of Abbreviations

HVAC	Heating, Ventilation, and Air Conditioning
FDD	Fault Detection and Diagnosis
AHU	Air Handling Unit
VAV	Variable Air Volume
CDD	Cooling Degree Day
MK	Mann-Kendall Test
PMK	Partial Mann-Kendall Test
MBBS	Moving Block Bootstrap
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

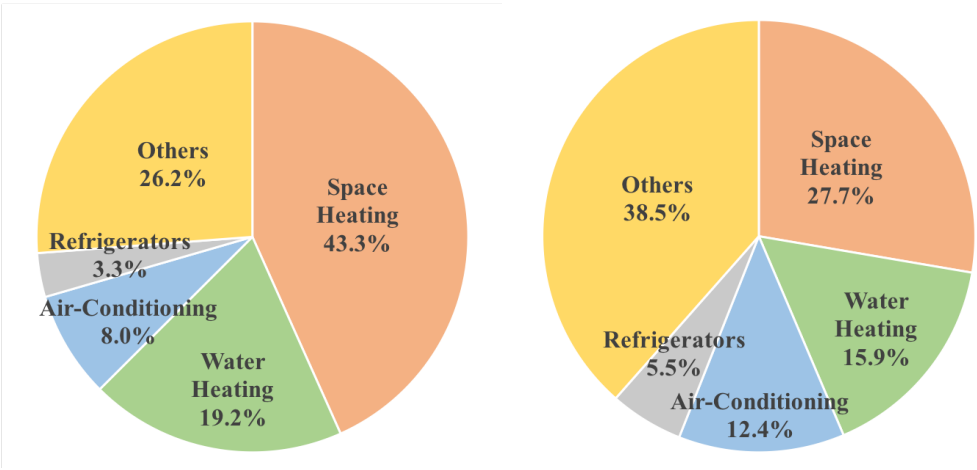
LIST OF TABLES

TABLE Page

1. INTRODUCTION

1.1 Background and Motivation

Residential Heating, Ventilation, and Air Conditioning (HVAC) systems have gradually become indispensable to install for residents, because it maintains the indoor environment with appropriate temperature and humidity levels. In 2009, 87% of households in the U.S. own air conditioning systems, which is a 31% increase compared to the year of 1978, and 99% of households own space heating devices [8]. Meanwhile, HVAC systems account for a large share of the residential households' total energy consumption. According to the U.S. Energy Information Administration (EIA) report in 2015 [1], the total household site end-use energy consumption in the U.S. is 3,671 TWh (319.34 billion dollars) each year. Figure 1.1 shows the percentage breakdown of the energy usage and expenditures. For a typical home, on average space heating consumes 43.3% of total energy and 27.7% of energy expenditures, and air conditioning consumes 8.0% of total energy and 12.4% of energy expenditures.



(a) Breakdown of energy usage (in TWh). (b) Breakdown of expenditures (in $).

Figure 1.1: Percentage breakdowns of residential total annual energy usage [1].

Although a large portion of household energy consumption and expenditures are devoted to space heating and air conditioning, often the energy is not used efficiently. Due to operational faults that may occur, residential systems on average function at least 17% below their rated efficiencies [9], causing unnecessary energy expenditures. [10] shows that, by fixing installation issues of residential HVAC systems including indoor airflow rate and refrigerant charge, 20.7 TWh per year (2.5 billion dollars) can be saved. Note that this situation is similar for commercial buildings. For example, [11] states that faulty systems waste an estimated 15% to 30% of energy.

Predictive and preventative maintenance through fault detection and diagnosis (FDD) plays an important role for systems to achieve the optimal operating efficiency and be fault-free. FDD ensures the systems are properly commissioned, and is able to identify specific operational problems, predict gradual degradation, prompt necessary maintenance, and reduce the time required to diagnose and repair systems. In this way, residential homes have lower energy bills every month and residents enjoy improved comfort.

Additionally, manufacturers, service companies, and the society benefit from FDD as well. Manufacturers establish a brand image for reliability and efficiency. Service companies gain customer loyalty by anticipating equipment problems and repairing them quickly and correctly. The society has reduced energy consumption and the corresponding emissions.

1.1.1 Classification and Processes of FDD Methods for Building Energy Systems

General FDD methods for building energy systems are classified into three categories by [2]: quantitative model-based, qualitative model-based, and process history-based (Figure 1.2). Similar classifications can also be found in review articles [12, 13]. Although sometimes the boundaries between these three categories can be blurred, generally quantitative and qualitative model-based methods is based on *a priori* knowledge and the models are built on first principles. In contrast, process history-based methods rely on historical data for training, testing, and analysis, and can be further classified into black-box models and grey-box models. The black-box models are driven completely empirically by process data, while the grey-box models require both first principles to build the mathematical form of terms and the process data to determine parameters.

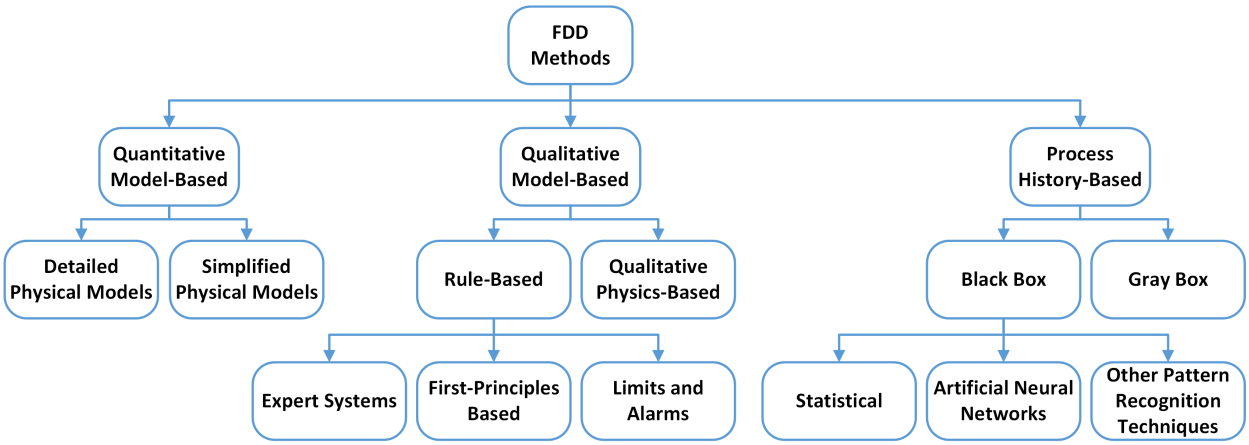


Figure 1.2: Classification of FDD methods for building energy systems [2].

Similar to the above review article, [14] classifies the FDD methods for building energy systems into data driven-based and knowledge driven-based. The data driven-based methods are the techniques used in this project for smart thermostat data. They detect faults through finding patterns of the extracted features automatically from historical data, similar to the definition of process history-based methods. The authors of [14, 15] reviewed the artificial intelligence (AI) methods from 1998 to 2018, which includes all data driven-based methods and a partial of knowledge driven-based methods. The data driven-based methods are abundant, accounting for 79% of the publications. The authors categorized the data driven-based methods into three groups:

- **Classification-based methods** such as support vector machine and artificial neural network, which are used when both faulty and fault-free data with true labels are available. These methods first train a model to classify operational data into a normal group and several fault groups (there are also one-class classification-based methods that only identify data in one group), and then apply the trained model to the monitored building energy system.

- **Unsupervised learning-based methods** such as principle component analysis (PCA), clustering methods, and association rule-based methods, which are used when true labels are not available. These methods recognize the patterns in both faulty and fault-free conditions directly from the unlabeled data and may need domain expertise for further diagnosis.

- **Regression-based methods** such as linear regression, artificial neural network, and support vector regression, which can be used when only fault-free data is available. These methods first build benchmark models using fault-free data, and then in application they compare the residuals between the monitored operational feature values and the values calculated from the benchmark models to determine whether a fault exists.

The author of [3] reviewed 197 publications until 2016, and found that process history-based methods are abundant as well. In total 62% of the FDD studies are process history-based methods (data driven-based methods), and in particular the black-box models account for 72% among all process history-based methods, which has become the most common technique. [3] also counted the applications of FDD methods on building energy systems. Shown in Figure 1.3, most published FDD methods for HVAC systems focus on large commercial systems. Approximately 42% of FDD methods are developed for variable air volume air handling units (VAV-AHUs), 17% developed for chillers and cooling towers, and 12% developed for overall buildings. Only about 16% of FDD studies are associated with air conditioners and heat pumps in small commercial and residential buildings, but the majority still focuses on rooftop packaged units (RTU). Research effort on residential HVAC systems is insufficient.

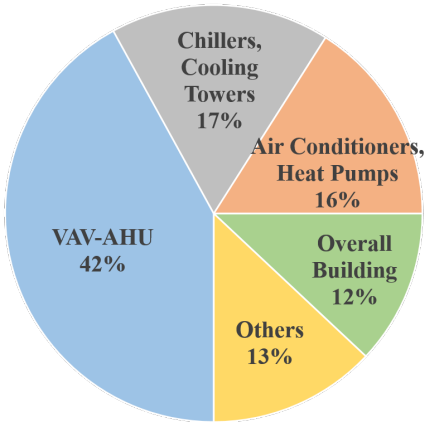


Figure 1.3: Applications of FDD methods on building energy systems [3].

[16] reviewed the FDD methods for residential air conditioning systems. In particular, the author summarizes the general FDD processes using data-driven and knowledge-driven models for air conditioning systems, which consists of the following five sequential steps:

1. Features are extracted from the measured data. A feature could be directly measured, or derived from the difference between measurement and models. For example, the superheat is a feature directly measured from temperature and pressure sensors, and the residual between measured superheat and modeled fault-free superheat is a feature derived from the differences.

2. A steady state detector is applied and steady-state operational data is obtained. This is because most FDD models are developed for steady-state operating conditions.

3. Fault-free FDD models are established to estimate the data features when the system has no fault. The model can be either data driven-based or knowledge driven-based, and the fault-free feature values are often used to calculate the residual.

4. A fault detection classifier is applied to label faulty data. For example, a common classifier could be a simple threshold on residuals.

5. A fault diagnosis classifier is applied to further classify the detected faults into several categories. Note that sometimes detection and diagnosis could be combined to a single step if features are decoupled such that each of them corresponds to a specific fault.

Similarly, [12, 13] summarized the general FDD procedures for building energy systems as data preprocessing (e.g. data cleansing), feature extraction, fault detection, and fault diagnosis. The majority of publications focus on the fault detection and diagnosis methods, but often fail to mention the preprocessing procedures that are also critical.

1.1.2 FDD Opportunities on Residential HVAC Systems Using Thermostat Data

Although FDD-related researches on building energy systems are increasing rapidly, relatively few published research efforts have been focused on residential split systems, mainly because FDD

generally requires installation of additional sensors on the equipment [17]. For mass-produced residential units with small profit margins, adopting additional sensors may not be cost-effective. Therefore, researchers have tried to perform FDD with limited sensor information to lower the cost. For example, [18] developed an FDD method and obtain key operating parameters such as airflow rate, cooling capacity, system efficiency, and refrigerant mass flow by only five temperature sensors, one humidity sensor and one current sensor all installed on the indoor unit. [19] detected the refrigerant leakage problem only using five temperature sensors installed on the air side and the heat exchangers.

An alternative data source for analysis and diagnostics in residential HVAC systems is from smart thermostats. In recent years, smart thermostats have been widely adopted by home owners, because they offer remote control and monitoring options (e.g. via smart phones), include learning algorithms to improve occupant comfort, and provide feedback to users regarding energy use. Meanwhile, the huge amount of data recorded by smart thermostats provides distinctive opportunities for FDD approaches, or even promotes a potential paradigm shift.

Process history-based (or data driven-based) models are especially suitable for the large-scale smart thermostat data analysis. Figure 1.4 illustrates the whole process of fault detection and diagnosis using thermostat data [4]. Smart thermostats in every residential home continuously record and upload real-time data to a server, forming an extensive database on the cloud. Then on a remote platform, the FDD methods process the data by extracting useful features and using statistical methods or artificial intelligence methods for comparative analysis. Next, systems with detected faults are ordered based on the severity of faults. Dealers and customers can be notified of faulty systems, and service technicians can be dispatched to repair systems based on priority and with a preliminary diagnosis in hand. By this means, anomalous systems are identified quickly by FDD algorithms and serviced to maintain occupant comfort and energy efficiency.

Generally, a smart thermostat database can store configuration parameters and operating data of thousands of residential HVAC systems. In this project, one cloud-based smart thermostat database is provided by the Trane Technologies Inc. Until April 2021, the database stores the data of approx-

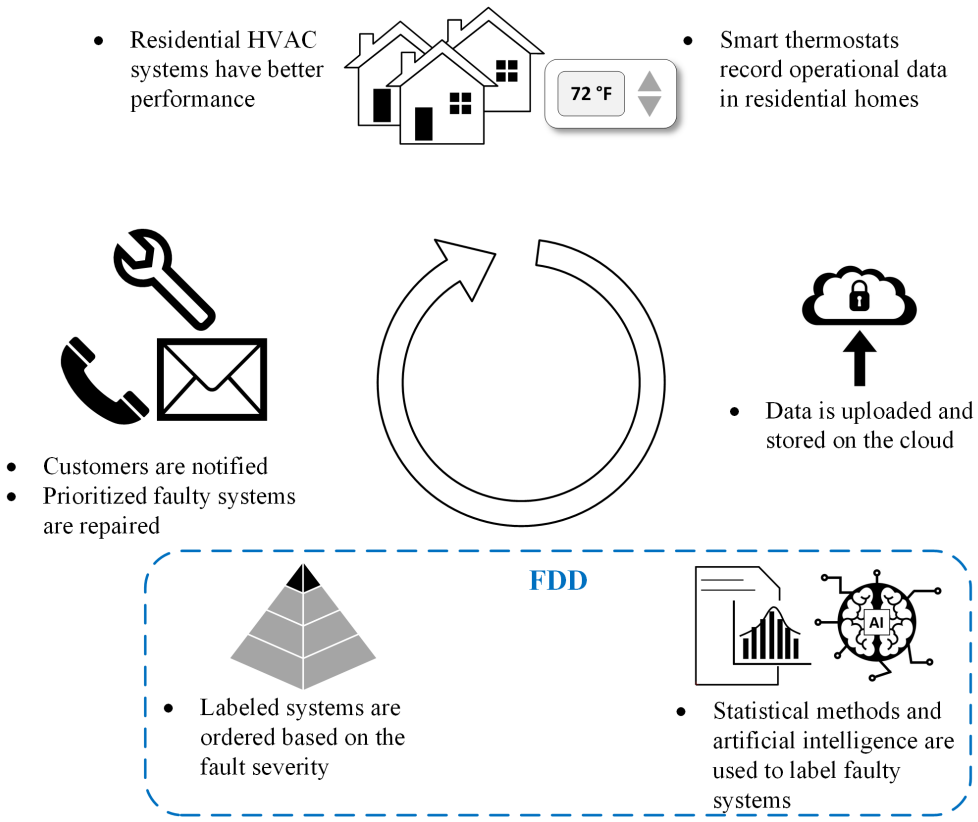


Figure 1.4: General process of fault detection and diagnosis with smart thermostat thermostat data [4].

imately 560,000 systems, and it is ever-expanding because the Trane Technologies is continuously adding more systems. The essential configuration parameters for each device include indoor and outdoor unit types, unit stages, and device locations. The essential operating parameters for each device include the indoor temperature, cooling/heating temperature setpoints, heating/cooling operating stages, and the outdoor temperature obtained from a third party. Table 1.1 shows a full list of useful configuration parameters and operating parameters.

System repair and maintenance ensures systems to operate normally. In the warranty claim database from the Trane Technologies, in total 39,967 repair claims from 22,245 systems are reported from 2015-01-01 to 2020-03-31. Note that the warranty claim data does not include refrigerant recharge records. Figure 1.5 classifies the reported faults into four main categories and an unknown group (i.e. no fault description). The most common type of fault comes from the

Table 1.1: List of useful parameters in the smart thermostat database provided by the Trane Technologies.

Configuration parameters	Device identity, Outdoor unit type*, Outdoor unit tonnage*, Outdoor unit stages*, Indoor unit type**, Indoor unit stages**, Indoor blower type**, Air-conditioned zone count, Customer state, Customer zip code, Dealer name.
Operating parameters	Cooling/Heating setpoint, Humidity cooling/heating setpoint, Indoor/Outdoor relative humidity, Indoor/Outdoor temperature, Fan on†, Airflow percentage††, Outdoor cooling stage†, Outdoor heating stage†, Delivered percent speed††, Indoor heating stage†, Delivered modulation percentage††.

* The outdoor unit refers to the compressor, heat exchanger, and compressor fan that located outside of the house.
** The indoor unit refers to the heat exchanger, expansion valve, indoor fan, and electrical/gas furnace/hydronic heating equipment that located inside of the house.

† This parameter is for non-variable speed equipment.

†† This parameter is for variable speed equipment.

electrical failures, then from the mechanical failures and pipe leaks. Compared to the electrical failures which cause sudden system breakdowns, researchers are more interested in the mechanical faults and pipe leaks, because they are difficult to detect, usually remain intact for a long period of time imperceptible to occupants, but often have signs that are detectable before complete system breakdowns.

However, FDD using smart thermostat data is still a relatively new field of research, and analyzing data for thousands of systems simultaneously is also challenging. Aiming to fill this gap, this study researches into possible solutions for using smart thermostat data in fault detection and diagnosis of residential HVAC systems.

The next section reviews the FDD-related research using thermostat data. Because relatively few studies in this area are published, the section will detail each method, and also include some process history-based models for purposes other than fault detection.

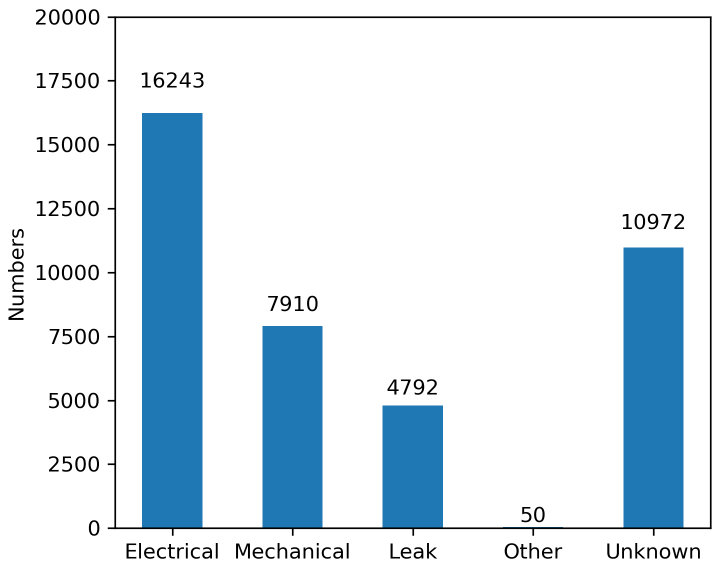


Figure 1.5: Type of faults reported for residential HVAC systems in the Trane warranty claim database.

1.1.3 FDD-related Research Using Thermostat Data

Although a lot of FDD methods have been proposed in recent years, only a few studies using thermostat data for FDD of residential HVAC systems are found in the literature. Most of them are process history-based with statistics models applied. In practical applications, these methods have the advantage of no need to install additional sensors.

AutoRegressive (AR) models and AutoRegressive Moving Average (ARMA) models are common process history-based approaches to fault detection using thermostat data. The models are often used to predict the indoor temperature using its own previous observations as well as other covariates such as the outdoor temperature and the status of HVAC systems. [20] developed an AR model for refrigerant leakage detection of refrigeration units in retail outlets, and leakage was flagged when the measured room temperature went beyond the estimated decision boundary obtained by the AR model for consecutively 36 hours. The author also mentioned that the flagged fault could help detect other issues in the refrigeration system beyond refrigerant leakage. In the AR and ARMA models developed by [21], the coefficients obtained by the recursive least-squares

estimator are compared between fault conditions and normal operating conditions, and change in coefficients is observed in fault conditions. Although this study investigates into both hard faults (e.g., total system failure) and soft faults (e.g., control malfunction, ventilation fan blockage), the developed models are only validated by simulated faults.

However, AR and ARMA models have drawbacks for FDD when generalizing to thousands of HVAC systems. Thermostat data retrieved from a fraction of systems may not have a satisfying fitting results due to several reasons. For example, the existence of occupant behavior is difficult to predict and can strongly affect the accuracy of the fitted models. Also, all kinds of time-series models need fault-free data of each particular residential home for training. However, fault-free data is often hard to collect without *in situ* testings to validate.

Except fault detection, AR models are also found in occupant behavior analysis. [22] developed an AR model aimed at calculating the percentage reduction in heating costs when the heating temperature setpoint is decreased or timer settings are changed, in order to provide personalized home heating advice to households.

Other process history-based thermostat data analysis includes [23, 24, 25], which focus on occupant behavior, and [26], which focuses on system runtimes. [23] used discrete-time and discrete-event Markov logistic regression models to relate probability of a setpoint change to the indoor temperature, outdoor temperature and relative humidity. The data is collected in two office buildings, and the regression model as well as other statistics provides suggestions on proper setpoint adjustments in the buildings during all seasons. [24] performed a statistical analysis on thousands of thermostats, investigating into 1) thermostat behavior changes based on seasonal variation, climate zones and utility pricing, and 2) discernible user types in thermostat operation. [25] studied the energy consequences of setting programmatic smart thermostat schedules but with frequent manual overrides. The author concluded that frequent manual overrides do not lead to poor energy performance for most residential HVAC systems, and are not as detrimental as commonly believed. [26] performed runtime estimations on over 7,000 systems and concluded that runtimes can vary

considerably between homes resulting from covariates besides climate, such as system sizing and user operation.

Although most analyses are process history-based, physics-based methods can be found as well. As an example, [27] used simple physical relationships to estimate the characteristics of a house, namely thermal resistance (R), thermal capacity (C), and thermal constants (RC). By comparing these features between homes, homeowners would be aware of the insulation level of their houses and whether the insulation should be improved.

Among all these studies, [23, 24, 26, 27] conducted multiple comparisons between systems using features extracted from history data records. Still, they focused on either identifying inefficient home construction or analyzing occupant behavior, rather than trying to detect mechanical faults of HVAC systems. Only [20, 21] proposed fault detection methods using thermostat data, but still the former one focused on refrigeration units and the latter one lacked field data validation.

One reason for the limited researches on thermostat data is that large thermostat datasets are not easily accessible to many researchers. Notice that [24, 26] used the cloud-based large scale 'Donate Your Data' dataset provided by the thermostat manufacturer Ecobee Inc., while the rest of them used relatively small scale datasets. As the 'Donate Your Data' dataset having been released to various research partners including academic labs, non-government organizations and research institutes, we believe plenty of other studies on FDD using thermostat data will be available in the next few years.

1.2 Outline of the Dissertation

So far, this introduction first presented the motivations behind developing fault detection and diagnosis (FDD) methods for building energy systems. Then, in Section 1.1.1 the categories and processes of FDD methods used in relevant researches for building energy systems are discussed. As to FDD methods developed for residential HVAC systems, Section 1.1.2 discussed the obstacles for using data collected from installed additional sensors, and proposed to use smart thermostat as an alternative data source. Based on this new idea, Section 1.1.2 then briefly introduced the cloud-based smart thermostat database being analyzed in this project. After that, Section 1.1.3

conducted a full literature view of the FDD-related research using thermostat data. Currently, large-scale thermostat data analysis for fault detection and diagnosis of residential HVAC systems is insufficient and deserves further studies. This dissertation will discuss the effective methods for thermostat data analysis and detail the developed FDD algorithms in this project.

The second chapter summarizes all of the contributions. It first presents the fundamental ideas in this study to perform fault detection and diagnosis. Then, the chapter discussed data selection, preprocessing, and each statistics-based fault detector. In total five fault detectors have been developed in this whole project, and four of them are major research products for the author. These four fault detectors will be presented from Chapter 3 to Chapter 7. Each of the five chapter is one research article, and the synopsis of each article is presented at the first section of each chapter. Finally, the last chapter will summarize the main contributions of this research and discuss possible directions for future work.

2. SUMMARY OF CONTRIBUTIONS

This chapter summarizes the contribution from the six papers (Papers A to E) included in this dissertation. These papers propose a few practical fault detection and diagnosis (FDD) methods for residential air conditioning systems. Paper A is published, Paper B, C and D are under review, and Paper E will be submitted to appropriate journals soon after. This chapter is organized as follows. First of all, two fundamental ideas in the field of thermostat data fault detection and diagnosis are introduced, namely multi-system feature comparison and single-system change detection. Then, a few recommendations for data selection considering climate zones, time period, quantity and type of systems are discussed. Next, the general approaches to preprocess selected data are presented, including data cleansing, mode labeling, and feature extraction, etc. After the preprocessing procedures, only clean and useful data is left, so in the following step FDD methods can be applied effectively. In total four statistics-based FDD methods are developed in this study, and the remainder of this chapter discusses their algorithms and the detectable types of faults. Note that case studies and validation of each method through real faulty systems are not included in this chapter, but readers can refer to the respective papers for details.

2.1 Fundamental Ideas of Performing FDD with Smart Thermostat Data

As described in the first chapter, the attributes of smart thermostat data is different from data collected from installed sensors on the equipment mainly in three aspects: (1) the smart thermostats only record limited information such as indoor/outdoor temperature, temperature setpoint, and heating/cooling operational mode, while the sensor data includes specific temperature, pressure, and volume flow rate information in the vapor compression cycle that can be used to calculate more important operational features (e.g., cooling capacity, coefficient of performance); (2) since smart thermostats are connected to the online server and upload data continuously, the server can receive huge amount of data from a large number of HVAC systems for analysis, while the sensor data are

often collected from one single system; (3) true labels of thermostat data is often unavailable, while the sensor data in faulty conditions can be obtained by manually imposing faults on the equipment.

Based on the above pros and cons of thermostat data compared with the sensor data, the procedures of traditional FDD methods discussed in Section 1.1 are difficult to apply to thermostat data. Therefore, some new research ideas that are able to analyze big data meanwhile avoiding the needs of true labels and detailed sensor data are required. This study proposes two basic principles suitable for thermostat data, which are multi-system feature comparison and single-system change detection. All of the fault detectors developed in this study follow one of the two approaches. The most important characteristic of these two approaches is that, because true labels are unknown, they try to detect faults through finding anomalous operational behaviors. Shown in Table 2.1, the anomalous behaviors that closely related to faults are identified by domain expertise, and each operational behavior is characterized by features extracted only from the thermostat data. In this way, each type of faults can be quantified by corresponding features. Although real faulty systems may not exhibit any abnormal behavior, systems exhibiting abnormal behaviors are likely to be faulty. That is to say, these two approaches flag systems that have high possibility to be faulty, but unfortunately not all faulty systems can be detected.

Table 2.1: Examples of abnormal behaviors associated with faults and examples of features characterizing the behaviors.

Fault type	Abnormal behaviors	Possible features
Control faults	High setpoint error	Setpoint error
	High compressor cycle frequency	Cycle frequency
Inadequate Capacity	Frequent setpoint tracking failures	Severity of tracking failures
	High duty cycle	Duty cycle
Refrigerant Leakage	Continual degradation of heating/cooling performance	Change of duty cycle
		Change of daily operating hours

Multi-system feature comparison takes advantages of the accessibility of thermostat data from a large number of residential systems. The FDD methods following this idea can be categorized to unsupervised outlier analysis. In this approach, most systems are assumed to be operating normally and are fault-free. Devices in the same group are compared with each other by using features. Common behaviors occurring among the majority of systems are considered as normal behaviors, but otherwise a few abnormal behaviors are associated with faulty behaviors. The systems exhibiting these abnormal behaviors in operation are labeled as faulty. Detectors following the idea of multi-system feature comparison are represented in Subsection 2.4.1 and 2.4.2. It can be summarized to five steps:

1. Researchers are supposed to select a large group of systems with the same unit type and operating under similar climate conditions. This step is called data selection, and will be detailed in Section 2.2.

2. From each system, crucial features that can characterize the system operational behavior are extracted. Examples include the setpoint error, cycle frequency, duty cycle shown in Table 2.1. Each individual feature can be treated as a univariate, and their combinations can be treated as multivariate.

3. The statistical distribution of each univariate and multivariate is generated using the feature values from all systems. The distributions of these features represent their frequency of occurrence in this large group of systems.

4. From the statistical distributions, the abnormal behaviors that related to faults are identified. To place a definite boundary between the faulty and fault-free behaviors, a pre-specified threshold should be provided. This study recommends to use a relatively low pre-specified threshold (e.g., top 1% of the univariate values are most likely to be related to faults) to avoid falsely labeling normal systems to be faulty.

5. Finally, systems identified with different faulty behaviors are classified as faulty. The faulty behaviors can also provide a preliminary diagnosis of the specific type of faults.

In contrast, single-system change detection takes advantages of the long historical records of thermostat data. The current performance of a system is compared with its own performance before. If degradation of performance is detected, then the system will be flagged as faulty. Because the system performance is characterized by features, generally statistical tests are required to study the variations of feature values. Detectors that follow this idea are represented in Subsection 2.4.3 as well as literature [28]. The general processes are as follows:

1. Same as the multi-system feature comparison, the first step is data selection. But since this approach only processes one system each time and does not consider the relationships between systems, researchers can select systems in arbitrary numbers and types.

2. Features that can characterize the change of system performance are extracted. Examples include the daily cooling effort and the steady-state cooling effort shown in Table 2.1.

3. Statistical tests are performed to detect whether the extracted features have changes or trends considering all covariates that have correlation with the extracted features. The detected changes or trends often indicate gradual capacity degradation caused by refrigerant leakage.

Figure 2.1 shows the whole FDD process of the above two approaches, including data selection, data preprocessing, and applying statistics-based FDD methods. Data selection is discussed in Section 2.2, and the retrieved data being used to exemplify the methodology of each detector can be found in Paper A, B, D, and E respectively. Data preprocessing is summarized in Section 2.3. It mainly consists of three steps, including data cleansing, partitioning and classification, and feature extraction. Section 4.3.2 in Paper B details a rule-based filter that is able to cleanse anomalous data due to communication and sensor integrity issues. Previous work by Rogers et al. [6] presents a mode labeling algorithm to partition and classify the raw thermostat data into five behavioral modes, such that more useful features can be extracted. The specific features applied in each detector are detailed in Paper A, B, D, and E respectively. Finally, statistics-based FDD methods using the idea of multi-system feature comparison and single-system change detection are presented in Section 2.4.

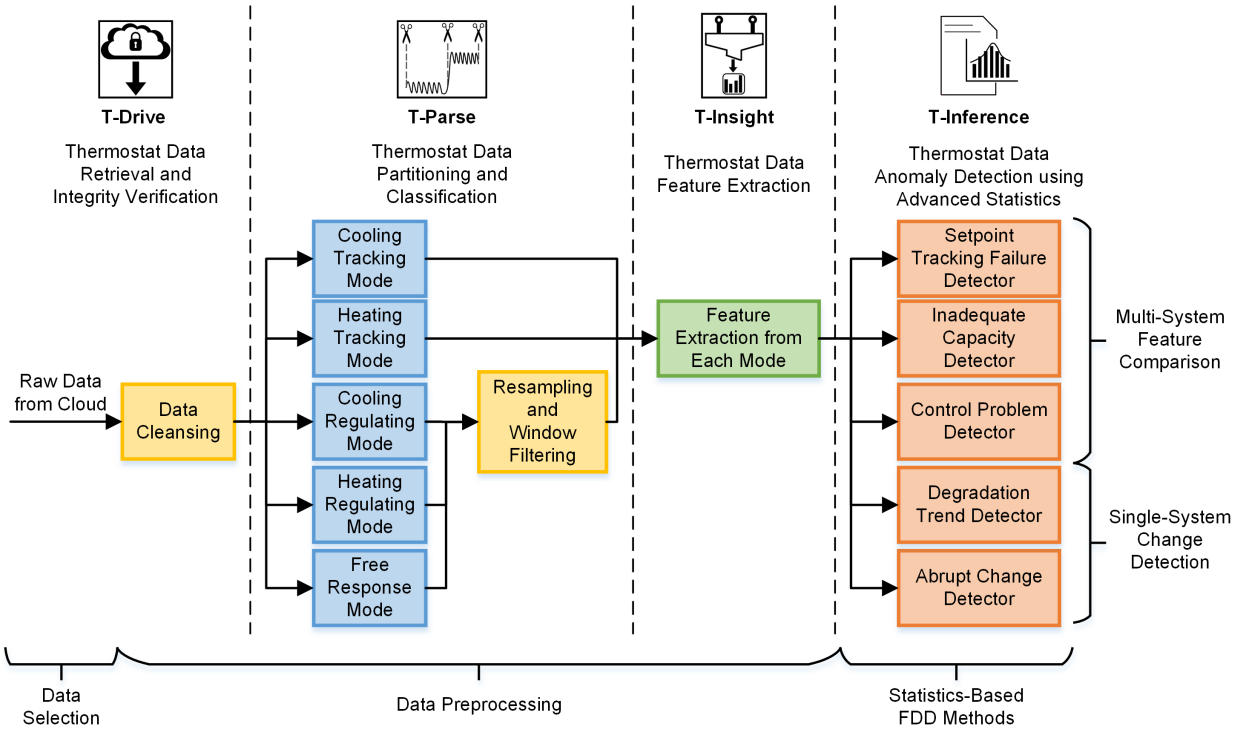


Figure 2.1: Flowchart of the whole FDD process.

2.2 Recommendations for Data Selection

This section provides some recommendations to retrieve data from the database before performing analysis. As mentioned in the above section, selecting appropriate data for analysis is crucial in the multi-system feature comparison methods, but not important in the single-system change detection. This is because in the multi-system feature comparison methods, features are compared *between* a large group of systems instead of *within* systems, so systems in the group must share enough similarity and be comparable. Generally, three aspects are considered to ensure the comparability, namely: climate zones, time periods, and system configurations. Some details can also be found in Section 3.2 of Paper A and Section 4.3.1 of Paper B.

First, systems should be selected within similar climate zones because the local climate and house size determines the size of a system. Figure 2.2 shows the International Energy Conversion Code climate regions and moisture regimes [5]. The climates in the U.S. are classified to seven zones and three moisture regimes. Comparing the performance of systems between two much

different climate zones will lead to wrong conclusions. For example, the air-conditioning systems installed in Zone 2A generally have larger rated capacity compared to the systems installed in Zone 7A for houses with similar sizes, because the summer season in Zone 7A is much cooler and shorter than in Zone 2A. Thus, when systems between these two climate zones are compared together, the systems with smaller capacity in Zone 7A tend to work much harder at the same ambient conditions and therefore seem undersized. However, this conclusion is wrong because in Zone 7A comfort can be adequately maintained with smaller capacity systems.

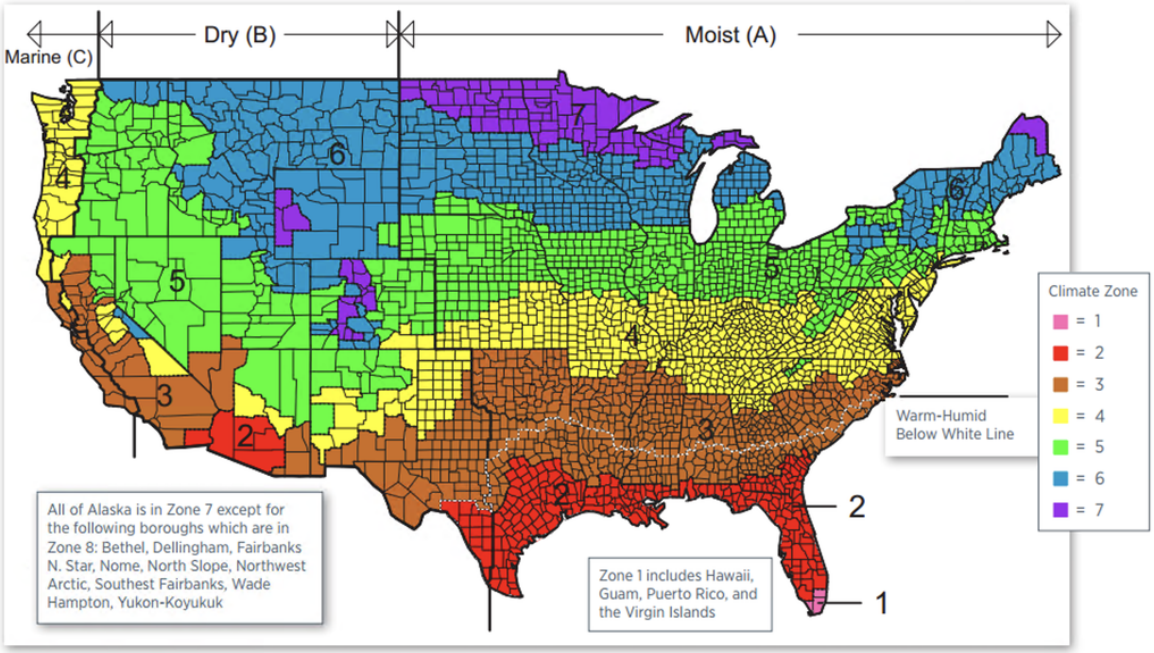


Figure 2.2: International Energy Conservation Code climate zones and moisture regimes [5]. Climate zones: 1-7 in different colors. Moisture regimes: marine, dry, and moist from west to east.

Second, the appropriate time scale needs to be determined. Using only one or two weeks of data can be computationally inexpensive, but has significant drawbacks because it generally will not accurately capture the full range of ambient conditions. Additionally, during short periods the statistical analysis is liable to be affected by some irrelevant covariates. For example, if an occupant left a window open, the air conditioner may seem not to be working properly. These

types of occupant or behavioral faults are not the primary focus of this paper, and their effects can be lessened using data from a relatively long time period. Shorter time periods may also limit the amount of useful data for analysis (e.g. a two-week period where the system operates infrequently). Conversely, time periods exceeding one year may also not be appropriate, as the operating condition of a system may have changed significantly due to degradation or repairs. A time window lasting for around one season (i.e. three months) is preferred in this research, since the data is abundant for analysis, and meanwhile the weather usually does not change too much within one season. If cooling units are analyzed, then the summer season will be the best choice; if heating units are analyzed, then the winter season will be the best choice; if the home thermal characteristics are analyzed, then the spring and fall seasons when the cooling/heating units are off in the majority of time will be good choices.

Third, the database contains air conditioning systems with plenty of different configurations. For example, the compressors of heat pumps can be single-stage, dual-stage and variable-speeds. Also, the electrical and gas furnace heating units can have from one to three stages, or even modulating between zero and full capacity. Because the operational behaviors and many data features depend on system type (e.g. on/off cycle frequency), each system configuration should be analyzed independently. This dissertation focuses on single-stage cooling systems, but the proposed methods are also applicable to other system configurations.

2.3 General Data Preprocessing Approaches

This section discusses the general preprocessing procedures for raw thermostat data. The general procedures are cleansing, partitioning and classification, and feature extraction. Relevant contents can be found in Section 3.3 of Paper A, Section 4.3 of Paper B, Section 6.5 of Paper D, and Section 7.3.1 of Paper E. Sometimes additional procedures are also necessary in a particular analysis. For example, Paper B presents the data separation and feature transformation that are indispensable for feature comparison between systems.

2.3.1 Data Cleansing

After raw thermostat data is retrieved from the online database, it first goes through the data cleansing process. This preprocessing step is necessary before applying all fault detection methods.

The raw smart thermostat data is event-based, which means that the database is updated only when a new event occurs. For example, when the indoor/outdoor temperature changes more than a quantized amount, the cooling/heating setpoint is changed, or the system starts/stops. Each event update records the timestamp and the changing variable. Sometimes, however, in this event-based database sensor integrity issues and network communication issues can occur (e.g. the sensor is unplugged, the thermostat goes offline or otherwise fails to communicate data). These issues could strongly affect the analytical process. For example, if a setpoint change event is missed, then the thermostat may seem to control the indoor temperature to value different from the desired setpoint; if a cooling off event is missed, then the system may appear to continue running for a long time. Data cleansing is performed by a rule-based filter that is capable of removing data with the following faults:

- The indoor/outdoor temperature climbs above 130°F or goes below 0°F (occasionally the thermostat will list the temperature as $2^{15} = 32,768$°F when the thermostat goes offline). The anomalous temperature data is removed.

- The indoor/outdoor temperature increases or decreases by more than 20°F between two consecutive updates. Or the indoor/outdoor temperature has a positive/negative spike of more than 5°F in three consecutive timestamps. This is often caused by sensor integrity issues. The anomalous temperature data is removed.

- The indoor temperature has not updated for more than four hours, or the outdoor temperature has not updated for more than five hours. If the sensor fails to communicate for a period of time and then suddenly resumes, then all data in these communication failure periods is removed.

- The system maintains a constant temperature with little variance (tightly), but the setpoint error is large. In this case the reported setpoint is likely incorrect, and a setpoint correction algorithm is applied to find the real setpoint values.

- The system runs for more than 24 hours. In most cases this is caused when the system "off" signal is not received. Because the exact system runtime is unknown, all data in the on-cycle is discarded.

2.3.2 Data Partitioning and Classification

After erroneous data is removed, but before extracting specific data features for analysis, a critical process termed mode labeling is applied to partition raw thermostat data and classify the data sections. The mode labeling algorithm for thermostat data is proposed in [6], and is used in all of the proposed fault detectors except the daily analysis of degradation trend detector. This algorithm transforms the *operational* modes of systems (i.e. cooling, heating, off) into the following three *behavioral* modes:

- A system is said to be in a regulating mode when the system cycles on and off and maintains a relatively constant temperature. The system may have both cooling and heating regulating modes.

- A system is said to be in a tracking mode when the system operates continuously to reach the desired setpoint. Again, both the cooling and heating functions have tracking modes. Tracking modes are typically activated after a decrease in cooling setpoint or an increase in heating setpoint.

- A system is said to be in a free response mode when neither setpoint is active. A free response mode may occur after a setpoint change (i.e. a rise in cooling setpoint or a drop in heating setpoint) or when the cooling and heating loads are low such that the indoor temperature remains between heating and cooling setpoints for an extended period.

Figure 2.3 provides an example of the behavioral modes [6]. In Figure 2.3a, the system is in cooling tracking modes when the indoor temperature drops significantly, and the system is in cooling regulating modes when the indoor temperature is maintained relatively constant. In Figure 2.3b, the system is in a free response mode when indoor temperature rises above the heating setpoint and then re-enters a heating regulating mode when the indoor temperature drops to the setpoint.

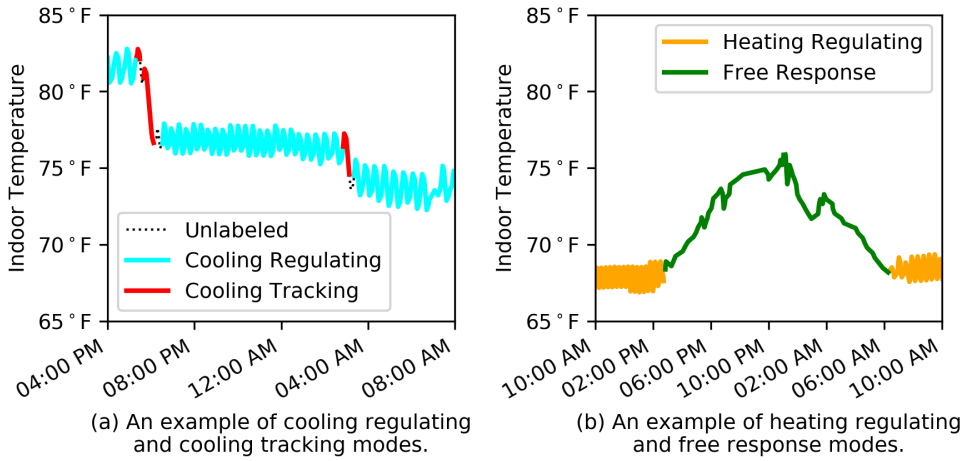


Figure 2.3: An example of the three behavioral modes [6].

After this transformation, more useful features can be extracted and analyzed in the fault detectors, which will be discussed in the next subsection. Note that [6] only provides an algorithm to label single-speed air conditioning systems with behavioral modes, and future work still needs to focus on generalizing this method to multi-speed and variable-speed systems.

2.3.3 Feature Extraction

Table 2.2 shows the features that extracted from each behavioral modes, and Table 2.3 provides the definition of each feature. Due to the distinct behavior in each mode, some features are useful in every mode, while some others are only applicable for a specific mode. The last column in Table 2.2 shows the features that use data directly from all modes (no mode labeling is needed).

Table 2.2: Features extracted from behavioral modes.

Symbol	Feature name	Behavioral modes			
		Tracking	Free Response	Regulating	All
ΔT_{oi}	Temperature difference	X	X	X	X
$\Delta \omega_{oi}$	Humidity difference			X	
ΔT_{ii}	Indoor temperature rise	X	X		
Δt	Cycle duration	X			
D	Duty factor			X	
L	Load factor	X		X	
E_c, E_h	Cooling/Heating effort			X	X
F	Cycle frequency			X	
T_{spe}	Setpoint error	X		X	
DH	Degree hour	X			
$T_{i,max}$	Max indoor temperature	X			

2.4 Proposed Statistics-Based FDD Methods

Through data preprocessing, raw thermostat data is partitioned and classified into three behavioral modes, and important features are extracted from each mode. After that, statistics-based methods can be applied to identify faulty systems using the behavioral features. In this project, five detectors have been developed by Dr. Rasmussen's research group, namely: setpoint tracking failure detector (Paper A), control problem detector (Paper B and C), inadequate capacity detector (Paper B and C, and improvement in Paper E), degradation trend detector (Paper D and E), and abrupt change detector (proposed by Dr. Rogers in literature [28]). This section first introduces the setpoint tracking failure detector, which studies the system transient behavior in the cooling/heating tracking mode. Then, a multivariate fault detection method which performs the inadequate capacity detector and the control problem detector altogether is discussed. This multivariate method studies the cooling/heating regulating mode. Finally, the degradation trend detector is presented with three different approaches, including hourly analysis, daily analysis, and a metric to monitor long-term changes. Both hourly analysis and the long-term monitoring analyzes fea-

Table 2.3: Definition of Features.

Symbol	Unit	Definition
ΔT_{oi}	[°F]	Temperature difference between home (indoor) and ambient (outdoor); defined as $\Delta T_{oi} = T_o - T_i$.
$\Delta \omega_{oi}$	[kg H_2O/ kg dry air]	Humidity difference between home and ambient; defined as $\Delta \omega_{oi} = \omega_o - \omega_i$.
ΔT_{ii}	[°F]	Indoor temperature rise (or drop) in a given time period: $\Delta T_{ii} = T_i(t_2) - T_i(t_1)$ in a time period from t_1 to t_2.
Δt	[hr]	Cycle duration. The amount of time that a system remains on.
D	[%]	The portion of time that the system is running, and this is typically defined for a given time period. For example, a duty factor of 40% in a 1-hour period denotes that the system ran for 24 minutes in that hour. However, a variable speed system may have a duty factor close to 100%.
L	[%]	The portion of full capacity at which the system runs. When defined over a given time period, this is the average portion of full capacity that the system runs at *when the system is on*. A single-stage system essentially always has a load factor of 100%.
E_c, E_h	[%]	Defined by the product of D and L, which denotes the overall portion of capacity for the given time period. Variable speed systems will typically have a high duty cycle and a low load factor, while single stage systems will have high load factor and low duty factor.
F	[cycle/hr]	The number of start/stop cycles per hour.
T_{spe}	[°F]	The error between indoor temperature and setpoint; defined as $T_{spe} = T_i - T_{spc}$ for cooling and $T_{spe} = T_{sph} - T_i$ for heating.
DH	[°F·hr]	The integral of setpoint error, during a time period from t_1 to t_2; defined as: $DH = \int_{t_1}^{t_2} \max(0, T_{spe}) dt$.
$T_{i,max}$	[°F]	$T_{i,max}$ is defined as the maximum indoor temperature during a cooling tracking period, and correspondingly $T_{i,min}$ is defined as the minimum indoor temperature during a heating tracking period.

tures in the regulating mode, but the daily analysis considers all of the behaviors. Note that this dissertation only analyzes cooling systems, and researches on heating systems are conducted by Arvind, another researcher in this group.

2.4.1 Setpoint Tracking Failure Detector

The research project starts from analyzing the system transient behavior in the cooling tracking mode, and the details can be found in Paper A. Figure 2.4 shows four types of behaviors that are common in the cooling tracking mode. Recall that a cooling tracking period is just one single cooling on-cycle. Figure 2.4(a) shows normal transient behavior (the most common one), where the cooling system starts after a drop in setpoint and remains on until reaching the new setpoint. Figure 2.4(b) shows poor transient behavior, where the indoor temperature increases over 10°F before returning, despite the cooling system operating continuously for nearly 24 hours. Both Figure 2.4(c) and 2.4(d) show interesting transient behavior where the indoor temperature first increases precipitously after the cooling system turns on. The behavior in 2.4(c) may be caused by lack of insulation on the return air duct in the attic, such that warm air is initially blown into the house when the fan starts after a long downtime; the behavior in 2.4(d) may be caused by the internal heat load of the thermostat itself on the sensor, such that indoor temperature seems to rise as soon as the fan starts (after cooling starts) and the air circulation direction of the room changes.

The setpoint tracking failure detector focuses on the time periods that exhibit poor transient behavior in Figure 2.4(b), also known as poor tracking periods or setpoint tracking failure periods. In cooling, it is defined as a single cooling on-cycle that usually occurs during high cooling load, and during which the indoor temperature can increase a few degrees above the cooling setpoint, unable to be controlled although the system is operating continuously. Note that the term 'high cooling load' is regarded as relative to the system capacity. If the capacity of a system is far below its rated capacity, even normal loads may well be incapable for it to deal with and thus will be considered as high loads. Therefore, this situation is unlikely to happen for a properly sized fault-free system, and is often related to the capacity degradation fault.

Table 2.4 shows all features extracted from the cooling tracking mode used in this analysis. Among all of the five features, $T_{i,inc}$, Δt_{inc}, and $\dot{T}_{i,inc}$ are calculated for each cooling tracking period, and then applied to separate the poor transient behavior in Figure 2.4(b) from the other

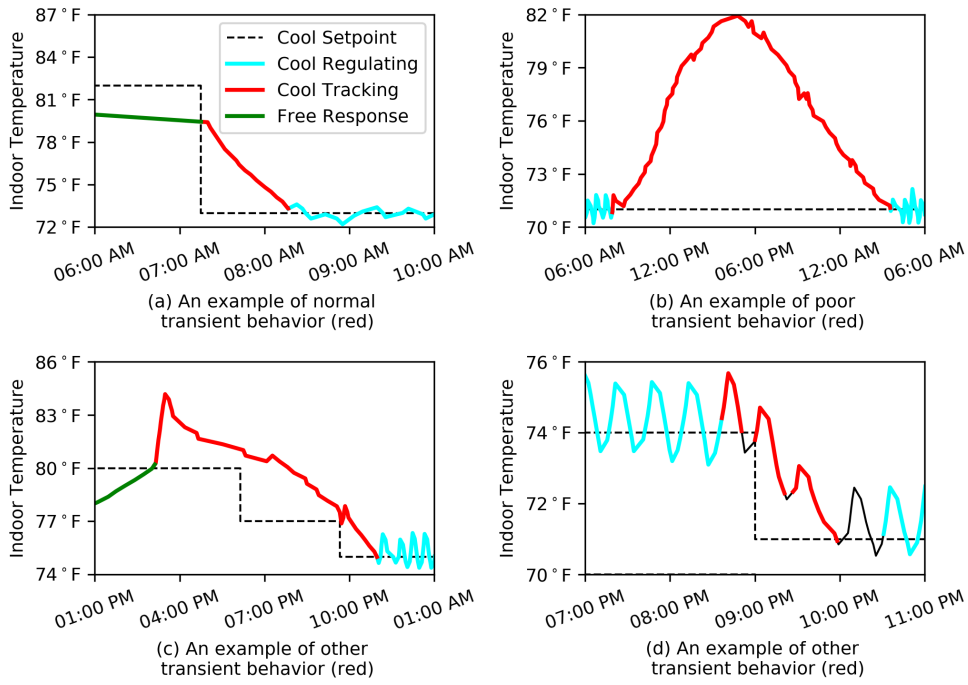


Figure 2.4: Four main types of transient behavior exhibited in the cooling tracking mode (red color) [7].

three types of transient behaviors. After that, features $\overline{DH}$ and $\bar{T}_{i,max}$ are used to characterize the problem severity.

The poor transient behavior is recognized from other transient behaviors in two steps. First, from all cooling tracking periods, the normal tracking periods in Figure 2.4(a) are removed by a simple range filter, $T_{i,inc} > 1$°F, because normal periods should not have large indoor temperature increase when the cooling system is switched on. After that, the kernel density estimation of the remaining tracking periods are shown in Figure 2.5(a). Second, the poor transient behavior are differentiated from the undesirable transient behaviors shown in Figure 2.4(c) and 2.4(d). This process leverages the main difference between these two behaviors: poor transient behavior exhibits long Δt_{inc} and low $\dot{T}_{i,inc}$, but on the contrary the undesirable transient behaviors exhibit short Δt_{inc} and high $\dot{T}_{i,inc}$. Shown in Figure 2.5(b), the cluster on the left could only correspond to the poor tracking periods, and the other one on the bottom could only correspond to undesir-

Table 2.4: Supporting features used in the setpoint tracking failure detector [7].

Symbol	Feature name	Definition
$T_{i,inc}$	Indoor temperature increase [°F]	The maximum indoor temperature increase*/** in a cooling tracking period.
Δt_{inc}	Increase time [hr]	The time required to reach the maximum indoor temperature increase.
$\dot{T}_{i,inc}$	Indoor temperature increase rate [°F/hr]	The rate at which indoor temperature increases; defined as $\dot{T}_{i,inc} = T_{i,inc}/\Delta t_{inc}$.
$\overline{DH}$	Average degree hour [°F·hr]	The integral of the difference between the indoor temperature and the cooling setpoint, during a time segment exhibiting poor transient behavior, averaged over the whole cooling season; defined as: $\overline{DH} = \frac{1}{N}\sum_{n=1}^{N}\int_{0}^{24} I \cdot \max(0, T_{spe})dt$; where N is the number of days in a cooling season; I equals 1 in poor tracking periods, and 0 in other cases.
$\bar{T}_{i,max}$	Average maximum indoor temperature [°F]	The average of ten highest $T_{i,max}$ in all poor tracking periods** after removing five most extreme instances.

Notes: * A searching algorithm is applied to find the maximum increase in a time series.
** If a poor cooling tracking period is longer than 24 hours, then it will be segmented with one period a day. Generally, there is at most one poor tracking period each day, and the maximum indoor temperature occurs in the afternoon when cooling load is the highest.

able behaviors. Detailed in Section 3.3.4, finally the Ward's agglomerative hierarchical clustering method is chosen to separate the data and flag the poor tracking periods.

Once all poor tracking periods are identified, a rule-based method is applied to label systems with severe tracking failure faults. The method uses $\overline{DH}$ and $\bar{T}_{i,max}$ (see Table 2.4), two comprehensive features averaged over all poor tracking periods for each system. These two features quantify the severity of the tracking failure problem. Normal system should have $\overline{DH}$ close to zero or $\bar{T}_{i,max}$ lower than approximate 80°F, but systems with severe tracking failure faults may have abnormally high values of $\overline{DH}$ and $\bar{T}_{i,max}$. Thus, after calculating a threshold from all poor

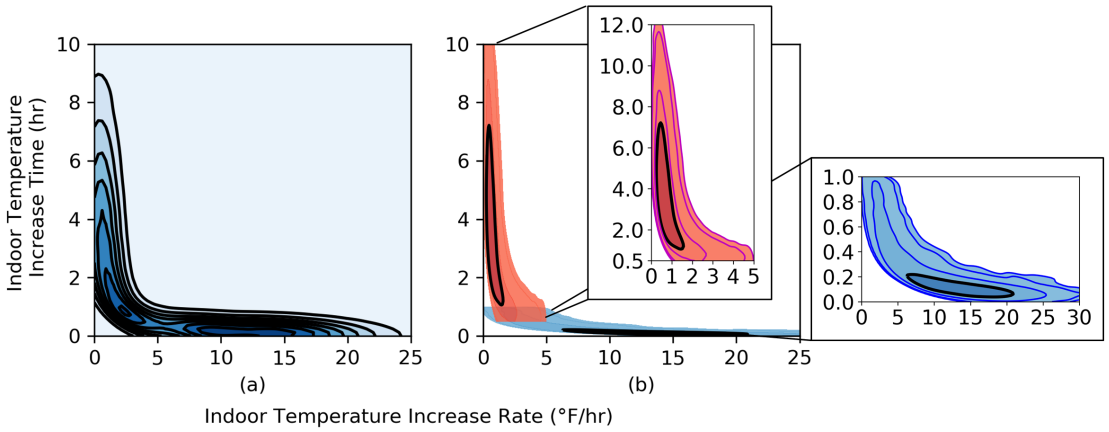


Figure 2.5: (a) Kernel density estimation of the remaining tracking periods after the normal ones are removed. (b) The data forms two groups. The group in red corresponds to poor transient behavior, and the group in blue corresponds to other transient behaviors [7]. The black circles show the core region (50% of mass) of each group.

tracking periods for each of these two features, a system with feature values of $\overline{DH}$ and $\bar{T}_{i,max}$ both higher than the calculated thresholds are possibly faulty and can be labeled.

As an example of this rule-based method, the kernel density estimation of $\overline{DH}$ and $\bar{T}_{i,max}$ for approximately 10,000 systems in Florida having poor tracking periods are shown in Figure 2.6. The method takes 80% quantile of $\overline{DH}$ and $\bar{T}_{i,max}$ as thresholds, respectively 12.4°F·hr and 80.0°F, and 68 systems are flagged as faulty with both of these two features greater than the thresholds.

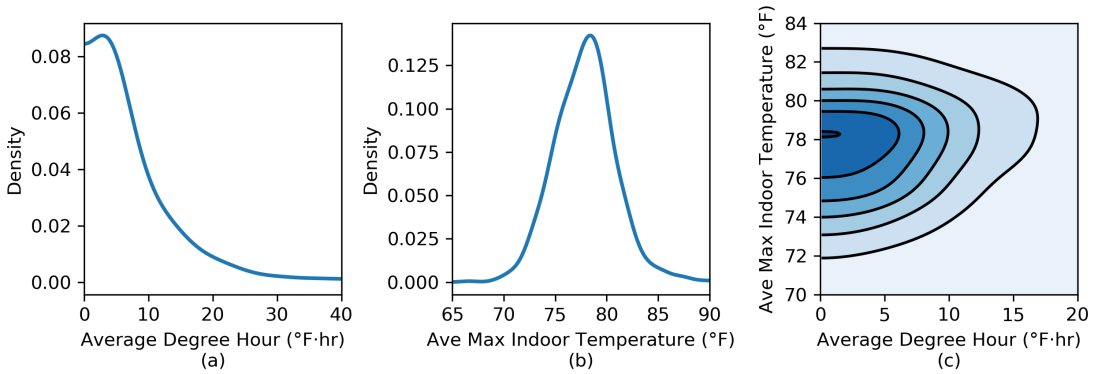


Figure 2.6: An example of the kernel density estimation of (a) the average degree hour above setpoint, (b) average maximum indoor temperature, and (c) both features together [7].

2.4.2 Inadequate Capacity and Control Problem Detector Through Multivariate Analysis

The multivariate analysis extracts features from the cooling regulating mode and detects inadequate capacity and control problems. Figure 2.7 represents the outline of the whole fault detection processes. The statistical analysis is composed of two parts: univariate fault detection and multivariate fault detection. The methodology is detailed in Paper B, and the case studies are detailed in Paper C.

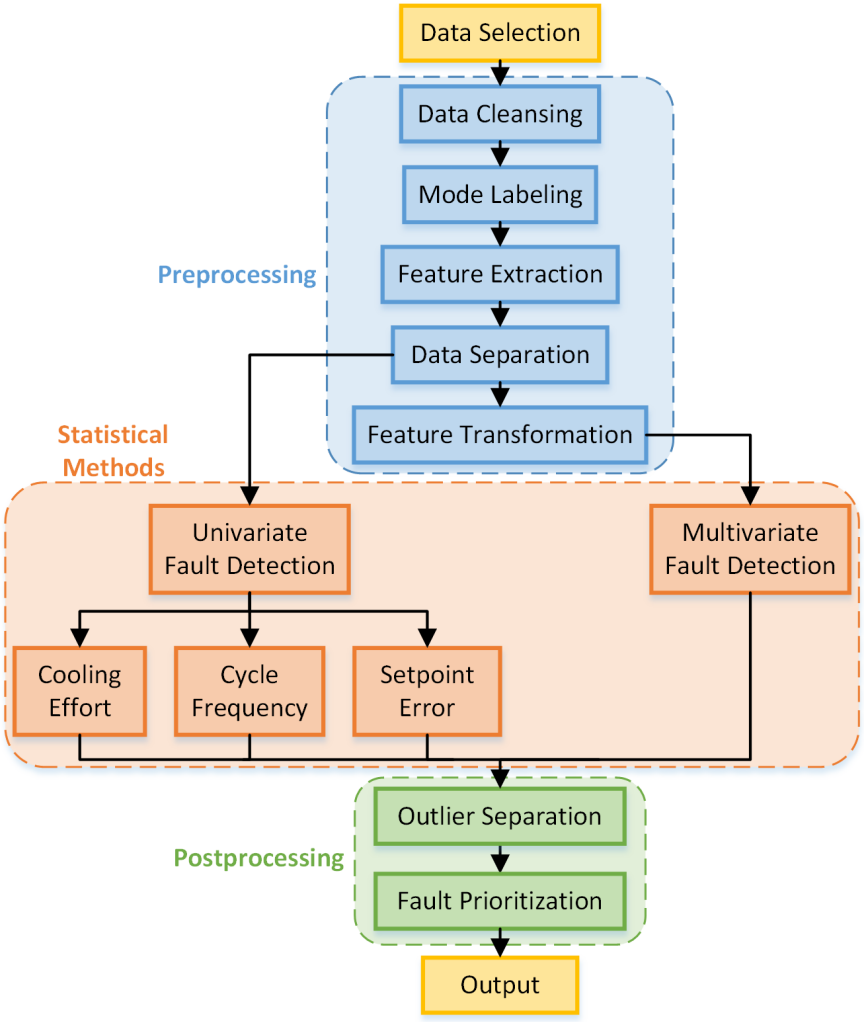


Figure 2.7: An outline of the multivariate analysis [4].

In the preprocessing procedures, data cleansing, mode labeling, and feature extraction has been discussed in Section 2.3. Table 2.5 shows all features extracted from the cooling regulating mode in this multivariate analysis. For these five features, E_c characterizes the cooling capacity of a system relative to the load value, and thus could be used to detect inadequate capacity faults. F and T_{spe} characterize the control behavior of a system, and thus could be used to detect control faults. ΔT_{oi} and $\Delta \omega_{oi}$ define the operational conditions, which strongly affect the behavior of a system, but do not indicate any anomalous behavior.

Data preprocessing also includes two additional procedures, data separation and feature transformation. In data separation, E_c, F, and T_{spe} are categorized into subsets of similar operational conditions according to ΔT_{oi} and $\Delta \omega_{oi}$. Only E_c, F, and T_{spe} are compared between systems. In feature transformation, data features are transformed to individually normally distributed to fit the statistical models in the multivariate tests.

Table 2.5: Supporting features used in the multivariate analysis [4].

Symbol	Feature name	Definition
ΔT_{oi}	Temperature difference [°F]	Temperature difference between home (indoor) and ambient (outdoor) $\Delta T_{oi} = T_o - T_i$
$\Delta \omega_{oi}$	Humidity ratio difference [kg H_2O/ kg dry air]	Humidity ratio difference between home (indoor) and ambient (outdoor) $\Delta \omega_{oi} = \omega_o - \omega_i$
E_c	Cooling effort [%]	The product of cooling duty factor (D) and cooling load factor (L)
F	Cycle frequency [cycle/hr]	The number of start/stop cycles per hour
T_{spe}	Cooling setpoint error [°F]	The error between indoor temperature and cooling setpoint $T_{spe} = T_i - T_{sp}$

Univariate fault detection examines each feature separately and does not consider possible correlations between features. In this analysis, abnormally high or low values of each feature are considered outliers and may be associated with faults. To identify the outliers, the kernel density estimation (KDE) is applied to estimate the distribution of each feature under similar operating conditions. From the distribution, the percentage of systems that has higher feature values (i.e. the upper percentage) for each system can be determined. As an example, the KDE of average cooling effort ($\bar{E}_c$) for one of the operating condition is shown in Figure 2.8. One system has $\bar{E}_c$ value of 45%. Because the whole area under the KDE curve is one, the shaded area shows the percentage of systems that has cooling effort higher than 45%. Generally abnormally high values of a feature will have percentages close to zero, while abnormally low values of a feature will have percentages close to unity. The process for labeling faulty systems from the calculated percentages will be discussed later in the outlier separation and fault prioritization process.

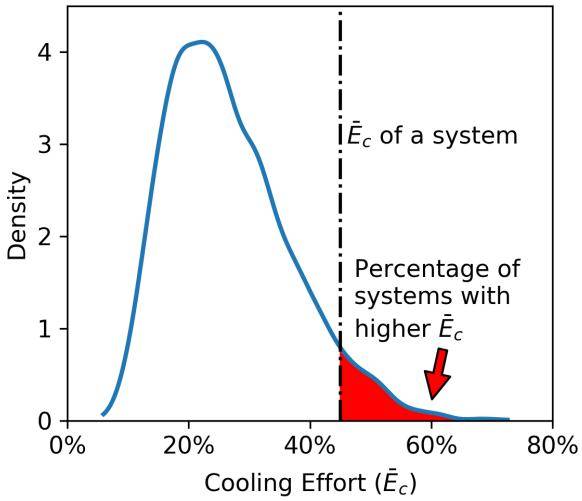


Figure 2.8: The KDE plot illustrates how univariate test of cooling effort is applied [4]. The system has $\bar{E}_c$ value of 45%, and the shaded area shows the percentage of systems that has cooling effort higher than 45%.

In comparison, the multivariate fault detection combines E_c, F, and T_{spe} as three bivariate statistics and one trivariate statistic to identify systems with uncommon combination of feature values. The method applies a metric termed the Mahalanobis distance (D_M^2) [29] that can capture the pattern of the multivariate distribution. Shown in Figure 2.9, for two highly correlated features X and Y, the Mahalanobis distance considers that point A is further away from the center than point B. In essence, point A is much more of an outlier. The multivariate fault detection first calculates the Mahalanobis distance of each feature value to the population distribution, and then applied kernel density estimation to calculate the percentage of systems that has higher D_M^2 (i.e. the upper percentage) for each system. Similar to the univariate fault detection process, abnormally high values of the percentage are considered outliers and may be associated with faults.

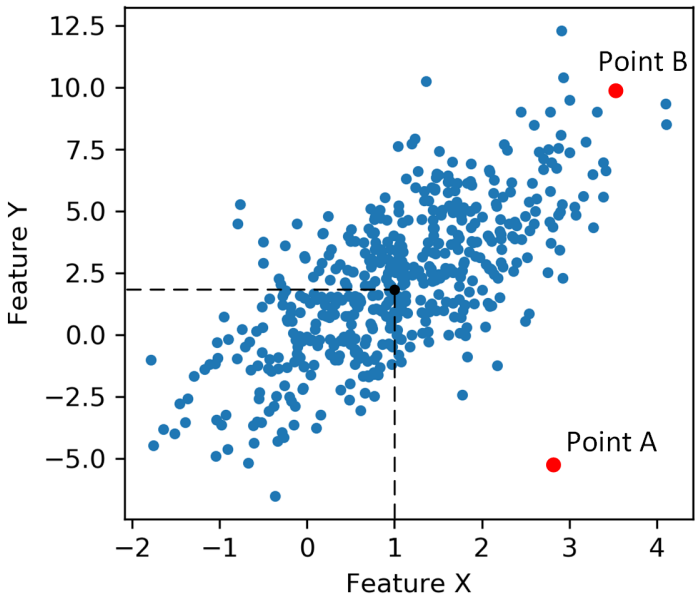


Figure 2.9: Multivariate outlier analysis with Mahalanobis distance [4].

Having obtained the upper percentages of each univariate and multivariate for all systems, finally the postprocessing procedures are required to separate true outliers and prioritize faulty systems. The outlier separation process differentiates real malfunctioning systems from all the

others. Table 2.6 shows the subgroups of outlier systems that are considered faulty or oversized. In general, abnormally high cooling effort is associated with undercharge, condenser/evaporator fouling, airflow restriction, etc.; abnormally low cooling effort is associated with the oversizing issue; abnormally high cycle frequency or setpoint error is associated with malfunctioning control systems or poor thermostat placement (e.g. close to supply air outlets); but low setpoint error is never related to faults. Note that by separating systems into subgroups, the multivariate analysis can detect both inadequate capacity and control problem.

With the criteria shown in Table 2.6, in the univariate tests only abnormally high values of each feature are considered faulty, so the upper percentage for each system indicates the fraction of systems in the population that exhibit worse performance. In the multivariate tests abnormally high values of the D_M^2 are considered faulty because they are far away from the distribution center, so the upper percentage of D_M^2 for each system also indicates the fraction of systems in the population that exhibit worse performance.

Table 2.6: Outlier separation: subgroups of outlier systems that are considered as faulty or oversized [4].

Type of test	Typical behaviors	E_c	F	T_{spe}
Univariate	Near-continuous operation	↑		
	High cycling		↑	
	Poor control			↑
Bivariate	Near-continuous operation	↑	↓	
	High cycling	↑	↑	
	Poor control	↑		↑
	Poor control		↑	↑
	Oversized	↓	↑	
Trivariate	Mixed faulty behaviors	↑	↑	↑
	Mixed faulty behaviors	↑	↓	↑

↑—abnormally high, ↓—abnormally low

The above analysis obtains the upper percentages that indicate faulty behaviors for each system. However, the upper percentage values are calculated within each operational condition, and there are in total eight operational conditions specified in data separation. In order to quantify the operational performance and prioritize faulty systems, a metric that aggregates over the full range of operational conditions is proposed. This metric is denoted by $\bar{\alpha}$ in Paper B, and for each system in each univariate and multivariate test, it is the weighted average of all upper percentages in all operational conditions. By using $\bar{\alpha}$, occupants and technicians know how a system performs compared to similar systems. Finally, each test labels faulty systems separately by comparing $\bar{\alpha}$ with pre-determined thresholds. In this research, the threshold for the univariate tests is 1%, and for the multivariate tests is 5%. Inadequate capacity fault is labeled to a system if the $\bar{\alpha}$ is smaller than the threshold in any one test using E_c as a feature, and control fault is labeled to a system if the $\bar{\alpha}$ is smaller than the threshold in any one test using F or T_{spe} as a feature.

Paper C presents five case studies in detail to illustrate the effectiveness of this multivariate analysis. Each of them represents a particular category of common behavior. The first case study shows a normal device to serve as a basis for comparison with subsequent faulty systems. The second case study shows a system with severe setpoint tracking failure caused by capacity degradation. The third case study also shows a system with inadequate capacity, but in contrast to the previous example, this fault is imperceptible to the occupants as comfort levels are maintained despite the fault. The fourth case study illustrates control problems, showing that the multivariate analysis can find faults for a broad range of behavioral faults. The final case study shows an oversized system, which demonstrates the versatility of the multivariate analysis.

2.4.3 Degradation Trend Detector

The degradation trend detector belongs to the single-system change detection, which compares the performance within each system through time to detect gradual capacity degradation. This issue is typically caused by refrigerant undercharge due to an active leak or partially blocked air flow (e.g. crushed air duct), and is often imperceptible to the occupants for months until eventually causing total system failure.

Equation 2.1 shows the energy balance of a residential home in a time period:

$$C \cdot E_c = UA \cdot \Delta T_{oi} + \dot{Q}_{sol} + \dot{Q}_{int} \tag{2.1}$$

where C is the total cooling capacity [W], E_c is the cooling effort (%), U is the overall heat transfer coefficient for the residential home [W/m^2·°F], A is the overall heat transfer area [m^2], ΔT_{oi} is the temperature difference [°F], $\dot{Q}_{sol}$ is the solar irradiance heat gains [W], and $\dot{Q}_{int}$ is the internal heat gains [W]. From the energy balance, under similar operating conditions (i.e. similar ΔT_{oi}, $\dot{Q}_{sol}$, and $\dot{Q}_{int}$), the continual increase of cooling effort (E_c) through weeks and months can reflect gradual degradation of cooling capacity (C) for a system.

In order to detect an increasing trend of the cooling effort while controlling the variations of ΔT_{oi}, $\dot{Q}_{sol}$, and $\dot{Q}_{int}$, the degradation trend detector proposes an hourly analysis method, an daily analysis method, and one long-term monitoring metric. The hourly analysis is effective in detecting trends for systems with long regulating periods. The daily analysis accounts for the whole operating conditions and is suitable for all systems, but may not be effective in shoulder seasons. In practice, both methods are applied to an operating system. Finally, a long-term monitoring metric denoted by $\Delta \bar{E}_c$ is proposed to actively evaluate the system performance through months and even years. The $\Delta \bar{E}_c$ indicates the extra percentage of time a system need to operate compared to its expected performance, when the system is actively controlling the indoor temperature around the desired setpoint values.

2.4.3.1 Hourly Analysis

The hourly analysis is proposed in Paper D. In this analysis, the smart thermostat data is resampled into two-hour averages, and the cooling effort (E_c) is extracted from the cooling regulating mode as a critical feature. To differentiate with features in the daily analysis, the E_c is termed *steady-state cooling effort*. To account for seasonal weather effects, solar irradiance effects, and daily schedules in internal heat gains, E_c is divided into subsets based on ΔT_{oi} and time of day.

To detect if an increasing trend of E_c exists, a popular trend detection method called the Mann-Kendall test is applied. The Mann-Kendall test is a non-parametric statistical test commonly employed to detect monotonic trends [30, 31]. The null hypothesis, H_0, is that the data comes from a population with independent and identically distributed realizations. The alternative hypothesis, H_A, is that the data follows a monotonic trend. The Mann-Kendall test statistic is calculated as:

$$S = \sum_{k=1}^{n-1} \sum_{j=k+1}^{n} sgn(X_j - X_k) \tag{2.2}$$

where X is the variable, n is the number of observations, and $sgn(\cdot)$ is the signum function that determines the positive/negative nature of the function argument.

The S-statistic is approximately in normal distribution with the mean $E[S]$ and the variance σ^2 calculated as:

$$E[S] = 0 \tag{2.3}$$

$$\sigma^2 = \frac{n(n-1)(2n+5) - \sum_{m=1}^{n} t_m m(m-1)(2m+5)}{18} \tag{2.4}$$

where t_m is the number of ties of length m.

Then, with the following transformation,

$$Z = \begin{cases} \frac{S-1}{\sigma} & \text{if } S > 0 \\ 0 & \text{if } S = 0 \\ \frac{S+1}{\sigma} & \text{if } S < 0 \end{cases} \tag{2.5}$$

the Z-statistic is approximately in standard normal distribution. In a one-side test for monotonic increasing trend, H_0 should be rejected if $Z > Z_\alpha$, where α is the significance level for the test.

Additionally, the E_c time-series is autocorrelated instead of independent. For example, the current E_c is positively correlated to the E_c several hours before due to the occupant behavior, building thermal capacity, thermostat control strategy and the weather. This positive correlation can cause the test more likely to show evidence of a trend while no trend actually exists [32].

Therefore, a modification of the Mann-Kendall test termed the Mann-Kendall test with moving block bootstrap (MK-MBBS) is selected to account for this issue through random resampling. Procedures of the moving block bootstrap (MBBS) are summarized as follows:

1. The Z-statistic of the test data is calculated by the original Mann-Kendall test.
2. The test data is randomly resampled with replacement for N times. In order to keep the autocorrelations between data points, the data is subdivided into $n - L + 1$ overlapping blocks, where n is the length of the test data and L is the length of each block. Resampling is performed on the block level. Figure 2.10 illustrates this step with $n = 5$ and $L = 3$.
3. The Z-statistics of all N resampled dataset are calculated by the original Mann-Kendall test. They form a distribution of the Z-statistic.
4. Percentile of the Z-statistic (P) of the test data in the distribution is calculated.
5. In a one-side test for monotonic increasing trend, if $P > 1 - \alpha$, where α is the significance level of the statistical test, then the null hypothesis should be rejected.

In Paper D, the performance of the MK-MBBS statistical test is proved using Monte Carlo simulations.

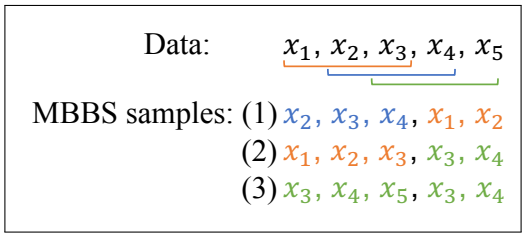


Figure 2.10: An illustration of MBBS procedures with $n = 5$ and $L = 3$. The data is split into $n - L + 1 = 3$ blocks marked by different colors, and samples are randomly chosen from the blocks with replacement. Presented in Chapter 6.

To perform the hourly analysis using the MK-MBBS test, the steady-state cooling effort (E_c) is processed in the following procedures:

Table 2.7: Time-series division according to temperature difference (ΔT_{oi}) and time of day. Presented in Chapter 6.

Subset number		ΔT_{oi} range [°F]					
		$-4 \sim 0$	$0 \sim 4$	$4 \sim 8$	$8 \sim 12$	$12 \sim 16$	$16 \sim 20$
Time of day [hr]	8:00 $\sim$ 20:00	No. 1	No. 2	No. 3	No. 4	No. 5	No. 6
	20:00 $\sim$ 8:00	No. 7	No. 8	No. 9	No. 10	No. 11	No. 12

1. The E_c values for each system are divided into subsets based on ΔT_{oi} and time of day. Twelve subsets are created and their ranges are shown in Table 2.7.

2. The MK-MBBS test is performed on each time-series independently. Percentiles of the Z-statistics (P) calculated by the MK-MBBS test are recorded.

3. The final decision on whether a system has an increasing trend is based on the second maximum percentile and the median percentile as a supplementary condition among all time-series of a system. The second maximum percentile should satisfy $P_{second\_max} > 99.9\%$ (i.e. under 0.1% significance level) and at the same time the median percentile should satisfy $P_{median} > 95\%$ (i.e. under 5% significance level).

2.4.3.2 Daily Analysis

The daily analysis is also proposed in Paper D. In this analysis, the smart thermostat data is first resampled into daily averages without mode labeling, and then the daily cooling effort (E_c) is extracted and analyzed. To consider the variations of ΔT_{oi}, the cooling degree day (CDD) is selected as a covariate that being analyzed simultaneously with E_c. Note that the cooling degree day can be defined as the daily temperature difference between indoor and outdoor space, or the daily temperature difference between cooling setpoint and outdoor space. This study adopts both covariates so that the method is adaptable to a wider range of system operational behaviors. However, the variations of $\dot{Q}_{sol}$ and $\dot{Q}_{int}$ are not considered, because they are supposed not to have significant impacts on daily changes of E_c.

In order to detect the general trend of E_c meanwhile considering the variations of cooling degree days, the study applies the partial Mann-Kendall test [33]. Same as the Mann-Kendall test, the null hypothesis, H_0, is that E_c comes from a population with independent and identically distributed realizations. The alternative hypothesis, H_A, is that E_c has a monotonic trend. The calculation of the partial Mann-Kendall test is relatively complicated, and therefore this subsection only shows the Z-statistic:

$$Z = \frac{S_1 - \hat{a}_2 S_2 - \hat{a}_3 S_3}{\sqrt{\hat{\sigma}_{11} - \hat{a}_2 \hat{\sigma}_{12} - \hat{a}_3 \hat{\sigma}_{13}}} \tag{2.6}$$

where subscript "1" denotes E_c, subscript "2" denotes the covariate of indoor and outdoor daily temperature difference, subscript "3" denotes the covariate of daily setpoint and outdoor temperature difference, S is the S-statistic (see Equation 2.2), $\hat{\sigma}_{11}$ is the variance of the S-statistic for E_c, $\hat{\sigma}_{12}$ and $\hat{\sigma}_{13}$ are the conditional covariances showing the relationships between E_c and covariates, $\hat{a}_2$ and $\hat{a}_3$ are derived variables calculated from the above variances and covariances. The first term in the numerator and denominator represents the contribution of E_c, and the followed subtraction terms represent the contribution of covariates. The detailed derivation of Equation 2.6 can be found in Paper D.

The Z-statistic is proved to weakly converge to the standard normal distribution [33]. In a one-side test for monotonic increasing trend, H_0 should be rejected if $Z > Z_\alpha$, where α is the significance level.

Similar to the problem in the hourly analysis, the E_c and covariate time-series is autocorrelated. Thus, a modification of the partial Mann-Kendall test termed the partial Mann-Kendall test with moving block bootstrap (PMK-MBBS) is used. The procedures for applying moving block bootstrap is the same as the hourly analysis.

The performance of the PMK-MBBS statistical test is proved using Monte Carlo simulations in Paper D. The procedures for degradation trend detection is shown as follows:

1. Systems without data for at least 30 days are excluded from analysis to ensure an unbiased result.

2. The cross-correlations between E_c and the cooling degree days are checked. If positive correlations do not exist, then the corresponding cooling degree day will be removed. If both features of the cooling degree days are removed, then the original Mann-Kendall test with moving block bootstrap will be performed to E_c without any covariates.

3. The PMK-MBBS test is performed to remaining systems and percentiles of the Z-statistics (P) are recorded.

4. The significance level (α) is set to 0.1%. An increasing trend of E_c (i.e. capacity degradation) is detected if $P > 1 - \alpha$.

In practice, both hourly and daily trend detection methods are implemented using either historical data or real-time data. The historical trend detection performs testings after the peak operating season (i.e. the summer season for a cooling system, or the winter season for a heating system) has passed and identifies degradation of a system during this peak season. This approach is applied by service companies to schedule off-season maintenance for degraded systems. In comparison, the real-time trend detection identifies degradation within the peak operating season and performs testings every day with a fixed-length moving window. In this approach, service companies can actively monitor the behavioral change of operating systems and allocate repair resources in advance for faulty systems.

A couple of case studies of verified faulty systems with the active refrigerant leakage problem are presented in Paper D to show the effectiveness of the hourly and daily methods.

2.4.3.3 Long-Term Monitoring Metric

The above hourly and daily methods provide statistical tests to detect whether capacity degradation trend occurs for a system. The detection results are qualitative, and have only addressed the *statistical* significance of the trend. However, engineers are more interested in addressing the *practical* significance of the trend by quantifying the actual capacity loss.

Paper D proposes to use the Sen's slope [34, 35] to estimate the change of steady-state cooling effort and daily cooling effort. The Sen's slope is a robust and unbiased estimate of the slope of

the trend calculated by:

$$d_k = \frac{X_j - X_i}{t_j - t_i} \text{ for all } 1 \leq i \leq j \leq n \tag{2.7}$$

where X is the cooling effort value, t is time, and d_k's are the slopes between each pair of cooling effort.

Then, the Sen's slope is calculated as the median of all the slopes (d_k):

$$\text{Sen's slope} = \text{median}(d_k) \tag{2.8}$$

After the slope of change is estimated, the difference of cooling effort between two timestamps can be measured by taking the product of the Sen's slope and the time length of the cooling effort data. Thereby the capacity degradation can be quantified.

However, the Sen's slope cannot monitor the change of cooling capacity on each single day or compare the changes with other systems under similar operational conditions to determine the severity of degradation. Paper E fills this gap by first introducing a concept named as the weighted average difference of cooling effort ($\Delta\bar{E}_c$) to compare the relative capacity of each system to the expected levels, and then proposing to use the *change of* $\Delta\bar{E}_c$ as a metric to monitor the system operational performance in the long term. That is, the change of $\Delta\bar{E}_c$ incorporates the multi-system feature comparison and the single-system change detection.

The $\Delta\bar{E}_c$ is extracted from the cooling regulating mode using the characteristic curve method, which consists of the following four steps:

1. Kernel density estimation (KDE) is performed to estimate the probability density function of the joint distribution between E_c and ΔT_{oi}, using selected data which follows the recommendations in Section 2.2. The KDE treats each data point as a Gaussian kernel, and sums up all kernels to obtain the full probability density function. An illustration is shown in Figure 2.11a.

2. From the estimated density field, the midpoint of accumulated density at every given ΔT_{oi} is calculated. In another word, the sum of density above the midpoint equals to the sum of

density below the midpoint. By this means, the midpoints at all slices of ΔT_{oi} connect to a curve termed the characteristic curve, which can be viewed as the *average performance* of the selected group of systems, or the *expected* E_c of these systems at every level of ΔT_{oi}. An illustration is shown in Figure 2.11b.

3. From the characteristic curve, the difference of each cooling effort data (ΔE_c) to the curve is calculated. Illustration of the data features from one system is shown in Figure 2.11c.

4. The weighted average of all differences of cooling effort ($\Delta \bar{E}_c$) for each system is calculated, using the probability density function of ΔT_{oi} as weight. The probability density function estimated by the KDE is shown in Figure 2.11d. The reason of weighting is that most data is distributed in the high-density region of ΔT_{oi}, such that estimate of the characteristic curve in the high-density region is more accurate.

In order to monitor the long-term changes, the $\Delta \bar{E}_c$ is calculated by averaging ΔE_c in a one-month moving window shifting on each single day. The absolute value of $\Delta \bar{E}_c$ shows how a system performs in comparison with other systems under similar operating conditions. For example, a system with $\Delta \bar{E}_c$ equals to 30% indicates that typically it spends 30% more effort on cooling than the average level under similar operating conditions. Further, the change of $\Delta \bar{E}_c$ values indicates changes of system cooling capacity. For example, an increase of $\Delta \bar{E}_c$ by 20% over two months can mean the system degraded significantly in the past two months, and a decrease of $\Delta \bar{E}_c$ by 20% after maintenance often means the fault was successfully repaired. This metric automates the process of evaluating the system performance continuously, and can alarm the occupants and service companies when significant increase of $\Delta \bar{E}_c$ is detected.

2.5 Future Work Plans

After presenting the developed statistical-based fault detectors, the author also discusses some future work in Chapter 9. The future work includes field validation of detected faulty systems, applying deep learning models, and extracting additional features. Especially, the author proposes two deep learning methods, which are the recurrent neural network (RNN) and the Siamese

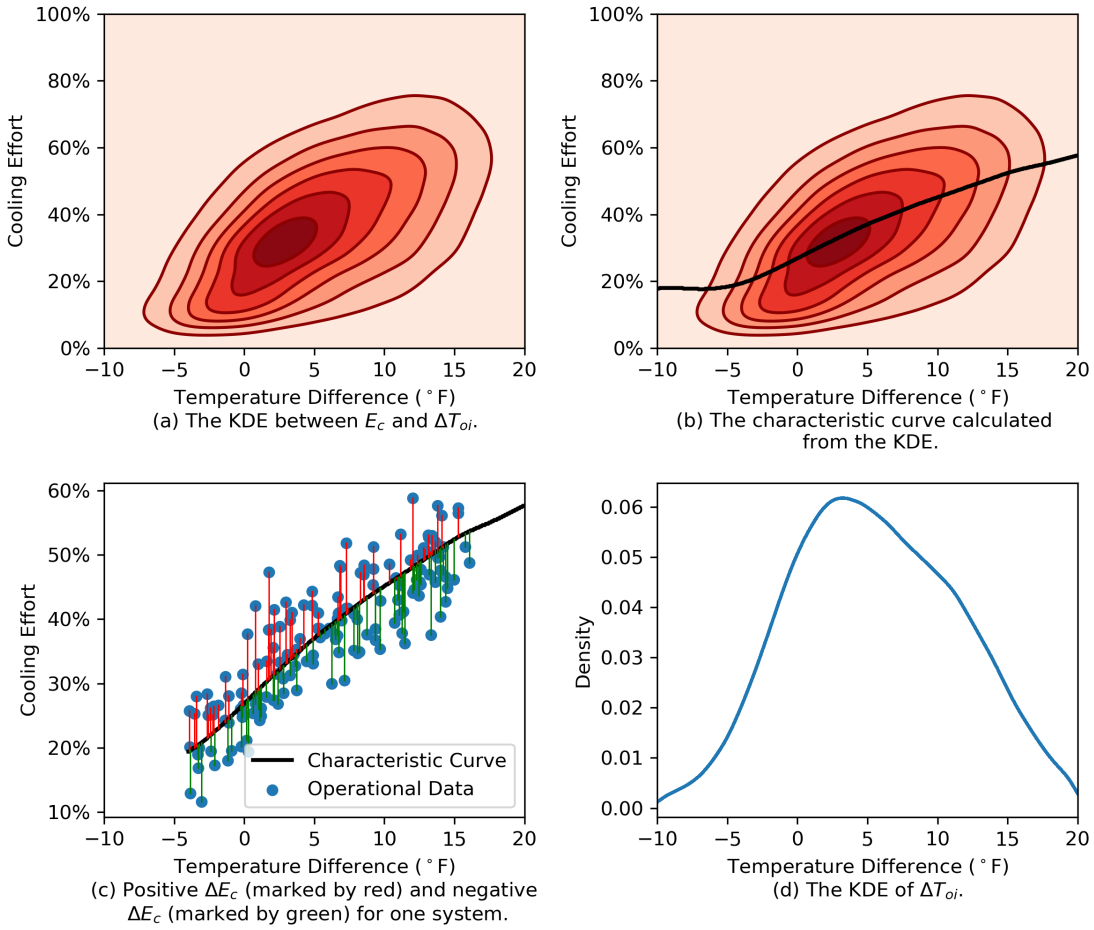


Figure 2.11: Illustrations of the procedures to calculate the $\Delta \bar{E}_c$ for each system. Presented in Chapter 7.

network. The RNN model performs multi-system feature comparison, and the Siamese network model performs single-system change detection. These two methods are promising with more labeled data and additional features.

3. PAPER A: FAULT DETECTION AND DIAGNOSIS FOR RESIDENTIAL HVAC SYSTEMS USING TRANSIENT CLOUD-BASED THERMOSTAT DATA[1]

3.1 Synopsis

Fault detection and diagnosis (FDD) using aggregated smart thermostat data is a relatively new research field, but one with immediate practical application to residential indoor climate control. This paper analyzes a cloud-based dataset which contains thermostat history records of nearly 370,000 distinct residential HVAC systems in the U.S. The large, diverse, and growing dataset enables novel methods for detecting and diagnosing faults on systems with limited sensor data. This paper proposes a statistics-based FDD method for non-variable speed heat pump and air conditioning units, and demonstrates the effectiveness with several case studies. The proposed method identifies systems within a similar climate region, and then segments and classifies the time series data based on operational mode and behavior. Various data features are then extracted from the time series segments to identify systems that exhibit poor transient behavior. Additional features are used to refine and classify the problem severity. Statistical methods are then used to compare system performance to the entire population and identify outlier behavior due to operational faults that affect system efficiency and occupancy comfort. The resulting algorithm demonstrates the potential of big data fault detection for air conditioning systems using limited cloud-based sensor information.

3.2 Introduction

Space heating and air conditioning accounts for a large amount of residential households' energy expenditures. From the Residential Energy Consumption Survey [1], the average annual energy expenditure for a household is $1,856, among which $543 (29.3%) spent on space heating and $265 (14.3%) spent on air conditioning. In order to reduce residential households' energy consumption on space heating and air conditioning, researchers have been working on improving

[1]F. Guo, and B.P. Rasmussen, "Fault Detection and Diagnosis for Residential HVAC Systems Using Transient Cloud-based Thermostat Data," in *6th International High Performance Buildings Conference at Purdue*, 2020.

system efficiency, employing better control strategies, and improving preventive maintenance using fault detection and diagnosis (FDD). By detecting operational faults caused by system aging or improper commissioning early, failures can be prevented and system inefficiencies avoided.

HVAC system faults have a considerable impact on system efficiency and occupant comfort. Studies report that the average performance of residential HVAC systems is at least 17% below their design performance [9]). Correcting only refrigerant charge and air flow problems has an estimated 17% demand savings potential and 7% peak demand savings potential [36]. Incorporating FDD into the preventive maintenance routines has the potential for increasing occupant comfort and reducing energy expenditures. Additionally, manufacturers could improve brand loyalty by ensuring high-performance and reliable systems; service companies would be able to provide better maintenance to their customers. With regard to FDD performed on building systems, [3] pointed out that only 16% studies are associated with small commercial and residential buildings, such as rooftop packaged units and split systems, while the majority of research focus is on variable air volume air handling units (VAV-AHUs), chillers, cooling towers, and overall building diagnosis. Thus traditional FDD methods are usually designed for one large system and require installing additional sensors [16]. But residential split systems are produced in large quantities, with limited sensors, and with a variety of system types and custom installation. For many systems, the addition of the sensors required for traditional FDD techniques is prohibitively expensive.

Cloud-based data from smart thermostats provides an opportunity for alternative FDD methods that do not require additional sensors, but utilize historical records of basic system operation. The popularity of smart thermostats that upload data to a remote server is increasing, allowing for computationally intensive fault detection on distributed platforms. Many manufacturers are establishing a cloud database, remotely monitoring thousands of installed residential systems. Typically, this cloud-based thermostat data includes indoor temperature and relative humidity measured by built-in sensors, system mode status, cooling and heating temperature setpoints, and outdoor temperature acquired from a third party. The resulting big data creates an opportunity for novel FDD approaches.

FDD methods for residential HVAC systems with smart thermostat data is still a relatively new research field, and few published studies are currently available. [27] proved the capability of thermostat data to evaluate the thermal characteristics of a house, such as insulation levels, thermal capacity, and thermal time constants. With these indicators, houses showing inferior thermal response characteristics, known as construction-level faults, could be detected. [21] developed a recursive least-squares model to detect faults on cooling, heating and ventilation equipment. However, the method was only validated by simulated indoor and outdoor temperature data, and focused mainly on total equipment failures. [20] trained a linear model to learn the normal behavior, and then applied the model to monitor capacity degradation faults such as refrigerant leakage. Although the model was originally developed for cold rooms in retail outlets, the author mentioned that its application could be extended to residential buildings. Besides, combining smart thermostat data with building assessor's data and energy data, [37] showed the potential to predict the electricity demand of residential HVAC systems, and then perform fault detection by comparing the predicted demand with the actual demand. In short, these studies all focused on a small group of systems, training models and performing fault detection for individual systems. In contrast, this paper seeks to leverage large-scale cloud-based thermostat data of thousands of systems to determine hard and soft faults. Service companies are generally more interested in finding soft faults, such as capacity degradation and control problems, rather than hard faults or total failures, since soft faults are generally difficult to detect and may not be recognized by occupants for an extended period of time [16].

We propose a statistics-based FDD method that finds anomalous operation behavior and preliminarily categorizes fault types by comparing fault indicators (known as features) among multiple systems. In this way, systems having soft faults would be identified as outliers exhibiting anomalous operational behavior. The general analytical process is described as follows. First, the method identifies systems within a similar climate region by relating the system location information to the U.S. climate zones, according to the International Energy Conversion Code climate regions and moisture regimes shown in Figure 3.1 [5]. Second, from a particular climate region,

thermostat time-series data of a group of systems is queried from the database and preprocessed. From the resulting data, raw system cycling mode data (i.e., cooling on, heating on, system off) are transformed to more specific categories of operating modes, namely regulating, tracking and free response modes. The time-series data is segmented and classified with these new mode labels [6]. Based on the type of abnormal behavior to be identified, multiple features will be selected and extracted from the segmented data. Generally, the features characterize system performance, indicate faults, and problem severity, or describe the overall condition of a house and its ambient environment. Third, statistics methods are applied to compare the features among all systems in the selected group to identify outliers. Finally, labeled systems with operational faults affecting system efficiency and occupancy comfort are outputted.

Using this general approach, the performance of a system can be deeply inspected, including both the pseudo steady-state and transient behavior in cooling and heating modes. This paper focuses on the time periods that exhibit poor transient behavior, known as poor tracking periods. Both cooling mode and heating mode can exhibit this type of behavior, but this study only focuses on the cooling mode. A poor tracking period is defined as a single cooling on-cycle that usually occurs during high cooling load, and during which the indoor temperature can increase a few degrees above the cooling setpoint, unable to be controlled although the system is operating continuously. Note that the term 'high cooling load' is regarded as relative to the system capacity. If the capacity of a system is far below its rated capacity, even normal loads may well be incapable for it to deal with and thus will be considered as high loads. Therefore, this situation is unlikely to happen for a properly sized fault-free system, and is often related to the capacity degradation fault.

In this paper, Section 3.3.1 and 3.3.2 discuss two necessary and innovative data preprocessing steps: data cleansing and mode labeling, which are applied prior to every statistics-based FDD algorithm we proposed. After this, the algorithm will analyze the poor tracking periods specifically. Section 3.3.3 will introduce a list of features which help to find this poor transient behavior and are able to indicate its severity. Section 3.3.4 will discuss an unsupervised machine learning method, hierarchical clustering, to identify the poor tracking period using some of the extracted features.

Then, in Section 3.4, statistics tests on the features indicating problem severity will be performed among systems. In order to provide further implications of the features and recommend thresholds to label faulty systems, the statistical distribution of the features are estimated by the kernel density estimation. Finally, Section 3.5 will show statistics of the fault detection results as well as a couple of case studies, illustrating the effect of operational faults on system efficiency and occupancy comfort.

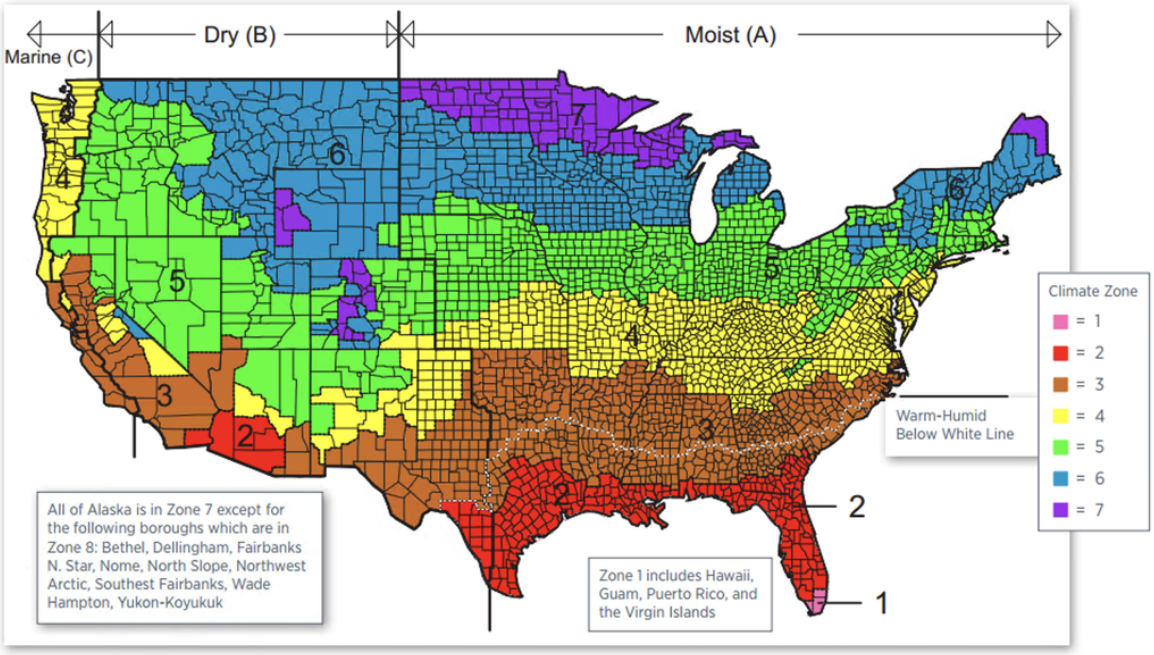


Figure 3.1: International Energy Conservation Code climate zones and moisture regimes [5]. Climate zones: 1-7 in different colors. Moisture regimes: marine, dry, and moist from west to east.

3.3 Innovative Data Preprocessing Procedures

Preprocessing procedures are indispensable in order to perform correct and accurate statistical tests for fault detection. Once raw thermostat data is queried from the database, four preprocessing procedures will be conducted sequentially, namely data cleansing, labeling modes of operation, feature extraction, and identification of poor tracking periods. In brief, data cleansing removes data points with sensor faults and time periods where the thermostat is offline. Then the mode

labeling transforms raw data into a more useful dataset by defining three system operation modes. Afterwards multiple useful data features are extracted from each operational mode. Finally, poor tracking periods are identified from the cooling tracking mode, one of the three operation modes, by using an unsupervised machine learning method.

3.3.1 Integrity Verification: Data Cleansing

The raw thermostat data is event-based, which means that the database is updated only when a new event occurs. For example, when the indoor/outdoor temperature changes more than a quantized amount, the cooling/heating setpoint is changed, or the system starts/stops. The event-based dataset records the event and corresponding time, and takes up less digital storage space than the uniformly sampled time series data. Sometimes, however, sensor integrity issues and network communication issues can occur (e.g. the sensor is unplugged, the thermostat goes offline or otherwise fails to communicate data). These issues could strongly affect the analytical process. For example, if a setpoint change event is not recorded, then the thermostat may seem to control the indoor temperature to value different from the supposed setpoint; or if a cooling off event is missed, then the system may appear to continue running for a long time. Therefore, after the raw data is queried, the first preprocessing step is to cleanse the data and remove these issues. This algorithm uses a combination of filters and logical tests to identify and remove these data.

3.3.2 Data Partitioning and Classification: Labeling Modes of Operation

After the raw data is cleansed, the mode labeling algorithm is applied to transform the raw operational data (cooling, heating, off) into a more useful dataset based on mode: regulating mode, tracking mode, and free response mode [6]. Note that both heating and cooling have respective regulating and tracking modes.

- In the regulating mode, a system cycles on and off to maintains a relatively constant indoor temperature around the setpoint. The space being cooled or heated is assumed to be in a pseudo steady-state condition.

- In the tracking mode, a system remains powered on to reach the setpoint and exhibits transient behavior. Typically, the cooling tracking mode is activated after a drop in cooling setpoint or during high cooling load; the heating tracking mode is activated after a rise in heating setpoint or during high heating load.

- In the free response mode neither setpoint is active. A free response mode may occur after a rise in cooling setpoint or a drop in heating setpoint, or when both cooling and heating loads are low such that the indoor temperature remains between setpoints for an extended period.

Figure 3.2 provides an example of labeling modes of operation. In Figure 3.2a, the cyan color shows the cooling regulating mode when the indoor temperature is fluctuating and maintained relatively constant, and the red color shows the cooling tracking mode when the indoor temperature drops significantly. In Figure 3.2b, the green color shows the free response mode when indoor temperature rises above the heating setpoint and the orange color shows the heating regulating mode when the indoor temperature drops to the heating setpoint.

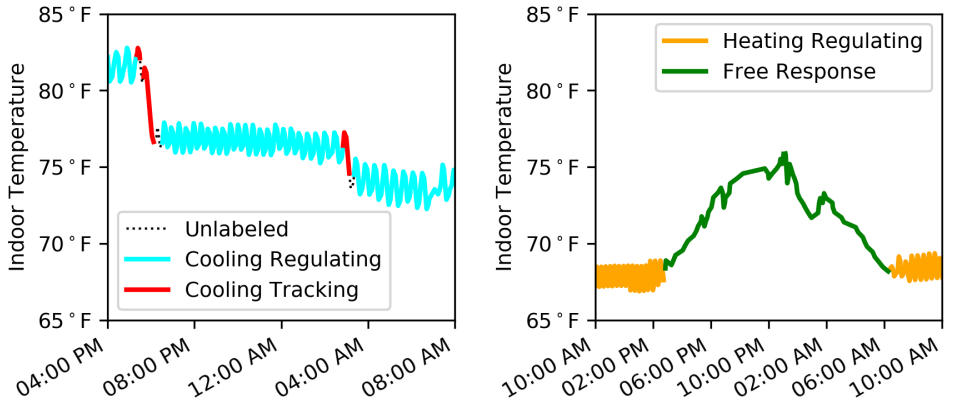


Figure 3.2: An example of labeled modes of operation. Subplot (a) shows the cooling regulating and cooling tracking modes, and subplot (b) shows the heating regulating and free response modes.

3.3.3 Feature Generation: Extracting Features from the Cooling Tracking Mode

Within the operating modes introduced above, this paper studies the poor transient behavior, corresponding to the cooling tracking mode activated during high cooling load. Note that even when the cooling tracking mode is activated by a drop in cooling setpoint, a system can sometimes exhibit poor transient behavior if the cooling load is too large for the system. The features listed in Table 3.1 are extracted from the cooling tracking periods. Among all of the features, the first three are calculated within each cooling tracking period, and will be applied to identify the poor transient behavior; the fourth and fifth features are comprehensive indicators averaged over all poor tracking periods for each system, and will be used to characterize the problem severity.

Table 3.1: Supporting features for the cooling tracking mode.

Symbol	Feature name	Definition
$T_{i,inc}$	Indoor temperature increase [°F]	The maximum indoor temperature increase*/** in a cooling tracking period.
Δt_{inc}	Increase time [hr]	The time required to reach the maximum indoor temperature increase.
$\dot{T}_{i,inc}$	Indoor temperature increase rate [°F/hr]	The rate at which indoor temperature increases; defined as $\dot{T}_{i,inc} = T_{i,inc}/\Delta t_{inc}$.
$\overline{DH}$	Average degree hour [°F·hr]	The integral of the difference between the indoor temperature and the cooling setpoint, during a time segment exhibiting poor transient behavior, averaged over the whole cooling season; defined as: $\overline{DH} = \frac{1}{N}\sum_{n=1}^{N}\int_0^{24} I \cdot \max(0, T_{spe})dt$; where N is the number of days in a cooling season; I equals 1 in poor tracking periods, and 0 in other cases.
$\bar{T}_{i,max}$	Average maximum indoor temperature [°F]	The average of ten highest $T_{i,max}$ in all poor tracking periods** after removing five most extreme instances.

Notes: * A searching algorithm is applied to find the maximum increase in a time series.

** If a poor cooling tracking period is longer than 24 hours, then it will be segmented with one period a day. Generally, there is at most one poor tracking period each day, and the maximum indoor temperature occurs in the afternoon when cooling load is the highest.

During normal cooling tracking periods activated by a decrease in cooling setpoint, the indoor temperature should decrease (or increase a very small amount) after cooling starts, such that both the value of the indoor temperature increase and the temperature increase time should be very small. For the poor tracking periods, however, both of these two features could show very high values.

3.3.4 Behavior Recognition: Identification of Poor Tracking Periods

In this step, the features extracted above will be analyzed to identify all poor transient behavior of a system. In total four types of transient behavior are common in the cooling tracking mode. Recall that a cooling tracking period is just one single cooling on-cycle. Figure 3.3a shows normal transient behavior (the most common), where the cooling system starts after a drop in setpoint and remains on until reaching the new setpoint. Figure 3.3b shows poor transient behavior, where the indoor temperature increases over 10°F before returning, despite the cooling system operating continuously for nearly 24 hours. Both Figure 3.3c and 3.3d show interesting transient behavior where the indoor temperature first increases precipitously after the cooling system turns on. The behavior in Figure 3.3c may be caused by lack of insulation on the return air duct in the attic, such that warm air is initially blown into the house when the fan starts after a long downtime; the behavior in Figure 3.3d may be caused by the internal heat load of the thermostat itself on the sensor, such that indoor temperature seems to rise as soon as the fan starts (after cooling starts) and the air circulation direction of the room changes.

Three features, namely indoor temperature increase ($T_{i,inc}$), increase time (Δt_{inc}), and increase rate ($\dot{T}_{i,inc}$)), are used together to separate the different causes of poor transient behavior. As a first step, normal transient behavior should exhibit $T_{i,inc} \approx 0$, and hence these normal periods can be eliminated by a simple range filter, e.g., $T_{i,inc} > 1$°F. Distinguishing between the types of undesirable transient behavior, however, is more difficult. Poor transient behavior could exhibit long Δt_{inc} and low $\dot{T}_{i,inc}$), or vice versa. Figure 3.4a shows the kernel density estimation of all tracking periods after removing those identified to be normal, with darker regions having more data. The data forms two groups, shown by Figure 3.4b. The cluster on the left could only correspond to

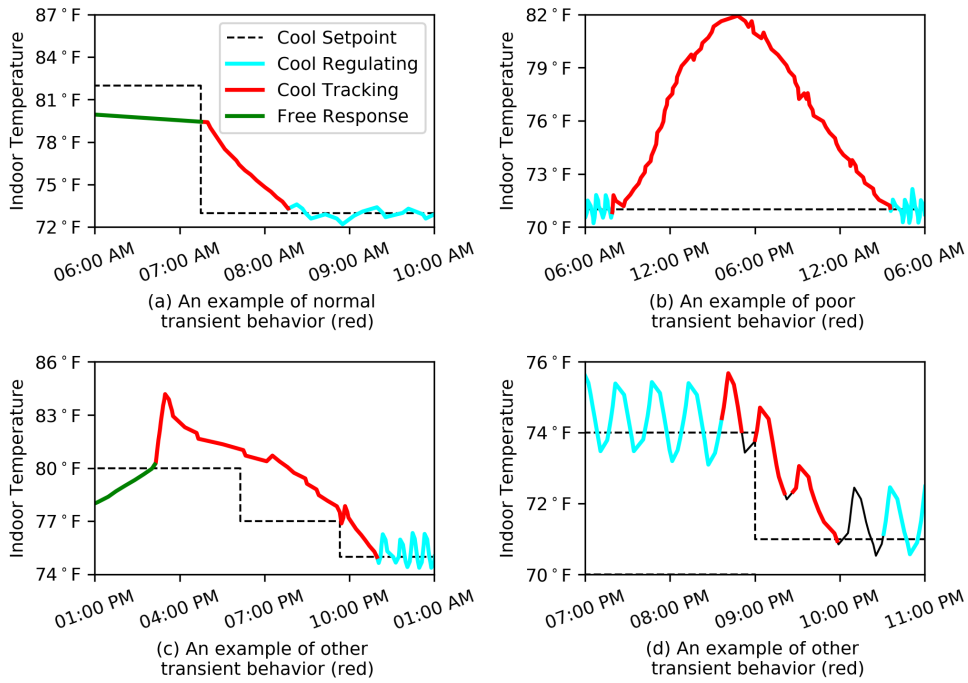


Figure 3.3: Four main types of transient behavior exhibited in the cooling tracking mode (red color).

the poor tracking periods due to its long increase time, and the other on the bottom could only correspond to other types of behavior due to its fast increase rate. Only data between two groups in the lower left region has some ambiguity. However, based on an assumption that if a system has some confirmed poor tracking periods, then the ambiguous ones are more likely to be poor tracking periods as well, such that the hierarchical clustering method can be applied to group the two behavior for each system.

In this study, Ward's agglomerative hierarchical clustering method [38] is applied to build a hierarchy of clusters from the transient data of each system. This method starts with each data point in its own cluster. Then as the algorithm moves up the hierarchy, it seeks to merge pairs of clusters which can minimize the variance of the clusters. In order to exactly find the two clusters of behavior, some generated data from the two black circles in Figure 3.4b is added into the real data. Note that the black circles correspond to the core regions of each cluster, taking up 50% of the mass, with known true labels. Also, features on both axes are normalized such that the Euclidean

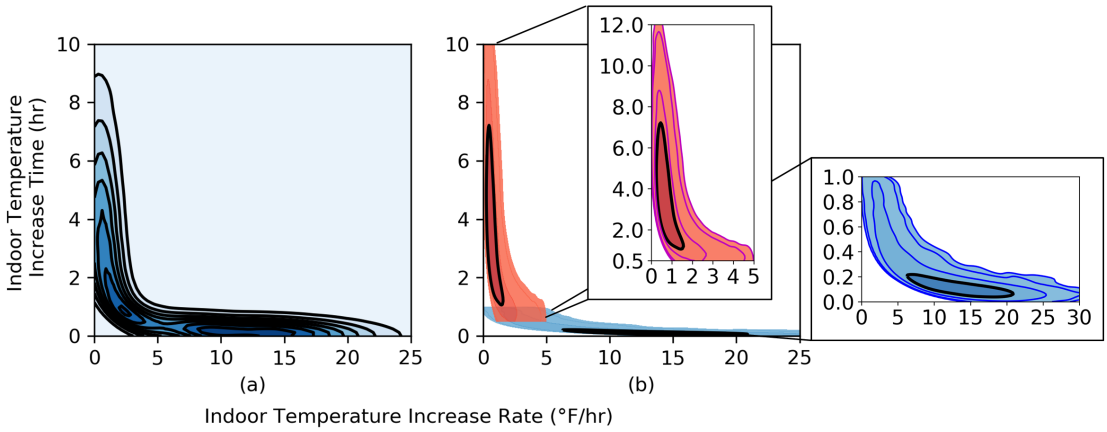


Figure 3.4: (a) Kernel density estimation of the remaining tracking periods after the normal ones are removed. (b) The data forms two groups. The group in red corresponds to poor transient behavior, and the group in blue corresponds to other transient behaviors. The black circles show the core region (50% of mass) of each group.

distance is applicable. By this means, the algorithm iterates over each system to label the tracking periods mixed with generated data. Since the two clusters are defined by expertise and experience, this approach can be regarded as setting soft thresholds between the two behaviors. After checking the results, the method is found to be stable and robust, and can correctly label the data of either type of behavior. Therefore, all cooling tracking periods labeled with poor transient behavior can be separated from the other behavior.

3.4 Statistics-Based Fault Detection Method

This section presents the proposed statistics-based fault detection method. In order to compare system performance to the entire population and identify outlier behavior, the statistical distribution of two features, namely average degree-hour above setpoint and average maximum indoor temperature, will be investigated among all systems by means of kernel density estimation. These two features could strongly indicate the existence and the severity of operational faults, and thresholds for identifying faulty systems will be recommended.

3.4.1 Statistical Distributions of Features

Once all poor tracking periods are identified, a statistics-based fault detection algorithm can be applied using two additional features, namely average degree hour above setpoint and average maximum indoor temperature. The average degree hour above setpoint ($\overline{DH}$, unit: [°F·hr]) averages the area between indoor temperature and cooling setpoint for only all poor tracking periods (see Figure 3.3b as an example) each day in the whole summer cooling season, which represents the system effectiveness and occupant comfort. For instance, a system with $\overline{DH} = 20$°F·hr may on average have indoor temperature 4°F above setpoint lasting for 5 hours every day in the summer, even though cooling runs in full power. The average maximum indoor temperature ($\bar{T}_{i,max}$) is calculated by averaging ten highest indoor temperatures during poor tracking periods, after first removing the five most extreme instances in the cooling season. Thus, this feature represents the most severe instances of occupant discomfort. For instance, a system with $\bar{T}_{i,max} = 85$°F will have indoor temperature rising over 85°F on average for at least 15 days in the summer, even though cooling runs in full power. These two metrics characterize the severity of the detected problem.

To exemplify the distributions of the two features, approximately 10,000 residential single-stage cooling systems in 2A IECC climate zone (see Figure 3.1) in Florida are queried with data from June 1st to September 30$^{\text{th}}$ in 2019. Among them, 904 systems have been identified with more than 15 (a hard threshold) poor tracking periods. The kernel density estimation of the average degree hour above setpoint and average maximum indoor temperature are shown in Figure 3.5. From Figure 3.5a and 3.5b, most systems have average degree hour above setpoint less than 20°F·hr and average maximum indoor temperature less than 85°F. From the 2-D distribution in Figure 3.5c, one can find that generally $\overline{DH} > 15$°F·hr along with $\bar{T}_{i,max} > 80$°F is abnormal.

3.4.2 Thresholds Recommendation for Fault Detection

Although high $\overline{DH}$ and high $\bar{T}_{i,max}$ are not equivalent to faults, they can strongly indicate the existence or severity of faults. We hope to use the distributions of these two metrics to set thresholds for fault and estimate fault severities. Practically the real numbers of faulty systems are

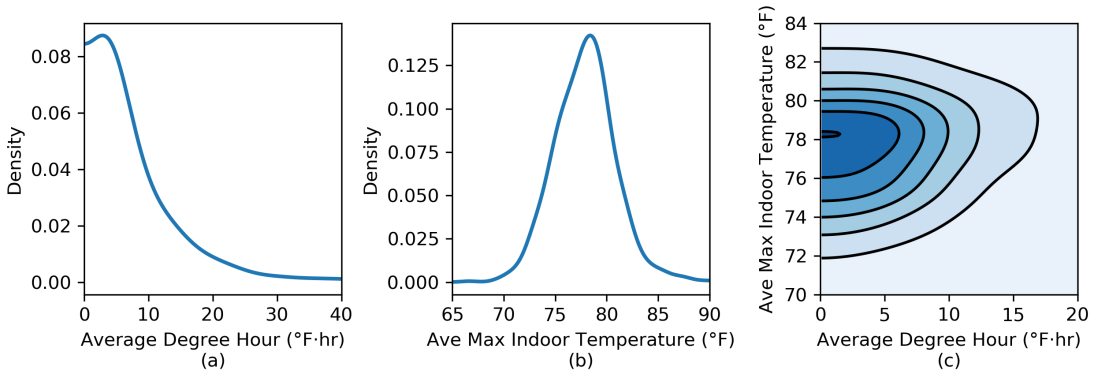


Figure 3.5: Kernel density estimation of (a) the average degree hour above setpoint, (b) average maximum indoor temperature, and (c) both features together, for all 904 systems identified with more than 15 poor tracking periods.

unknown and the exact boundary between faulty and fault-free systems is indeterminate. Furthermore, all systems exhibiting poor transient behavior could have some level of operational faults. Nevertheless, to be conservative, in this study the thresholds are adjusted such that about 5% systems with the poor transient behavior are classified as faulty. Note that approximately 9% systems are identified with poor transient behavior, so this threshold will label about 0.5% among all systems to be faulty. The threshold is flexible and subject to change based on the service companies' preferences. The study finally takes 80% quantile of the $\overline{DH}$ and $\bar{T}_{i,max}$ in the dataset as thresholds, respectively 12.4°F·hr and 80.0°F, calculated from the estimated probability density function in Figure 5a and 5b. Any systems with both features above thresholds are flagged. Two case study examples are shown in the next section.

3.5 Results and Discussion

Using the approach outlined above, systems that exhibit faulty behavior are flagged. These faults could be related to either capacity degradation faults or undersizing issues. In this section, general fault detection results are summarized, and two typical cases are presented, with each representing one type of common behavior. The results demonstrate the potential of large data fault detection methods for residential HVAC systems using cloud-based thermostat data.

3.5.1 Summarization of Fault Detection Results

Figure 3.6 briefly summarized the fault detection results. Figure 3.6a shows how faulty systems are flagged layer by layer. The Venn diagram in Figure 3.6b shows a total of 181 systems are flagged by each feature using the proposed thresholds, and the 68 systems at the intersection of the two groups are labeled as faulty. The next three scatterplots illustrate the behavior of these 68 labeled faulty systems in the 2019 summer cooling season. Figure 3.6c shows that in most of these homes the indoor temperature often reaches to around 84°F although the cooling setpoint is usually set to around 74°F, and there are some extreme systems having very high indoor temperature and degree hour above setpoint. Figure 3.6d compares the previously defined degrees hours with the degrees hours when the indoor temperature is 2°F above setpoint. The latter feature can be calculated by the equation in Table 3.1 assuming the setpoint increased by 2°F. The result shows strong linear relationship, indicating higher degree hours means both higher temperature and longer time. Occupants living in a high-degree-hour conditioned space have to endure high indoor temperature well above the setpoint value they selected for several hours every day. Finally, Figure 3.6e shows that most faulty systems spend between 400 and 2000 hours with poor tracking during the cooling season. No particular relationship can be found between maximum indoor temperature and total time on poor tracking due to the difference in setpoint values for each system.

3.5.2 Case Study I: Significant Time above Setpoint, High Maximum Indoor Temperature

Figure 3.7 displays a part of daily operation of a selected faulty system, recorded from June 10th to June 14th. Note that this is the most common behavior among all labeled faulty systems. The first plot records the indoor temperature variation and cooling setpoint changes with modes labeled in red, cyan and green color. The second plot records the cooling load factor (i.e. on/off status) of this single-stage cooling unit. The third plot records the outdoor temperature variation. From this figure, the system frequently shows long periods of poor transient behavior, and the indoor temperature can spike from 75°F to above 85°F when the outdoor temperature approaches 90°F in the afternoon. Typically, every day the system operates all the time from the sunrise to sunset,

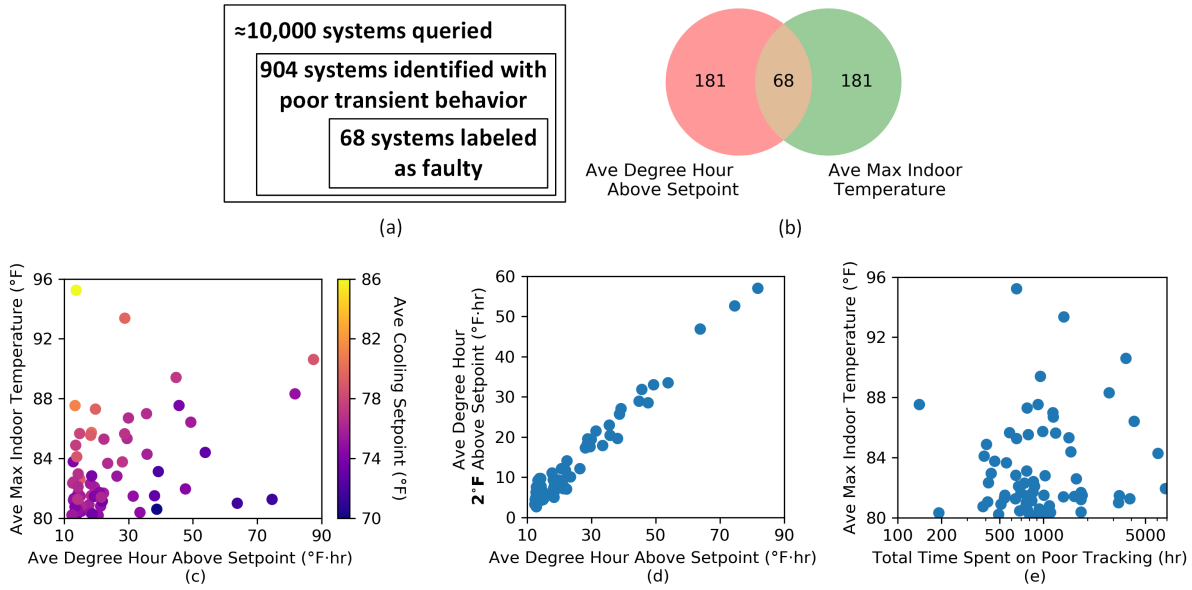


Figure 3.6: Illustration of the statistics-based fault detection results. (a): General statistics. (b): A Venn diagram showing the numbers of flagged systems. (c) to (e): Statistics of the 68 faulty systems in the summer cooling season.

until the outdoor temperature drops below 80°F. It works very little during the night, because the cooling setpoint is above the outdoor temperature and the house is cooled passively by the ambient environment.

Statistics shows the average degree hour above setpoint is 45.7°F·hr and average maximum indoor temperature is 87.5°F. These numbers could be understood in the following way: the occupants in the house on average experience 5°F above setpoint every day for 9 hours, and the maximum indoor temperature in a day approaches 87.5°F on average for at least 15 days in summer. Clearly the comfort level is very low, and the system capacity does not meet the cooling load.

3.5.3 Case Study II: Significant Time above Setpoint, Low Maximum Indoor Temperature

Another example shown in Figure 3.8 also has many long periods of poor tracking and even have one poor tracking period lasting for several days (e.g., the period from September 8th to September 14th). The average degree hour above setpoint is 83.5°F·hr, much higher than the former

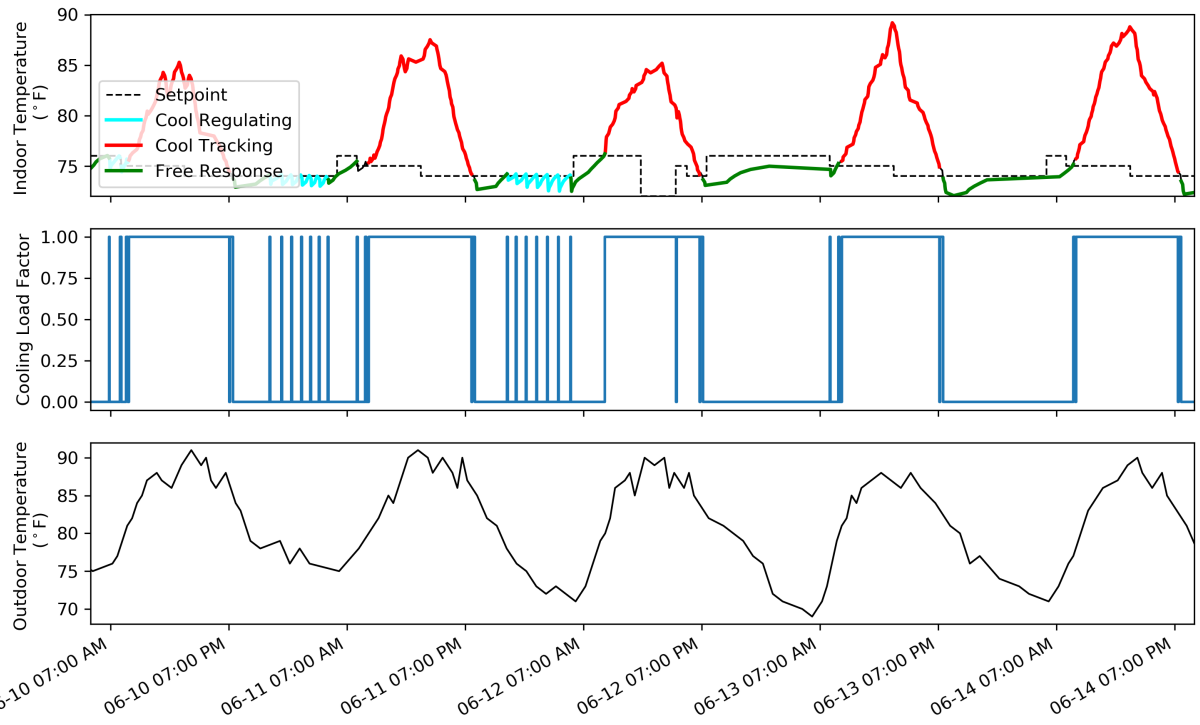


Figure 3.7: Daily operating condition of system I with labeling modes of operation, from June 10th to June 14th. Red: poor cooling tracking periods; cyan: cooling regulating periods; green: free response periods; dotted line: cooling setpoint.

example. This case, however, is very different from the former one in that the occupant comfort is not totally lost. The average maximum indoor temperature is 77.5°F, and the cooling setpoint is adjusted 67.1°F on average when the system exhibits poor transient behavior. Because of the low indoor temperature, this example is *not* labeled as faulty. Nonetheless, this behavior is very common, characterized by an occupant driven issue in addition to possible capacity degradation issues.

Figure 3.8 captures a segment of the occupant driven issue. On the evening of September 8th, the occupant deliberately lowered the cooling setpoint to 63°F in order to receive more cooling, even though he or she understood that the indoor temperature might never reach the setpoint. From the figure, the system precools the home at night to approximately 67°F, and during the daytime the indoor temperature can be well-controlled below 77°F. In other words, the occupants were

forced to decrease the setpoint because he or she had realized that the system capacity was not large enough, and if the system did not operate during the nighttime, the indoor temperature would probably rise above 80°F in every afternoon. Additionally, we speculate that nobody was at home before and after the displayed poor tracking period, so the occupant could adjust the setpoint at 80°F to reduce cooling and the associated electric bill. However, this precooling method is not always effective and is symptomatic of faulty equipment. A better solution for the home owner will be to contact service companies and request maintenance.

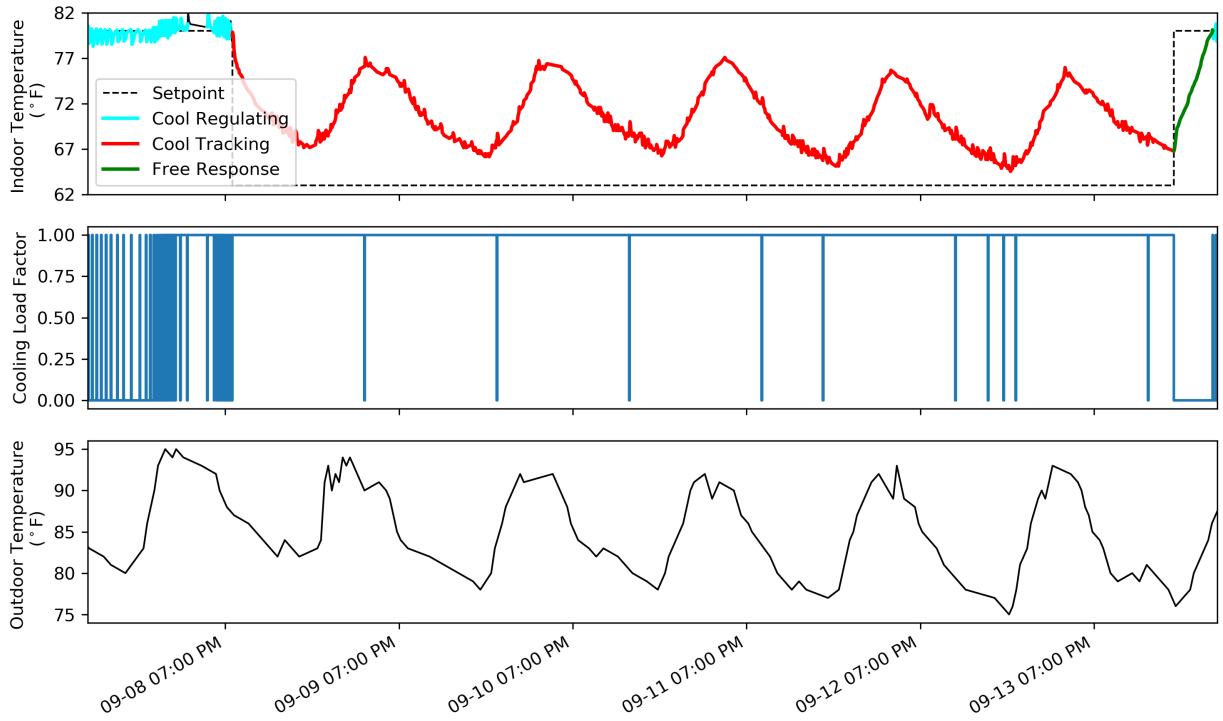


Figure 3.8: Daily operating condition of system II with labeling modes of operation, from September 8th to September 14th. Red: poor cooling tracking periods; cyan: cooling regulating periods; dotted line: cooling setpoint.

3.6 Conclusions

By means of cloud-based smart thermostat data of thousands of residential HVAC systems, the proposed fault detection method can successfully identify faulty systems from analysis of their

transient behavior, and infer the problem severity using specific data features. The flagged faulty systems exhibit low cooling capacity compared to the load, which is the result of system degradation, improper sizing, or improper commissioning, and the occupants either experience discomfort or are forced to compensate by artificially adjusting the cooling setpoint.

This paper only introduces two features indicating severity of capacity degradation fault. Other features extracted from the cooling tracking mode can potentially indicate additional faults observed only from the transient behavior. For instance, indoor temperature increase ($T_{i,inc}$) and increase time (Δt_{inc}) can be applied to diagnose problems such as lack of insulation in the house attic, thermostat misplacement issues, and smart thermostat internal heat load issues. The potential of large data fault detection methods requires both domain expertise and sound statistical analysis.

4. PAPER B: MULTIVARIATE FAULT DETECTION FOR RESIDENTIAL HVAC SYSTEMS USING CLOUD-BASED THERMOSTAT DATA, PART I: METHODOLOGY[1]

4.1 Synopsis

Fault detection and diagnosis (FDD) using smart thermostat data is a relatively new field of research, but one with immediate practical application to residential indoor climate control. This two-part paper proposes a statistics-based method for using aggregated data from many residential HVAC systems to identify outliers. These outlier systems can then be related to operational problems, such as inadequate capacity, malfunctioning control strategies, or other soft faults. The first part of this paper presents the analytical methodology. For this study historical and real-time data from thousands of cloud-based thermostats is used to extract features during the cooling mode of operation. First, the data is parsed to identify time segments that are characteristic periods of pseudo-steady state operation. Then, a multistage fault detection method compares the features among all systems to find statistical outliers. Each feature is analyzed independently using kernel density estimation (KDE) methods. Additionally, multiple features are analyzed simultaneously using multivariate statistics and the Mahalanobis distance (D_M) to determine outliers. Anomalous systems are sorted heuristically, and information from univariate and multivariate methods are combined to more precisely classify systems exhibiting specific faulty behavior. The second part of this paper provides some typical case studies, illustrating how this method can be used to identify aberrant behavior across a large number of monitored systems.

4.2 Introduction

Heating, ventilating, and air conditioning (HVAC) equipment maintains the indoor environment with appropriate temperature and humidity levels. However, it accounts for a large share of residential households' total energy consumption. According to the Residential Energy Consumption Survey (RECS) in 2015 [1], the annual average heating costs is $850 per household in

[1]F. Guo, A.P. Rogers, and B.P. Rasmussen, "Multivariate Fault Detection for Residential HVAC Systems Using Cloud-based Thermostat Data, Part I: Methodology," *Science and Technology for the Built Environment*, **In Review**.

the Northeastern United States (37.5% of energy costs). Households in the southern United States pay $392 annually for air conditioning (20.4% of energy costs). With a large portion of household expenditures devoted to climate control, 31% of US households face challenges to pay household energy costs or maintain comfortable indoor conditions [1].

Improving the operation and efficiency of a residential HVAC system can be achieved by 1) properly sizing the unit to the customer needs, 2) improving component performance efficiencies, 3) improving and optimizing system control algorithms, and 4) implementing predictive and preventative maintenance strategies. Fault detection and diagnosis (FDD) methods can assist in each of these areas by ensuring the system is properly commissioned, identifying specific operational problems, predicting gradual degradation and prompting necessary maintenance, and reducing the time required to diagnose and repair systems.

These types of diagnostic and predictive maintenance algorithms benefit residents by improving comfort and lowering energy costs, but also benefit manufacturers and service companies. Manufacturers establish a brand image for reliability and efficiency. Service companies gain customer loyalty by anticipating equipment problems and repairing them quickly and correctly. The society also benefits by reducing energy use and the corresponding emissions.

Most published fault detection and diagnosis (FDD) methods for HVAC systems focus primarily on large commercial systems, such as variable air volume air handling units (VAV-AHUs), chillers and cooling towers. About 16% of FDD studies are associated with small commercial and residential buildings, but the majority focused on rooftop packaged units (RTU) [3]. Relatively few published research efforts have been focused on residential split systems, mainly because FDD generally requires installation of additional sensors on the equipment [17]. For mass-produced residential units with small profit margins, adopting additional sensors may not be cost-effective.

An alternative data source for analysis and diagnostics in residential HVAC systems is smart thermostat data, which records important, but limited, information including indoor/outdoor temperature, temperature setpoint, and heating/cooling operational mode. In recent years, smart thermostats have been widely adopted by home owners, because they offer remote control and monitor-

ing options (e.g. via smart phones), include learning algorithms to improve occupant comfort, and provide feedbacks to users regarding energy use. Meanwhile, the huge amount of data recorded by smart thermostats provides distinctive opportunities for FDD approaches, or even promotes a potential paradigm shift. [27] presents two physics-based models and shows the thermostat data can detect building construction faults such as poor thermal insulation. Similarly, [22] focuses on optimizing occupant behavior and develops an algorithm to provide personalized home heating advice to households. While the above studies use experimental data, simulated thermostat data is also used for model validation. For example, [21] proposes a recursive least-squares model and identifies airflow restrictions and total system failures.

The previous studies leverage data from multiple thermostats to generate models that can be applied to a particular system to detect and diagnose faults. In contrast, the fault detection methods presented here are based on a paradigm of large-scale data analysis of the operational behavior of thousands of systems for comprehensive and simultaneous FDD of cloud-based residential HVAC systems. Figure 4.1 illustrates the role of this multi-system FDD method in the iterative process of predictive maintenance. Smart thermostats in every residential home continuously record and upload real-time data to a server, forming an extensive database on the cloud. Then on a remote platform, the FDD methods process the data by extracting useful features and using statistical methods or artificial intelligence for comparative analysis. Due to the large number and diversity of systems, comparing aggregate features is computationally and practically feasible. Next, systems with detected faults are ordered based on the severity of faults. Dealers and customers can be notified of faulty systems, and service technicians can be dispatched to repair systems based on priority and with a preliminary diagnosis in hand. By this means, anomalous systems are identified quickly by FDD algorithms and serviced to maintain occupant comfort and energy efficiency.

An overview of the proposed FDD algorithm is depicted in Figure 4.2. The remainder of this paper will describe each element in the methodology, and illustrate these with simple statistical examples or representative thermostat data. First, raw data is queried from the cloud and preprocessed (Section 4.3). This procedure first removes sensor and communication errors, and then

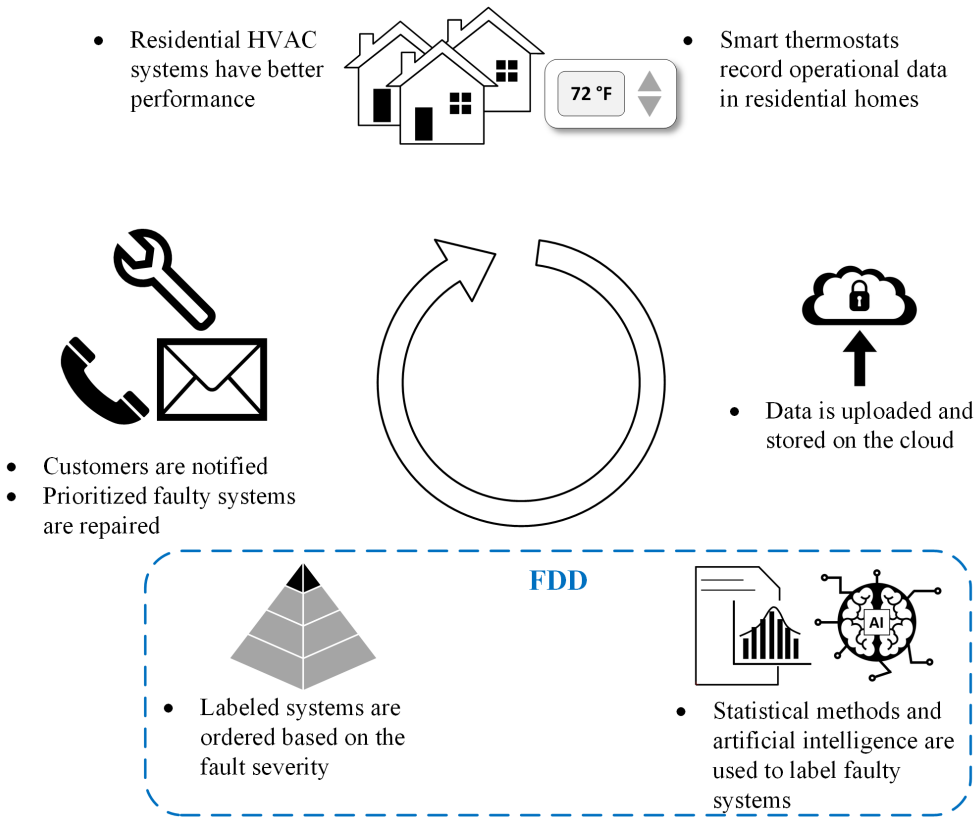


Figure 4.1: The process of predictive maintenance and the role of FDD methods.

parses the cleansed time-series data based on operating mode to extract key data features. The algorithm then uses univariate and multivariate statistical methods to identify anomalous system behavior (Section 4.4). Finally, postprocessing separates true faulty systems from among all outlier systems and performs preliminary prioritization based on severity (Section 4.5). While part one of this paper focuses exclusively on the proposed methodology, part two will present detailed case studies exemplifying the various capabilities of the proposed algorithm, using data from over 9,000 residential air conditioning systems.

4.3 Data Selection and Preprocessing

Data preprocessing is a crucial requirement before applying fault detection algorithms. The baseline models used for fault detection will not be effective unless the input data is clean and representative of the operating condition of the systems. In this section, suggestions regarding

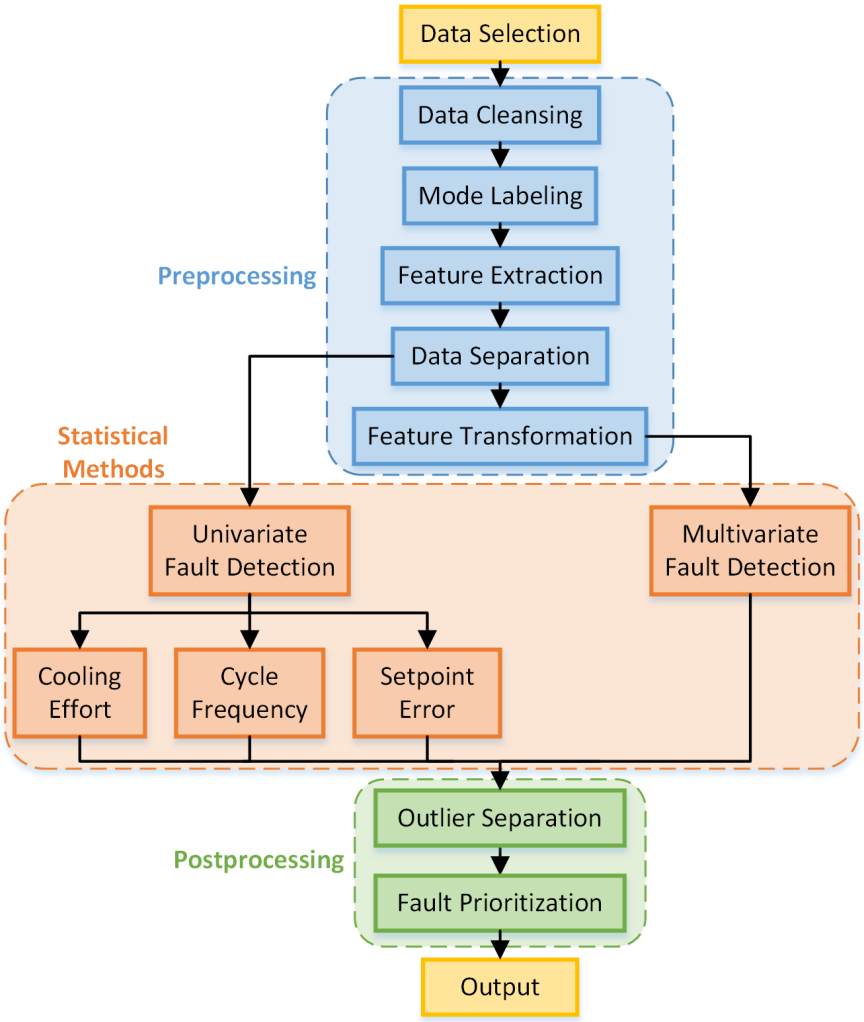


Figure 4.2: An outline of the proposed fault detection method.

data selection will be briefly discussed. Then, five major preprocessing procedures are presented, namely: data cleansing, mode labeling, feature extraction, data separation, and feature transformation. Data cleansing removes anomalous data related to sensor faults and offline thermostats; mode labeling partitions raw data, followed by then extraction of useful features to be investigated; data separation ensures comparability of features within each subgroup; and lastly transformation of features enables unbiased parameter estimations in the statistical models. A graphical representation of the preprocessing procedures is shown in Figure 4.3.

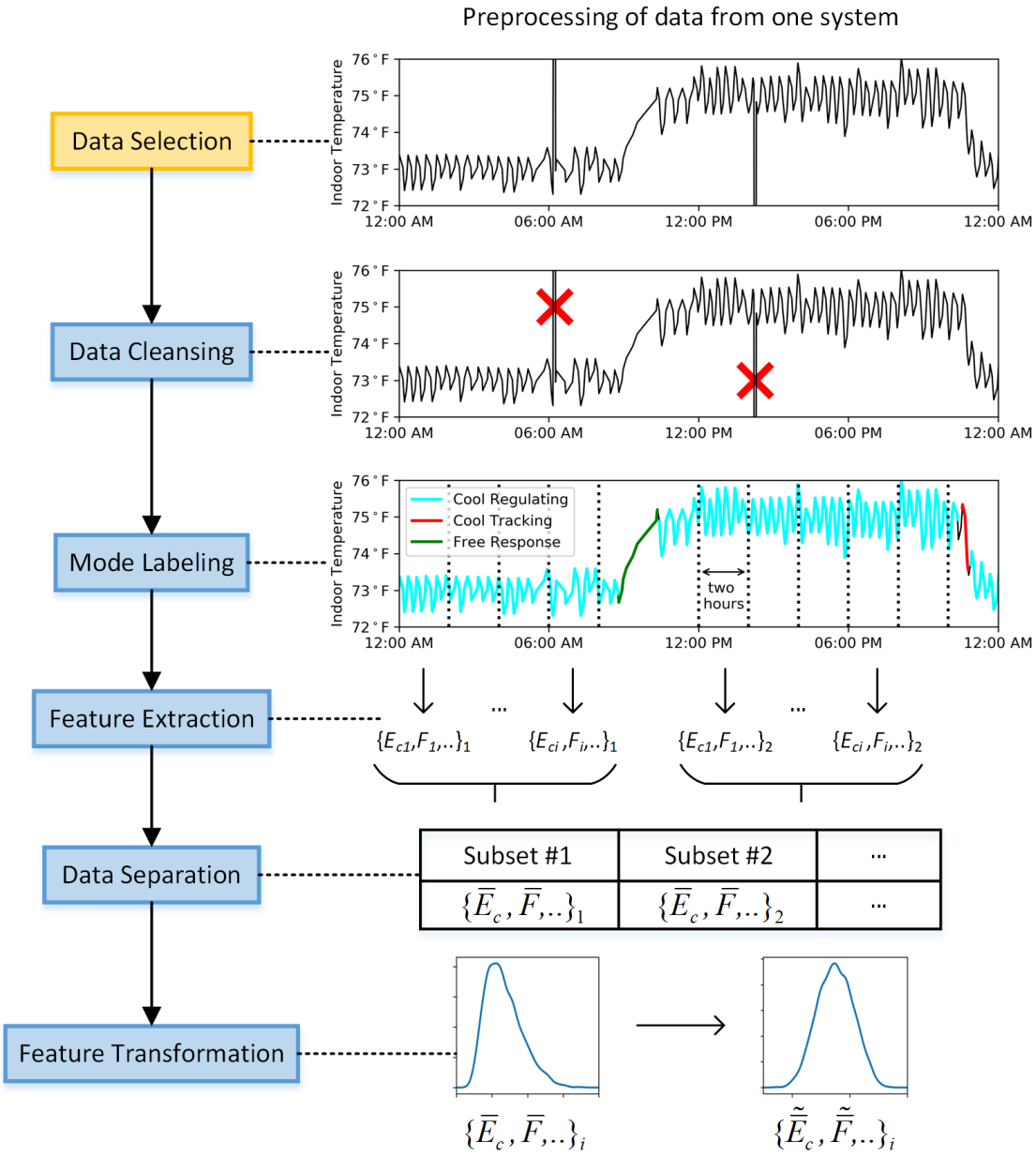


Figure 4.3: An illustration of the data preprocessing procedures. Thousands of systems are processed in the same way.

4.3.1 Data Selection

This study utilizes raw smart thermostat data drawn from a database with historical records of hundreds of thousands of systems, most with more than one year's worth of data and located throughout the United States. Since the proposed FDD method relies on comparisons of system

operating in similar conditions, data selection should consider climate zones, time period, and quantity and type of similar systems.

First, systems within similar climate zones ought to be grouped because the local climate and house size determines system size. For example, air-conditioning systems in Texas are generally larger in capacity compared to those in Minnesota. Thus, for the same indoor and outdoor temperature, the smaller system in Minnesota tends to work much harder and therefore seems undersized. However, because the summer season in Minnesota is much cooler and shorter than in Texas, comfort can be adequately maintained with a smaller capacity system.

Second, the appropriate time scale needs to be determined. Using only one or two weeks of data can be computationally inexpensive, but has significant drawbacks because it generally will not accurately capture the full range of ambient conditions. Additionally, during short periods the statistical analysis is liable to be affected by some irrelevant covariates. For example, if an occupant left a window open, the air conditioner may seem not to be working properly. These types of occupant or behavioral faults are not the primary focus of this paper, and their effects can be lessened using data from a relatively long time period. Shorter time periods may also limit the amount of useful data for analysis (e.g. a two-week period where the system operates infrequently). Conversely, time periods exceeding one year may not be appropriate, as the operating condition of a system may have changed significantly due to degradation or repairs. Therefore, in this study, thermostat data spanning a 3-month period is chosen to carry out a relatively comprehensive investigation for each system.

Third, the database contains a variety of air conditioning system types, including systems with single-stage, dual-stage and variable-speed compressors. Because the operational behavior and many data features depend on system type (e.g. on/off cycle frequency), each system type should be analyzed independently. The results presented in this paper are based on data collected between June and August in 2018 from approximately 9,000 residential single-stage cooling units located in Texas, Georgia, North Carolina, and Florida.

4.3.2 Data Cleansing

The raw thermostat data in the database is event-based. In other words, new data is recorded only when a change occurs: indoor temperature setpoint is updated, the system cycles on or off, or a measured temperature changes more than a defined amount. However, cloud-based thermostats rely on home internet connections, and may go offline and fail to communicate. Thus, when the data is selected, the first preprocessing step is cleansing data to remove the communication and sensor integrity issues, as these anomalous data strongly interfere with the effectiveness of the statistics-based models. Data cleansing is performed by a rule-based filter that is capable of removing data with the following faults:

- The indoor or outdoor temperature climbs above 130°F or goes below 0°F (occasionally the thermostat will list the temperature as 2^{15}=32,768°F when the thermostat goes offline).
- The temperature increases or decreases by more than 20°F between two consecutive updates. This is often caused by sensor integrity issues.
- The indoor or outdoor temperature has not updated for more than four hours. If the sensor fails to update for a period of time and then suddenly resumes, then removing these communication failure periods is necessary.
- The system maintains a constant temperature with little variance (tightly), but the setpoint error is large. In this case the reported setpoint is likely incorrect.
- The system runs for more than 24 hours. In most cases this is caused when the system "off" signal is not received.

Fortunately, the thermostat system is generally robust and requires applying only these simple filters to ensure data validity. The remaining data is presumed to be an accurate record of system behavior.

4.3.3 Mode Labeling

After erroneous data are removed, but before extracting specific data features for analysis, a critical process termed *mode labeling* is applied to partition raw thermostat data and classify the data sections [6]. A typical cycling HVAC system has three *operational* modes: cooling, heating, and off. However, these modes are less useful for analysis than differentiating between time periods when the system is cycling about a fixed temperature setpoint, adjusting to meet a sudden increase in cooling demand, or simply allowing the home temperature to respond passively to changes in ambient temperatures. Within each of these *behavioral* modes, specific data features can be subsequently extracted and analyzed.

- A system is said to be in a ***regulating*** mode when the system cycles on and off and maintains a relatively constant temperature. The system may have both cooling and heating regulating modes. Critical data features could include cycle frequency and duty factor.

- A system is said to be in a ***tracking*** mode when the system operates continuously to reach the desired setpoint. Again, both the cooling and heating functions have tracking modes. Tracking modes are typically activated after a decrease in the cooling setpoint or an increase in the heating setpoint. Critical data features for this mode could include temperature decay rate.

- A system is said to be in a ***free response*** mode when neither setpoint is active. A free response mode may occur after a setpoint change (i.e. a rise in cooling setpoint or a drop in heating setpoint) or when the cooling and heating loads are low such that the indoor temperature remains between heating and cooling setpoints for an extended period. Critical data features for this mode could include the thermal time constant of the house.

The mode labeling algorithm first determines whether the temperature during a given on/off cycle is similar to the cycles before and after it. If the indoor temperature fluctuations of a cycle are similar to the preceding and following cycles, it will be considered part of a regulating mode.

Otherwise, if the temperature is not similar to either the preceding or the following cycles, then the indoor temperature is assumed not to be maintaining a constant temperature, and the mode may be labeled as tracking (if the system is on) or free response (if the system is off). After this initial rule-based filter, several additional conditions are applied to improve the accuracy of the labeling algorithm. For example, regulating modes must contain a minimum number of on/off cycles to ensure that the temperature is being maintained for an extended period. Tracking modes must contain a minimum change in temperature during the run cycle, and free response modes must persist for a minimum duration. For more complete details regarding the mode-labeling procedure, see [6].

Figure 4.4 provides an example of labeled modes of operation. In Figure 4.4a, the system is in cooling tracking modes when the indoor temperature drops significantly, and the system is in cooling regulating modes when the indoor temperature is maintained relatively constant. In Figure 4.4b, the system is in a free response mode when indoor temperature rises above the heating setpoint and then re-enters a heating regulating mode when the indoor temperature drops to the setpoint.

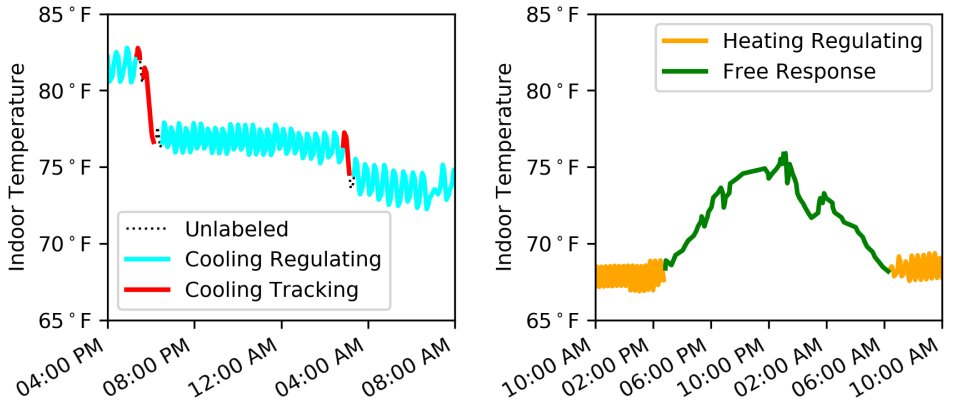


Figure 4.4: An example of labeled modes of operation using the proposed preprocessing methods.

4.3.4 Feature Extraction from Mode Labeling

The mode labeling algorithm sections raw data into regulating, tracking and free response modes, and after that various features can be extracted from each mode. The proposed fault detection method specifically focuses on the cooling regulating mode, and the extracted features are listed in Table 4.1.

Table 4.1: Supporting features for cooling regulating mode.

Symbol	Feature name	Definition
ΔT_{oi}	Temperature difference [°F]	Temperature difference between home (indoor) and ambient (outdoor) $\Delta T_{oi} = T_o - T_i$
$\Delta \omega_{oi}$	Humidity ratio difference [kg H_2O/kg dry air]	Humidity ratio difference between home (indoor) and ambient (outdoor) $\omega_{oi} = \omega_o - \omega_i$
E_c	Cooling effort [%]	The product of cooling duty factor (D) and cooling load factor (L) (see Equation 4.1)
F	Cycle frequency [cycle/hr]	The number of start/stop cycles per hour
T_{spe}	Cooling setpoint error [°F]	The error between indoor temperature and cooling setpoint $T_{spe} = T_i - T_{spc}$

Among these features, temperature difference and humidity ratio difference are the most basic descriptors of the conditions of a house and its ambient environment. Cooling effort describes how hard a system is working by denoting the overall portion of capacity for the given time period. In a period Δt, the time-averaged cooling effort is calculated by

$$E_c = \begin{cases} \frac{1}{\Delta t} \int_{t=0}^{t=\Delta t} L(t)dt & \text{for variable-speed systems} \\ D \cdot L & \text{for multi-speed systems} \\ D & \text{for single-speed systems} \end{cases} \tag{4.1}$$

where duty factor (D) is defined as the fraction of time that the system is running, and load factor (L) is defined as the fraction of full capacity at which the system runs. A high cooling effort during moderate climate conditions may indicate an undersized or faulty system, while an abnormally low cooling effort during typical conditions may indicate an oversized system.

However, cooling effort alone does not fully capture the operational behavior because the same duty factor can occur with short or long on/off periods. Cycle frequency successfully captures this distinction and can be a strong indication of the system sizing relative to the load.

Cooling setpoint error shows the degree of deviation of indoor temperature from temperature setpoint as a system cycles on and off in the regulating mode. A properly sized and well-maintained system will generally have an average setpoint error close to zero in regulating conditions. (If the system is operating in a dehumidification mode, the setpoint error may be below zero, as the temperature is deliberately controlled lower than the occupant-defined setpoint).

For these five features cooling effort, cycle frequency, and setpoint error characterize the operational *behavior* of a system, while temperature difference and humidity ratio difference define the operational *conditions*. Temperature and humidity strongly affect the behavior of a system, but do not indicate any anomalous behavior. Thus, in the following process they will serve to separate data features into subsets of similar ambient conditions, rather than being indicators of faulty systems.

4.3.5 Data Separation

The previous section presented the mode labeling algorithm and listed the features extracted from the cooling regulating mode. This section proceeds to further split the extracted features based on operational conditions. As shown in Table 4.2, the data is separated into eight subsets according to temperature difference (ΔT_{oi}) and humidity ratio difference ($\Delta \omega_{oi}$). This allows the comparison of data features with similar temperature and humidity conditions.

Additionally, extreme or outlier data subsets are excluded based on the following criteria:

- Operational data under extreme weather conditions (i.e. outside of the ranges given in Table 4.2).

Table 4.2: Data separation according to temperature and humidity differences.

Subset number		ΔT_{oi} range [°F]			
		$0 \sim 4$	$4 \sim 8$	$8 \sim 12$	$12 \sim 16$
$\Delta\omega_{oi}$ range [kg H_2O/kg dry air]	$0.002 \sim 0.007$	#1	#2	#3	#4
	$0.007 \sim 0.012$	#5	#6	#7	#8

- Data with cooling setpoint error lower than -0.5°F. This condition occurs when systems are operating in a dehumidification mode.
- A subset of data for a single system is less than 25. This is to ensure an unbiased analysis.

After this step, for each system in each subset the average of cooling effort, cycle frequency and temperature setpoint error is calculated. This data is a compact and accurate summary of the operating behavior of each system at different weather conditions.

4.3.6 Transformation of Features

In the previous data separation process, the dataset is separated into eight subsets, and in each subset the average value of each feature is calculated. Figure 4.5a shows the kernel density estimation of cooling effort data (E_c) for all systems in subset #1. Importantly, the distribution is skewed, due to inherent characteristics in the operating behavior. Similarly, the distributions of cycle frequency and setpoint error in each subset are also skewed. The statistics-based algorithms to be introduced in Section 2.4 include both univariate and multivariate fault detection. While the univariate method is unaffected by the skewed distribution, the proposed multivariate technique uses a method called the minimum covariance determinant (MCD) estimator [39], which has a better estimation of the feature mean and variance if each feature is individually normally distributed. Hence, the multivariate method necessitates a data transformation procedure to normalize the distribution.

In this study, the Box-Cox transformation [40] is applied to yield datasets which are approximately normally distributed. In the Box-Cox transformation, datasets must only have positive values. Here, this rule is satisfied for cooling effort and cycle frequency, as these two features

only have positive values. Because the cooling setpoint error is greater than -0.5°F, the values are shifted to positive by adding 0.5°F to each data point.

Suppose the data X_1, X_2, ... X_n, is independent and identically distributed random variables which has skewed probability density function. Then, Box-Cox transformation is defined by

$$\widetilde{X}(\lambda) = \begin{cases} \frac{X_\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(X) & \text{if } \lambda = 0 \end{cases} \tag{4.2}$$

By using maximum likelihood estimation to optimize the value of λ, it can often transform datasets into an approximate normal distribution. For any λ, the transformation function is monotonically increasing with respect to X, which guarantees the order of all data samples are unchanged.

After the transformation (shown in Figure 4.5b), the KDE is converted to a minimally skewed normal distribution, which is applicable for further analysis. To test the performance of the transformation, a quantile-quantile plot is drawn for comparing the transformed quantiles against the quantiles in the normal distribution. Figure 4.5c shows one plot for cooling effort (E_c) in subset #1. The fitted line has an R^2 value of 0.998, which indicates that the transformed data is very close to the normal distribution. Applying this transformation technique to the remaining data features and subsets yields similar results, with R^2 value always greater than 0.9. In contrast, for data which has not undergone a transformation, R^2 value is typically 0.8.

4.4 Statistics-based Methods

After the preprocessing, the statistics-based multi-system outlier detection method is applied. In this approach, most systems are assumed to be operating normally and are fault-free. Devices in the same categories are compared with each other by using features. Common behavior among most systems is considered as normal. Otherwise, it is considered as abnormal behavior. A device with no abnormal behavior is flagged as fault-free. A device with abnormal behavior could be improperly sized, malfunctioning, installed at a non-residential location, or have above

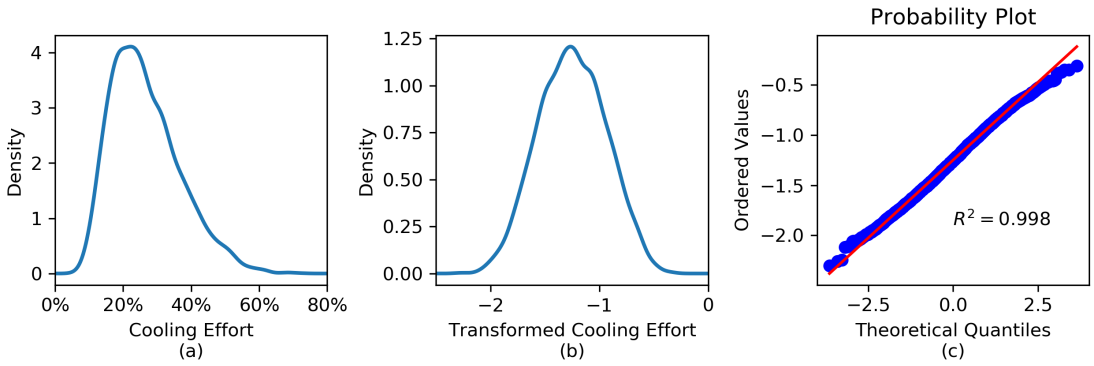


Figure 4.5: Subplot (a) shows the KDE of actual cooling effort in subset #1, subplot (b) shows the KDE of transformed cooling effort in subset #1, and subplot (c) is the quantile-quantile plot of transformed cooling effort in subset #1.

average performance. In other words, malfunctioning systems are a subset of the outliers, and need to be separated from all identified outliers during postprocessing (to be discussed in section 4.5.1) by analyzing a combination of data features. Once a malfunctioning device is identified, expert technicians or artificial intelligence algorithms can evaluate the specific data features and system behavior to provide a preliminary diagnosis of the fault. The purpose of the statistics-based methods is to identify potentially faulty systems and provide quantifiable metrics for determining relative severity, thus allowing prioritization in repair and maintenance.

4.4.1 Univariate Fault Detection

Univariate fault detection examines each feature separately and does not consider possible correlations between features. After separating the data into eight subsets (note that applying a feature transformation is not required), a non-parametric method called kernel density estimation (KDE) is applied to estimate the probability density function of each feature in every subset, respectively. Since outliers behave differently from normal systems, their features are located in the low density regions in the distribution. Moreover, the distribution is generally unimodal, so features with either abnormally high or abnormally low values are classified as outliers. However, as will be discussed in the outlier separation process (Section 4.5.1), only abnormally high values of cooling effort, cycle frequency, and setpoint error will be considered as faulty behavior.

As an example, the KDE of average cooling effort ($\bar{E}_c$) for the operating conditions defined data subset #1 is shown in Figure . One system has $\bar{E}_c$ value of 45%. Because the whole area under the KDE curve is one, the shaded area shows the percentage of systems that has cooling effort higher than 45%. In other words, the area indicates the systems that exhibit worse performance in this operating condition. The process for deciding whether this is sufficient evidence to indicate a fault will be discussed later in the fault prioritization process (Section 4.5.2). Note that the study calculates this percentage through KDE instead of just finding the upper ranks of $\bar{E}_c$ among the sampled several thousand of systems, because the KDE can map the distribution of the selected systems to the population distribution of all similar systems under similar conditions.

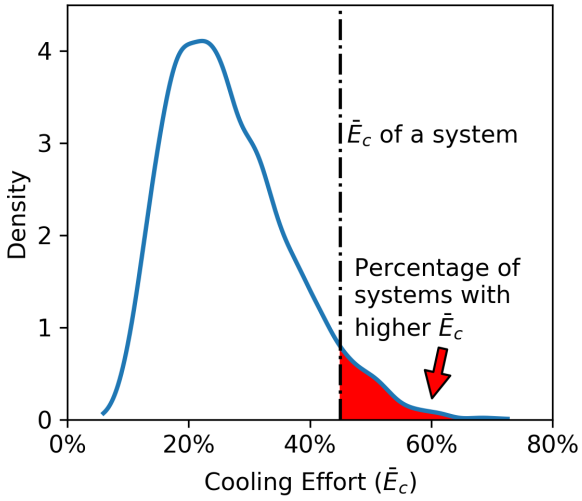


Figure 4.6: The KDE plot illustrates how univariate test of cooling effort is applied. The system has $\bar{E}_c$ value of 45%, and the shaded area shows the percentage of systems that has cooling effort higher than 45%.

4.4.2 Multivariate Fault Detection

As discussed above, univariate fault detection investigates one feature in isolation and does not consider possible correlations between features into account. However, a system with mild faults may not necessarily have extreme values in any one feature, but may have an uncommon

combination of feature values. This section discusses a multivariate fault detection method, in which cooling effort, cycle frequency, and setpoint error are combined into multivariate statistics (i.e. three bivariate statistics and one trivariate statistic).

The same KDE method can be applied for multivariate statistics, but in a high dimensional space with thousands of systems, the computational load is very high. Therefore, *the KDE of the squared Mahalanobis distance* is applied as a metric to find outliers within each subset of operational conditions. The Mahalanobis distance (D_M) is a measure of the distance between a point P and its mean of the multi-dimensional distribution [29]. The distance is zero if P is exactly at the mean of the distribution and grows as P moves away along each principal component axis. Compared to the Euclidean distance, the Mahalanobis distance is scale-invariant and takes the correlations of the data set into consideration as well. It is defined as follows:

$$D_M = \sqrt{(\vec{x} - \vec{\mu})^T \Sigma^{-1} (\vec{x} - \vec{\mu})} \tag{4.3}$$

where $\vec{\mu}$ and Σ are respectively the vector of means and the covariance matrix of the distribution. If Σ is the identity matrix, the Mahalanobis distance will reduce to the Euclidean distance.

Figure 4.7 shows an example of the advantage of Mahalanobis distance. From a Euclidean perspective, point A and point B are equidistant from the center of the distribution. However, since features X and Y are highly correlated, point A is considered to be further away from the center, as measured by Mahalanobis distance. In essence, point A is much more of an outlier. Using a Euclidean metric would fail to capture the pattern of distribution.

But the Mahalanobis distance itself can be very sensitive to outliers ([41]. Several extreme observations located far away from the data center can bias the arithmetic mean and sample covariance matrix, thereby affecting the measurement of Mahalanobis distance. One solution to this issue is that the mean and covariance be estimated in a manner resistant to potential outliers. Currently, a few robust estimators have been proposed, such as the Fast-PCS [42], the Fast-MVE [43],

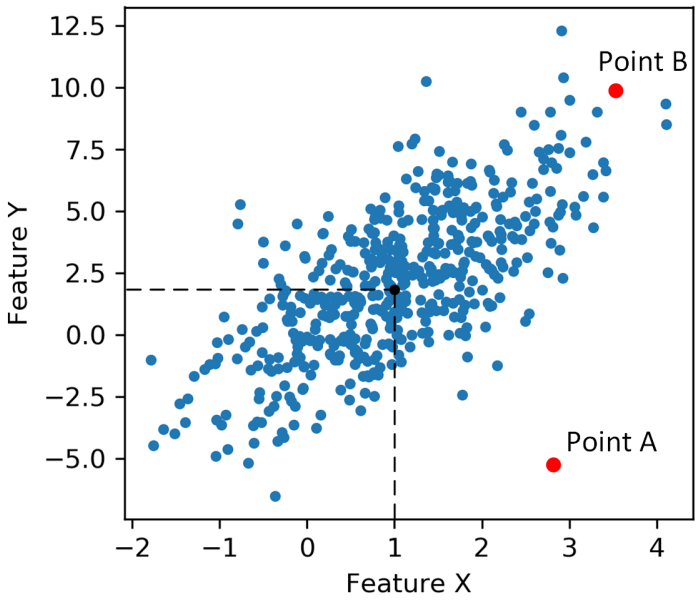


Figure 4.7: Multivariate outlier analysis with Mahalanobis distance.

the Fast-MCD [44], and the SDE [45]. These estimators all try to select a subset of size h among the whole sample by minimizing a single criterion [42].

This paper employed the minimum covariance determinant (MCD) estimator [39]. In this method, the mean and covariance are computed by a subset of observations of size h, which can minimize the determinant of the covariance matrix. The resulting breakdown value is approximately $(n-h)/n$, where n is the sample size. In another word, as long as the fraction of outliers is below the breakdown value, this estimator will offer an unbiased estimation. The subset size h is a tuning hyperparameter chosen based on the distribution of the sample. In this study, since most systems are assumed to be normally behaved, h is set to $0.95n$.

The MCD estimator is also widely used because of its computational efficiency. Rather than using brute force to compute the covariance matrix of each subset of size h, [44] developed the FAST-MCD algorithm. The procedures are summarized as follows [46]:

1. Choose a subset of size h randomly from the sample.

2. Compute the mean vector and the sample covariance matrix.

3. Compute the Mahalanobis distance of all n observations using the mean and covariance obtained in step 2.

4. Choose the h smallest Mahalanobis distances and create a new subset.

5. Repeat steps 2–4 until the difference of the determinants of the covariance matrices between the two newly updated subsets is smaller than a specified threshold. The iteration is convergent.

6. Repeat steps 1–5 and choose the subset with the smallest determinant of covariance matrix.

Other researchers have proved that the MCD estimator is reliable and sound. For instance, [46] showed the superiority of the MCD over the Mahalanobis distance by comparing the detected outliers using various sizes and correlation of simulated multivariate data, and also compared several breakdown points for MCD.

With the help of the MCD estimator, the squared Mahalanobis distance (D_M^2) can be calculated for each system in each operational subset. Note that the transformed features are used in order to better estimate the mean and covariance in Equation 3. The next step is to label outliers using D_M^2. If the combined features follow the multivariate normal distribution, then D_M^2 follows the χ^2 distribution, and the statistics of the χ^2 distribution can be used as detection metrics. But in this case, due to the nonlinear relationship between features, the multivariate is not necessarily jointly normally distributed even if each feature is individually normally distributed. Thus, after calculating the D_M^2 using the MCD estimator for the transformed features, KDE is performed to estimate the distribution of D_M^2 in each subset, and the percentage of systems that has a larger D_M^2 is calculated for each system in each subset as well. Similar to the univariate case, not all outliers are faulty systems, so the outlier separation process is also required for the multivariate.

4.5 Data Postprocessing

This section continues to postprocess the preliminary fault detection results obtained from the univariate and multivariate methods. Postprocessing is necessary to differentiate between faulty

and non-faulty outlier behavior. Systems that are identified as outliers include systems that are: malfunctioning, improperly sized, installed at a non-residential location, performing significantly above average, or simply unrelated to faults. Thus, outlier separation is required to differentiate real malfunctioning systems from the other four categories. Unfortunately, undersized systems are not easily distinguishable without system rated capacity and building thermal characteristics, and unusual system applications cannot be identified without installation records. These are both handled using proprietary data only available from the owner or installer, and are beyond the scope of this paper. However, systems with above average performance (usually oversized) or just unrelated to faults can be separated from malfunctioning systems in postprocessing.

Finally, this section provides a metric to prioritize detected faulty systems after performing the statistics-based FDD method. This prioritization metric reconciles the separate results (i.e. faulty or fault-free) identified in different operating conditions for each system. Additionally, the metric could assist technicians to quantify the severity of faults and determine the priority for system repairs.

4.5.1 Outlier Separation

To separate systems that are oversized or otherwise not faulty, the univariate fault detection methods eliminate outliers with abnormally low values of cooling effort, cycle frequency, or set-point error for the following reasons:

- **High cooling effort is a strong indication of faults in a system, while low cooling effort is not.** High cooling effort is often caused by inadequate capacity, owing to undercharge, condenser/evaporator fouling, airflow restriction, etc. Low cooling effort is particularly the behavior of oversized systems.

- **High cycle frequency is problematic, while low cycle frequency alone is not informative.** The cycle frequency is supposed to be a controllable parameter in operation, normally with two to three start/stop cycles per hour. High cycle frequency itself could indicate malfunctioning control systems, poor thermostat placement (e.g. close to supply air outlets),

or poor home construction (e.g. low home thermal capacity). Low cycle frequency is only conclusive when coupling with other features, which will be discussed next in the outlier separation of multivariate fault detection method.

- **Setpoint error is always expected to be low during a regulation period, if not operating in a dehumidification mode.** Since indoor temperature is controlled by the thermostats to meet the temperature setpoint, high setpoint error is indicative of a control system failure, poor thermostat placement, etc. In unusual circumstances, a system may be operating to meet specific dehumidification requirements, even if the indoor temperature is below the desired setpoint.

The separation of outliers in the multivariate fault detection method is performed by evaluating the combined features. The squared Mahalanobis distance finds all outliers that behave much differently from the average performance, but only the subgroups of outliers shown in Table 4.3 are considered as faulty or oversized. In general, abnormally high cooling effort is associated with faults, abnormally low cooling effort is associated with the oversizing issue, but low setpoint error is never related to faults.

4.5.2 Fault Prioritization

Finally, to prioritize faulty systems, this method proposes a metric which quantifies the performance of a system in terms of its position in the feature spaces of population distributions. In this way, occupants and technicians know how a system performs compared to similar systems under similar conditions. For each system, the metric first calculates the fraction of systems in the population that exhibit worse performance, denoted by α, within each subset of the univariate and multivariate tests. This calculation uses the fault detection result obtained from kernel density estimations (see Section 4.4). Then, the weighted average of α, denoted by $\bar{\alpha}$, of a system in one test is calculated by taking the average of α using the number of data in each subset as weight. Finally, each test labels faulty systems separately by comparing the weighted fraction $\bar{\alpha}$ with pre-determined thresholds. In this research, the threshold for the univariate tests is 1%, and

Table 4.3: Outlier separation: subgroups of outlier systems that are considered as faulty or oversized.

Type of test	Typical behaviors	E_c	F	T_{spe}
Univariate	Near-continuous operation	↑		
	High cycling		↑	
	Poor control			↑
Bivariate	Near-continuous operation	↑	↓	
	High cycling	↑	↑	
	Poor control	↑		↑
	Poor control		↑	↑
	Oversized	↓	↑	
Trivariate	Mixed faulty behaviors	↑	↑	↑
	Mixed faulty behaviors	↑	↓	↑

↑—abnormally high, ↓—abnormally low

for the multivariate tests is 5%. The larger threshold for multivariate tests is intended to identify mild faults with uncommon combination of feature values. In practice, the thresholds should be selected by the user based on the number of systems in the population, and the number of potentially faulty systems to be identified.

As an example, the statistics of System #00C0E364 are shown in Table 4.4, which was labeled faulty by the proposed procedure. The $\bar{\alpha}$ for univariate E_c is 0.5%, which indicates that on average only 0.5% of other systems have higher cooling efforts compared to this system (see Figure 4.6). Also, the $\bar{\alpha}$ for multivariate (E_c, F) is 0.3%, which indicates that 0.3% of other systems locate further away from the center in the joint distribution of cooling effort and cycle frequency (see Figure 4.7). In general, lower values of $\bar{\alpha}$ imply worse performance and are stronger signs of fault, and thus the corresponding systems will have higher priority. In this example, the $\bar{\alpha}$ for univariate E_c is lower than the univariate threshold (1%), and the $\bar{\alpha}$ for multivariate (E_c, F) is lower than the multivariate threshold (5%), and therefore both indicate a faulty system. A similar criterion is used for the other five tests. Note that due to outlier separation (see Table 4.3), not all $\bar{\alpha}$'s in

the multivariate tests lower than the threshold are classify the system as faulty. For example, even the $\bar{\alpha}$ for multivariate (F, T_{spe}) is lower than 5%, a fault is not presumed because this system has abnormally low cycle frequency and abnormally high setpoint error, which is not a subgroup of outliers considered as faulty in Table 4.3.

Table 4.4: Statistics of a faulty system after fault prioritization.

System #00C0E364		$\bar{\alpha}$	Threshold	Faulty	Description
Univariate tests	E_c	0.5%	1%	True	Abnormally high E_c
	F	94%	1%	False	Low F
	T_{spe}	0.1%	1%	True	Abnormally high T_{spe}
Multivariate tests	(E_c,F)	0.3%	5%	True	Near-continuous operation
	(E_c,T_{spe})	0.7%	5%	True	Poor control
	(F,T_{spe})	0.08%	5%	False*	-
	(E_c,F,T_{spe})	0.2%	5%	True	Mixed faulty behaviors

E_c – cooling effort, F – cycle frequency, T_{spe} – setpoint error

$\bar{\alpha}$ – percentage of other systems that exhibits worse performance than the presented one

* Due to outlier separation, not all outliers are considered as faults. In this case, the system has abnormally low cycle frequency but abnormally high setpoint error, which is not a typical case of fault.

4.6 Conclusions

The statistics-based fault detection method presented in this paper analyzes cloud-based thermostat data for thousands of residential HVAC systems based on knowledge of the underlying physics. The principal benefit of this method is detecting faults by comparing features *between* systems rather than *within* systems. This leverages the advantages of large data sets to overcome the limited sensor information available from a single thermostat. Both preprocessing and post-processing methods are non-trivial. For example, mode labeling transforms system operational modes into behavioral modes, such that more useful features can be extracted; data separation process enables the comparison of features free from the effects of temperature and humidity ratio;

in postprocessing, outlier separation helps to find true faulty systems among all outliers. After combining the results from both univariate and multivariate methods, experts can easily prioritize systems for further evaluation and repair, significantly simplifying a large-scale diagnostic problem of monitoring hundreds of thousands of systems.

Although the proposed statistics-based method requires relatively little computational power even for large data sets, there are still several opportunities for improvement. First, additional meta-data about the individual systems and homes (e.g. cooling/heating capacity, home size, etc.) would enable more specific fault diagnosis. Second, a compiled data set of verified faulty *in situ* systems would facilitate the selection of statistical thresholds and fault prioritization metrics, to further automate the process.

The second part of this paper will demonstrate the effectiveness of the proposed methods with several case studies drawn from a database of residential thermostat data for more than 9,000 systems. The case studies will illustrate the detected faulty systems with inadequate capacity and control problems, along with oversizing issues in the practical application.

5. PAPER C: MULTIVARIATE FAULT DETECTION FOR RESIDENTIAL HVAC SYSTEMS USING CLOUD-BASED THERMOSTAT DATA, PART II: CASE STUDIES[1]

5.1 Synopsis

Fault detection and diagnosis (FDD) using smart thermostat data is a relatively new field of research, but one with immediate practical application to residential indoor climate control. In this two-part paper, a statistics-based FDD method is proposed to identify anomalous residential air conditioning systems using aggregated thermostat data. The proposed method extracts features that characterize system operational behaviors, and determines outliers by comparing the features between thousands of systems. Those outliers are subsequently classified into several categories of anomalous operational behaviors such as near-continuous operation, poor control, etc. which are closely related to soft mechanical faults. While the first part of the paper details the methodology, this second part provides five case studies to demonstrate the effectiveness of the proposed method. Although these cases should not be considered a complete representation of all kinds of anomalous behaviors, the examples illustrate typical problems including: performance degradation faults with or without affecting occupant comfort, control problems with slight degradation, and oversized systems. Each case study exemplifies one type of behavior that is common among a group of systems, rather than a single outlier that behaves much differently from all others. The case studies illustrate how the proposed fault detection and diagnosis method is capable of identifying malfunctioning systems from a population of residential air conditioning systems.

5.2 Introduction

As discussed in part one of this paper, fault detection and diagnosis (FDD) plays an important role in the predictive maintenance for residential heating, ventilating, and air conditioning (HVAC) systems [3, 14, 16]. Increasingly, smart thermostat data is becoming an effective resource for developing advanced FDD algorithms because operational data from hundreds of thousands of

[1]F. Guo, A.P. Rogers, and B.P. Rasmussen, "Multivariate Fault Detection for Residential HVAC Systems Using Cloud-based Thermostat Data, Part II: Case Studies," *Science and Technology for the Built Environment*, **In Review**.

residential systems can be easily collected and tracked with relatively low cost [22, 27, 21]. Figure 5.1 shows how the FDD method using cloud-based smart thermostat data proposed in part one of this paper is used in the process of predictive maintenance. The FDD method first queries raw data from a cloud server, which receives and stores the data uploaded from smart thermostats in residential homes. Then the main algorithm extracts features from the behavioral modes, and applies statistical methods and artificial intelligence to label systems with anomalous feature values. This process takes advantage of the large scale of data by estimating the population operational behavior from thousands of systems and comparing the operating performance between them. Finally, the FDD method classifies outliers into groups of faults, and assigns severity metrics to order the labeled systems. With the list of faulty systems, maintenance companies can notify the relevant customers and dispatch service technicians to repair the systems based on priority and the initial diagnosis. In this way, predictive maintenance helps residential systems ensure occupant comfort and maintain a high level of energy efficiency.

The proposed algorithm in part one of this paper relies on a preprocessing method known as mode labeling, which sections and classifies raw thermostat data into behavioral modes so that valuable features can be extracted [6]. The mode labeling algorithm defines three behavioral modes: cooling/heating regulating mode, cooling/heating tracking mode, and free response mode. In the regulating mode, a system cycles on and off and maintains a relatively constant indoor temperature; in the tracking mode, a system operates continuously to reach the target setpoint; in the free response mode, a system is off for a period of time and the indoor environment freely responses to the ambient conditions. Illustrations of these modes can be found in the operation plots of the five case studies to be shown in Sections 5.3.1 to 5.3.5. Table 5.1 shows the extracted features from the cooling regulating and tracking modes. Only features in the regulating mode are used for fault detection, but features from both regulating and tracking modes are used for diagnosis and classification.

The FDD method consists of univariate tests and multivariate tests. In the univariate tests, each feature in the cooling regulating mode (i.e., cooling effort, cycle frequency, and setpoint

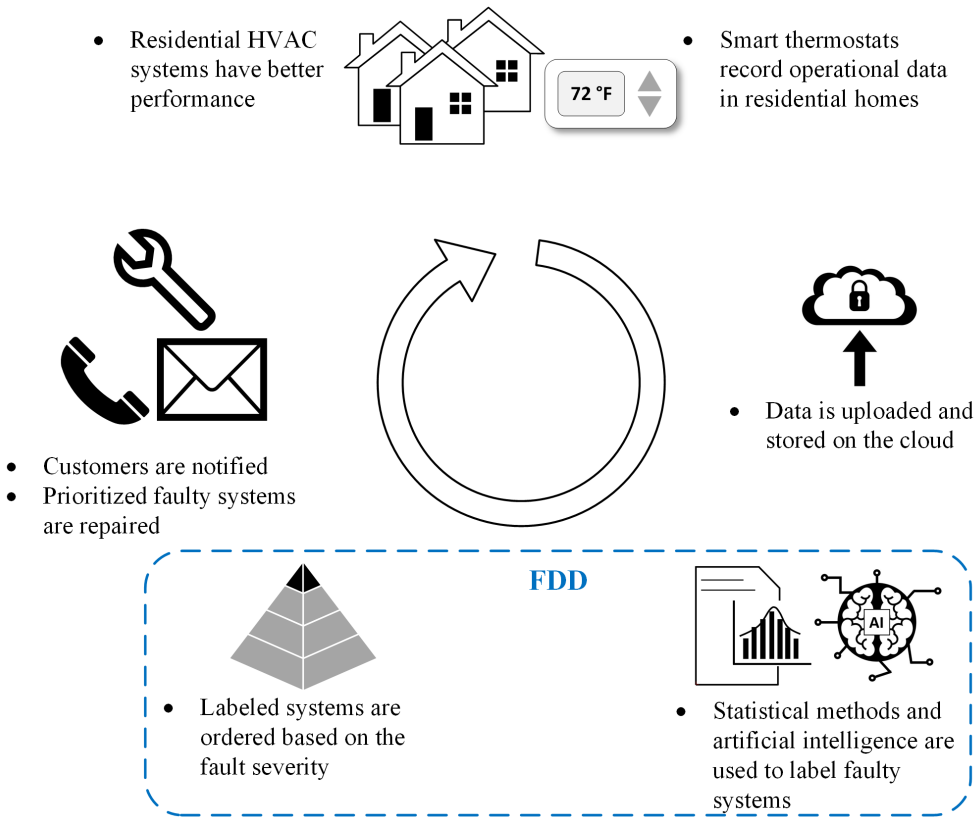


Figure 5.1: The process of predictive maintenance and the role of FDD methods.

error) is compared separately between systems. The tests estimate the percentage of systems that exhibits worse performance than each individual system under similar operating conditions. These indicators are aggregated over the full range of operational conditions using a weighted average, denoted by $\bar{\alpha}$. Thus, the method identifies systems with consistently anomalous behavior, rather than systems exhibiting a single abnormal event. In the multivariate tests, those three features are combined (bivariate and trivariate) and analyzed simultaneously between systems. For each group of features, the test measures the Mahalanobis distance [29] in the joint distribution, and estimates the percentage of systems that locates further away from the distribution center for each system under similar operating conditions, also denoted by $\bar{\alpha}$. Based on the detection results of the FDD method, the authors selected normal and faulty systems for comprehensive case studies. Each example system represents a common faulty or fault-free behavior, not simply a single anomalous event. Because this research focus on detecting soft faults that are more challenging to detect, the

displayed faulty systems are primarily related to inadequate capacity and malfunctioning control strategies.

Table 5.1: Supporting features for cooling modes.

Mode	Symbol	Feature name	Definition
Cooling regulating mode	ΔT_{oi}	Temperature difference [°F]	Temperature difference between home (indoor) and ambient (outdoor) $\Delta T_{oi} = T_o - T_i$
	$\Delta \omega_{oi}$	Humidity ratio difference [kg H_2O/kg dry air]	Humidity ratio difference between home (indoor) and ambient (outdoor) $\omega_{oi} = \omega_o - \omega_i$
	E_c	Cooling effort [%]	The product of cooling duty factor (D) and cooling load factor (L)
	F	Cycle frequency [cycle/hr]	The number of start/stop cycles per hour
	T_{spe}	Cooling setpoint error [°F]	The error between indoor temperature and cooling setpoint $T_{spe} = T_i - T_{spc}$
Cooling tracking mode	ΔT_{oi}	Temperature difference [°F]	Temperature difference between home (indoor) and ambient (outdoor) $\Delta T_{oi} = T_o - T_i$
	$\Delta \dot{T}_{ii}$	Indoor temperature drop rate [°F/hr]	Decrease of indoor temperature per hour (indoor) and ambient (outdoor) $\Delta \dot{T}_{ii} = (T_i^{initial} - T_i^{final})/t$ where t is the total tracking hours

5.3 Case Studies

For the selected case studies, smart thermostat data was collected from June to August 2018 for approximately 9,000 residential single-stage cooling systems located in Texas, Georgia, North Carolina, and Florida, representing systems from climate zones 2A and 3A [5]. After preprocessing, 8,309 systems are selected for the univariate and multivariate tests. As shown in Table 5.2, in total 257 systems (3.1%) are flagged as faulty, with 164 flagged by at least one of the univariate tests and 172 flagged by at least one of the multivariate tests. The faults are related to inadequate capacity and control problems, as well as oversizing. Note that the algorithm is able to find faulty

systems exhibiting anomalous behavior, but cannot guarantee to find all existing faulty systems. For instance, an oversized system with degradation may perform identical to a normal system and not be detected based on its operational behavior.

Table 5.2: Typical behaviors that is considered faulty in all univariate and multivariate tests.

Type of test	Typical behaviors	E_c	F	T_{spe}	Number of systems
Univariate	Near-continuous operation	↑			56
	High cycling		↑		59
	Poor control			↑	58
Bivariate	Near-continuous operation	↑	↓		63
	High cycling	↑	↑		7
	Poor control	↑		↑	104
	Poor control		↑	↑	47
	Oversized	↓	↑		7
Trivariate	Mixed faulty behaviors	↑	↑	↑	35
	Mixed faulty behaviors	↑	↓	↑	36

E_c – cooling effort, F – cycle frequency, T_{spe} – setpoint error
↑—abnormally high, ↓—abnormally low

In order to illustrate the effectiveness of the proposed fault detection algorithm, five systems are analyzed in detail, with each of them representing a particular category of common behavior. First, a normal device is presented to serve as a basis for comparison with subsequent faulty systems. Second, a system with severe setpoint tracking failure is discussed, showing that the proposed fault detection method is able to detect systems with inadequate capacity. The third example system is also identified as having inadequate capacity, but in contrast to the previous example, this fault is imperceptible to the occupants as comfort levels are maintained despite the fault. Fourth, one system exhibiting control problems is discussed, illustrating that the fault detection method can

find faults for a broad range of behavioral faults. Lastly, an oversized system is shown. Although this system is not considered as faulty, it demonstrates the versatility of the proposed method.

5.3.1 System I: Normal Operation

The case studies begin with an example of normal system, that is not identified as an outlier using the statistical methods proposed in part one of this two-part paper. As an example of the typical behavior of this system, Figure 5.2 displays the operation of the compressor, indoor and outdoor temperature variations, and cooling temperature setpoint changes, recorded by the smart thermostat from September 10^{th} 8:00 PM to September 12^{th} 4:00 PM in 2018. The behavioral modes are labeled on the variation of the indoor temperature, where the system exhibits periods in both the cooling regulating mode and the free response mode. For this particular subset of data, the cooling tracking mode is not observed. The plot does not show any abnormal behavior. The indoor temperature is maintained around the desired temperature setpoint, with fluctuation amplitude less than 1°F, as occupants adjust the cooling setpoint, even as the outdoor temperature exceeds 90°F every day. The cooling effort is neither excessively high or low, but appropriate for the cooling load. The cycle frequency is also well-controlled with two or three cycles per hour.

The first step in applying the proposed statistical methods is to estimate the probability density function of each data feature. Figure 5.3 shows the kernel density estimation (KDE) contours for different pairs of data features for all 8,309 systems employed in this study. Then in Figure the same contours are overlaid with the data from system I shown as a scatterplot. Note that each data point for system I in blue is a two-hour average, and the quadratic curve in Figure b is the least squares fitting considering that the cycle frequency is necessarily zero when cooling effort is 0% and 100%. In all subplots of Figure , the data points for system I are located close to the center of the red contours, indicating normal behavior under most operational conditions.

Next, the data features associated with the cooling tracking periods are presented. Similar to the analysis above, Figure 5.5a displays the KDE between temperature difference and indoor temperature drop rate for all systems. In Figure 5.5b, the corresponding data for system I is shown as a scatterplot, with each data point representing a single cooling down period. The data demon-

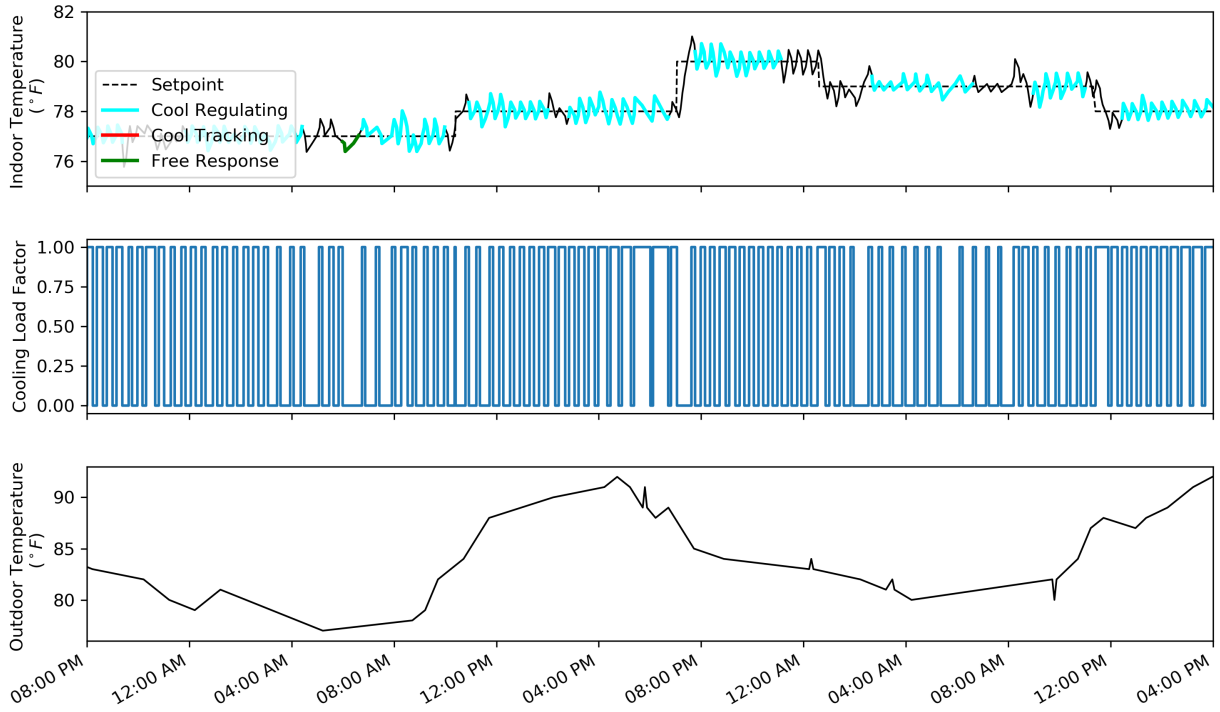


Figure 5.2: Daily operating condition of system I from September 10th 8:00 PM to September 12th 4:00 PM with behavioral modes labeled. Cyan: cooling regulating mode; red: cooling tracking mode (not captured in this plot); green: free response mode; black: unlabeled mode of operation; dotted line: cooling temperature setpoint.

strates that the behavior of system I is consistent with other similar systems. This system is able to cool the house and track the setpoint with proper indoor temperature drop rate, so it is considered neither oversized nor undersized.

The figures above provide visual confirmation that the system behavior is normal. However, when applying the proposed methods to large populations of systems, the fault detection method relies on quantifiable metrics for identifying and labeling faulty systems. Table 5.3 gives the detection statistics that identify system I as a normal system. In the table, the symbol $\bar{\alpha}$ denotes the fraction of systems that exhibits worse performance within each test, and is compared with a determined threshold. A fault is labeled when $\bar{\alpha}$ is lower than the threshold, which means the system performs poorly compared to the population. In this research, the threshold for all univariate tests is 1%, and for all multivariate tests is 5%. The multivariate tests have a larger threshold to identify

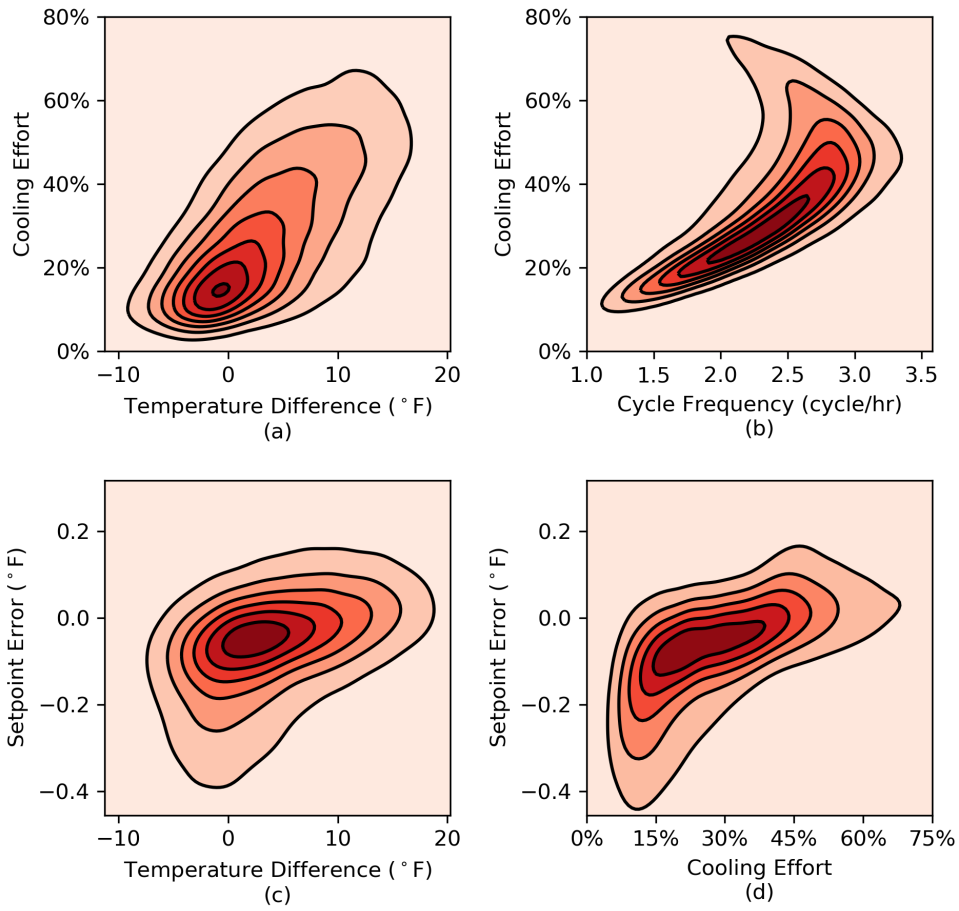


Figure 5.3: Kernel density estimation (KDE) of the population distribution of similar systems. The contours show the relationships between cooling effort, cycle frequency, temperature difference and setpoint error for all systems in cooling regulating periods in the queried sample (8,309 systems).

mild faults with uncommon combination of feature values. In practice, the thresholds should be selected by the user based on the number of systems in the population and the number of potentially faulty systems to be identified. For system I, no fault is detected and all $\bar{\alpha}$ are much greater than the threshold. Therefore, the system is labeled as normal.

5.3.2 System II: Degradation Fault Affecting Occupant Comfort

In contrast to the normal system discussed above, system II is a typical example of a faulty system. From the representative period of operation shown in Figure 5.6, the indoor environment

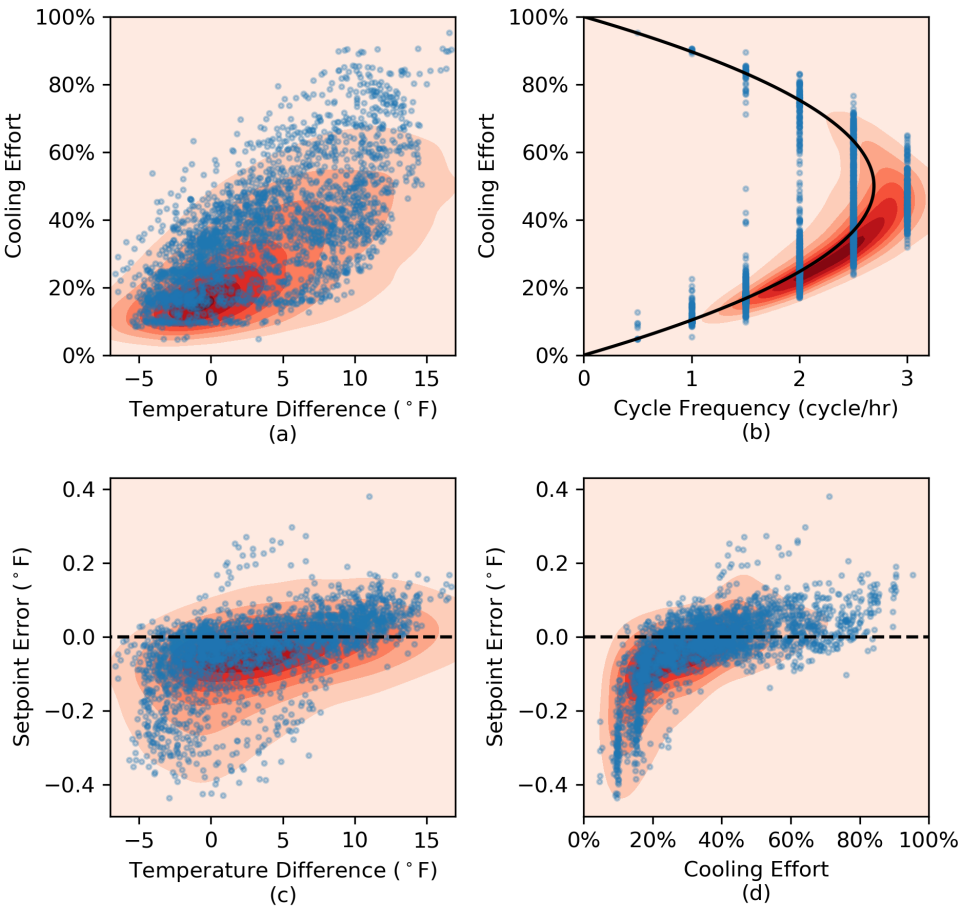


Figure 5.4: The scatterplots show the relationships between four pairs of data features for system I in cooling regulating periods. The contours show the population distributions, same as Figure 5.3. The curve in subplot (b) is the quadratic fitting of the individual system data points.

is not properly controlled. Both cooling regulating and cooling tracking modes are observed. During hot summer days when the outdoor temperature approaches 90°F, the system operates nearly continuously. But even at full capacity, the indoor temperature reaches almost 80°F in the afternoon, 9°F above the cooling setpoint. Additionally, because the night of August 17th is relatively warm, the system operates at full capacity through the night. As a result, the indoor environment is primarily determined by the outdoor environment and the air conditioner has little ability to maintain a comfortable environment.

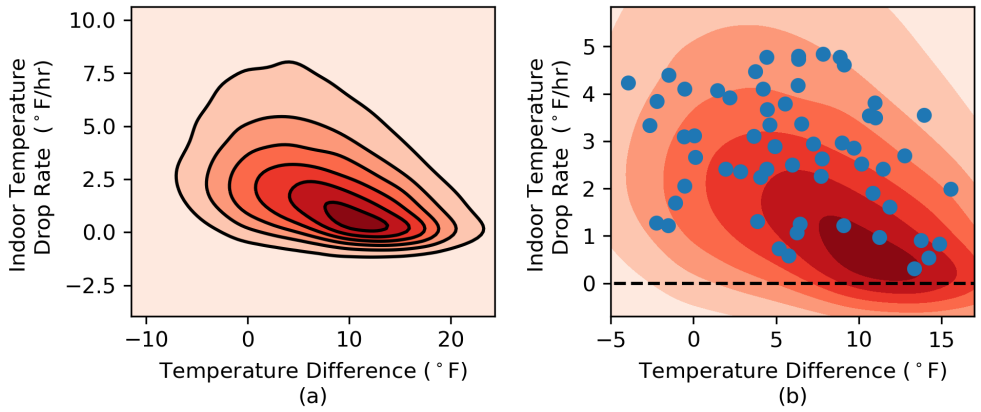


Figure 5.5: Subplot (a) is the KDE of the population distribution. The contour shows the relationship between temperature difference and indoor temperature drop rate in cooling tracking periods for all systems. Subplot (b) shows the relationship between the same pair of features for system I in cooling tracking periods.

Although the inadequacies of the cooling system are visually evident from the data shown in Figure 5.6, this faulty behavior can be detected automatically using the proposed fault detection algorithm. Figure 5.7 shows the distribution of data features in the cooling regulating mode. The following should be noted:

- **In Figure 5.7a, cooling effort is very high during the regulating period.** When the temperature difference is positive (outdoor conditions are warmer than indoor conditions), the system operates more than 50% of the time. Almost all data points fall in the lightest region of the density contour, which indicates the system is an obvious statistical outlier, when compared to similar systems under similar conditions.

- **In Figure 5.7b, the cycling frequency is normal.** Usually the system starts two or three times per hour, which is considered appropriate.

- **In Figure 5.7c and 5.7d, the setpoint error is very high.** If the indoor environment is well controlled, the setpoint error should be approximately zero and insensitive to the change of outdoor temperature and cooling effort. But in this case, the majority of data points locate far

Table 5.3: Fault detection output of system I.

System I		$\bar{\alpha}$	Threshold	Faulty	Description
Univariate tests	E_c	35.7%	1%	False	Average E_c
	F	60.0%	1%	False	Average F
	T_{spe}	40.7%	1%	False	Average T_{spe}
Multivariate tests	(E_c,F)	80.4%	5%	False	-
	(E_c,T_{spe})	93.6%	5%	False	-
	(F,T_{spe})	94.6%	5%	False	-
	(E_c,F,T_{spe})	88.4%	5%	False	-

E_c – cooling effort, F – cycle frequency, T_{spe} – setpoint error

$\bar{\alpha}$ – percentage of other systems that exhibits worse performance than the presented one

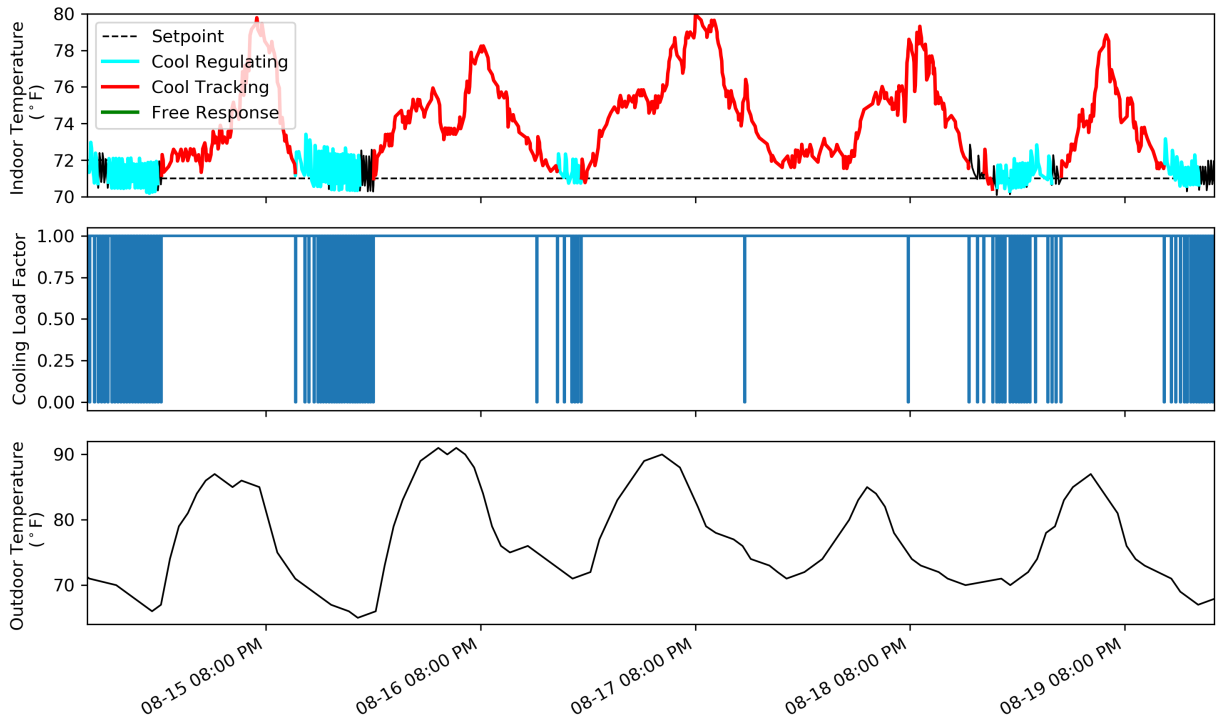


Figure 5.6: Daily operating condition of system II from August 15th to August 29th with behavioral modes labeled. Cyan: cooling regulating mode; red: cooling tracking mode; black: unlabeled mode of operation; dotted line: cooling temperature setpoint.

above from the center of the contour, and the average setpoint error over a two-hour period can even approach 1.5°F. The high setpoint error means the system struggles to maintain the indoor conditions at the desired temperature.

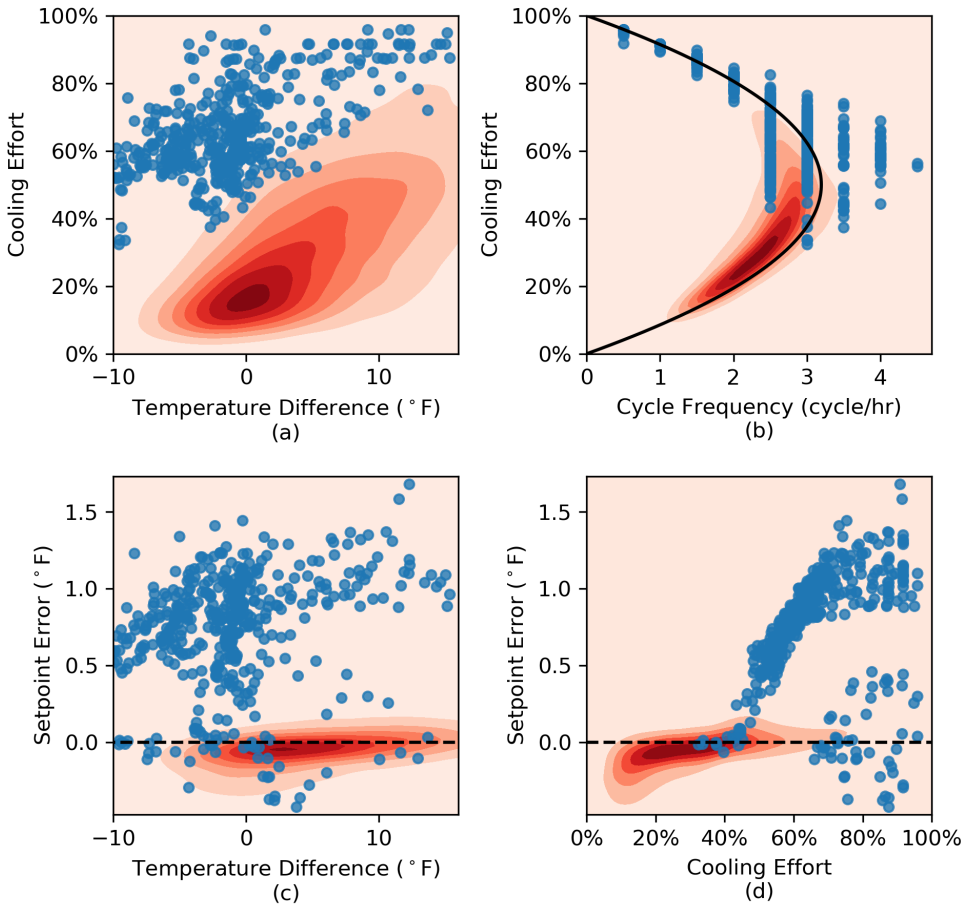


Figure 5.7: The scatterplots show the relationships between four pairs of data features for system II in cooling regulating periods. The contours show the population distributions. The curve in subplot (b) is the quadratic fitting of the individual system data points.

Figure 5.8 shows data features for all cooling tracking periods of this system. In total there are 59 data points, and each of them is an on-cycle. The majority of the data indicates an indoor temperature drop rate of zero, which means the system runs for a long time, but is unable to lower the

indoor temperature. The cooling tracking periods shown in Figure 5.6 are good examples. Those periods occur when the system exceeds its capacity and fails to maintain the indoor temperature around setpoint.

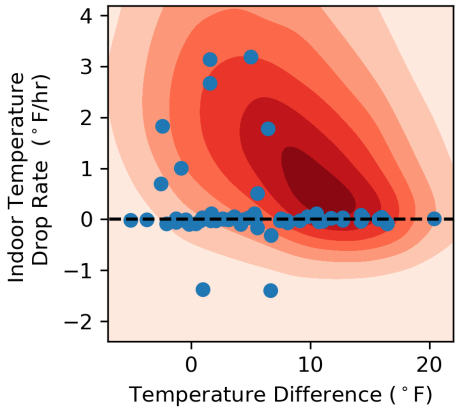


Figure 5.8: The scatterplot shows the relationship between temperature difference and indoor temperature drop rate for system II in cooling tracking periods, and the contour is the population distribution.

Finally, Table 5.4 shows how system II is identified as a faulty system by the detection algorithm. Faults are detected in cooling effort and setpoint error using univariate tests, as well as a bivariate test using these two features. Based on these data, the following conclusions are drawn:

- **Cooling effort is higher than most of the systems.** The reader should note that the $\bar{\alpha}$ value of the cooling effort is extremely low: approximately only 0.08% of systems operate in a larger proportion of time on average when regulating around the cooling setpoint. The cooling load is much larger than the actual cooling capacity.

- **Compressor cycling frequency is just above average.** The $\bar{\alpha}$ value is close to 33%, which is not considered an outlier when compared to other systems.

- **Temperature setpoint error is very high.** The $\bar{\alpha}$ value is 0.29%, which means the indoor environment is not well-regulated.

Table 5.4: Fault detection output of system II.

System II		$\bar{\alpha}$	Threshold	Faulty	Description
Univariate tests	E_c	**0.08%**	1%	**True**	Abnormally high E_c
	F	60.0%	33.3%	False	Average F
	T_{spe}	**0.16%**	1%	**True**	Abnormally high T_{spe}
Multivariate tests	(E_c,F)	-	5%	False	-
	(E_c,T_{spe})	**0.29%**	5%	**True**	Poor control
	(F,T_{spe})	-	5%	False	-
	(E_c,F,T_{spe})	-	5%	False	-

E_c – cooling effort, F – cycle frequency, T_{spe} – setpoint error

$\bar{\alpha}$ – percentage of other systems that exhibits worse performance than the presented one

- **The squared Mahalanobis distance of bivariate (E_c,T_{spe}) is much higher than most of the systems.** Note that a smaller $\bar{\alpha}$ value means higher squared Mahalanobis distance and greater possibility of an outlier. For this system, the $\bar{\alpha}$ value of (E_c,T_{spe}) is 0.29%, indicating this bivariate is significantly far from the center of the distribution.

Based on the high cooling effort and high setpoint error, the statistics show strong evidence of capacity degradation that significantly affects occupant comfort.

5.3.3 System III: Degradation Fault without Affecting Occupant Comfort

The former example is a system exhibiting difficulty controlling the indoor environment. This abnormal behavior is very likely to be noticed by the occupants. The following example, however, shows that the proposed method can also detect abnormal behaviors not easily to be noticed by the occupants.

An example of the typical operational behavior of system III is displayed in Figure 5.9. The plots show clear evidence that this system operates abnormally. Every morning as the outdoor temperature rises, the cooling unit struggles to maintain the setpoint. The compressor operates almost all the time in the regulating mode until evening comes and the outdoor temperature drops below the setpoint. Occasionally the system changes from the regulating mode to the tracking mode in

operation because the indoor temperature nearly could not be controlled. Compared with system II, however, the indoor temperature does not have significant increase above cooling setpoint, which would cause occupant discomfort.

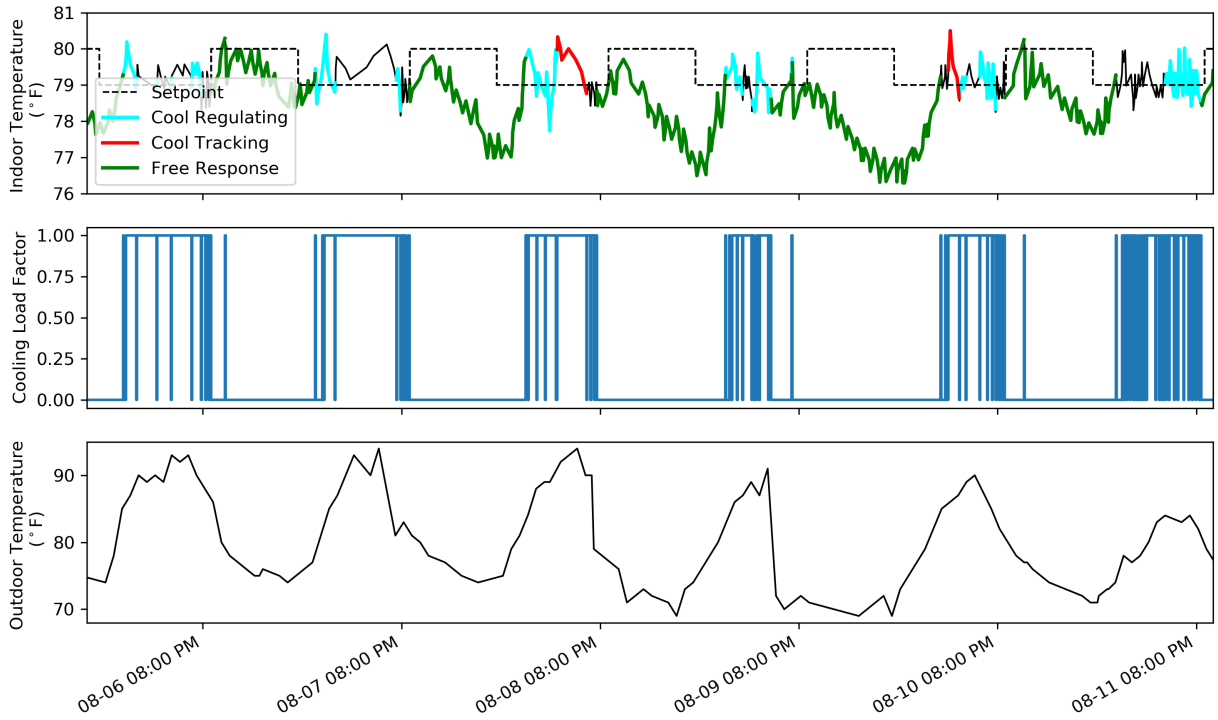


Figure 5.9: Daily operating condition of system III from August 6th to August 12th with behavioral modes labeled. Cyan: cooling regulating mode; red: cooling tracking mode; green: free response mode; black: unlabeled mode of operation; dotted line: cooling temperature setpoint.

Figure 5.10 presents four pairs of features in the cooling regulating mode for system III. As is clear from the plots, the distribution of operating data from system III is distinctly different from the population distribution of similar systems, indicating that this system operates quite differently from most others. In particular, please note:

- **In Figure 5.10a, cooling effort surges as the temperature difference increases during the regulating period.** Cooling effort nearly approaches 100% even though the outdoor temperature is only about 5°F above indoor temperature. Compared with the background

contour, cooling effort greater than 70% is uncommon at a temperature difference between 5°F and 10°F.

- **In Figure 5.10b, the cycle frequency is lower than normal.** The quadratic curve shifts to the left of the background contour, and a few periods have relatively low cycle frequency at moderate cooling effort.

- **In Figure 5.10c and 5.10d, the setpoint error is high when temperature difference and cooling effort are high.** From both plots, as the temperature difference (or cooling effort) increases, the setpoint error gradually increases. However, the deviation is always within ± 0.5°F, a difference hard to be noticed by occupants.

Figure 5.11 shows the operational behavior during the cooling tracking periods. The system is able to lower the indoor temperature and track the setpoint at relatively low temperature difference, but the indoor temperature drop rate is lower than average. As the temperature difference increases, the capability for the system to decrease indoor temperature is further weakened and nearly exceeds cooling capacity.

Finally, Table 5.5 shows how system III is identified as a faulty system by the detection algorithm.

- **Cooling effort is significantly higher than most systems.** The $\bar{\alpha}$ value of the cooling effort means approximately only 0.34% of other systems operating in a larger proportion of time on average when regulates around the cooling setpoint.

- **Compressor cycling frequency is below the average.** The $\bar{\alpha}$ value is 92.6%, which means that most systems have greater cycling frequency values. This is due to the high cooling effort that leads to nearly continuous operation.

- **Temperature setpoint error is higher than the average.** The $\bar{\alpha}$ value is 9.2%, a little bit high among all systems. But because setpoint tracking failures do not happen, the occupants may still feel comfortable and consider the air conditioner operates just well.

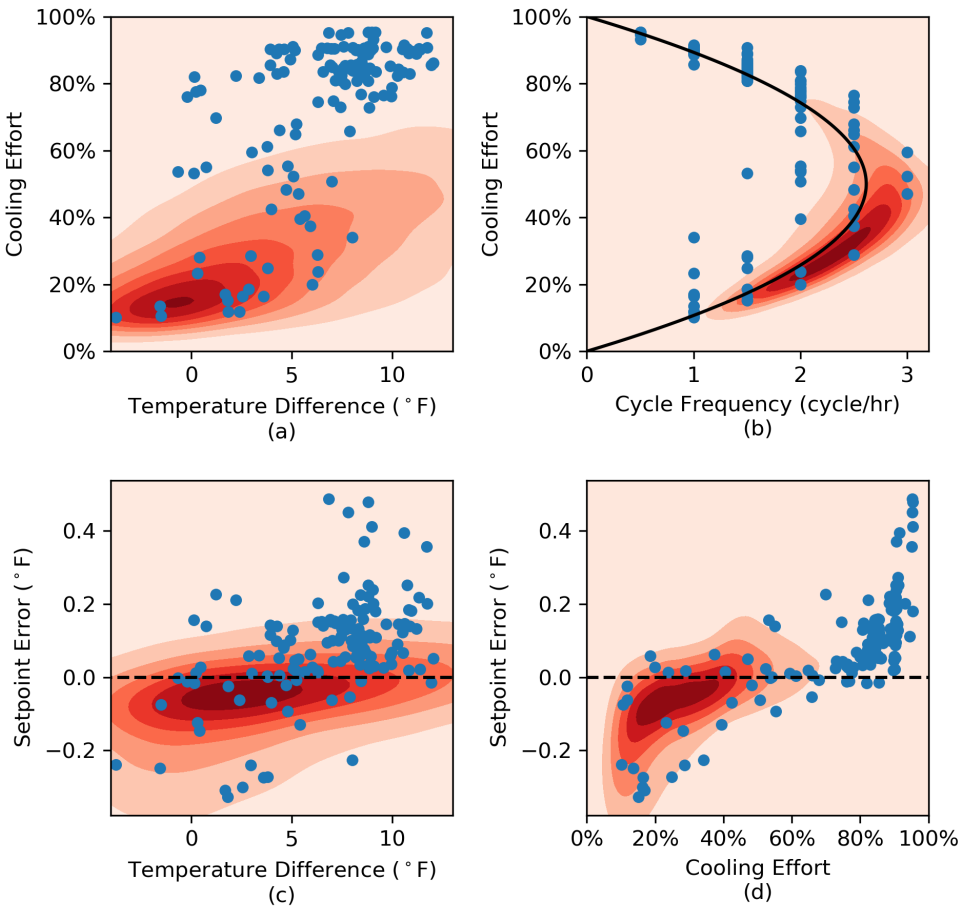


Figure 5.10: The scatterplots show the relationships between four pairs of data features for system III in cooling regulating periods. The contours show the population distributions. The curve in subplot (b) is the quadratic fitting of the individual system data points.

- **The squared Mahalanobis distances of (E_c,F) and (E_c,F,T_{spe}) are much higher than most systems**, which provides further evidence of the abnormality.

In conclusion, this example shows this algorithm can identify capacity degradation problems that may not be noticeable by occupants. Compared to hard faults such as equipment failures, these types of soft faults are generally much more difficult to be detected.

5.3.4 System IV: Control Problem

System II and III are two case studies identified with degradation faults. This section shows the proposed method can detect other types of system faults, such as control problems. Figure 5.12

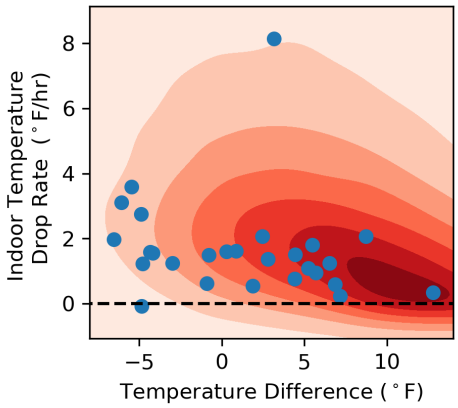


Figure 5.11: The scatterplot shows the relationship between temperature difference and indoor temperature drop rate for system III in cooling tracking periods, and the contour is the population distribution.

displays a period of typical operational behavior. During a period of fourteen hours, the system continuously regulates around the setpoint, but the behavior seems abnormal. The compressor switches on and off quickly, and the cycle frequency is very high. Also, when the temperature difference is high, the indoor temperature always stays above the temperature setpoint up to 4°F, not well-controlled by the thermostat. Moreover, the indoor temperature rises and drops extremely fast as the system cycles on and off.

In Figure 5.13, four pairs of data features for system IV in the cooling regulating mode are shown. Please note:

- **In Figure 5.13a, cooling effort is somewhat higher than the population distribution.** But because the pattern is still within a normal range, there is no significant evidence for severe system degradation.

- **In Figure 5.13b, the cycle frequency is extremely high.** The quadratic curve representing the individual system data does not match the population distribution contour. When cooling effort is around 50%, cycle frequency reaches an average of 5 cycles per hour, and sometimes can be up to 7.5 cycles per hour (i.e. 8 minutes for an on and off cycle). This problem is known as "short cycling". Not only does this increase energy consumption due

Table 5.5: Fault detection output of system III.

System III		$\bar{\alpha}$	Threshold	Faulty	Description
Univariate tests	E_c	**0.34%**	1%	**True**	Abnormally high E_c
	F	92.6%	1%	False	Low F
	T_{spe}	9.2%	1%	False	High T_{spe}
Multivariate tests	(E_c,F)	**0.39%**	5%	**True**	Near-continuous operation
	(E_c,T_{spe})	-	5%	False	-
	(F,T_{spe})	-	5%	False	-
	(E_c,F,T_{spe})	**0.96%**	5%	**True**	Mixed faulty behaviors

E_c – cooling effort, F – cycle frequency, T_{spe} – setpoint error

$\bar{\alpha}$ – percentage of other systems that exhibits worse performance than the presented one

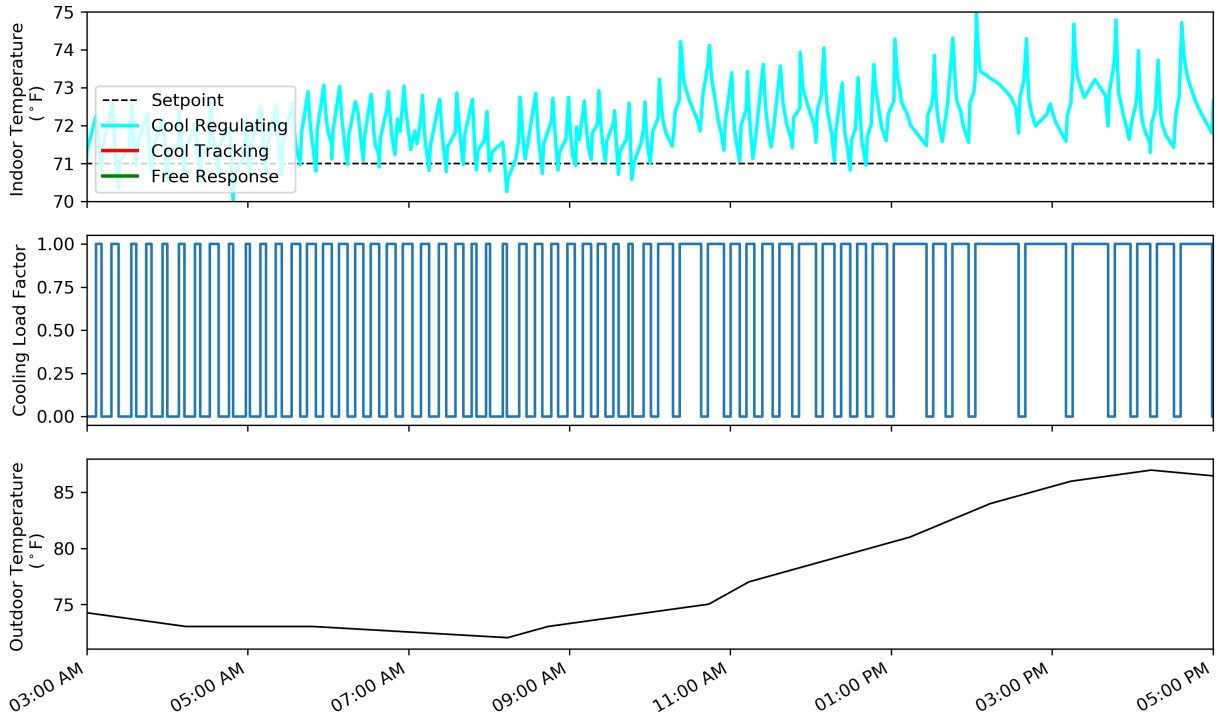


Figure 5.12: Daily operating condition of system IV on September 28th with behavioral modes labeled. Cyan: cooling regulating mode (the only mode shown in this figure); dotted line: cooling temperature setpoint.

to transient efficiency losses during start-up, but also negatively impacts the system in terms of component wear.

- **In Figure 5.13c and 5.13d, the setpoint error is also very high, and has a strong linear relationship with cooling effort.** For a normal system, setpoint error should fluctuate around zero and have a weak correlation with either temperature difference or cooling effort, as the trend shown by the population distribution. Here, the abnormally high setpoint error and strong correlation could be caused by the control problems.

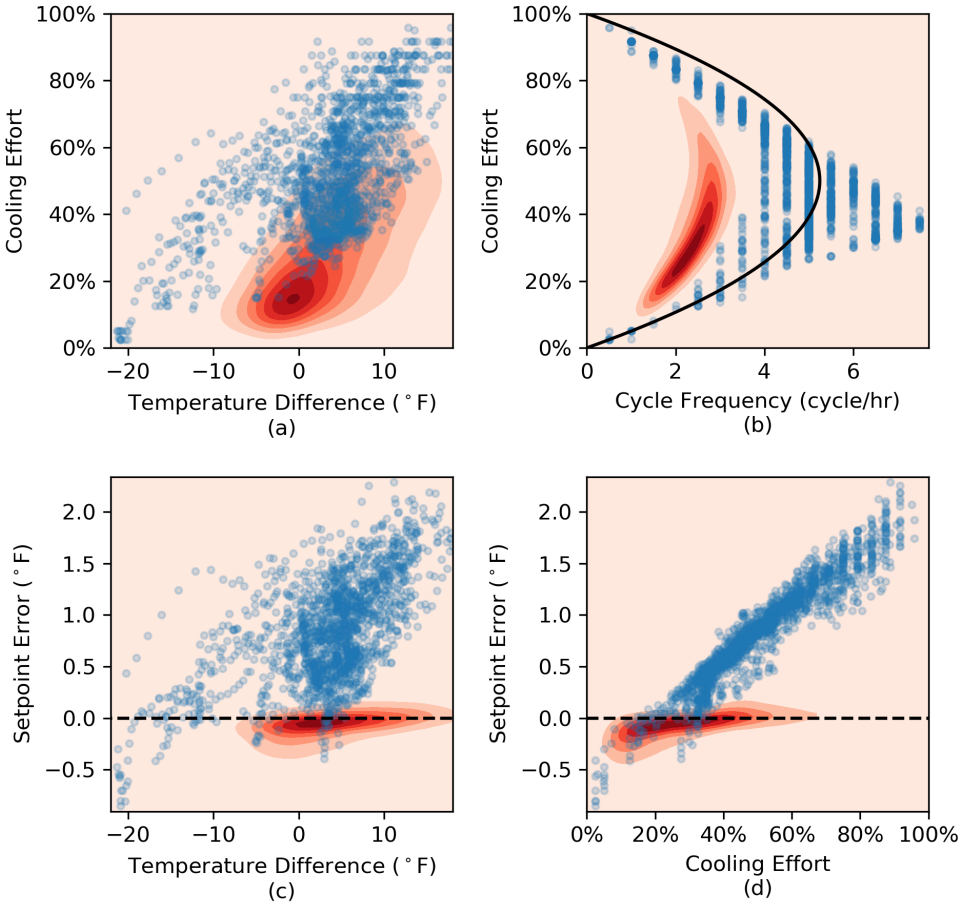


Figure 5.13: The scatterplots show the relationships between four pairs of data features for system IV in cooling regulating periods. The contours show the population distributions. The curve in subplot (b) is the quadratic fitting of the individual system data points.

Figure 5.14 shows the system behavior in cooling tracking periods. Consistent with the visual indications in Figure 5.12, indoor temperature drops very fast every time the system turns on. Since cooling effort in the regulating mode is not very low, this system cannot be considered oversized, but is more likely a control problem.

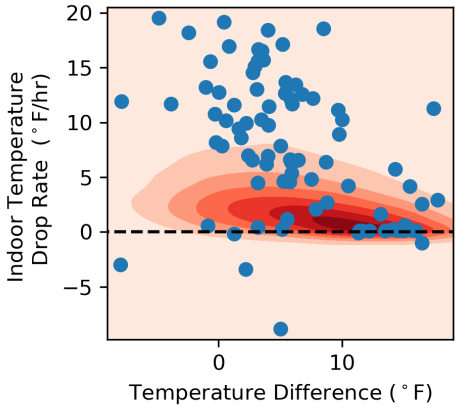


Figure 5.14: The scatterplot shows the relationship between temperature difference and indoor temperature drop rate for system IV in cooling tracking periods, and the contour is the population distribution.

Finally, Table 5.6 shows how system IV is identified as a faulty system by the detection algorithm:

- **Cooling effort is higher than average.** Approximately 16.7% of other systems operate in a larger proportion of time on average when regulates around the cooling setpoint.
- **Compressor cycling frequency is significantly higher than the average.** In the univariate test this system is flagged as faulty, because only 0.04% of other systems operate with a higher compressor cycling frequency on average.
- **Temperature setpoint error is significantly higher than the average.** The $\bar{\alpha}$ value is only 0.18%, showing that the setpoint error is abnormally high.

Table 5.6: Fault detection output of system IV.

System IV		$\bar{\alpha}$	Threshold	Faulty	Description
Univariate tests	E_c	16.7%	1%	False	Above average E_c
	F	**0.04%**	1%	**True**	Abnormally high F
	T_{spe}	**0.18%**	1%	**True**	Abnormally high T_{spe}
Multivariate tests	(E_c,F)	**0.10%**	5%	**True**	High cycling
	(E_c,T_{spe})	**0.30%**	5%	**True**	Poor control
	(F,T_{spe})	**0.15%**	5%	**True**	Poor control
	(E_c,F,T_{spe})	**0.06%**	5%	**True**	Mixed faulty behaviors

E_c – cooling effort, F – cycle frequency, T_{spe} – setpoint error

$\bar{\alpha}$ – percentage of other systems that exhibits worse performance than the presented one

- **The squared Mahalanobis distances of all four multivariate tests are extremely high.** This gives further evidence that this system is an outlier.

In conclusion, this system is a typical example that has relatively high cooling effort, abnormally high cycle frequency, and abnormally high setpoint error. Perhaps the system has control problems caused by software error, control interrupt, or thermostat misplacement. Even though no mechanical part is malfunctioning, this behavior is identified in a group of systems and deserves further investigation and repair.

5.3.5 System V: Oversize Problem

Lastly, this paper discusses the behavior of a system with oversized cooling capacity. Figure 5.15 displays its operational behavior in May. When the system regulates around 74°F, it always cools the house quickly and then shuts off. The duty cycle is very low even when the indoor and outdoor temperature difference exceeds 10°F. When the setpoint drops to 71°F, the system starts and stops quite frequently. Also, the reader should note that the setpoint error is high. This is because the control algorithm ensures at least five minutes of off-cycle between two on-cycles to protect the equipment, but the indoor temperature increases quickly during the off-cycle .

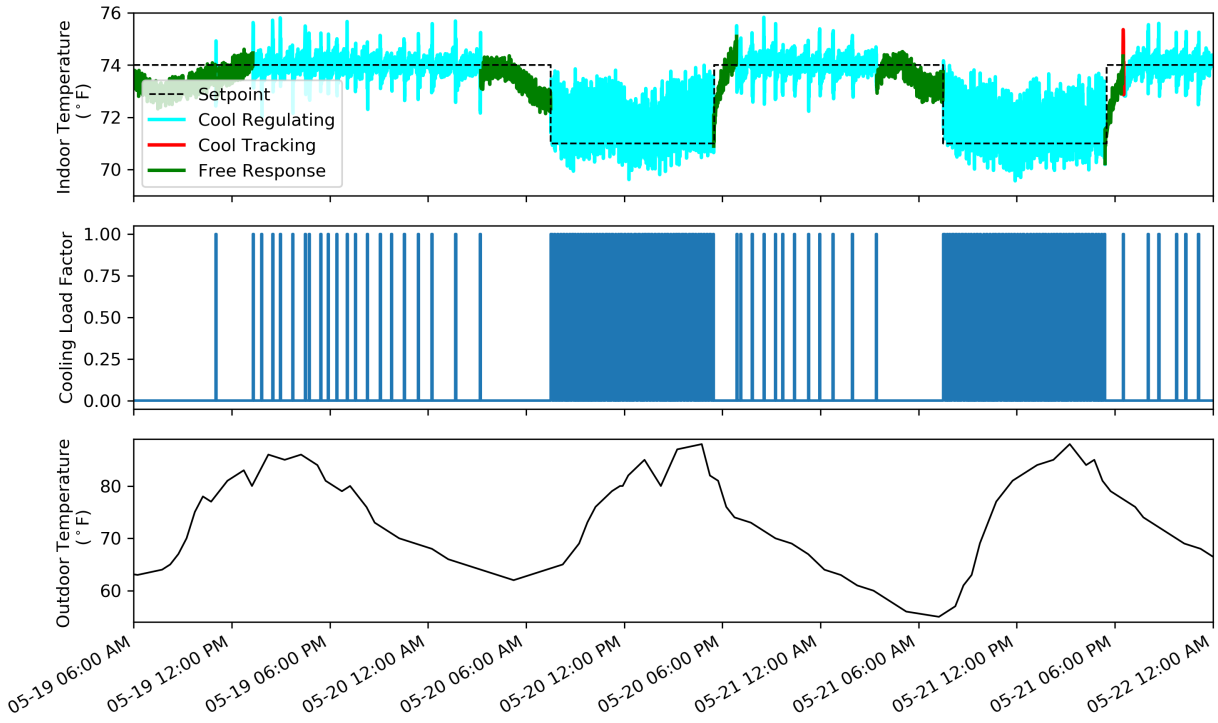


Figure 5.15: Daily operating condition of system V from May 19th to May 22nd with behavioral modes labeled. Cyan: cooling regulating mode; green: free response mode; dotted line: cooling temperature setpoint.

In Figure 5.16, four pairs of data features for system V in the cooling regulating mode are shown. Please note:

- **In Figure 5.16a, cooling effort is only 40% even when the outdoor temperature rises 20°F above the indoor temperature.** Note that all data in the cooling regulating mode from January to October in 2018 is included in this figure. The low percentage of cooling effort indicates this system is oversized.

- **In Figure 5.16b, the cycle frequency is extremely high and up to 7 cycles per hour.** Similar to system IV, this system has the problem of "short cycling", but the reason for this problem in this case is the large system size.

- **In Figure 5.16c and 5.16d, the setpoint error is usually positive and increases with temperature difference and cooling effort.** This situation occurs because the thermostat control algorithm balances the delay time of off-cycles and the dead-band width of the setpoint.

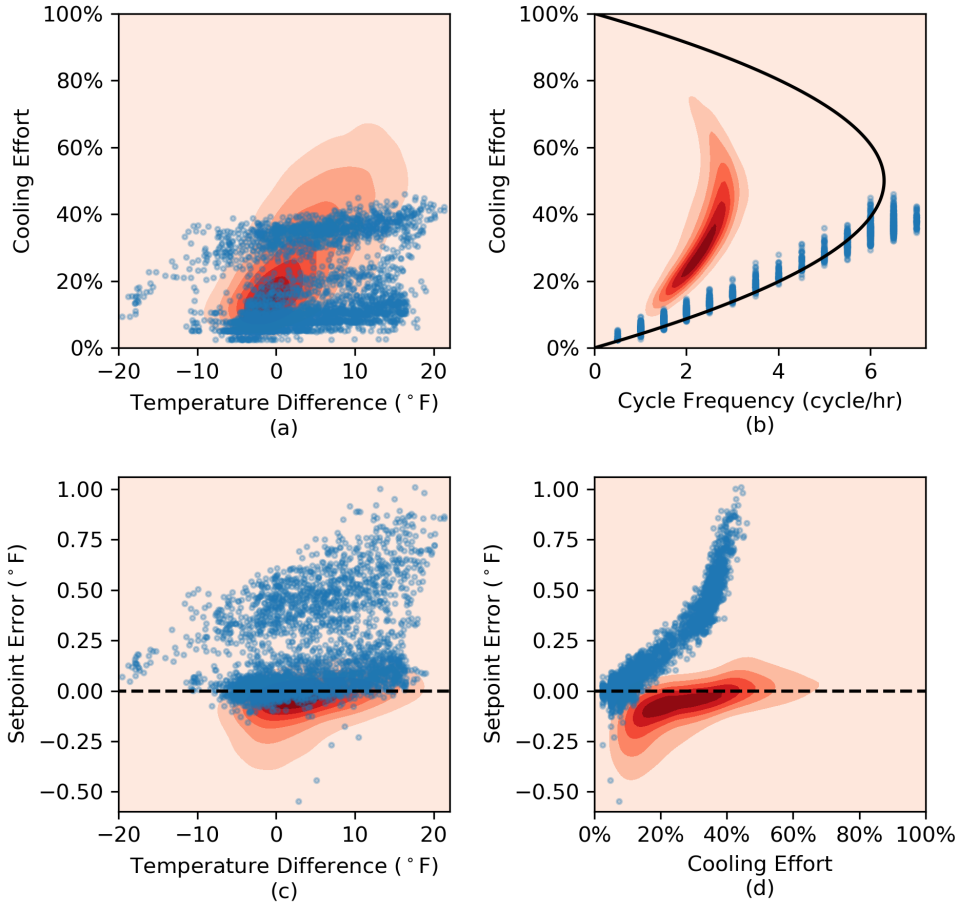


Figure 5.16: The scatterplots show the relationships between four pairs of data features for system V in cooling regulating periods. The contours show the population distributions. The curve in subplot (b) is the quadratic fitting of the individual system data points.

Figure 5.17 shows data features for all cooling tracking periods of this system. Since the system usually operates in the regulating mode, tracking periods are quite few. Nevertheless, the data features show the indoor temperature drop rate is usually greater than 10°F/hr, much higher

than most systems. Therefore, the system capacity is much greater than the total cooling load of the house.

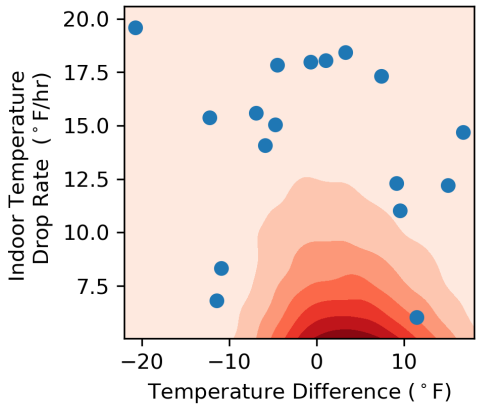


Figure 5.17: The scatterplot shows the relationship between temperature difference and indoor temperature drop rate for system V in cooling tracking periods, and the contour is the population distribution.

Finally, Table 5.7 shows how system V is identified as an oversized system by the detection algorithm:

- **Cooling effort is much lower than average.** Only 6.2% of other systems operates in a lower proportion of time on average when regulates around the cooling setpoint. This statistic is a strong indication of an oversizing issue.
- **Compressor cycling frequency is much higher than the average.** Though not flagged as faulty, only 8.2% of other systems operate with a higher compressor cycling frequency on average.
- **Temperature setpoint error is also much higher than the average.** Approximately only 7.2% of other systems have higher setpoint error.

Table 5.7: Fault detection output of system V.

System V		$\bar{\alpha}$	Threshold	Faulty	Description
Univariate tests	E_c	93.8%	1%	False	Low E_c
	F	8.2%	1%	False	High F
	T_{spe}	7.2%	1%	False	High T_{spe}
Multivariate tests	(E_c,F)	**2.7%**	5%	**True**	Oversized
	(E_c,T_{spe})	-	5%	False	-
	(F,T_{spe})	-	5%	False	-
	(E_c,F,T_{spe})	-	5%	False	-

E_c – cooling effort, F – cycle frequency, T_{spe} – setpoint error

$\bar{\alpha}$ – percentage of other systems that exhibits worse performance than the presented one

- **The squared Mahalanobis distance of cooling effort and cycling frequency (E_c,F) is significantly higher than most systems.** This metric successfully captured the abnormal behavior of this system.

In conclusion, this is a typical example of a system with low cooling effort, high cycle frequency, and high setpoint error, and this behavior is associated with the oversizing issue. Continued operation with this type of behavior has several drawbacks: the compressor and motors use more power due to unnecessarily high numbers of start; there is increased wear on the equipment due to frequent cycling; and the system does not operate for long enough periods to effectively remove moisture in the house.

5.4 Conclusions

The second part of this paper selects five representative normal and faulty systems, and presents a detailed case study of their operational behavior, and the application of the proposed fault detection method. First, a typical period of the operational data is shown to help the reader intuitively understand the general behavior. Then, extracted data features for each system in the cooling regulating mode and the cooling tracking mode are visually compared to the population distributions. Finally, the output statistics from the proposed fault detection method are analyzed to show the

reasons why these systems are automatically identified as normal or faulty, and the potential type of faults based on the statistics are discussed.

The case studies demonstrate that the proposed fault detection method is effective in identifying systems with abnormal behaviors by using only smart thermostat data. The detected faulty systems are generally related to soft faults that are usually difficult to detect without extensive sensors or time-intensive review by an experienced technician. Moreover, many of these faults will go unnoticed by the occupant, as comfort conditions are maintained, albeit with degraded system efficiency. These include inappropriate capacity (e.g., undersized/oversized), capacity degradation, and control problems. Although the proposed method is not intended to explicitly diagnose exact faults, the detection results can provide a preliminary diagnosis for technicians. Moreover, the service companies no longer need to dispatch technicians to check every residential system manually, but instead focus on a smaller group of systems that are very likely to be faulty as identified by the proposed method. Future research is anticipated to include field studies to confirm the fault diagnosis and verify that repairs effectively corrected the identified problems.

6. PAPER D: DETECTION OF GRADUAL CAPACITY DEGRADATION FOR RESIDENTIAL HVAC SYSTEMS USING SMART THERMOSTAT DATA[1]

6.1 Synopsis

Smart thermostats have enabled powerful methods for detecting faults in residential HVAC systems with limited sensor information. In particular, the detection of gradual or soft faults, such as the degradation of heating or cooling capacity, with no additional sensors would allow technicians to identify and prioritize systems for repair prior to catastrophic failure. Degradation of capacity is often caused by refrigerant leakage, which goes largely unnoticed by occupants. Traditional methods rely on installing additional sensors and monitoring changes of key operational temperatures. In contrast, smart thermostats are widely adopted by residential home owners, with data streamed to the cloud for potential analysis. This paper presents two methods for extracting key data features from smart thermostat data, and then applying modified Mann-Kendall statistical tests to identify significant trends in cooling capacity. The first method analyzes hourly averaged data and has higher possibility to detect trends for systems spending the majority of time in effective regulating, while the second method analyzes daily averaged data and is suitable for all kinds of system operational behaviors with a couple of limitations. These two methods are applied to data collected from approximately 10,000 residential air conditioning systems for historical trend detection and real-time condition monitoring, with case studies selected from a few verified faulty systems to validate the approach.

6.2 Overview

A critical task in fault detection and diagnosis (FDD) for residential air conditioning systems is the capability to detect gradual system capacity degradation problems, typically caused by refrigerant undercharge due to an active leak or partially blocked air flow (e.g. crushed air duct). Because air conditioning systems are designed to be inherently robust, this capacity degradation may be

[1]F. Guo and B.P. Rasmussen, "Detection of Gradual Capacity Degradation for Residential HVAC Systems Using Smart Thermostat Data", *Applied Thermal Engineering*, **In Review**.

imperceptible to the occupants for months, but could eventually cause total system failure. Since an underperforming system wastes energy and eventually cause occupant discomfort, the ability to detect such faults early would benefit home owners, manufacturers, and service technicians.

6.2.1 Introduction to Capacity Degradation

As an example, Figure 6.1 shows a single-speed cooling system that suffered from capacity degradation from August to September. During these two months, the outdoor temperature did not increase, but the system runtime steadily increased. Eventually this problem began to affect indoor temperature, and the system was unable to regulate the room temperature to the desired cooling setpoint. The occupant did not notice this problem until late September, when s/he reduced the setpoint in an effort to compensate for the loss of comfort.

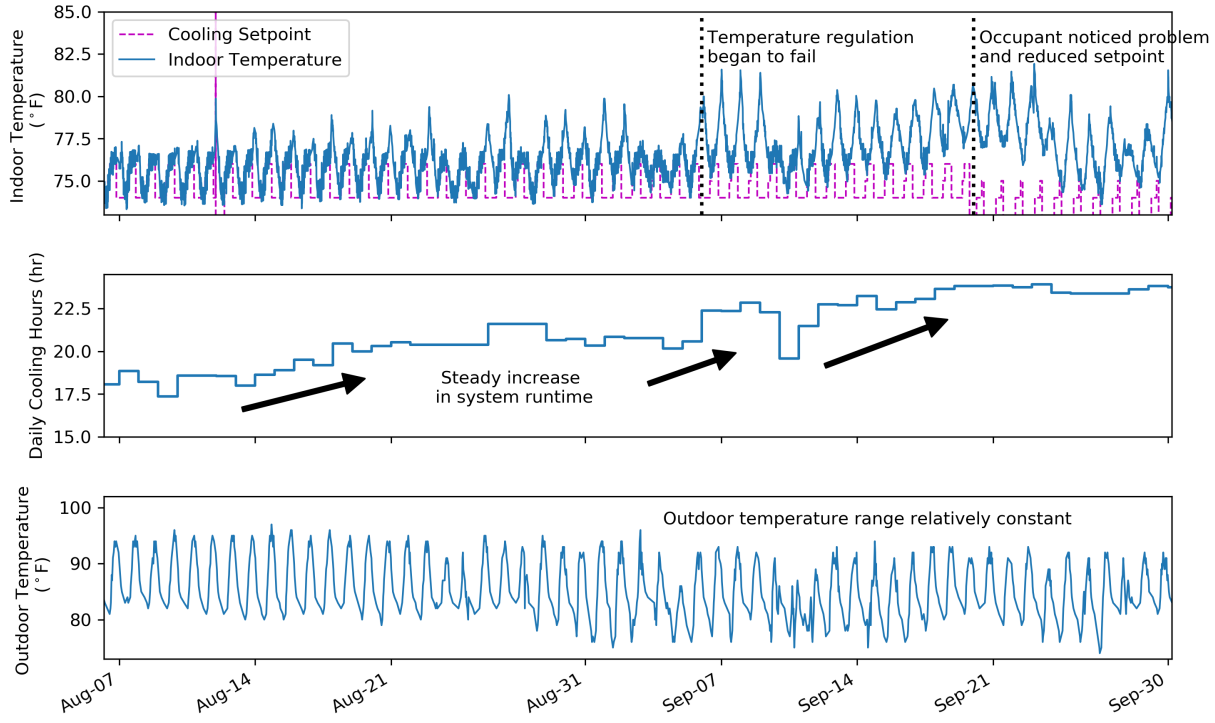


Figure 6.1: A single-speed cooling system that suffered from capacity degradation from August to September, and finally lost all of its cooling capacity. The occupant did not notice this problem until late September.

A conventional approach to diagnosing this issue would be to install pressure and temperature sensors to measure heat exchanger performance and determine if the system is undercharged, or install flow meters to estimate the air flow in the duct system. [47] summarized the experimental studies that focus on detecting degradation faults of residential split systems, including undercharge, reduced airflow, and liquid line restriction. To detect abnormality, usually these methods need to extract multiple features calculated from sensor records at the heat exchanger coils/ducts, compressor, liquid line and suction line, and compare them to the fault-free values. These inspection methods are accurate, but because residential HVAC systems are produced in large quantities, installing multiple sensors to every system is generally cost-prohibitive. Since this issue only exists in a small fraction of the systems and usually occurs long after initial system installation, this solution has not been widely adopted by manufacturers.

An alternative method is to use smart thermostat data. Currently, data collected from smart thermostats has shown broad potential for fault detection and diagnosis (FDD) of residential HVAC systems. This is a fundamentally different paradigm for FDD; instead of installing multiple sensors on a single system and comparing performance to general baseline, this approach uses relatively few sensors and uses historical data stored in the cloud. This results in several distinct advantages. First, thermostat data adds no initial cost. The indoor temperature, relative humidity, system cycling status, and setpoint all come from the thermostat built-in sensors, and the outdoor temperature may come from a third party or online sources. Second, smart thermostats record system operation conditions continuously and upload them to a remote database instantly. Engineers are able to access all the data from when the smart thermostat was installed and perform a full analysis for a specific system. Third, as smart thermostats have been widely adopted residential homes, practical FDD methods using thermostat data could be used by manufacturers or dealers to simplify and automate FDD efforts.

The principal challenge resulting from using thermostat data is correlating a large amount of transient data from multiple inputs to detect degradation trends. For example, one intuitive fault indicator that is correlated with capacity degradation of a cooling system, is *cooling effort*. Con-

sider a home where the internal heat loads remain unchanged, outdoor temperature is constant, and the indoor temperature is well-controlled at the desired cooling setpoint. In these conditions, if the residential cooling system's operating time increases, the system likely suffers from gradual loss in cooling capacity. Thus, cooling effort could be a useful fault indicator for air conditioning systems (and heating effort would be a useful indicator for heat pumps). However, in reality, the outdoor temperature does not remain constant, internal heat loads vary, and temperature setpoints are often changed by occupants or pre-programmed schedules. Under these conditions detecting gradual capacity degradation visually from the data is not trivial, and these dynamic variables must be considered in any FDD analysis.

This paper focuses on analyzing the *steady-state cooling effort* and the *daily cooling effort* as two primary indicators of capacity degradation. In order to mathematically define the steady-state cooling effort, another concept termed the *cooling regulating period* needs to be introduced first [6]. As shown in Figure 6.2a, a cooling regulating period is defined as a time period when a system cycles on-and-off to maintain a relatively constant indoor temperature around a constant cooling setpoint. The home being cooled is assumed to be in pseudo steady-state conditions. In a regulating period, cooling effort is defined as the product of the duty factor and the load factor for multi-speed systems, where duty factor is the percentage of time that a system operates, and load factor is the fraction of full capacity being used. For single-speed cooling systems with on-off control, cooling effort is equivalent to duty factor. Figure 6.2b exemplifies the calculation of cooling effort in a one-hour time window. During this hour, the duty factor is 40% and the load factor is unity when the system is on, so the cooling effort is 40%.

The daily cooling effort is defined as the percentage of operating time in a single day for a single-speed cooling unit with on/off control. For a multiple-speed or variable-speed cooling unit, this definition is extended to the percentage of equivalent full capacity operating time.

To account for the effects of conditional variables such as internal heat load and outdoor conditions, the proposed methods apply the following techniques and procedures. A complete description is given in Section 6.5.

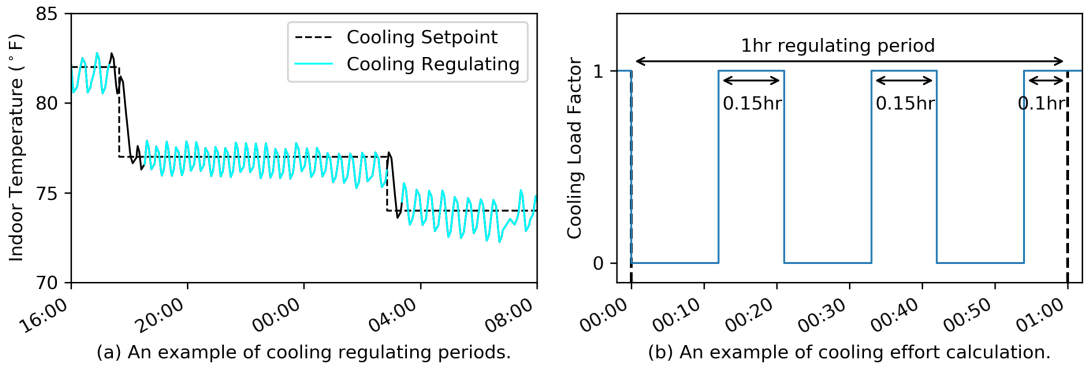


Figure 6.2: An illustration of the cooling regulating periods and the calculation of cooling effort.

- The variation of internal heat load is modeled as a normally distributed random variable.
- For hourly analysis, the steady-state cooling effort for a system is divided into several separate time series data, according to the indoor and outdoor temperature difference and time of day. The trend detection is performed for each time series, and then reconciled to determine if capacity degradation is statistically likely. Using data subsets with similar operating conditions, changes in steady-state cooling effort are more likely caused by system degradation.
- For daily analysis, the daily cooling effort is treated as a response variable, and other variables such as the daily average outdoor and indoor temperature, daily average cooling setpoint, are treated as covariates that can have a large impact on the response variable. Changes in daily cooling effort are more likely caused by system degradation when covariate effects are eliminated.

6.2.2 Proposed Degradation Trend Detection

Once the thermostat data or other available sensor data is collected, capacity degradation can be monitored by identifying sudden changes or detecting trends in these fault indicators. For example, [28] proposed a change point detection method using a moving data window and the statistical T-test, and successfully applied this method to identify sudden persistent changes in the airflow rate and steady-state cooling effort from smart thermostats.

While there are several other available change point detection methods intended for identifying sudden persistent changes, this study specifically proposes a method for detecting various gradual changes. Real operational data for residential heating and cooling systems is anticipated to exhibit a mixture of gradual changes and small abrupt changes. A key element in the overall detection scheme is the Mann-Kendall test, a nonparametric trend detection method. The Mann-Kendall test is commonly used in other disciplines, such as hydrology, climate change, rainfalls and streamflow, etc. [48] One advantage of its application to thermostat data is the nonparametric nature, such that the fault indicators are not necessarily assumed to follow a particular statistical distribution. (In fact, they probably do not follow a common statistical distribution, considering the different behavior of thousands of systems operating in a variety of conditions.)

When covariates exist, there are two types of techniques for analyzing trends of the response variable [33]. The first is a stagewise procedure: a regression model is created between the response variable and covariates, and the regression residuals are tested for trends using any acceptable trend detection method. The second is a multivariate approach, where trend detection and correction for covariates are performed simultaneously. [49, 50, 51] argue that the multivariate approach is better at analyzing trends, in that the stagewise procedure can yield misleading results especially if the covariates also have trends. This paper chooses to apply the partial Mann-Kendall test, a multivariate approach accounting for covariates, to test the increasing trends of the daily cooling effort.

The Mann-Kendall test is originally designed for *independent* time-series data, however *autocorrelated* (or *serially correlated*) time-series data is very common in practice. For example, the cooling effort of a system at the current moment in time is positively correlated to the cooling effort several hours prior, due to the occupant behavior, building thermal capacity, thermostat control strategy, and the weather. This positive correlation can cause the Mann-Kendall test to show evidence of a trend while no trend actually exists [32]. According to this study, the lag-one autocorrelation coefficient (i.e. the correlation in a time series between two sequential values) of the

cooling effort typically ranges from a value of 0 (i.e. no correlation) to 0.6 (i.e. highly correlated), which shows the necessity of considering the autocorrelation effect for this application.

There are four main approaches to modifying the Mann-Kendall tests to account for the autocorrelation, namely pre-whitening [32], trend-free pre-whitening [52], variance correction [53, 54], and block bootstrapping [55]. For this particular application, the authors determined that the block bootstrapping method is more effective and is also applicable to the partial Mann-Kendall test. For a detailed comparison of the performance of each method, see Appendix A.1. Bootstrapping is a method of random sampling with replacement used in statistical tests. Through shuffling a given dataset many times and performing the test for each generated sample, a distribution of the test statistic can be obtained. The testing result of the given dataset is based on the percentile of the test statistic in its distribution. For correlated data, at the beginning the given dataset is subdivided into several blocks to keep the autocorrelation, and shuffling is performed on the block level. Therefore, the method is called block bootstrapping (BBS).

The following section introduces the theoretical background of the original Mann-Kendall test, the partial Mann-Kendall test, and their bootstrapping modifications that can be applied to the smart thermostat data. Section 6.4 uses a Monte Carlo simulation to quantify the relative performance of these tests, and demonstrate that the Mann-Kendall tests with moving block bootstraps have a low false positive/negative detection rate. Finally, Section 6.5 applies this process to the cooling effort data from residential HVAC systems validates the approach using actual case studies of known faulty systems suffering from refrigerant leakage.

6.3 Theoretical Background

In this section, the theoretical background of the Mann-Kendall statistical trend test is introduced. The original Mann-Kendall test (MK) is shown first, which is able to detect monotonic trends for independent time series data. Then, the partial Mann-Kendall test (PMK) which incorporates the effect of covariates on trend is introduced. When autocorrelation effect exists, however, the Mann-Kendall test result will be biased [32]. Therefore, modified Mann-Kendall tests using a

moving block bootstrap (MBBS) approach are used. Finally, a robust and unbiased estimate of the slope of trend, known as the Sen's slope, is introduced.

6.3.1 The Original Mann-Kendall Trend Test (MK)

The Mann-Kendall test is a non-parametric statistical test commonly employed to detect monotonic trends [30, 31]. The null hypothesis, H_0, is that the data comes from a population with independent and identically distributed realizations. The alternative hypothesis, H_A, is that the data follows a monotonic trend. The Mann-Kendall test statistic is calculated as:

$$S = \sum_{k=1}^{n-1} \sum_{j=k+1}^{n} sgn(X_j - X_k) \tag{6.1}$$

where X is the variable, n is the number of observations, and $sgn(\cdot)$ is the signum function that determines the positive/negative nature of the function argument.

The S-statistic is approximately in normal distribution with the mean $E(S)$ and the variance $\sigma^2(S)$ calculated as:

$$E(S) = 0 \tag{6.2}$$

$$\sigma^2(S) = \frac{n(n-1)(2n+5) - \sum_{m=1}^{n} t_m m(m-1)(2m+5)}{18} \tag{6.3}$$

where t_m is the number of ties of length m.

Then, with the following transformation,

$$Z = \begin{cases} \frac{S-1}{\sigma} & \text{if } S > 0 \\ 0 & \text{if } S = 0 \\ \frac{S+1}{\sigma} & \text{if } S < 0 \end{cases} \tag{6.4}$$

the Z-statistic is approximately in standard normal distribution. In a one-side test for monotonic increasing trend, H_0 should be rejected if $Z > Z_\alpha$, where α is the significance level for the test.

6.3.2 The Partial Mann-Kendall Trend Test (PMK)

The partial Mann-Kendall test (PMK) is proposed by [56] and examined by [33]. As an improvement of the original test, the PMK test also considers the effect of covariates on the testing variable. The PMK test analyzes a multivariate time series X with a total of p variables ($g = 1, 2, \ldots, p$) and each variable has n observations. X can be written as an $n \times p$ matrix where $g = 1$ is the response variable (i.e. testing variable) and from $g \geq 2$ are the covariates.

$$X = \begin{bmatrix} X_{11} & \ldots & X_{1p} \\ \vdots & \ddots & \vdots \\ X_{n1} & \ldots & X_{np} \end{bmatrix} \tag{6.5}$$

The S-statistic for each variable X_g is calculated as:

$$S_g = \sum_{k=1}^{n-1} \sum_{j=k+1}^{n} sgn(X_{jg} - X_{kg}) \tag{6.6}$$

Same as the original test, the S-statistic of the partial test is approximately normally distributed. The mean $E(S_g)$ and the variance $\sigma^2(S_g)$ are calculated as:

$$E(S_g) = 0 \tag{6.7}$$

$$\sigma^2(S_g) = \frac{n(n-1)(2n+5) - \sum_{m=1}^{n} t_m m(m-1)(2m+5)}{18} \tag{6.8}$$

where t_m is the number of ties of length m for variable g.

The conditional covariance σ_{gh} between the S-statistic of two variables S_g and S_h for a dataset is estimated as [57]:

$$\hat{\sigma}_{gh} = \frac{1}{3}(K_{gh} + 4\sum_{j=1}^{n} R_{jg}R_{jh} - n(n+1)^2) \tag{6.9}$$

where

$$K_{gh} = \sum_{1 \leq i < j \leq n} sgn((X_{jg} - X_{ig})(X_{jh} - X_{ih})) \tag{6.10}$$

$$R_{jg} = \frac{1}{2}(n + 1 + \sum_{i=1}^{n} sgn(X_{jg} - X_{ig})) \tag{6.11}$$

Note that the estimated conditional variance $\hat{\sigma}_{gg}$ is equal to the variance of the S-statistic $\sigma^2(S_g)$ in Equation (6.8).

Then, defining

$$\begin{bmatrix} \hat{a}_2 \\ \vdots \\ \hat{a}_p \end{bmatrix} = \begin{bmatrix} \hat{\sigma}_{22} & \dots & \hat{\sigma}_{2p} \\ \vdots & \ddots & \vdots \\ \hat{\sigma}_{p2} & \dots & \hat{\sigma}_{pp} \end{bmatrix}^{-1} \begin{bmatrix} \hat{\sigma}_{12} \\ \vdots \\ \hat{\sigma}_{1p} \end{bmatrix} \tag{6.12}$$

[33] prove that the following Z-statistic

$$Z = \frac{S_1 - \sum_{j=2}^{p} \hat{a}_j S_j}{\sqrt{\hat{\sigma}_{11} - \sum_{j=2}^{p} \hat{a}_j \hat{\sigma}_{1j}}} \tag{6.13}$$

weakly converges to a standard normal distribution as n approaches infinity, if the following matrix

$$\Sigma = \begin{bmatrix} 1 & \rho_{12} & \dots & \rho_{1p} \\ \rho_{21} & 1 & \dots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \dots & 1 \end{bmatrix} \tag{6.14}$$

is full rank. In Equation (6.14), ρ_{gh} is the unconditional correlation between S_g and S_h

$$\rho_{gh} = \frac{cov(S_g, S_h)}{\sqrt{\sigma^2(S_g)\sigma^2(S_h)}} \tag{6.15}$$

In a one-side test for monotonic increasing trend, H_0 should be rejected if $Z > Z_\alpha$ in Equation (6.13), where α is the significance level of the test.

6.3.3 The Mann-Kendall Test Using Moving Block Bootstrap (MBBS)

When autocorrelation exists in the time series, the Z-statistic will deviate from the standard normal distribution. Bootstrapping provides an alternative approach to obtain the distribution of the

Z-statistic by random sampling. This paper applies the moving block bootstrapping method [58] to both original Mann-Kendall test and partial Mann-Kendall test. The procedures are summarized as follows:

1. The Z-statistic of the test data is calculated by the Mann-Kendall test.

2. The test data is randomly resampled with replacement for N times. In order to keep the autocorrelations between data points, the data is subdivided into $n - L + 1$ overlapping blocks, where n is the number of observations and L is the length of each block. Resampling is performed on the block level. Figure 6.3 illustrates this step with $n = 5$ and $L = 3$.

3. The Z-statistics of all N resampled dataset are calculated by the Mann-Kendall test. They form a distribution of the Z-statistic.

4. Percentile of the Z-statistic (P) of the test data in the distribution is calculated.

5. In a one-side test for monotonic increasing trend, if $P > 1 - \alpha$, where α is the significance level of the statistical test, then the null hypothesis should be rejected.

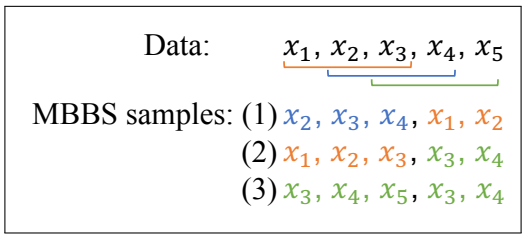


Figure 6.3: An illustration of MBBS procedures with $n = 5$ and $L = 3$. The data is split into $n - L + 1 = 3$ blocks marked by different colors, and three samples are randomly chosen from the blocks with replacement.

The number of bootstrap samples (N) to be generated depends on the significance level (α), and the selection of the block length (L) depends on the autocorrelation coefficient (ρ) and number of observations (n) in the time series.

6.3.4 Estimating the Slope and Magnitude of the Trend

The above trend test can only indicate whether there is *statistical* significance of a trend in the dataset. After a trend is detected, however, engineers are more interested in the *practical* significance of the trend, in other words, the actual magnitude of increase of the cooling effort. [34, 35] proposed a robust and unbiased estimate of the slope of the trend known as Sen's slope, given by:

$$d_k = \frac{X_j - X_i}{t_j - t_i} \text{ for all } 1 \leq i \leq j \leq n \tag{6.16}$$

and

$$\text{Sen's slope} = \text{median}(d_k) \tag{6.17}$$

In the equation, X is the variable, t is time, and d_k's are the slopes between each pair of variables. Then, the Sen's slope is calculated as the median of all the slopes d_k. The magnitude of the trend therefore is the product of the Sen's slope and the time length of the data series.

6.4 Monte Carlo Simulation of the Mann-Kendall Tests

A critical objective in fault detection and diagnosis (FDD) for residential HVAC systems is to not generate false positives, where fault-free systems are incorrectly identified as faulty. There is a low tolerance for false positives in practice, as it is a waste of time and resources for maintenance technicians. Additionally, the occupants may feel misled, which jeopardizes customer loyalty. Another possibility is that the FDD methods may generate false negatives, where faulty systems fail to be detected. However, false negatives, to some extent, are preferable to false positives, because faulty systems that degrade in performance or operation will eventually be detected.

In statistical theory, a false positive is defined as a Type I error, which means the rejection of a true null hypothesis. A false negative is defined as a Type II error, which means failure to reject a false null hypothesis. In this Mann-Kendall test the Type I error should be minimized, such that virtually every system with the accepted alternative hypothesis is guaranteed to have a capacity degradation trend. After achieving this primary objective, the prevalence of Type II errors

can be investigated to determine the most effective FDD strategy. The purpose of this section is to quantify the Type I and Type II error for the modified Mann-Kendall tests using Monte Carlo simulation of anticipated fault scenarios.

Before evaluating the Type I and Type II error of the Mann-Kendall tests, two key factors should be discussed: 1) the signal autocorrelation, and 2) the signal-to-noise ratio. Figure 6.4 compares two datasets with zero mean and no trend. The first dataset is uncorrelated, and the second dataset exhibits high positive autocorrelation. Compared to the random uncorrelated data in Figure 6.4a, the data with high autocorrelation in Figure 6.4b shows strong similarity between adjacent data points and thus seems to have a trend. The existence of positive autocorrelation adds difficulty in detecting a trend and often results in higher probability of obtaining a Type I error.

Similarly, Figure 6.5 compares two uncorrelated datasets with the same trend slope, but with high and low signal-to-noise ratios. The noise level in Figure 6.5b is five times greater than the noise in Figure 6.5a, such that detecting a trend in Figure 6.5a is much easier.

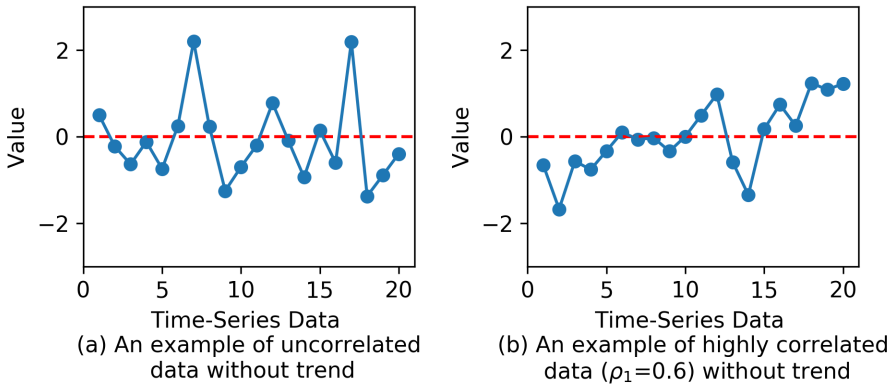


Figure 6.4: Two datasets with zero mean (dashed red line) and zero trend. Compared with the uncorrelated data in plot (a), Type I error occurs more frequently and trend detection is more difficult for the highly positively correlated data in plot (b).

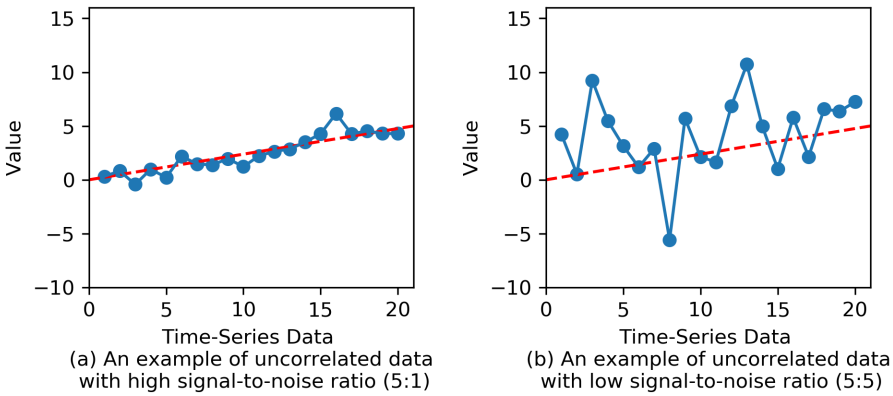


Figure 6.5: Two datasets with the same slope of trend (dashed red line) and without autocorrelation. Trend detection of the dataset with high signal-to-noise ratio in plot (a) is much easier than the dataset with low signal-to-noise ratio in plot (b).

6.4.1 Control of the Type I Error for the MK-MBBS Test

In the proposed degradation trend detection, the significance level α, also referred to as the upper bound of the Type I error, is set to 0.1%. That is to say, at most one out of a thousand fault-free systems will be mistakenly detected as faulty. At this significance level, the Type I error of the Mann-Kendall test with moving block bootstrap (MK-MBBS) is compared with the original test. To evaluate the impact of the autocorrelation coefficient and sample size on Type I errors, a Monte Carlo approach to simulate fault-free scenarios is used. 50,000 autocorrelated time series without a trend is generated in each level of lag-one autocorrelation coefficient (ρ_1) ranging from 0 to 0.8 and sample size (n) ranging from 30 to 100. The autocorrelated time series is generated by autoregressive model of order one (i.e. AR(1) model):

$$X_t = \rho_1 X_{t-1} + \epsilon_t \tag{6.18}$$

where X_t is data at time t, X_{t-1} is data one time-lag before, and ϵ_t is the random noise term. For each simulated time series, 50,000 bootstrap samples (N) are taken, and the block length (L) is 7 when the sample size is smaller or equal to 50, and is 10 otherwise. The algorithms proposed in

Table 6.1: Type I error under 0.1% significance level for the original Mann-Kendall test and the MK-MBBS test with different lag-one autocorrelation coefficient (ρ_1) and sample sizes (n). Each case is obtained by 50,000 Monte Carlo simulations.

Method	ρ_1	Type I Error (supposed to be lower than 0.1%) $n = 30$	$n = 50$	$n = 100$
Original MK	0	0.08%	0.09%	0.09%
	0.2	0.33%	0.41%	0.49%
	0.4	1.25%	1.61%	1.79%
	0.6	3.76%	4.63%	5.17%
	0.8	10.0%	12.0%	13.7%
MK-MBBS	0	0.02%	0.01%	0.01%
	0.2	0.03%	0.02%	0.03%
	0.4	0.04%	0.03%	0.05%
	0.6	0.09%	0.06%	0.06%
	0.8	0.21%	0.31%	0.25%

[59] are used to generate moving block bootstrap samples, and the a modified algorithm from [60] is used to perform the Mann-Kendall tests.

From the results shown in Table 6.1, the original Mann-Kendall test has problems maintaining the Type I error lower than 0.1% when autocorrelation exists ($\rho_1 > 0$). Especially, when $\rho_1 > 0.4$ (i.e. highly autocorrelated), the actual Type I error is above 1%. That is to say, at least one out of a hundred fault-free systems will be mistakenly detected as faulty, which is unacceptable considering the large quantity of residential HVAC systems. In contrast, the MK-MBBS test is capable of keeping the Type I error below the 0.1% significance level and generates much fewer false positives. When the data is slightly autocorrelated, the Type I error is even much lower than 0.1%. When $\rho_1 > 0.6$, the Type I error slightly rises above the significance level, but this study found that most cooling effort time-series has $\rho_1 < 0.6$. Therefore, the MK-MBBS test is anticipated to be effective in identifying the degradation trend issues of residential HVAC systems with low false positivity rate.

6.4.2 Control of the Type I Error for the PMK-MBBS Test

Similar to Section 6.4.1, at 1% and 5% significance level, the Type I error of the partial Mann-Kendall test with moving block bootstrap (PMK-MBBS) is compared with the partial test. Note that a higher significance level is used in this simulation because of the high computational cost to generate a large number of bootstrap samples, but the eventual application of the method to residential cooling systems uses a 0.1% significance level.

The simulation constructs a bivariate time series (X) containing one response variable (X_1) and one covariate (X_2) with length n:

$$X = \begin{bmatrix} X_1 & X_2 \end{bmatrix} = \begin{bmatrix} X_{11} & X_{12} \\ \vdots & \vdots \\ X_{n1} & X_{n2} \end{bmatrix} \tag{6.19}$$

The autocorrelation of X_1 and X_2 is modeled by the AR(1) model:

$$\begin{cases} X_{t1} = \rho_{1,1} X_{(t-1)1} + \epsilon_t \\ X_{t2} = \rho_{1,2} X_{(t-1)2} + \delta_t \end{cases} \tag{6.20}$$

where $\rho_{1,1}$ is the lag-one autocorrelation coefficient for X_1, and $\rho_{1,2}$ is the lag-one autocorrelation coefficient for X_2.

To ensure the variance of X_1 and X_2 are both unity, and the cross-correlation between X_1 and X_2 is ϕ, the covariance matrix of the Gaussian noise $cov(\epsilon_t, \delta_t)$ is calculated as:

$$cov(\epsilon_t, \delta_t) = \begin{bmatrix} 1 - \rho_{1,1}^2 & (1 - \rho_{1,1}\rho_{1,2})\phi \\ (1 - \rho_{1,1}\rho_{1,2})\phi & 1 - \rho_{1,2}^2 \end{bmatrix} \tag{6.21}$$

To simulate fault-free scenarios 5,000 one-month-long and three-month-long bivariate time series without trends are generated for each level of cross-correlation and autoregressive coefficients.

In the PMK-MBBS test, 5,000 bootstrap samples are taken from each simulated time series and the block length is 7.

Shown in Table A.2 in the Appendix A.2, the PMK test maintains the Type I error at varying levels of cross-correlation when no autocorrelation exists, but positive autocorrelation of the response variable causes the Type I error to increase significantly, and can be as high as 28% at a 5% significance level, which is intolerable. In contrast, the PMK-MBBS test performs much better at controlling the Type I error, particularly at the 1% significance level. Although the Type I error can still rise above the nominal level when the correlation is very high, it is acceptable considering that there is no perfect way to correct the effect of autocorrelation.

6.4.3 Evaluation of the Type II Error for the MK-MBBS Test

After verified that the proposed tests can ensure the Type I error at the chosen significance levels, the Type II error is evaluated. If the null hypothesis cannot be rejected even when the time series data has trends typical of residential cooling systems, then the MK-MBBS test will be ineffective. In this study, four types of possible degradation trends are tested for the MK-MBBS test as shown in Figure 6.6: a single step change indicating a sudden but persistent behavioral change, multiple step changes indicating stepwise capacity degradation, a linear trend indicating capacity degradation at the same speed, and a quadratic trend indicating accelerated degradation. All of the four trends are tested using autocorrelated data.

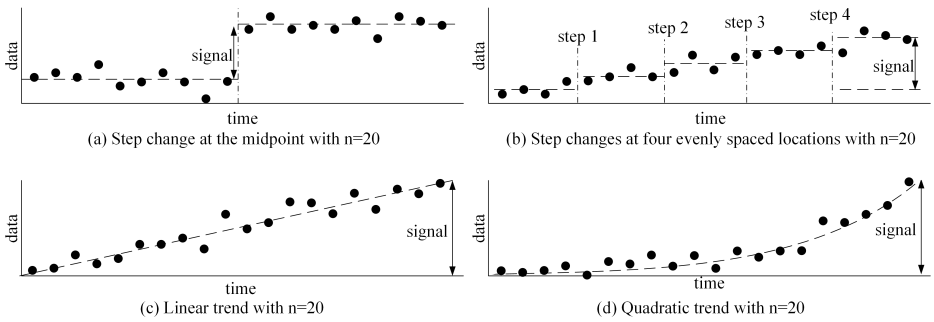


Figure 6.6: Four types of possible trends for HVAC systems studied by Monte Carlo simulations.

Figure 6.7 shows the probability of detecting a change or trend when it actually exists (i.e. the probability of avoiding a Type II error) for the MK-MBBS test. The sample size of each dataset is 100. The noise (variance) of the dataset is set to be one so the signal-to-noise ratio is equivalent to the magnitude of change. Note that to ensure the noise level of the simulated samples is unity, the variance of the random noise (ϵ_t) is calculated as $1 - \rho_1^2$. The results demonstrate that the MK-MBBS test is capable of detecting multiple step changes, linear trends, and quadratic trends with high detection probability, especially when the signal-to-noise ratio is higher than three. However, the MK-MBBS test cannot effectively detect a single persistent step change. To address this problem in practice, the authors recommend that the trend test be combined with a change point detection method, such as [28]. Additionally, similar to the original Mann-Kendall test, detecting a trend using the modified Mann-Kendall test becomes more difficult as ρ_1 increases.

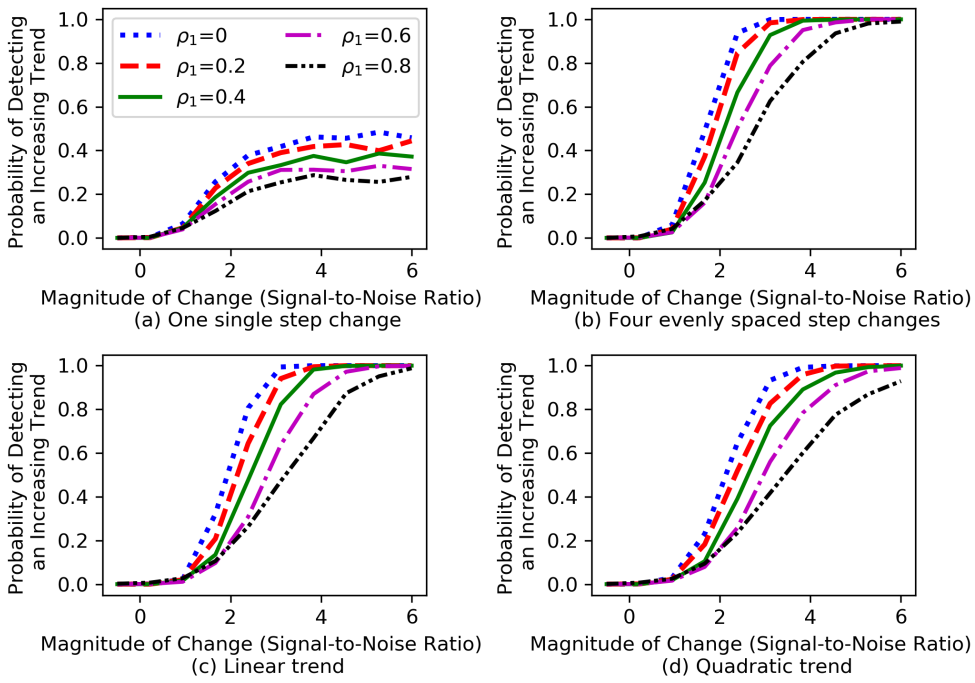


Figure 6.7: The probability of detecting four types of degradation trends when it actually exists. The sample size is 100 with ρ_1 ranging from 0 to 0.8. Trend is detected if percentile of the Z-statistic (P) satisfies $P > 99.9\%$.

6.4.4 Evaluation of the Type II Error for the PMK-MBBS Test

Similar to Section 6.4.3, the Type II error is evaluated for the PMK-MBBS test. Figure 6.8 and 6.9 show the probability of the PMK-MBBS test to detect a trend of the response variable at different levels of signal-to-noise ratio. In the simulations, the sample size is 30 and the significance level is 1%. In Figure 6.8, a trend is added only to the response variable with signal-to-noise ratio plotted on the horizontal axis, and the covariate exhibits no trend. In Figure 6.9, both response variable and covariate have trends. The signal-to-noise ratio of the response variable is plotted on the horizontal axis, and the signal-to-noise ratio of the covariate is represented as a ratio (η) of the response variable. The cross-correlation (ϕ) before adding trends is fixed at 0.8.

Shown in Figure 6.8, the PMK-MBBS test is able to detect monotonic increasing trends of the response variable and has better performance when the cross-correlation (ϕ) is high and the lag-one autocorrelation coefficients ($\rho_{1,1}$ and $\rho_{1,2}$) are low. Figure 6.9 shows that the probability to detect a trend of the response variable decreases as the trend of the covariate increases. This is reasonable because when the covariate has a trend, the increase of response variable can be partially explained by the increase of covariate, instead of being solely caused by the increasing trend of the response variable itself.

6.4.5 Situations When a Faulty System Is Repaired

In addition to the cases discussed above, false positives can also occur when a faulty system is recently repaired. Figure 6.10 shows a scenario where a linearly degraded system is fixed. A moving window is initially defined to include the data prior to the moment of repair. At first an increasing trend is likely to be detected, but as the window moves right, the probability of detecting an increasing trend declines until finally a decreasing trend is anticipated to be detected. Note that in practice this will be the primary reason why decreasing trends are detected, since systems themselves are unlikely to gradually improve in performance. Finally, as the window moves away from the data with the increasing trend, the detection algorithm is anticipated to identify no trend in the data.

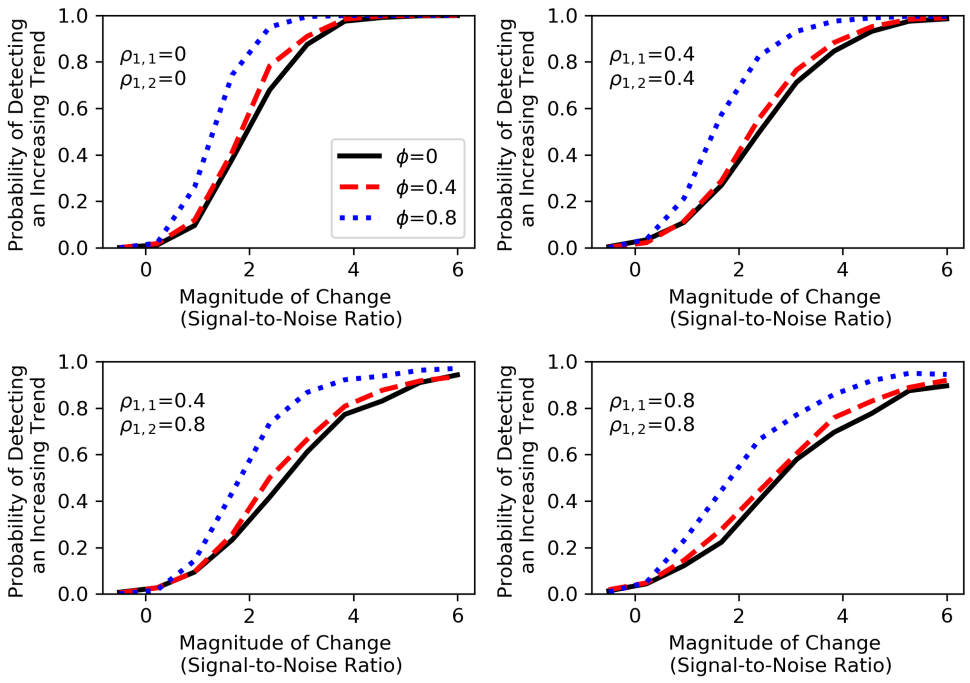


Figure 6.8: The probability of detecting an increasing trend of the response variable. The horizontal axis shows the signal-to-noise ratio of the response variable, and the covariate has no trend. Trend is detected if percentile of the Z-statistic (P) satisfies $P > 99\%$.

Figure 6.11 shows the simulation result of the MK-MBBS test for this scenario. For this simulation a size of the moving window is defined as 100 data points, with a signal-to-noise ratio of 5. In general, the MK-MBBS test will not report an increasing trend only after the moving window includes more than 10% of the data after the moment of repair, which indicates it can effectively avoid false positives.

6.5 Application to Smart Thermostat Data

The previous sections present the MK-MBBS and PMK-MBBS tests, and use simulated data to demonstrate their effectiveness for detecting the types of trends typical of faulty and repaired residential cooling systems. This section outlines two proposed methods, based on ***hourly*** analysis and ***daily*** analysis, to apply the tests to data collected from smart thermostats to detect capacity degradation problems.

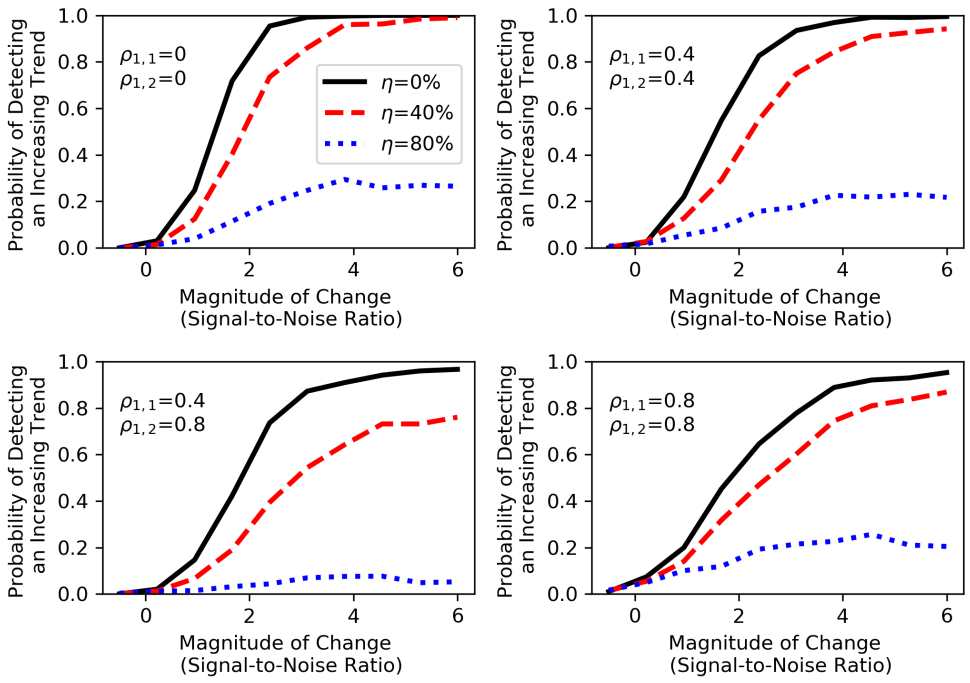


Figure 6.9: The probability of detecting an increasing trend of the response variable. Both response variable and covariate have trends. The horizontal axis shows the signal-to-noise ratio of the response variable, and the signal-to-noise ratio of the covariate is represented as a ratio (η) of the response variable. Trend is detected if percentile of the Z-statistic (P) satisfies $P > 99\%$.

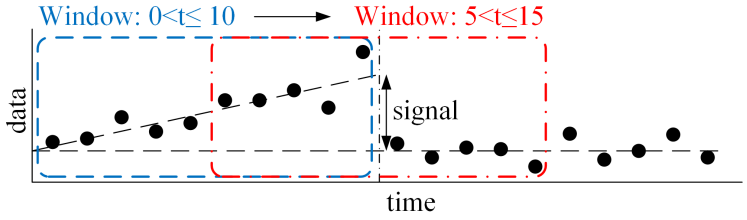


Figure 6.10: The situation when a linearly degraded system is repaired. The sample size is defined as the data points in the moving window, which is 10 in this example.

As discussed in the beginning of this paper, several conditional variables should be considered in the analysis residential cooling systems. For hourly analysis, the steady-state cooling effort is selected as the primary fault indicator. Equation 6.22 gives the energy balance of a residential

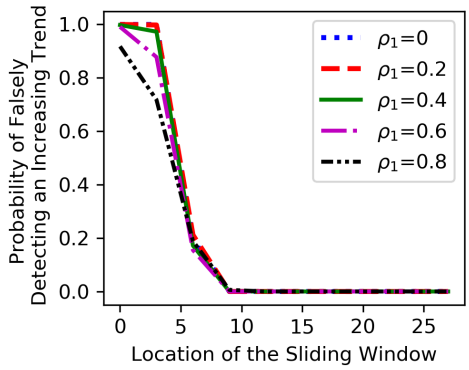


Figure 6.11: The probability of *falsely* detecting an increasing trend when actually a degraded system is fixed. The sample size is 100 with ρ_1 ranging from 0 to 0.8. The signal-to-noise ratio is 5. Degradation trend is detected if percentile of the Z-statistic (P) satisfies $P > 99.9\%$.

home operating in a pseudo steady-state condition:

$$C \cdot E_c = UA \cdot \Delta T_{oi} + \dot{Q}_{rad} + \dot{Q}_{int} \tag{6.22}$$

where C is the capacity of a cooling unit [W], E_c is the steady-state cooling effort in a time period (%), U is the overall heat transfer coefficient [W/m^2·°F], A is the area [m^2] for heat transfer through the building walls and roof, ΔT_{oi} is the indoor and outdoor temperature difference [°F] (i.e. the outdoor temperature minus indoor temperature), $\dot{Q}_{rad}$ is the solar irradiance heat transfer [W], and $\dot{Q}_{int}$ is the home internal heat generation [W]. The temperature difference (ΔT_{oi}) has a significant impact on the steady-state cooling effort (E_c). Therefore, the E_c data for a particular system should be divided into subsets based on the value of ΔT_{oi}. Moreover, to account for the daily variation of $\dot{Q}_{rad}$ and $\dot{Q}_{int}$ this paper further divides subsets using the time of day, knowing that during the daytime $\dot{Q}_{rad}$ and $\dot{Q}_{int}$ are generally higher than during the nighttime.

For daily analysis, the daily average cooling effort is selected as the fault indicator, and cooling degree day (CDD) is selected as the covariate due to its major effect on the daily cooling effort. The cooling degree day can be defined as the daily temperature difference between indoor and outdoor space, or the daily temperature difference between cooling setpoint and outdoor space [61]. This

study adopts both covariates so that the analysis is adaptable to a wider range of system operational behaviors.

Both the *hourly* and *daily* trend detection methods can be implemented using either *historical* or *real-time* data. Trend detection using ***historical*** data is primarily intended to evaluate system performance at the conclusion of peak cooling season, so that owners can schedule off-season maintenance. ***Real-Time*** trend detection is designed to actively monitor the system behavioral changes to identify urgent or imminent cooling capacity failures. The following subsections detail the applications of the proposed hourly and daily methods to residential cooling systems, using historical data (Section 6.5.1) and real-time condition monitoring (Section 6.5.2).

6.5.1 Historical Trend Detection

Since this study focuses on cooling systems, which operate primarily in the summer, performing a historical trend detection test after the summer season has passed is useful in identifying any changes in system performance during this peak season. Presumably, the historical trend detection would applied only once using the smart thermostat data for the whole summer. If significant system capacity degradation is identified after operating in the peak season, then service companies can schedule off-season maintenance afterwards to repair system faults and improve operating efficiency.

The data processing procedure is shown in Figure 6.12. The smart thermostat data is analyzed separately by the hourly and daily analysis methods. If both methods show practically significant evidence of a degradation trend, then the cooling system will be labeled as faulty and prioritized for maintenance. In some cases when only one method shows practically significant evidence of a degradation trend, the other method usually shows the same evidence, but at a lower confidence level (e.g. above 90% or 95%), and the system will still be labeled as faulty but categorized as a secondary priority for maintenance.

For hourly analysis, the cooling regulating periods are first labeled by the mode labeling algorithm [6], and the steady-state cooling effort is calculated for every consecutive two-hour time window in all cooling regulating periods. Second, the cooling effort data is divided into subsets

based on the indoor and outdoor temperature difference (ΔT_{oi}) and time of day. Twelve subsets are created and their ranges are shown in the Table 6.2. Subsets with less than 30 data points are excluded from analysis, as having insufficient data for analysis. Third, the MK-MBBS test is performed to each subset of the steady-state cooling effort time-series with $N = 50,000$ and $L = 7$, and the magnitude of trend in each subset is estimated by the Sen's slope. In general, a system has approximately four operational subsets to be tested. The percentile of the Z-statistics (P) and magnitude of trend (M) for each subset are recorded. Lastly, a statistically significant degradation trend is identified based on the second maximum percentile and the median percentile among all subsets (the total number of subsets must be greater than two). The second maximum percentile should satisfy $P_{second\_max} > 99.9\%$ (i.e. under 0.1% significance level), and as a supplementary condition the median percentile should satisfy $P_{median} > 95\%$ (i.e. under 5% significance level). Additionally, the trend is considered practically significant only if the second maximum magnitude of trend satisfies $M_{second\_max} > 0.125$.

For the daily analysis, first the daily cooling effort, the cooling degree day between setpoint and outdoor (covariate), and the cooling degree day between indoor and outdoor (covariate) are extracted from the smart thermostat data. Second, the algorithm checks whether the cross-correlation between daily cooling effort and cooling degree day is positive. If not, then the corresponding cooling degree day will be removed. A negative relationship can be caused by degradation, repair, or other occupant related issues, which often results in erroneous statistical results. Third, the Mann-Kendall test is performed to the daily cooling effort time-series if the time-series has data for at least 30 days, and the Sen's slope is applied to estimate the magnitude of trend of the daily cooling effort. If all covariates are removed, the MK-MBBS test is applied with $N = 50,000$ and $L = 7$; otherwise, the PMK-MBBS test is applied with $N = 50,000$ and $L = 7$. Finally, a practically significant degradation trend is detected if the percentile of Z-statistics satisfies $P_{daily} > 99.9\%$ and the magnitude of trend satisfies $M_{daily} > 0.125$.

To illustrate the performance of the historical trend detection, smart thermostat data collected from approximately 10,000 single-speed cooling systems located in Florida between June 1st and

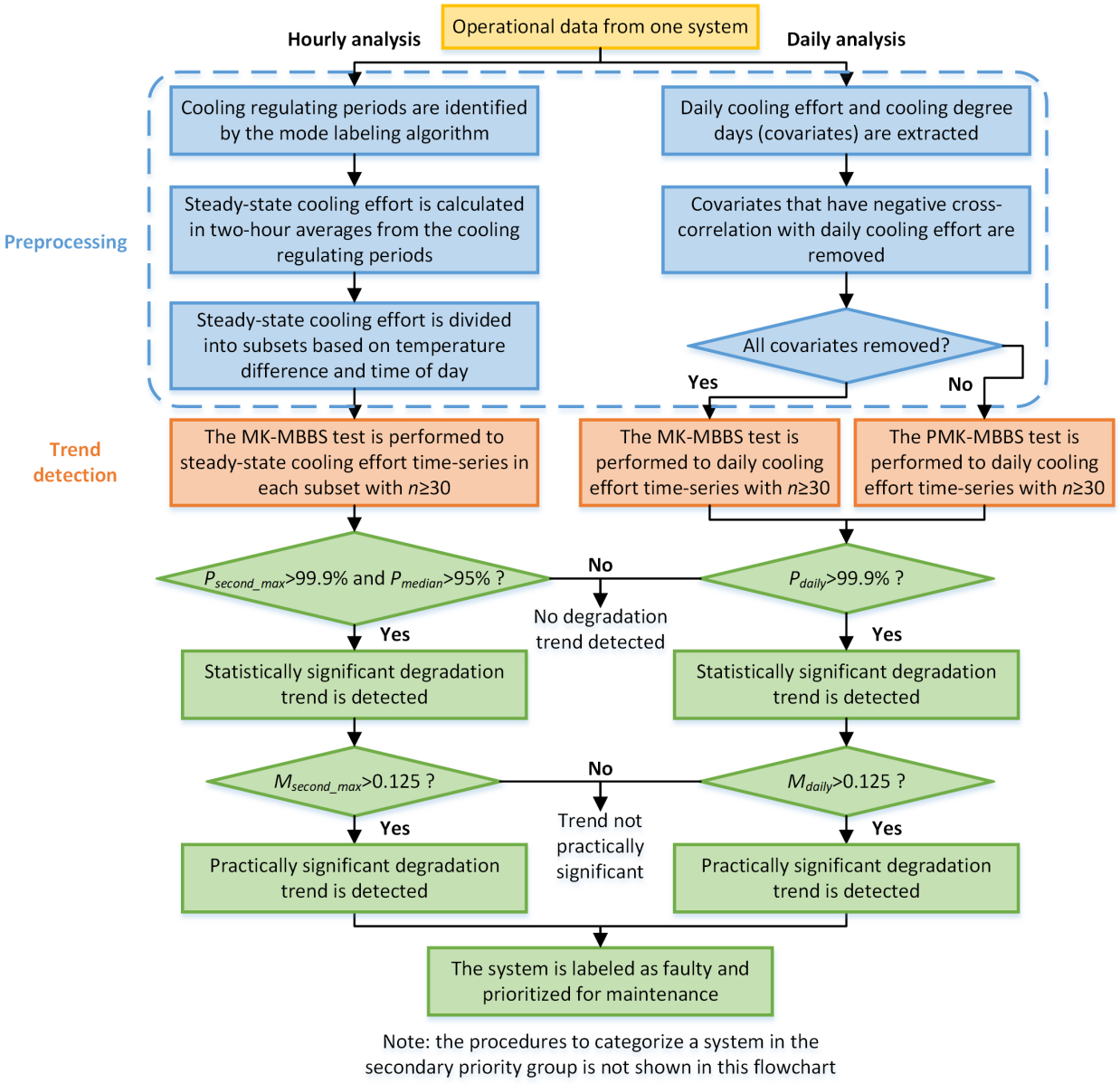


Figure 6.12: The data processing procedures for the historical trend detection.

September 30th in 2019 is used for analysis. Figure 6.13 shows the distribution of percentiles of the Z-statistics (P_{daily}) in the daily analysis for all systems. Most systems behave normally with no particular trends. A system with P_{daily} close to unity indicates an increasing trend of the daily cooling effort, and thus degradation possibly exists. On the contrary, a system with P_{daily} close to zero indicate decreasing trend of the daily cooling effort, which could be caused by mechanical

Table 6.2: Time-series division according to temperature difference (ΔT_{oi}) and time of day.

Subset number		ΔT_{oi} range [°F]					
		$-4 \sim 0$	$0 \sim 4$	$4 \sim 8$	$8 \sim 12$	$12 \sim 16$	$16 \sim 20$
Time of day [hr]	8:00 $\sim$ 20:00	No. 1	No. 2	No. 3	No. 4	No. 5	No. 6
	20:00 $\sim$ 8:00	No. 7	No. 8	No. 9	No. 10	No. 11	No. 12

faults being repaired or decrease of occupancy. The red bar shows the number of systems that has $P_{daily} > 99.9\%$. These 204 systems are very likely to have degradation issues.

Applying the hourly analysis method results in similar distributions of P_{second_max} with 108 systems likely to have degradation issues. The next step in the process is to show how many of these systems have degradation trends that are practically significant.

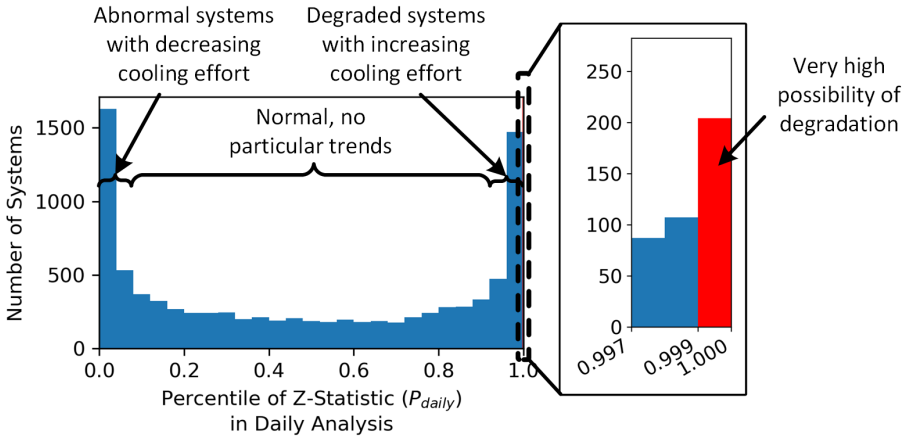


Figure 6.13: This histogram shows the distribution of percentiles of the Z-statistics (P_{daily}) in the daily analysis for 10,000 systems calculated by the historical trend detection.

After applying the criteria for P_{second_max}, P_{median}, and M_{second_max} metrics, a total of 69 systems are labeled with practically significant degradation faults (from among the initial 108 systems). Similarly, for the daily analysis, after applying the criteria for P_{daily} and M_{daily} metrics, 99 systems are labeled with practically significant degradation faults (from among the initial 204 systems). Figure 6.14 shows these faulty systems labeled by the hourly and daily analysis. A total

of 110 systems labeled as faulty had sufficient cooling effort data to be analyzed by both hourly and daily methods, with 30 systems in common. A few systems are only available for the hourly analysis because the thermostats have intermittent connectivity problems almost every day, and therefore daily averaged features cannot be extracted. But in this case, none of these systems were labeled faulty by the hourly analysis. In a similar way, 28 of the labeled systems were analyzed using the daily method because the systems spend little time operating in the cooling regulating mode, and the extracted steady-state feature values are insufficient to perform a statistical test.

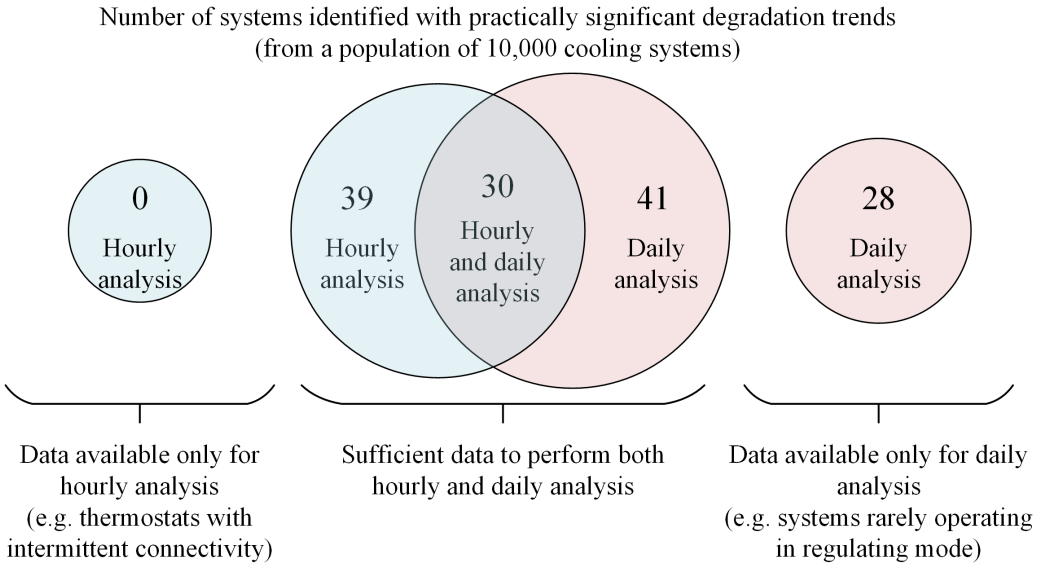


Figure 6.14: The Venn diagram shows the faulty systems labeled by the hourly analysis and the daily analysis.

Figure 6.15 clearly shows an increasing trend in cooling effort in all subsets of the 69 faulty systems identified by the hourly analysis. Remember that the steady-state cooling effort data in each subset is under similar operating conditions (i.e. seasonal weather effects has been removed). The solid line represents the daily average of the plotted data, and the dashed lines represent one standard deviation above and below the daily average. The lowest average cooling effort increment occurs in mid-June, which is around -5% (black dotted line), and after that it gradually in-

creases. The highest average cooling effort increment is 14% (black dotted line), which is reached in September. The increase of steady-state cooling effort data is significant, and therefore the labeled systems are possibly degraded after the summer peak season. The trend of daily cooling effort is not shown here because the seasonal weather effect cannot be removed.

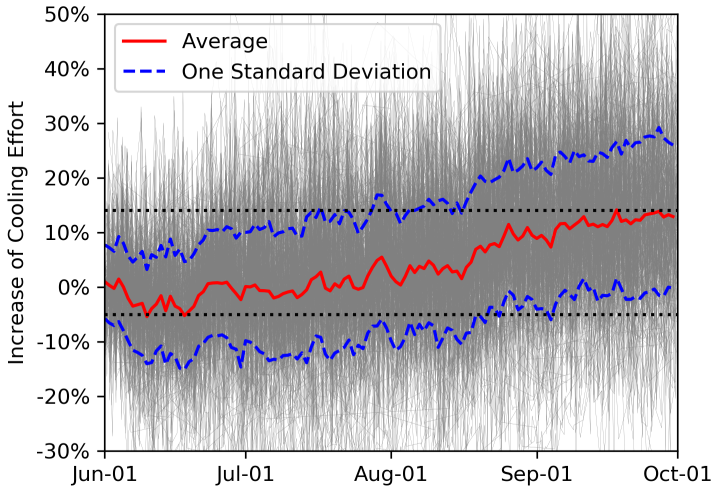


Figure 6.15: The plot shows the increase of steady-state cooling effort since June 1st in all subsets of the 69 faulty systems identified by the hourly analysis. In each subset the data is collected under similar operating conditions (i.e. seasonal weather effects have been removed).

From the detection results shown in Figure 6.14, a total of 28 faulty systems are labeled only by the daily method, because they do not have sufficient data in the regulating mode to use the hourly analysis method. To better understand the limitations of using only one detection method, the applicability of both methods to different system behaviors is investigated.

Figure 6.16 shows the percentage of time operating in the cooling regulating mode, respectively for: (a) the systems labeled faulty by the hourly analysis, (b) the systems in labeled by the daily analysis, and (c) all 10,000 systems. All systems labeled by the hourly analysis spend at least 65% of time in cooling regulating, while the daily analysis is less affected by the operational time spent in the cooling regulating mode and is comparable with the overall population. Additionally, as the percentage of cooling regulating time increases, the hourly analysis is more likely to correctly

identify systems with degradation trends. If a system spends the majority of time in cooling regulating, then the number of steady-state cooling effort data that can be extracted from the cooling regulating mode is greater, which enables the hourly analysis to perform a powerful statistical test to identify a trend. But if a system spends little time in the cooling regulating mode, the extracted data may be insufficient for a statistical hourly analysis, and only the daily analysis is feasible. Note that over 60% of systems spend more than 65 of time operating in cooling regulating, which means the hourly analysis is applicable to the majority of residential systems.

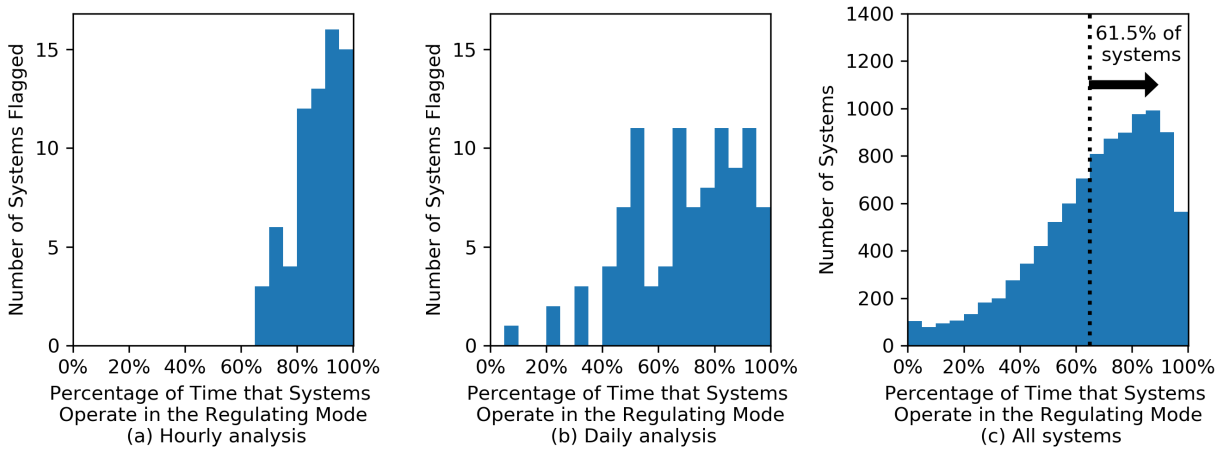


Figure 6.16: The histograms show the percentage of time operating in the cooling regulating mode, respectively for (a) all labeled systems in the hourly analysis, (b) all labeled systems in the daily analysis, and (c) all the approximately 10,000 systems.

Based on the above analysis and observations, the main advantages and disadvantages for each analytical method are summarized in Table 6.3.

Finally, as a case study a system labeled as faulty by both methods is shown in Figure 6.17. The first subplot shows the indoor temperature variation and the cooling regulating periods. The system essentially operates exclusively in regulating mode, and thus the slow degradation is likely imperceptible for occupants. The second subplot shows the change of steady-state cooling effort data over time in four subsets (i.e. four different operational conditions). The cooling effort data

Table 6.3: The advantages and disadvantages for the hourly and daily analysis methods.

	Hourly analysis	Daily analysis
Advantages	• The method has a high possibility to correctly detect a degradation trend for a system with relatively constant cooling setpoints, due to frequent regulating periods and sufficient steady-state cooling data, which can be used to perform a powerful Mann-Kendall trend test.	• The method is suitable to all operating conditions, independent of whether the system spends the majority of time in a regulating mode or not.
Disadvantages	• The method is not appropriate for a system with frequent setpoint changes or setpoint tracking failures, because the system will operate primarily in transient, with few regulating periods that are needed for the trend test. • The hourly averaged cooling effort data may have higher signal-to-noise ratio compared to the daily averaged data, which lowers the possibility for the Mann-Kendall test to detect a trend.	• The magnitude of trend of the daily cooling effort may not precisely characterize the severity of system degradation, because the estimation by the Sen's slope does not consider the covariate effects. • The method struggles to identify trends during shoulder seasons (i.e. spring and autumn) with quickly changing outdoor temperatures, because in this case the covariates in the partial Mann-Kendall test also exhibit trends (see Figure 6.9 in Section 6.4.4 as reference).

continuously increases in each subset over these four months, and overall exhibits large increase in magnitude. Additionally, from the third and fourth subplots, the daily cooling effort also exhibits an increasing trend while the cooling degree days (CDD) do not have a noticeable trend. According to the test statistics of the hourly and daily analysis, $P_{second\_max} = 1$, $P_{median} = 1$, $P_{daily} = 1$, $M_{second\_max} = 0.39$, and $M_{daily} = 0.43$.

This particular system also exemplifies the value of the proposed detection methods. By the end of the summer, both methods clearly indicated a degradation trend, indicating a significant

system fault. However, a problem was not reported by the occupants for an additional four months (January 28th), whereupon technicians reported a refrigerant leakage due to pinhole or puncture in tubes, and the system was repaired.

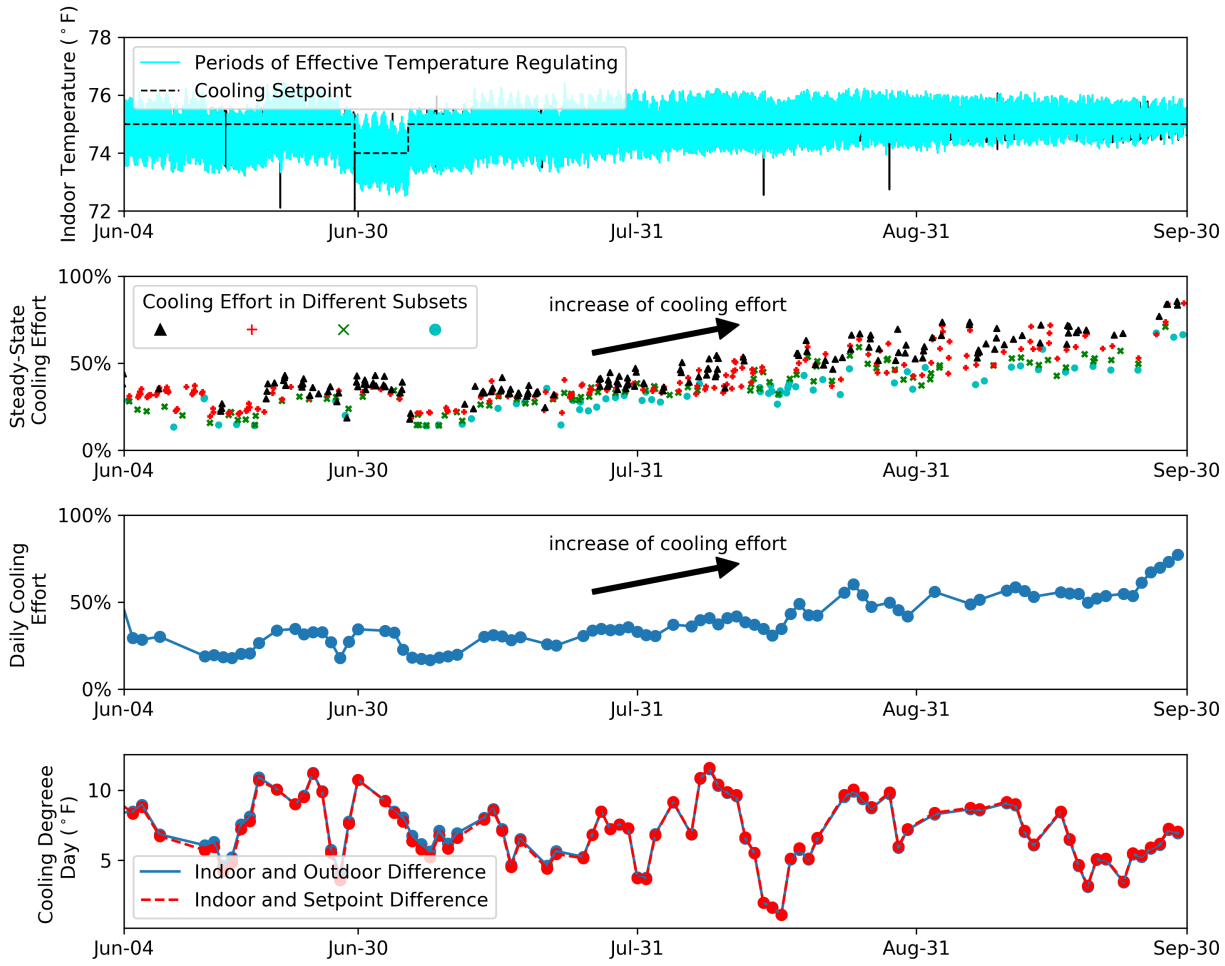


Figure 6.17: A system with refrigerant leakage identified by the historical trend detection method.

6.5.2 Real-Time Condition Monitoring

The above historical trend detection is applied to the data collected for an entire summer for a single system, and the corresponding analysis happens *after* the peak season has concluded. However, during the peak season real-time trend detection methods are particularly useful for

actively monitoring the behavioral change of a system and notifying technicians before complete failure. Using real-time condition monitoring, maintenance could be scheduled and prioritized for severely faulty systems. Service companies could allocate repair resources in advance and avoid technician overtime. Additionally, occupant comfort issues caused by complete failure could be prevented.

In general, the data processing procedures for real-time condition monitoring is similar to historical trend detection. A few minor modifications in the procedures are required, as described below:

- Real-time trend detection is performed every day with a fixed-length moving window of daily data.
- To reduce the high computational cost of daily analysis, the number of bootstrap samples (N) is reduced to 5,000. Appendix A.3 and A.4 show that this modified case of 5,000 samples and a 0.2% significance level results in a similar level of the Type II error and very little increase in the Type I error, when compared to using 100,000 samples and a 0.1% significance level.
- The detecting thresholds are changed to $P_{second\_max} > 99.8\%$ for the hourly analysis and $P_{daily} > 99.8\%$ for the daily analysis. Additionally, an alert will be reported when $P_{second\_max} > 95\%$ or $P_{daily} > 95\%$.
- The detecting thresholds of the M_{second_max} and M_{daily} are also reduced due to the shortened testing period.

Note that because the testing period in real-time trend detection is shorter, the magnitude of change and signal-to-noise ratio will be lower for the same trend. Therefore, detecting a trend in the real time is more difficult.

The following study performs real-time condition monitoring for the 110 labeled faulty systems with available data for both hourly and daily methods in Figure 6.14. The data from May to September 2019 is used, and the detection results are shown in Table 6.4. In general, the daily

Table 6.4: Detection results of 110 faulty systems using real-time condition monitoring.

Confidence level	Number of faulty systems that identified earlier by the hourly analysis	Median of the earlier time [days]	Number of faulty systems that identified earlier by the daily analysis	Median of the earlier time [days]
90%	27	12	67	7
95%	28	9	66	8
99%	28	10	45	9
99.8%	24	9	36	16

analysis identifies two thirds of systems earlier than the hourly analysis with a median of around 10 days. However, for the remaining one third systems the hourly analysis identifies a fault earlier with a median of around 10 days as well. Thus, depending on the operational behavior of a system, both methods are anticipated to be useful in practice, and can be implemented in parallel.

To illustrate the effectiveness of the real-time condition monitoring, three case studies are presented as below. All three cases are verified with active refrigerant leakage problems and are repaired on a known date. The hourly analysis detects a degradation trend of the first case earlier, while the daily analysis is more effective in the second and third cases. In these cases, real-time trend detection is performed with a two-month moving window. This window size provides sufficient data for analysis, while detecting degradation early when the fault occurs over a short period. Also, this window size can effectively avoid false positive detection results when a faulty system is recently repaired. Conversely, when the degradation develops gradually for months and the slope of trend is low, a bigger window size (e.g. a three-month-long moving window) may be preferable.

The first case study shows a typical faulty system that suffered from refrigerant leakage. The hourly analysis detects the degradation trend a few days earlier than the daily analysis because the system operates almost exclusively effective regulating the indoor temperature. Figure 6.18 shows the system operation and the detection results of the hourly analysis. Complete failure occurred on September 5th, but was not fixed until September 10th. From the first subplot, the system regulated

the indoor temperature to the desired cooling setpoint with minimal setpoint error up to the failure point. Occupants were unlikely to notice the anomalous system operation. The second subplot shows the steady-state cooling effort data through time. Although there is some noise in the data, the continual increase of the steady-state cooling effort in every subset of operational conditions is visible. The increase is clearly seen in the real-time testing results of the hourly analysis shown in the third subplot, where each point shows the confidence level (i.e. P_{second_max}) of detecting capacity degradation issues using a two-month moving window starting from June 1st. An early alert is observed on August 18th when the confidence level has arisen to more than 95%. After 10 days on August 28th, the real-time testing shows $P_{second\_max} > 99.8\%$ and $P_{median} > 95\%$ (P_{median} is not shown in the plot), which means there is over 99.8% confidence that degradation exists. The fourth subplot shows the magnitude of increase in steady-state cooling effort (M_{second_max}) over a two-month period. The increase is greater than 0.1 after the early alarm and rises to 0.4 at the time of system failure. Before the failure on September 5th, real-time condition monitoring could have notified service companies of the system faults with an alert two weeks earlier and a confirmed fault one week earlier, potentially avoiding the system failure and the 5 days without air conditioning.

When a system has frequent setpoint changes every day, the daily analysis is more effective than the hourly analysis. As a second case study, Figure 6.19 shows a residential system with continuous refrigerant leakage identified by the daily analysis. Repairing technicians have already diagnosed the leakage problem as early as in April, but because the leakage point was difficult to located, this system was only recharged. Then, due to active leakage, this system was recharged again on August 17th and finally repaired by re-brazing the duct joint a few days later. In Figure 6.19, the first subplot shows the variation of the indoor temperature and setpoint changes. The smart thermostat has setpoint schedules and the indoor temperature is effectively controlled around the desired setpoints. Thus, similar to the former case study, occupants themselves are unlikely to notice the degradation issue in the absence of repairing technicians. The second subplot shows the changes of daily cooling effort, with two covariates plotted in the third subplot. Before a fault is fixed, the daily cooling effort has an increasing trend in general, but because the covariates

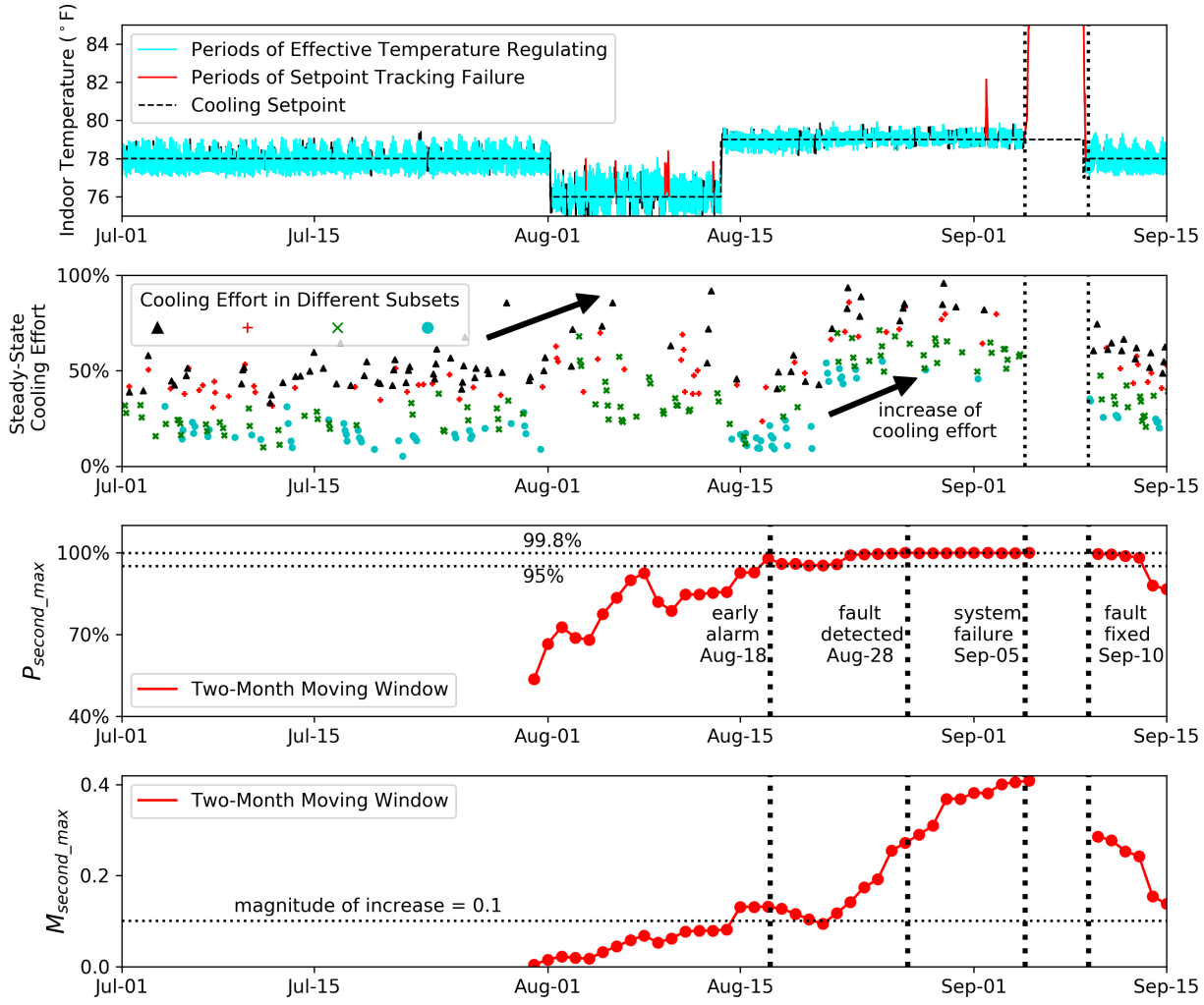


Figure 6.18: Case one: real-time trend detection results of a system with refrigerant leakage by the hourly analysis.

are also increasing, visual identification of the fault would be prone to errors. The confidence levels (i.e. P_{daily}) calculated from the real-time monitoring of the daily analysis are shown in the fourth subplot. With a two-month moving window starting in May, the method shows an over 90% confidence of degradation all the time. An early alarm is triggered on June 13th when the confidence level rises above 95%, and fault is detected on July 12th when the confidence level rises above 99.8%. The fifth subplot shows the magnitude of increase in daily cooling effort (M_{daily})

over a two-month period. Before the system fault is fixed, M_{daily} varies between 0.15 and 0.4, which can be considered significant increases.

After the system is recharged again on August 17th and subsequently repaired, the daily cooling effort declines significantly while the cooling degree days are relatively constant, which indicates the system performance is improved. Additionally, P_{daily} and M_{daily} drop down quickly, so false alarms are avoided.

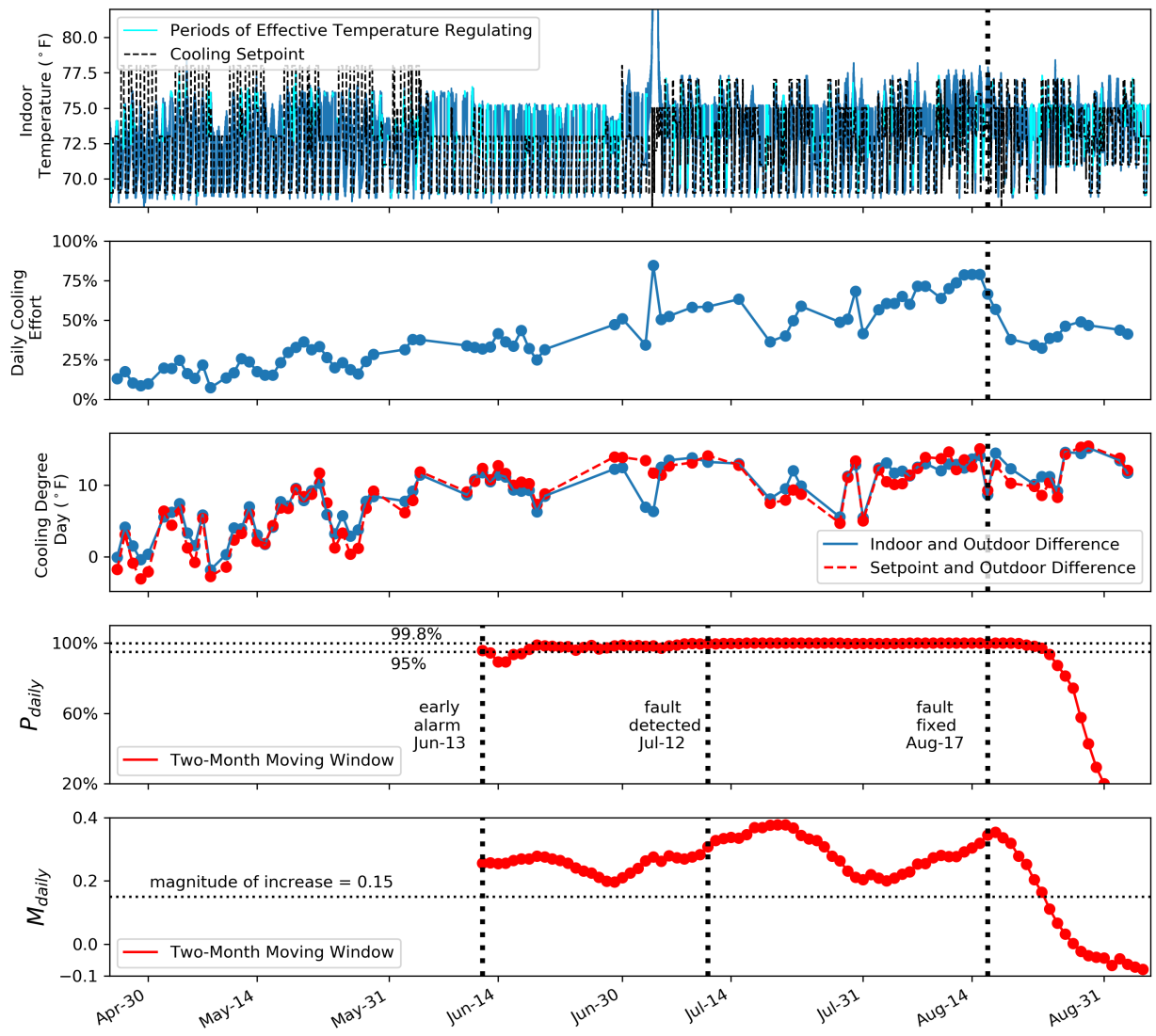


Figure 6.19: Case two: real-time trend detection results of a system with refrigerant leakage by the daily analysis.

Another scenario when the daily analysis is more effective is when a system has severe setpoint tracking failure problems and is not able to effectively regulate indoor temperature. Figure 6.20 shows a third case study with a refrigerant leakage problem reported on July 22nd. From the first subplot, the system is in effectively regulates temperature in May, but starting from June there are long periods where the system fails to track the desired setpoint. From the second and third subplot, the daily cooling effort increases gradually from 20% to nearly 100% before a fault is reported and repaired on July 22nd, while the covariates increase slightly. The fourth subplot shows that the daily analysis with a two-month moving window can detect the fault as early as June 26th, almost one month before a fault is actually reported. Finally, the fifth subplot shows the magnitude of increase in daily cooling effort (M_{daily}) over a two-month period. M_{daily} approaches 0.5 when the fault is detected by the daily method, which is extremely high.

After repairs, the daily cooling effort decreases to about 50%, and the system can effectively regulate the indoor temperature around the desired cooling setpoints. P_{daily} almost approaches zero, which indicates a decreasing trend of the daily cooling effort is detected. Additionally, M_{daily} plunges to -0.5, which means the cooling capacity of the system after repair is nearly the same as before the fault occurs.

6.6 Conclusions

This paper proposes hourly and daily analysis methods to detect degradation trends of a large number of residential HVAC systems using only smart thermostat data. The two methods both extract an essential fault indicator, termed the cooling effort, and identify degradation issues by detecting if the cooling effort has an increasing trend. A critical element in the trend detection is the original and partial Mann-Kendall statistical tests with moving block bootstraps. The performance of the two statistical tests are evaluated by Monte Carlo simulations, and results show the statistical tests have low false positive/negative rate.

In practice, two applications using the hourly and daily analysis on smart thermostat data are discussed: 1) historical trend detection, and 2) real-time condition monitoring. The historical trend detection evaluates the performance of a system at the end of peak cooling season, and if

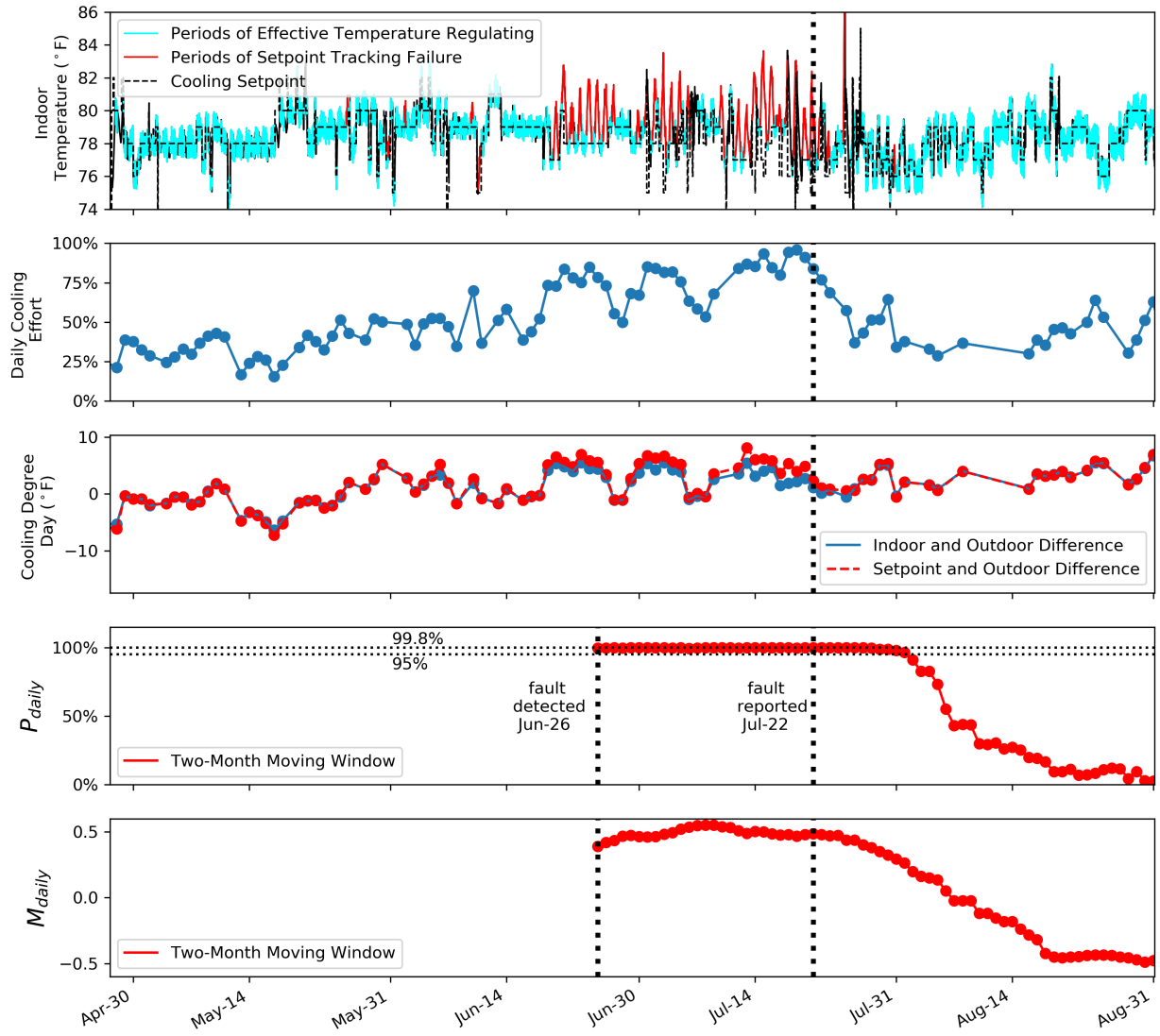


Figure 6.20: Case three: real-time trend detection results of a system with refrigerant leakage by the daily analysis.

degradation is identified, the home owner can schedule off-season maintenance. The real-time condition monitoring is applied to actively monitor system behavioral changes, and if urgent or imminent cooling capacity failures are identified, maintenance can be scheduled immediately to avoid system failure. Finally, several distinct case studies are presented to verify the effectiveness of the proposed hourly and daily analysis. The residential systems featured in these cases studies are actual, occupied homes with verified refrigerant leaks.

Additionally, the advantages and disadvantages for the hourly analysis and the daily analysis are demonstrated using the case studies. The hourly analysis is more effective to detect a trend for a system spending the majority of time in the cooling regulating mode, but is not applicable for the opposite situation. The daily analysis is suitable for almost all systems, but it is hard to quantify the magnitude of trend, and may not be as effective when used in shoulder seasons.

7. PAPER E: CHARACTERIZING THE PERFORMANCE OF RESIDENTIAL HVAC SYSTEMS USING SMART THERMOSTAT DATA[1]

7.1 Synopsis

Faulty residential Heating, Ventilation, and Air Conditioning (HVAC) systems waste billions of dollars of energy usage in the U.S. each year, meanwhile heavily affecting occupant comfort and causing complete operational failures. Fault detection and diagnosis (FDD) methods assist technicians in discovering and locating the faults as early as possible, thereby optimizing the system efficiency through predictive maintenance and benefiting the home owners as well. However, due to uneconomical initial cost of additional sensor installation, traditional FDD methods using sensor data are hard to apply in the residential sector. To solve this issue, this paper proposes an FDD method that only uses smart thermostats as the data source to detect inadequate system capacity issues from a large number of residential HVAC systems. The proposed method analyzes the pseudo steady-state behavior and compares a critical feature termed the weighted average difference of cooling effort ($\Delta\bar{E}_c$) both between systems and within each system. Three applications of the proposed method are detailed, namely: identification of system inadequate capacity, quantification of active system capacity degradation, and verification of system repair after maintenance. Within each application, real faulty systems are illustrated to demonstrate the usefulness of this method.

7.2 Introduction

Residential Heating, Ventilation, and Air Conditioning (HVAC) equipment has gradually become a requisite for people to install in their home. In 2009, 87% of households in the U.S. own air conditioning systems, which is a 31% increase compared to the year of 1978, and 99% of households own space heating devices [8]. With the increasing need of comfort indoor conditions, residential HVAC systems consume 214 TWh (27.28 billion dollars) of energy on air conditioning and 1156 TWh (60.74 billion dollars) on space heating each year [1]. However, due to operational

[1]F. Guo and B.P. Rasmussen, "Characterizing the Performance of Residential HVAC Systems Using Smart Thermostat Data", *ASHRAE Transactions*, **In Preparation**.

faults that may occur, residential systems on average function at least 17% below their rated efficiencies [9], causing extra and unnecessary energy expenditures. For example, [10] shows that 20.7 TWh per year (2.5 billion dollars) can be saved by fixing installation issues including indoor airflow rate and refrigerant charge.

Residents benefit from fault-free HVAC systems with reduced energy bills, improved thermal comfort, and avoidable complete failures. Fault detection and diagnosis (FDD) plays an important role for systems to achieve the optimal operating efficiency (free of fault), in that FDD ensures proper commissioning, optimized system control, effective predictive and preventive maintenance. [16] reviewed state-of-the-art FDD methods for split and rooftop HVAC systems, and described the general processes: feature selection, steady-state detection, fault-free modeling, fault detection, and fault diagnosis. Soft mechanical faults on the air-side and refrigerant-side are greater interests for researchers than total failures, because they do not cause immediate system breakdowns and can be unnoticed for an extended time. Though relatively few researches are focused on residential split systems, the FDD methods developed for small commercial rooftop systems are still applicable and heuristic to residential systems. [3] reviewed FDD methods published after 2004 in the commercial building sector. The authors classified FDD methods according to whether a priori knowledge is used in the detection algorithms, and found the process history-based methods that do not rely on a priori are the most popular. [14] reviewed the artificial intelligence methods used for building energy systems including split and rooftop air conditioners, and also found that process history-based methods are abundant. The authors categorized process history-based methods into three groups: classification-based, unsupervised learning-based, and regression-based.

The process history-based FDD methods usually needs large amounts of data collected from multiple sensors to train a model. However, since the marginal profit of repairing a residential system is comparatively small, the uneconomic initial cost of installing additional sensors in the vapor compression cycle becomes a major obstacle to apply these developed FDD methods widely to residential systems. Thus, researchers have tried to perform FDD with limited sensor information to lower the cost. For example, [18] developed an FDD method and obtained key operating parame-

ters such as airflow rate, cooling capacity, system efficiency, and refrigerant mass flow by only five temperature sensors, one humidity sensor and one current sensor all installed on the indoor unit. [19] detected the refrigerant leakage problem only using five temperature sensors installed on the air side and the heat exchangers.

An alternative source of data for FDD of the residential systems is smart thermostat data. As the development of internet of things and smart homes, smart thermostats are gradually adopted by home owners for the convenience of remote control, comfort monitoring, and schedules learning. Meanwhile, the smart thermostats can provide huge amount of important system operational data used for FDD methods without paying extra cost for installing more sensors. Also, under the permission of customers, smart thermostats can continuously upload real-time data to a database through Wi-Fi connection, allowing FDD methods to be performed on a remote platform. What's more, when numerous smart thermostats are connected to a remote database and the operational data is collected continuously since system installation, FDD is practically feasible either through comparing the operational features between large numbers of systems (multi-system feature comparison), or through identifying the change of operational features for each system in a long period of time (single-system change detection).

However, developing FDD methods using smart thermostat data has several challenges as well. First of all, only a few types of operational information are available, which mainly includes indoor/outdoor temperature, temperature setpoint, and cooling/heating operational modes. Thus, the operational features that can be extracted for analysis are limited. Also, true labels of operational data (i.e. fault-free or faulty in specific types) are often unknown or sometimes inaccurate, because usually the true labels come from warranty claim reports created by repairing technicians. These reports can only record the claim date and possible failure reasons for a system, but the exact faulty and fault-free periods are unknown, and the recorded failure reason may not be the real ones causing the system to fail. Additionally, warranty claims are usually filed when complete failure occurs, but the operational data of incipient mechanical faults is more useful to develop FDD methods because they are difficult to detect.

Considering the above characteristics of the smart thermostat data, this paper proposes a statistics-based FDD approach termed the characteristic curve method, which uses a metric termed the weighted average of cooling effort ($\Delta\bar{E}_c$) to quantify the system performance and monitor the behavioral change. The remainder of this article will first detail the preprocessing and main procedures for the characteristic curve method, and then introduce the applications of this method which include detecting inadequate system capacity problems and monitoring the system behavioral change over time. Case studies of real faulty systems are shown for each application to verify the effectiveness of the proposed characteristic curve method.

7.3 The Characteristic Curve Method

Figure 7.1 shows the processing procedures of the characteristic curve method and its applications. First of all, the raw thermostat data from a large number of residential HVAC systems is preprocessed and the cooling effort (E_c) for a system to remove cooling load is extracted as a critical feature. Then, using the data features from all systems, the expected cooling effort at each ambient condition for an operating system is calculated, termed the characteristic curve. From the characteristic curve, the extra amount of time that a system needs to operate to meet the cooling load at the given ambient condition is calculated. This extra time is termed the weighted average of cooling effort ($\Delta\bar{E}_c$), and has three applications: 1) systems spending anomalous large amount of time in operation can be identified by sorting $\Delta\bar{E}_c$ among large quantities of systems; 2) with the help of single-system change detection, systems having capacity degradation faults (e.g. active refrigerant leakage) can be identified with significant increase of $\Delta\bar{E}_c$ over time; 3) faulty systems being fully repaired can be verified by significant decrease of $\Delta\bar{E}_c$ after maintenance.

7.3.1 Preprocessing and Feature Extraction

The operation of residential air conditioning systems can be categorized into three behavioral modes, which are constantly regulating around a setpoint, actively tracking towards a new setpoint, and freely responding to the change of ambient conditions [6]. This paper investigates the regulating mode, during which a cycling system has already reached the setpoint and cycles on and

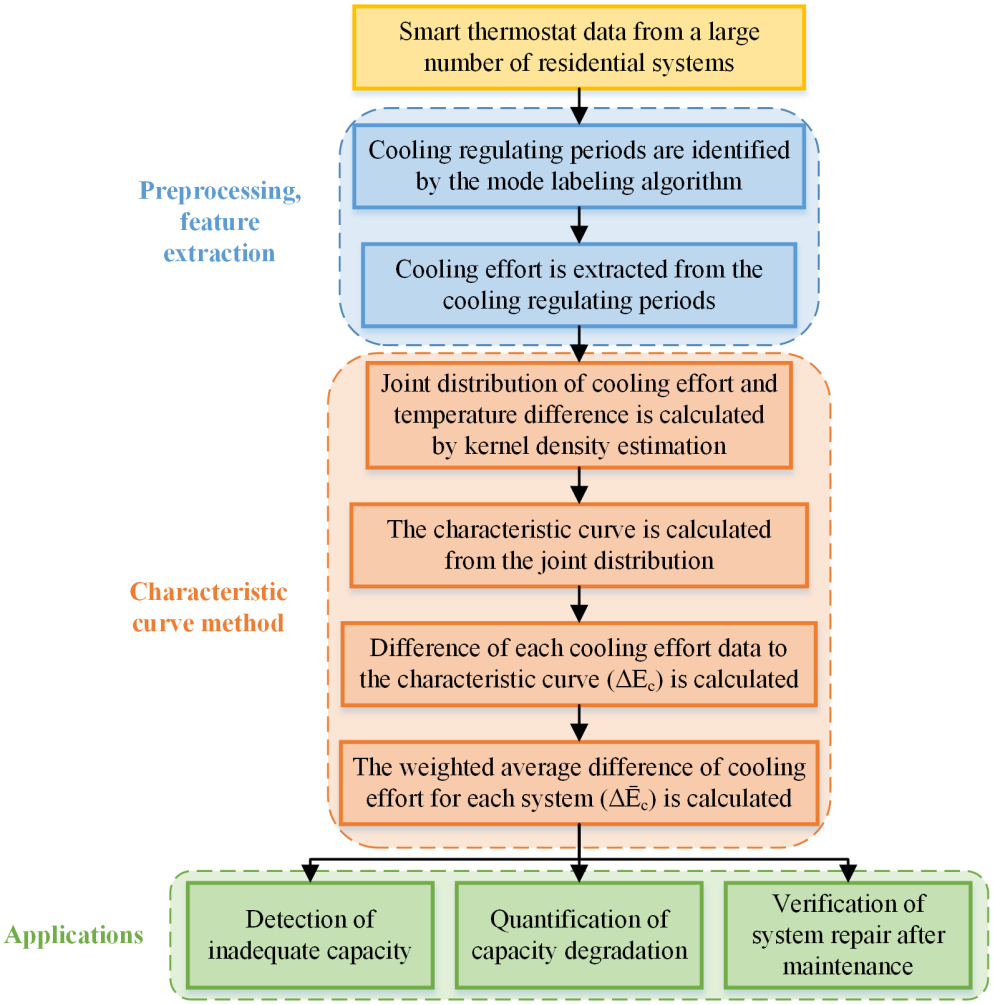


Figure 7.1: The flowchart shows the processing procedures and applications of the proposed characteristic curve method.

off to maintain the setpoint. It is often referred to as pseudo steady-state conditions because the air-conditioned home has relatively constant indoor temperature. A system often has both cooling regulating mode and heating regulating mode based on cooling/heating is operating, and this paper only focuses on the cooling regulating mode. Compared to the tracking mode, the behavior in regulating mode is much similar among a large group of systems, making it easier for the multi-system comparison.

From the cooling regulating mode, a useful feature termed *cooling effort* is extracted. For multi-speed cooling systems, the cooling effort is defined as the product of duty factor and load factor in

a specific period, where duty factor is the fraction of time that a system runs and load factor is the fraction of load when the system runs. For single-speed cooling systems, the load factor equals one during operation and hence the cooling effort equals duty factor. The cooling effort describes how hard a system works at given ambient conditions, and systems suffered from inadequate capacity usually has abnormally high cooling effort comparing to other systems. Equation 7.1 shows the energy balance equation in a cooling regulating period:

$$C \cdot E_c = UA \cdot \Delta T_{oi} + \dot{Q}_{sol} + \dot{Q}_{int} \tag{7.1}$$

where C is the total cooling capacity [W], E_c is the cooling effort (%), U is the overall heat transfer coefficient for the residential home [W/m^2·°F], A is the overall heat transfer area of building walls and roof [m^2], ΔT_{oi} is the temperature difference [°F] defined as outdoor temperature minus indoor temperature, $\dot{Q}_{sol}$ is the solar irradiance heat gains [W], and $\dot{Q}_{int}$ is the internal heat gains [W]. From Equation 7.1, the *temperature difference* must be chosen as another important feature when the cooling effort is compared among a large group of systems since they are highly correlated. Also, temperature difference affects the coefficient of performance (COP) of a system, thereby affecting the cooling capacity. Apart from that, high cooling effort seems also possibly resulting from low C, high UA, or high $\dot{Q}_{sol}$ and $\dot{Q}_{int}$. However, the proposed algorithm focuses on behavior with abnormally high cooling effort, and these factors could hardly be the cause for the following reasons: (1) the system capacity (C) is selected based on the home thermal property, and usually dealers are unlikely to install a much undersized system for customers; (2) UA is determined by the home thermal resistance level of the construction materials and the area of building walls and roof, but most residential homes are built with similar materials and a larger home is also fitted with a larger system; (3) most residential homes have normal heat gains mainly from solar irradiance, cooking, shortly opening windows and doors, and additionally, a few time periods having much higher heat gains (e.g. holding parties, leaving windows open for a whole day) does not affect the

analysis too much because the proposed FDD method evaluates system operating behavior on the scale of months.

Based on the above analysis, in the main algorithm of the characteristic curve method, the expected cooling effort at every temperature difference value is estimated. The line connecting all expected cooling effort values is termed the *characteristic curve*. The cooling effort data of each system is compared with the characteristic curve to evaluate its relative value above or below the curve at each temperature difference.

7.3.2 Fault Detection Algorithm

The characteristic curve method obtains the expected cooling effort at every temperature difference by estimating the joint distribution of cooling effort and temperature difference. In order to clarify all the calculating steps and exemplify the fault detection results, data from approximately 10,000 residential single-speed split cooling systems in Florida between May and September in 2019 is processed. The characteristic curve method consists of four procedures as follows:

First, kernel density estimation (KDE) is performed to estimate the probability density function of the joint distribution of cooling effort and temperature difference using data from a large number of systems. The KDE treats each data point as a Gaussian kernel, and sums up all kernels to obtain the full probability density function. In this case, since KDE is calculated by data from thousands of systems, which is very computationally expensive, one better idea is to randomly resample a few data points from every system and calculate KDE with this smaller dataset. Also, to minimize biased calculation results that may occur, KDE is calculated multiple times using different resampled datasets, and the final density field is the average of all calculated density fields. Figure 7.2 shows the probability density field of the joint distribution of cooling effort and temperature difference generated by 10,000 systems in Florida, where darker red color indicates more popular behavior of the air conditioners. Generally, temperature difference ranges from -10°F to 20°F in Florida's residential homes, and cooling effort ranges from 0% to 80%.

Second, from the estimated density field, the midpoint of accumulated density at every given temperature difference is calculated. In another word, the sum of density above the midpoint

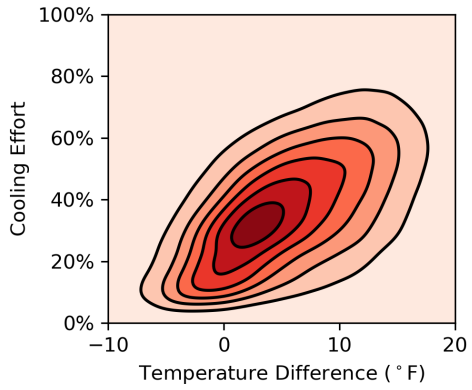


Figure 7.2: An example of Kernel density estimation (KDE) between cooling effort and temperature difference using approximately 10,000 systems in Florida. Darker red color indicates more popular behavior.

equals to the sum of density below the midpoint. The midpoints at all slices of temperature difference connect to a curve termed the characteristic curve, which can be viewed as the average performance of a single-speed residential cooling system in Florida, or the expected cooling effort of these systems at every given temperature difference. Figure 7.3 shows the characteristic curve in black. When the temperature difference is greater than -4°F, the curve is monotonically increasing (shown by the solid line), in consistent with the fact that a cooling system works harder in an ambient environment with higher temperature difference. In contrast, when the temperature difference is lower than -4°F the curve is flat (shown by the dashed line), because in this region a residential home is passively cooled by the outside and the internal heat gain usually dominates the cooling load. Therefore, only data with temperature difference between -4°F and 20°F is considered as representative of system performance in an ordinary residential home and will be used in the following steps.

Third, from the characteristic curve, the difference of each cooling effort data (ΔE_c) to the curve is calculated. Figure 7.4 illustrates the difference of every cooling effort data (in blue) to the characteristic curve (in black) for one system, where positive differences (i.e. data above the curve) are marked by red and negative differences (i.e. data below the curve) are marked by green.

A positive difference means that during the period the cooling system performs worse than the average level, and a negative difference means that during the period the cooling system performs better than the average level.

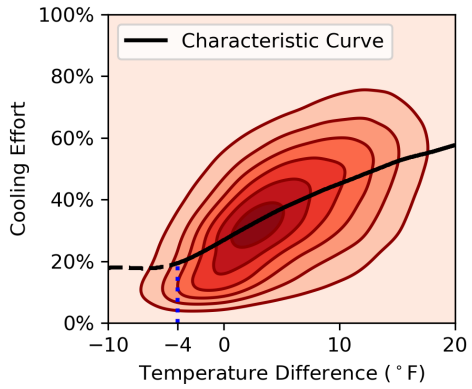


Figure 7.3: The calculated characteristic curve from the KDE. The curve above -4°F temperature difference is monotonically increasing, while below -4°F temperature difference is flat.

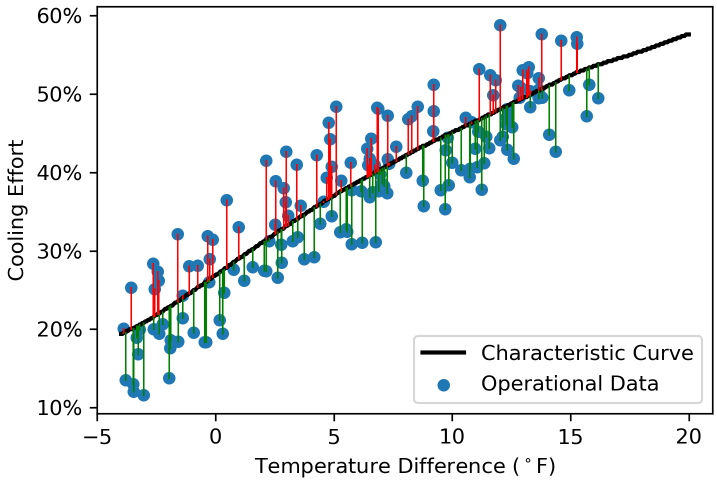


Figure 7.4: Illustration of differences of cooling effort data to the characteristic curve for a system. The data is shown in blue dots. Positive differences (i.e. data above the curve) are marked by red and negative differences (i.e. data below the curve) are marked by green.

Fourth, to obtain a single metric for comparison between systems, this method calculates the weighted average of all differences of cooling effort for each system, termed $\Delta\bar{E}_c$, using the probability density function of temperature difference as weight. Because the difference of cooling effort (ΔE_c) for a system still slightly depends on the temperature difference, by always giving higher weights to the majority of ΔE_c values in the region of moderate temperature difference, $\Delta\bar{E}_c$ is considered not strongly affected by the seasonal temperature difference changes. Estimated by one-dimensional KDE, Figure 7.5 shows the probability density function of temperature difference for the Florida systems. Similar to the two-dimensional KDE shown in Figure 7.2, multiple times of resampling is also used to generate this one-dimensional KDE.

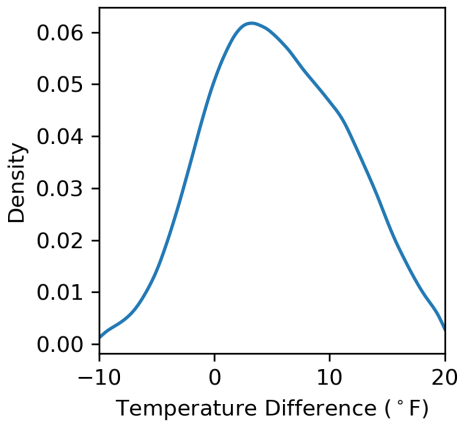


Figure 7.5: An example of kernel density estimation (KDE) of temperature difference for all approximately 10,000 systems in Florida.

The weighted average difference of cooling effort ($\Delta\bar{E}_c$) can be considered as the extra time that a system needs to operate in the cooling regulating mode compared to an imaginary system with average performance, which is also understandable for customers. For example, a system with $\Delta\bar{E}_c$ equals 0.1 means on average it spends additional 0.1 hour operating in full capacity in every hour of the cooling regulating mode.

Note that the characteristic curve method is computationally efficient after the initial analysis. This is because the probability density function of variables and the characteristic curve generated in the initial analysis are applicable to other single-speed split systems in the same climate region. For example, when smart thermostat data of another group of systems in Florida is extracted from the database, researchers can directly use the available characteristic curve to calculate ΔE_c, and then use the available probability density function of temperature difference to calculate $\Delta \bar{E}_c$. The characteristic curve and distributions require to be updated only when systems from another climate region are analyzed or a lot of new systems with higher efficiency are installed in the region as time passes.

7.4 Applications

In the characteristic curve method, after the weighted average difference of cooling effort ($\Delta \bar{E}_c$) is calculated for each system within a period, the values are compared between systems (i.e. multi-system feature comparison) and within each system (i.e. single-system change detection). This section shows three applications of the characteristic curve method for fault detection, preliminary diagnosis, and fault prioritization. The applications include detection of system inadequate capacity, quantification of system degradation trends, and verification of system repair after maintenance. Within each application, case studies of a few real faulty systems are provided to validate the effectiveness of the characteristic curve method. Note that the case studies are all related to active refrigerant leakage, because this type of fault will finally cause system breakdown and thus is easily noticed by home occupants. Some causes of the inadequate capacity issue such as commissioning faults and airflow restriction are hard to verify without field testing.

7.4.1 Detection of Inadequate Capacity

The first application of the characteristic curve method is to detect inadequate system capacity issues. Figure 7.6 shows a histogram of the $\Delta \bar{E}_c$ calculated by the characteristic curve method for all the selected approximately 10,000 systems in Florida. Typically, $\Delta \bar{E}_c$ ranges from -0.4 to 0.4, and the median is around zero. In Figure 7.7, the number axis shows the indications of different

ranges of $\Delta\bar{E}_c$. From the definition, $\Delta\bar{E}_c$ can only range from -1 to 1, and in reality ±1 can never be reached.

- $\Delta\bar{E}_c$'s much lower than zero indicate oversized systems, because they spend less effort than most of other systems to meet the cooling load.
- $\Delta\bar{E}_c$'s close to zero indicate either fault-free and properly sized systems, or oversized but degraded systems. The majority of systems have average performance and belong to the former category, though in some states like Texas, plenty of residential systems tend to be slightly oversized in model selection. The latter category of faulty systems is difficult to detect by using smart thermostat data, because they do not show any abnormal operational behavior.
- $\Delta\bar{E}_c$'s much higher than zero indicate faulty systems with inadequate capacity, which is often caused by degradation, undersized equipment, commissioning faults, or even unusual applications (e.g. large commercial buildings having considerable cooling load, server rooms generating enormous heat). Systems in this category have significantly higher cooling effort than other systems, and in practice are identified by $\Delta\bar{E}_c$ above a pre-defined detection threshold.

From the distribution of $\Delta\bar{E}_c$ in Figure 7.6, systems with abnormally high $\Delta\bar{E}_c$ are labeled as faulty. Because there is not an exact boundary between faulty and fault-free systems, the detection threshold is often determined by a percentage of systems. For example, this study set the detection threshold to be 1%, which means systems with $\Delta\bar{E}_c$ in the top 1% among all systems are considered as faulty. The faulty systems can be automatically prioritized by the $\Delta\bar{E}_c$ metric, such that systems with higher values of $\Delta\bar{E}_c$ are examined and repaired first.

In comparison with a previously proposed univariate analysis, which also uses cooling effort as a feature and performs comparison between a large group of systems to identify faulty systems, the characteristic curve method has an improvement in matching the detection threshold with the actual percentage of systems labeled as faulty. The Venn diagrams in Figure 7.8 show the number

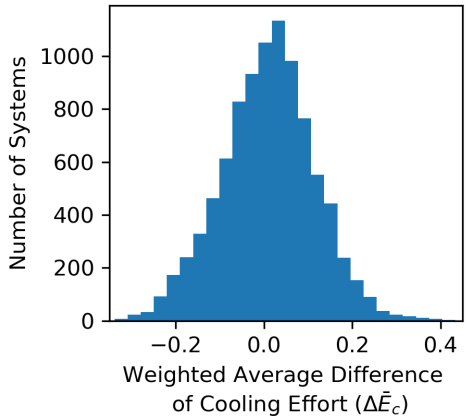


Figure 7.6: The histogram shows the distribution of the weighted average difference of cooling effort ($\Delta\bar{E}_c$) for the selected approximately 10,000 systems in Florida.

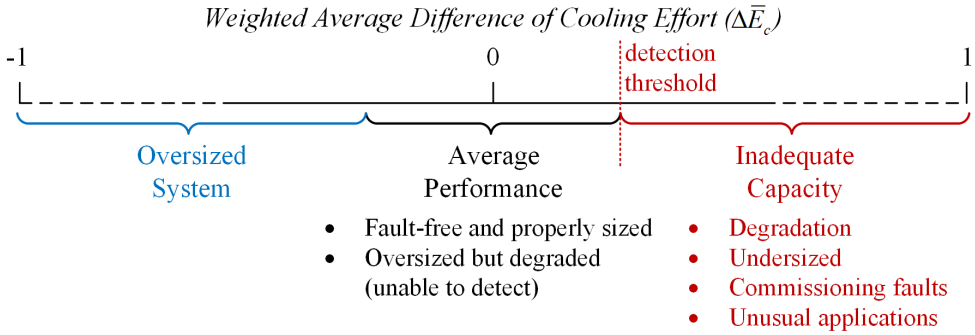


Figure 7.7: The number axis shows the indication of weighted average difference of cooling effort ($\Delta\bar{E}_c$) for a system.

of systems that labeled as faulty by the univariate analysis and the characteristic curve method when the detection threshold is set to different percentages. Approximately 10,000 Florida systems with data from June to August in 2019 are tested. Under 1% detection threshold, theoretically approximately 100 systems are supposed to be labeled as faulty. Although the two methods do identify a lot of common systems, the univariate analysis labels fewer systems. The following case study shows the performance of one labeled faulty system.

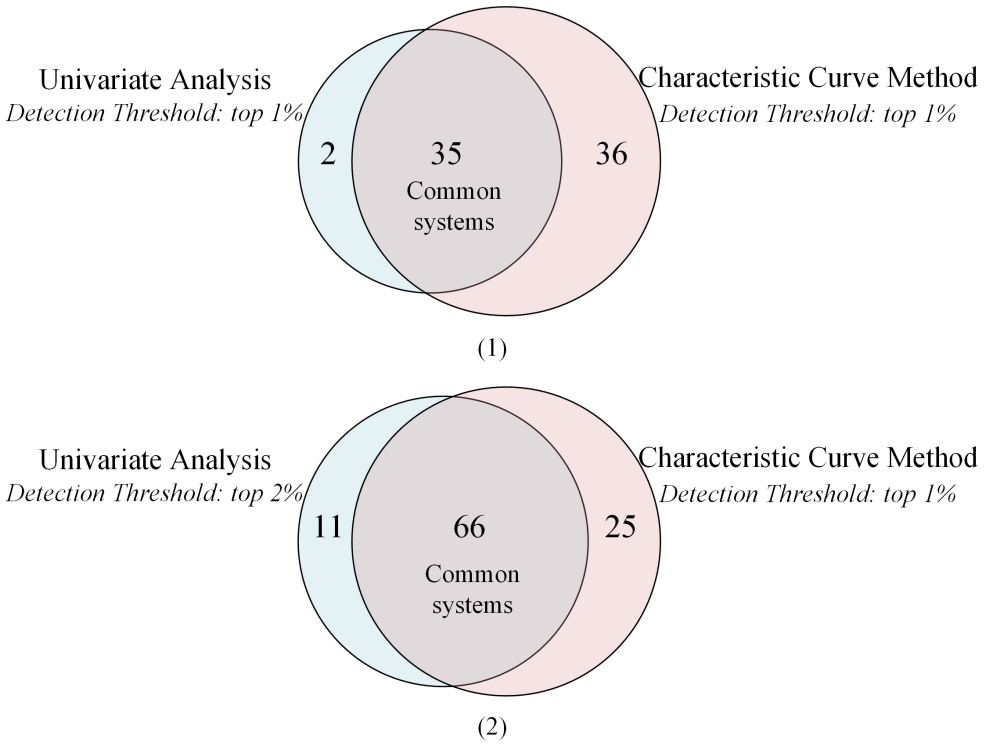


Figure 7.8: The Venn diagrams showing number of systems labeled by the univariate analysis and the characteristic curve method when the detection threshold is set to different percentages. Three-month data from June to August for all approximately 10,000 systems in Florida is used.

7.4.1.1 Case Study One: A System with Inadequate Capacity

As an example, the operational behavior of a labeled faulty system is illustrated in Figure 7.9. The cooling capacity of the system is lower than the cooling demand on most occasions, and thus abnormally high cooling effort in the whole range of temperature difference is detected. In the figure, the solid black line shows the characteristic curve, which is the expected performance of a system in the climate region. The blue points are cooling effort data of the faulty system in the cooling regulating mode from June to August in 2019. The weighted average difference of cooling effort ($\Delta\bar{E}_c$) equals 0.291, higher than the top 1% detection threshold which is 0.268, and thus the system is labeled as faulty by the characteristic curve method. The dashed line in the figure sketches the actual performance of this system, which is shifted upward from the characteristic curve by 29.1%. When the temperature difference is around 0°F, the system spends over 50% of

time regulating the indoor temperature to remove the internal heat load and the solar load. As the temperature difference surpasses 5°F, the system starts working hard with very short stopping time. Sometimes the cooling effort approaches 100%, which means the system is unable to remove all the cooling load anymore and the cooling setpoint temperature cannot be maintained. All the above evidence indicates that the system has inadequate capacity.

Four months later in December 2019, refrigerant leakage is detected for this system, and the low refrigerant charge level also causes the evaporator to freeze due to the low temperature and pressure inside the coil. This event proves the usefulness of the characteristic curve method, which enables the maintenance to be performed a few months earlier, avoiding the operational failure and increasing the indoor thermal comfort. After maintenance, $\Delta\bar{E}_c$ drops to less than 0.2 in the summer of 2020, much lower than the detection threshold.

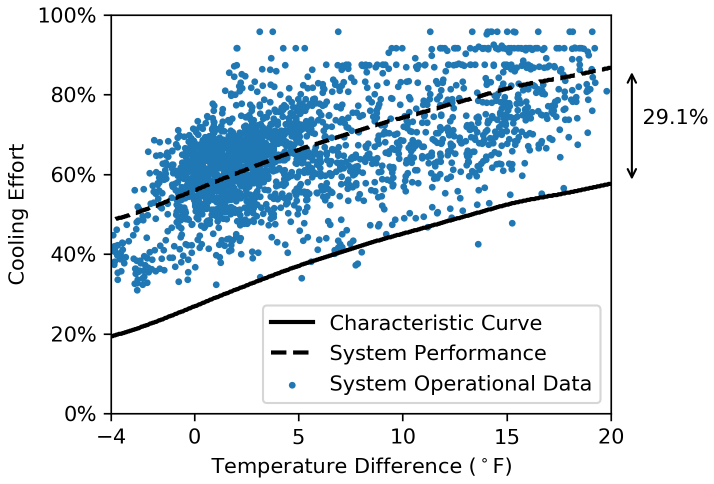


Figure 7.9: Case one: a system without enough cooling capacity. The solid black line shows the characteristic curve, the blue points show the operational data of the faulty system, and the dashed black line sketches the performance of the faulty system.

7.4.2 Quantification of Degradation Trends

The above analysis belongs to fault detection, since the only conclusion from the identified faulty systems is inadequate cooling capacity. However, inadequate capacity could be caused either by degradation faults occurred after installation such as refrigerant leakage or equipment aging, or by commissioning faults occurred during installation such as undercharge or non-condensable gas. To differentiate between these two issues, degradation trend analysis is required to trace back and analyze the historical data of these labeled systems. With the help of degradation trend analysis, the characteristic curve method can take a step further into preliminary fault diagnosis to identify degradation faults. Additionally, the weighted average difference of cooling effort ($\Delta\bar{E}_c$) can be calculated in a fixed-length moving window, and change of $\Delta\bar{E}_c$ through time can assist in quantifying the decline of system performance once a degradation trend is detected.

The degradation trend analysis detects the behavioral change of systems from when they are installed. Because the details are not a main scope of this paper, this paragraph only provides a brief description. The detection algorithm extracts the cooling effort from historical data, and then evaluates if an increasing trend of the cooling effort exists considering the change of ambient conditions. When an increasing trend is detected with high confidence level, the corresponding system is considered as suffering from gradual capacity loss in daily operation with great chance.

The following cases studies (the second and third cases) show two systems both with capacity degradation detected, and the change of $\Delta\bar{E}_c$ is applied to quantify the decline of system performance. The second case is already in the late stage of fault in May and continues to deteriorate in the summer season. The third case is in the early stage of fault in May and gradually loses its cooling capacity during the summer.

7.4.2.1 Case Study Two: Poorly-Performed System with Degradation

The faulty system in case one is analyzed only by the characteristic curve method. In comparison, this second case study shown in Figure 7.10 is diagnosed by the characteristic curve method incorporated with degradation trend analysis. This single-speed split system is labeled with inad-

equate capacity in every month from May to September, and degradation tend is detected as well in these five months. In the first subplot of Figure 7.10, upon the variations of indoor temperature, the cyan color shows the effective cooling regulating periods from which cooling effort data is extracted, and the red color shows periods when setpoint tracking failure happens. Usually tracking failure is caused by the cooling load of the house greater than the cooling capacity of the system, such that the system is unable to maintain the indoor temperature around cooling setpoint. As this system degrades, tracking failure occurs more and more often, and indoor temperature becomes increasingly harder to control. In May, tracking failure happens in less than 24% of system-on time, but in September, tracking failure happens over 70% of system-on time.

The second subplot of Figure 7.10 shows the change of daily cooling effort (i.e. percentage of time that the system is on during a day) in black solid line. The degradation trend analysis indicates this system has significant degradation. From May to September, the daily cooling effort increases from about 50% to nearly 100%, while the degree day (i.e. daily averaged indoor and outdoor temperature difference) shown in the third subplot is relatively constant.

The second subplot also displays the weighted average difference of cooling effort ($\Delta\bar{E}_c$) calculated by the characteristic curve method. The orange dotted line shows the change of $\Delta\bar{E}_c$ in a one-month moving window, and on top of that the red dashed step line shows the values of $\Delta\bar{E}_c$ in each month. The top 1% detection threshold of $\Delta\bar{E}_c$ is obtained by data of all 10,000 systems in Florida from May to September in 2019. In May, the system is labeled as faulty because $\Delta\bar{E}_c$ equals 0.28, already above the threshold. However, due to not being repaired, the system continues to deteriorate and finally reaches a $\Delta\bar{E}_c$ of 0.43, extremely high among all systems. Figure 7.11 shows the change of $\Delta\bar{E}_c$ from May (point a) to September (point b). In May, fewer than 1% of other systems have $\Delta\bar{E}_c$ higher than this system. As degradation continues, $\Delta\bar{E}_c$ has increased by 0.15 and this system gradually becomes one of the worst-performed systems.

In November 2019, large decrease of $\Delta\bar{E}_c$ from 0.45 to 0.14 is detected, which indicates that the system had a maintenance (to be discussed in the third application of the characteristic method). However, this system is still identified with inadequate capacity in the summer of 2020, and finally

a complete breakdown and repair is noticed in August 2020. This study speculates that this system suffers from active refrigerant leakage caused by pinhole or punctures in tubes, but in November 2019 the repairing technicians only recharges the system with more refrigerant (note that refrigerant recharge will not show up in the warranty claim records). However, because the pinhole or punctures is not brazed and sealed, active refrigerant leakage still exists and finally causes system breakdown again in August 2020.

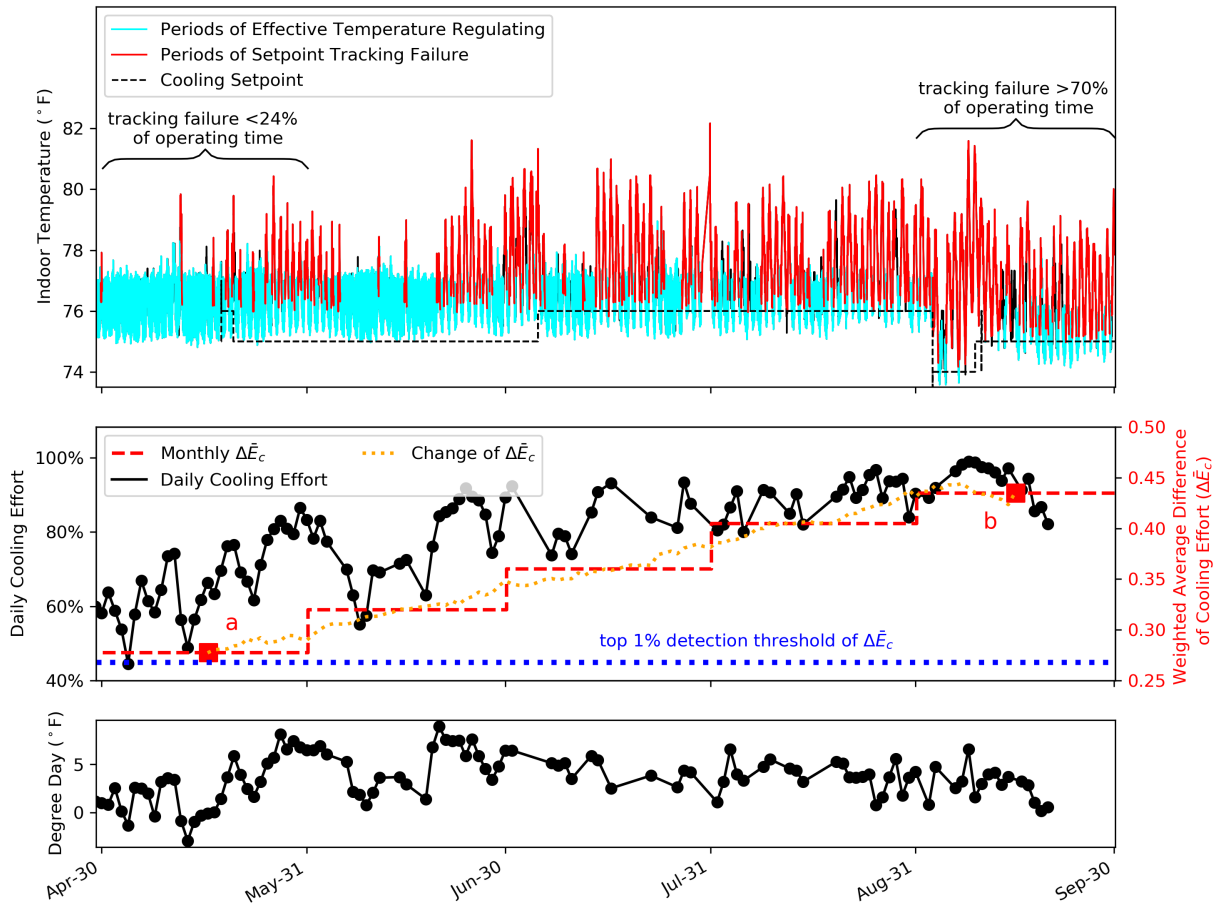


Figure 7.10: Case two: a system labeled with both inadequate capacity and degradation trend. Especially, inadequate capacity is labeled in all five months from May to September.

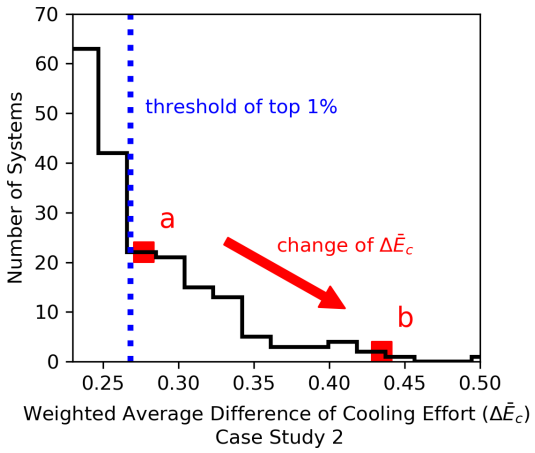


Figure 7.11: Case two: change of weighted average difference of cooling effort ($\Delta\bar{E}_c$) from May (point a) to September (point b). The histogram shows a section of the $\Delta\bar{E}_c$ distribution for all 10,000 systems.

7.4.2.2 Case Study Three: Normal System Gradually Losing Capacity Caused by Degradation

Similar to the second case study, the single-speed system shown in Figure 7.12 is also labeled with degradation trend. However, compared to the former one, this system is labeled with inadequate capacity only in September. As shown in the first subplot, during most of the time in May the system can effectively regulate around the cooling setpoint and tracking failure hardly ever occurs (less than 4% of operating time). But in September, setpoint tracking failure takes up more than 34% of operating time, which is a considerable increase compared with May.

In the second subplot, the daily cooling effort is detected as having significant increasing trend considering the variation of degree days. Initially the $\Delta\bar{E}_c$ value is much lower than the threshold, but eventually rises above the threshold in September due to its fast increase starting from July. Note that $\Delta\bar{E}_c$ decreases a little bit at the end of July, most possibly because occupants are not at home during that period and thus the internal heat gain is relatively low (this speculation is based on the increase of cooling setpoint). Figure 7.13 shows the change of $\Delta\bar{E}_c$ from May (point c) to September (point d). In May, $\Delta\bar{E}_c$ equals 0.17, which means the system is in the early stage of fault, not detectable by the characteristic curve method. But in September, $\Delta\bar{E}_c$ has increased

to 0.29, which means significant degradation occurs within the past months. In fact, the detection algorithm does not even need to wait until $\Delta\bar{E}_c$ rising above the threshold to label it as faulty. $\Delta\bar{E}_c$ increases sharply by over 0.1 between July and August, and meanwhile, degradation trend is also detected. Hence, in this case the change of $\Delta\bar{E}_c$ can provide convincing evidence that enables faults to be reported even earlier and before the system starts exhibiting anomalous behavior.

A complete failure of this system is noticed in June 2020, almost ten months after fault is detected. After maintenance, $\Delta\bar{E}_c$ decreases significantly from 0.27 to 0.05. Application of the proposed fault detection and diagnosis method allows an earlier repair such that home owners would not use a severely degraded system for months.

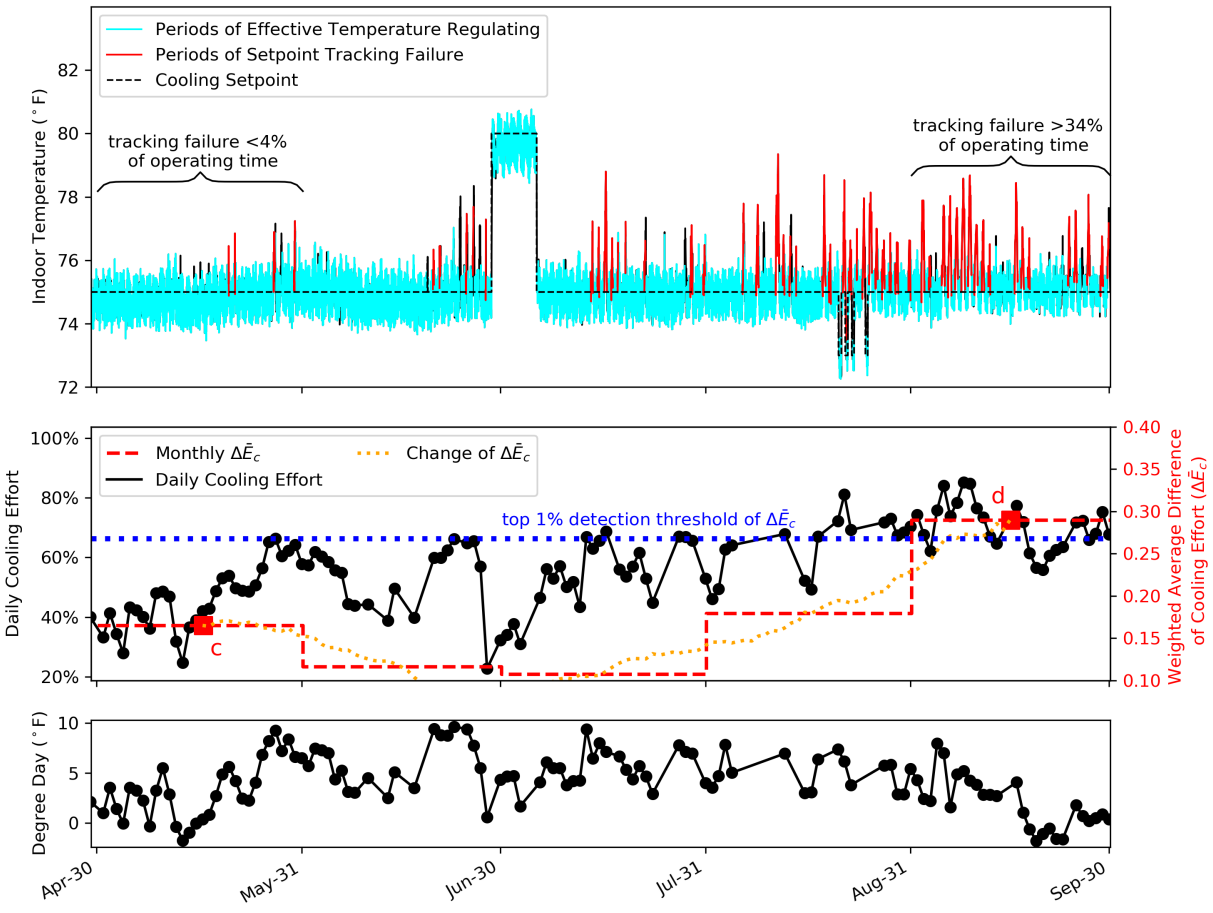


Figure 7.12: Case three: a system labeled with both inadequate capacity and degradation trend. Inadequate capacity is only labeled in September.

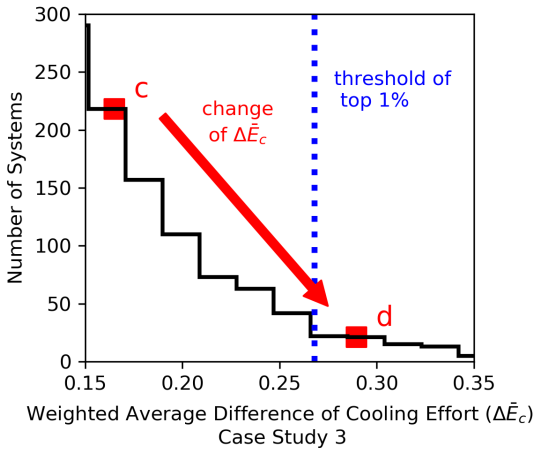


Figure 7.13: Case three: change of weighted average difference of cooling effort ($\Delta\bar{E}_c$) from May (point c) to September (point d). The histogram shows a section of the $\Delta\bar{E}_c$ distribution for all 10,000 systems.

7.4.3 Verification of System Repair After Maintenance

As discussed in the third case study, the increase of $\Delta\bar{E}_c$ is useful in quantifying the degradation of capacity. Conversely, the decrease of $\Delta\bar{E}_c$ can be used to verify improvement of system performance after maintenance. This is useful because sometimes repairing technicians are not able correctly locate the faults and successfully repair the system. As a result, a faulty system could still have the same fault after repair. Also, sometimes when repairing technicians find a system is undercharge, they just recharge the system with more refrigerant. However, the undercharge problem may be caused by leakage in pipes. In this case, $\Delta\bar{E}_c$ will decrease suddenly after maintenance, but slowly increase again. Additionally, verification of system repair is not only useful for service companies, but for researchers as well. Once improvement of system performance is verified, the operational data after maintenance is considered fault-free and can be used to train other process history-based FDD methods. As an example, the fourth case study compares the performance of a system before and after repair, and shows the characteristic curve method can assist in checking whether faults are really fixed.

7.4.3.1 *Case Study Four: Improvement of System Performance after Maintenance*

In Figure 7.14, on July 9th refrigerant leakage due to pinhole or puncture in tube is reported and the problem is fixed soon after by technicians. From the first subplot, effective temperature regulating is heavily affected before the repair. In May, setpoint tracking failure occurs more than 19% of operating time, and as the weather grows warmer in June, setpoint tracking failure periods increase to over half of the operating time. After the leakage problem is fixed, however, essentially the system can effectively regulate the indoor temperature around cooling setpoint and tracking failure scarcely occurs.

From the second subplot, the daily cooling effort in June often approaches 90%, which means that the system is switched on for over 21 hours in those days. But after maintenance daily cooling effort decreases to approximately 60%, much lower compared to June at the same level of degree day. The degradation trend detector shows that, before the fault claim date capacity degradation exists with 98% confidence level, but when combining the data before and after the fault claim date, capacity improvement exists with 99.9% confidence level.

Most importantly, as shown in Figure 7.15, the monthly$\Delta\bar{E}_c$ decreases significantly from 0.42 in June (point e) to 0.24 in August (point f) after maintenance. Note that point f is already lower than the top 1% detection threshold. Based on the observation that $\Delta\bar{E}_c$ does not increase from July after maintenance, the leakage problem is considered being successfully fixed.

7.5 Conclusions

This paper analyzes the pseudo steady-state operational behavior of residential split HVAC systems using only smart thermostat data, and proposes the characteristic curve method to label faulty systems without enough cooling capacity through both multi-system comparison and single-system change detection. The characteristic curve method extracts a feature termed the cooling effort from the pseudo steady-state operational behavior of each system, and derives a metric termed the weighted average difference of cooling effort ($\Delta\bar{E}_c$) to characterize the capability of a system

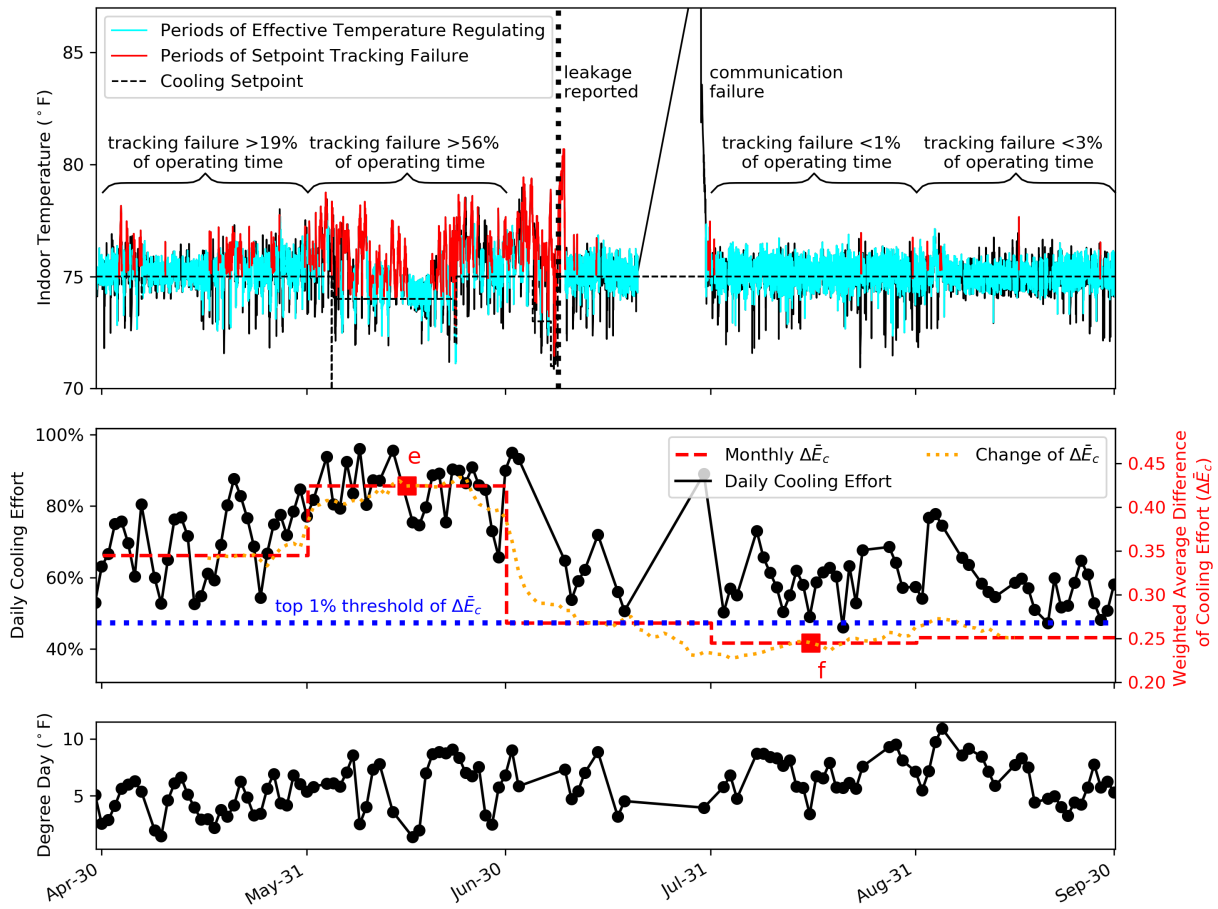


Figure 7.14: Case four: a system labeled with inadequate capacity before July. The problem is detected in July and then repaired.

to remove the cooling load in the house. Usually the detected faulty systems suffer from inadequate capacity issues, especially degradation due to active refrigerant leakage.

The characteristic curve method has at least three applications, including prioritizing faulty system with inadequate cooling capacity, monitoring the change of system cooling capacity over a long period of time, and verifying that a faulty system is fixed after repairing by technicians. For these applications, this paper also provides four case studies of real faulty systems installed in residential homes to demonstrate the usefulness of the characteristic curve method. With this FDD method applied, the service companies could identify refrigerant leakage problems a few months before complete failure happens, and then schedule maintenance to avoid future breakdowns in

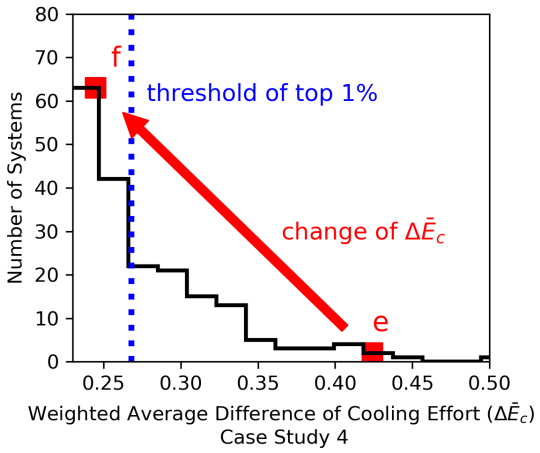


Figure 7.15: Case four: change of weighted average difference of cooling effort ($\Delta\bar{E}_c$) from June (point e) to August (point f). The histogram shows a section of the $\Delta\bar{E}_c$ distribution for all 10,000 systems.

system operation. After maintenance, service companies could also know whether a system fault is successfully repaired. In this way, the service companies benefit from customer loyalty and the home owners benefit from lower energy expenditures and increased indoor comfort levels.

8. SUMMARY AND CONCLUSIONS

This study proposes several fault detection and diagnosis methods for residential HVAC systems. The contributions are summarized in Chapter 2 and detailed from Chapter 3 to Chapter 8. The methods only use the smart thermostat data and detect faults from hundreds of thousands of systems by using large data analytics. The abnormal behavior of the labeled faulty systems are illustrated and the soft faults are validated by the warranty records from Trane Technologies. Finally, this chapter summarizes the current progress in this field of research and proposes some future work.

8.1 Development of FDD Methods Using Smart Thermostat Data

The fault detection methods (fault detectors) developed in this study are mainly statistics-based, with some procedures involving machine learning techniques. The setpoint tracking detector, the inadequate capacity detector, and the control problem detector follow the idea of multi-system feature comparison, while the degradation trend detector follows the idea of single-system change detection.

Both setpoint tracking detector and inadequate capacity detector identify systems that do not have enough cooling/heating capacity among a large number of systems operating under similar conditions. The setpoint tracking detector examines the tracking periods and searches for systems that could not control the indoor temperature around setpoint regularly. On the other side, the inadequate capacity detector examines the regulating periods and evaluates systems with excessive cooling/heating effort under specific operating conditions.

The control problem detector examines the regulating periods as well, and thus can be combined with the inadequate capacity detector through multivariate analysis. It identifies systems with non-working control strategies due to software error, control interrupt, or thermostat misplacement, using the data features of compressor cycling frequency and setpoint error.

The degradation trend detector monitors the behavioral change of a system and can identify whether the system suffers from gradual capacity loss. Three approaches are proposed. The hourly analysis examines the operating behavior in the regulating periods, the daily analysis examines the operating behavior using daily averaged data features, and the proposed monitoring metric (i.e. change of $\Delta\bar{E}_c$) can determine long-term capacity changes. Another functionality of the degradation trend detector is to check whether a mechanical fault is fixed after maintenance or repair. When a fault is actually fixed, the system capacity should increase (i.e. the opposite of a degradation trend) and not have a degradation trend thereafter.

8.2 Future Work

Fault detection and diagnosis for residential HVAC systems is a relatively new field of research. The research ideas are open and several new directions could be tried in the future.

8.2.1 Fault Validation

To validate and improve the performance of fault detectors, a compiled dataset of verified faulty *in situ* systems would be helpful. Currently the Trane Technologies provides a dataset with warranty claims information, which mainly includes the reported types of fault and the claim date for each system that went through the warranty. However, the provided information is still not complete and can be expanded. First of all, if a faulty system is only recharged with refrigerant or the real fault is not detected, then it will not show up in the warranty dataset. That is to say, a lot of faulty data series is not recorded. Moreover, the approximate time period when a system is faulty or fault-free is unknown. Thus, in this research half-month data before a claim date is considered faulty while one-month data one week after a claim date is considered fault-free. Exact periods of validated faulty and fault-free data can facilitate the training process of machine learning methods.

8.2.2 Incorporating Deep Learning Models

The FDD methods proposed in this study are statistics-based, but in the future as the validated data becomes abundant, deep learning methods are worth trying. Figure 8.1 shows the architecture of a general artificial neural network model, which consists of an input layer, a few hidden layers,

and an output layer. The parameters in each layer are optimized by the training data, and after that the trained model can be applied to make predictions for new data.

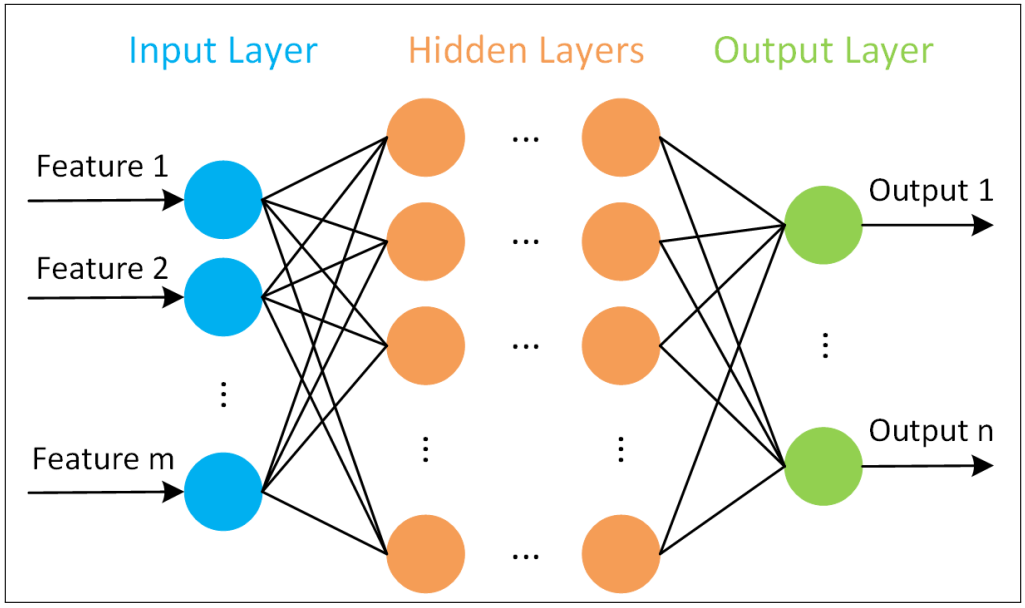


Figure 8.1: The architecture of an artificial neural network model.

This research task can be treated as supervised multi-class classification problems, where the data features are extracted from raw thermostat time-series and the true labels are either normal or faulty (or specific types of fault). Here the author proposes two possible solutions: recurrent neural network and Siamese network. As the development of deep learning models, there are definitely more solutions for this problem.

8.2.2.1 Recurrent Neural Network

The operational data features for each system in a period can be viewed as a multivariate time series. Thus, the recurrent neural network (RNN) model is an appropriate option. Figure 8.2 shows the architecture of an RNN model that consists of an input layer, one long short-term memory (LSTM) layer [62], a few dense layers, and an output layer. By using the LSTM layer, the model can learn both short and very long range connections in a time series. This functionality is very

useful because the thermostat data often has cyclical patterns due to outdoor temperature variations and occupant behaviors. In this study, the input features may include the time (starting from 0), time of day, day of week, temperature difference, setpoint error, and cooling/heating status. In the training process, the fault-free data is labeled '0', and the faulty data is labeled '1'. In the testing process, the detection output of each multivariate time series is the possibility of detecting a fault, a number between 0 and 1. This approach belongs to the multi-system feature comparison, because data features from all systems are compared and analyzed simultaneously.

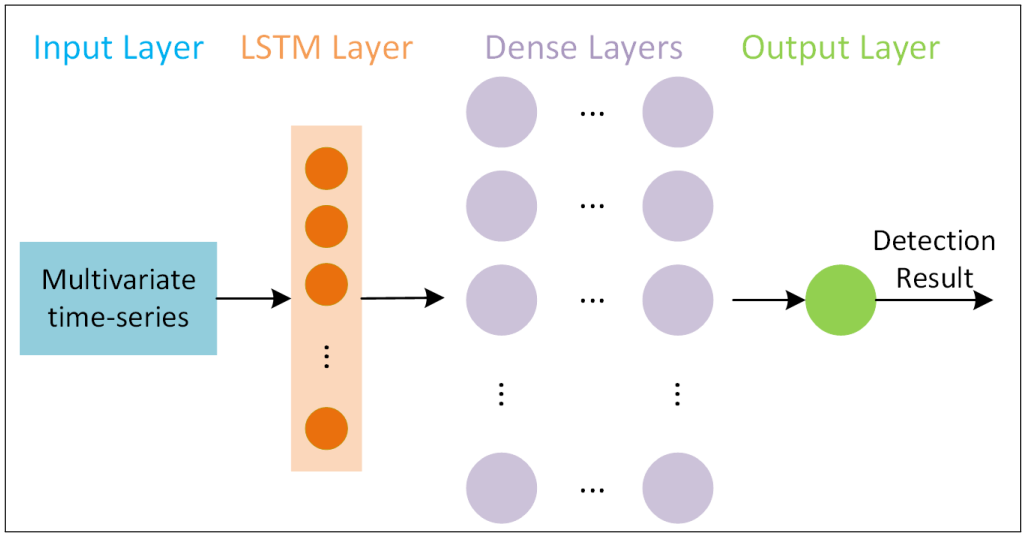


Figure 8.2: The architecture of the long short-term memory (LSTM) neural network model.

Preliminary results show that the model does not perform very well. In the future more features should be added for training, and the possible ones will be briefly discussed in Section 8.2.3.

8.2.2.2 Siamese Network

Besides multi-system feature comparison, another approach proposed in this project is single-system change detection. The Siamese network [63] is a method to compare the similarity between multiple time series, and therefore can be used to measure the change of the operational behavior for each system through time in a fixed-length moving window.

In the training process, each system should have three datasets: anchor, positive, and negative. Both anchor and positive datasets are fault-free data, and the negative dataset is faulty data. An LSTM neural network shown in Figure 8.3 is built to encode the input dataset into a vector. Then, the LSTM neural network is trained which can minimize the distance between the anchor vector and the positive vector for each system, and maximize the distance between the anchor vector and the negative vector for each system. The training process is achieved by minimizing the triplet loss function L:

$$L = \sum_{i=1}^{M} \max(\|f(A_i) - f(P_i)\|^2 - \|f(A_i) - f(N_i)\|^2 + \alpha, 0) \tag{8.1}$$

where M is the total number of systems for training, A_i is the anchor dataset for system i, P_i is the positive dataset for system i, N_i is the negative dataset for system i, $f()$ is the LSTM neural network which encodes a dataset to a vector, and α is a margin.

Then, in the testing process, each system should have an anchor (fault-free) dataset, which could be the operational data records when the systems are installed or repaired. The distance between the anchor vector and the testing data vector for the same system is calculated. If the distance is greater than a threshold, then change of operational behavior is detected for the corresponding system.

Preliminary results show that this approach is promising. In the future more data is needed to train an accurate model.

8.2.3 Fault Detection with Additional Data Features

Additional data features are very helpful for fault detection and diagnosis as well. Currently only data recorded by smart thermostats and the outdoor temperature data from a third party are available for analysis. In the future, we hope the solar irradiance data, home thermal characteristics, and sensor records can be obtained, because they provide valuable information. Note that although installing a dozen of sensors are impractical due to the high initial cost, limited number of low-cost sensors should be considered where important features are extracted.

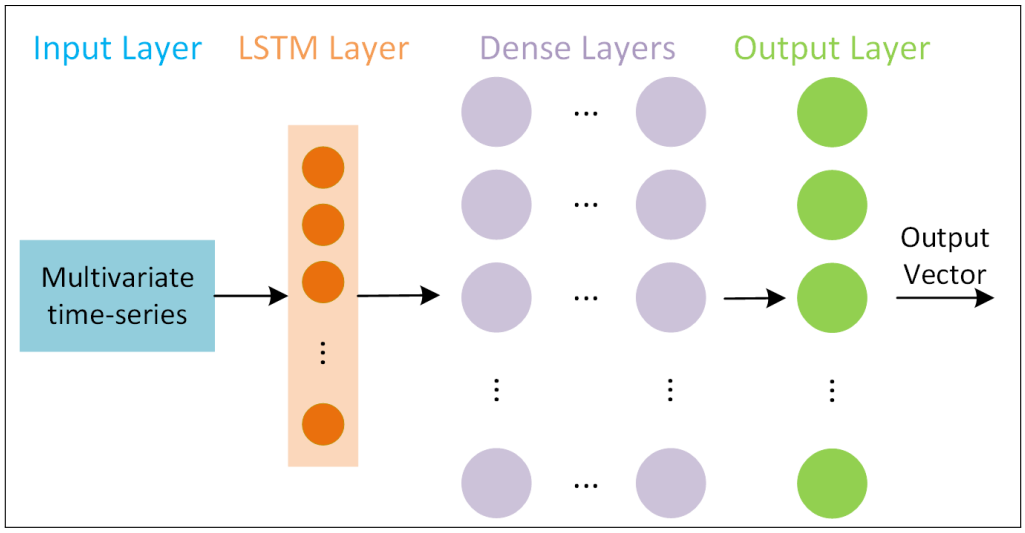


Figure 8.3: The architecture of the long short-term memory (LSTM) neural network used in the Siamese network.

Furthermore, advanced technologies to be applied in the next a few years can also change this field of research significantly. For example, the utility meters can record the home energy consumption, which helps quantify the internal heat load of a home; the occupancy sensors can detect the number of home occupants, and thereby the comfort level when occupants are at home can be accurately evaluated.

REFERENCES

[1] EIA, "Residential Energy Consumption Survey (RECS)," tech. rep., U.S. Energy Information Administration, 2015.

[2] S. Katipamula and M. R. Brambley, "Methods for fault detection, diagnostics, and prognostics for building systems—a review, part i," *Hvac&R Research*, vol. 11, no. 1, pp. 3–25, 2005.

[3] W. Kim and S. Katipamula, "A review of fault detection and diagnostics methods for building systems," *Science and Technology for the Built Environment*, vol. 24, no. 1, pp. 3–21, 2018.

[4] F. Guo, A. Rogers, and B. Rasmussen, "Multivariate fault detection for residential hvac systems using cloud-based thermostat data, part i: Methodology," *Science and Technology for the Built Environment*, 2021.

[5] *International energy conservation code*. International Code Council, Building Officials, Code Administrators International, International Conference of Building Officials, and Southern Building Code Congress International, 2012.

[6] A. Rogers, J. Martinez, F. Guo, and B. Rasmussen, "Labeling modes of operation and extracting features for fault detection with cloud-based thermostat data.," *ASHRAE Transactions*, vol. 126, no. 2, 2020.

[7] F. Guo and B. Rasmussen, "Fault detection and diagnosis for residential hvac systems using transient cloud-based thermostat data," in *Proceedings of the 6th International High Performance Buildings Conference at Purdue*, 2020.

[8] EIA, "Residential Energy Consumption Survey (RECS)," tech. rep., U.S. Energy Information Administration, 2012.

[9] J. Proctor and T. Downey, "Transforming routine air conditioner maintenance practices to improve equipment efficiency and performance," in *Proceedings of the 1999 international energy program evaluation conference*, 1999.

[10] J. Winkler, S. Das, L. Earle, L. Burkett, J. Robertson, D. Roberts, and C. Booten, "Impact of installation faults in air conditioners and heat pumps in single-family homes on us energy usage," *Applied Energy*, vol. 278, p. 115533, 2020.

[11] S. Katipamula and M. R. Brambley, "Methods for fault detection, diagnostics, and prognostics for building systems—a review, part ii," *Hvac&R Research*, vol. 11, no. 2, pp. 169–187, 2005.

[12] Z. Shi and W. O'Brien, "Development and implementation of automated fault detection and diagnostics for building systems: A review," *Automation in Construction*, vol. 104, pp. 215–229, 2019.

[13] M. S. Mirnaghi and F. Haghighat, "Fault detection and diagnosis of large-scale hvac systems in buildings using data-driven methods: A comprehensive review," *Energy and Buildings*, p. 110492, 2020.

[14] Y. Zhao, T. Li, X. Zhang, and C. Zhang, "Artificial intelligence-based fault detection and diagnosis methods for building energy systems: Advantages, challenges and the future," *Renewable and Sustainable Energy Reviews*, vol. 109, pp. 85–101, 2019.

[15] Y. Zhao, C. Zhang, Y. Zhang, Z. Wang, and J. Li, "A review of data mining technologies in building energy systems: Load prediction, pattern identification, fault detection and diagnosis," *Energy and Built Environment*, vol. 1, no. 2, pp. 149–164, 2020.

[16] A. Rogers, F. Guo, and B. Rasmussen, "A review of fault detection and diagnosis methods for residential air conditioning systems," *Building and Environment*, vol. 161, p. 106236, 2019.

[17] H. Li and J. E. Braun, "Decoupling features and virtual sensors for diagnosis of faults in vapor compression air conditioners," *International Journal of Refrigeration*, vol. 30, no. 3, pp. 546–564, 2007.

[18] A. Rogers, F. Guo, and B. Rasmussen, "Uncertainty analysis and field implementation of a fault detection method for residential hvac systems," *Science and Technology for the Built Environment*, vol. 26, no. 3, pp. 320–333, 2020.

[19] J. W. Yoo, S. B. Hong, and M. S. Kim, "Refrigerant leakage detection in an eev installed residential air conditioner with limited sensor installations," *International Journal of Refrigeration*, vol. 78, pp. 157–165, 2017.

[20] M. Jain, M. Gupta, A. Singh, and V. Chandan, "Beyond control: Enabling smart thermostats for leakage detection," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 3, pp. 14:1–14:21, Mar. 2019.

[21] W. Turner, A. Staino, and B. Basu, "Residential hvac fault detection using a system identification approach," *Energy and Buildings*, vol. 151, pp. 1–17, 2017. Exported from https://app.dimensions.ai on 2019/02/22.

[22] A. Rogers, S. Ghosh, R. Wilcock, and N. R. Jennings, "A scalable low-cost solution to provide personalised home heating advice to households," in *Proceedings of the 5th ACM Workshop on Embedded Systems For Energy-Efficient Buildings*, pp. 1–8, 2013.

[23] H. B. Gunay, W. O'Brien, I. Beausoleil-Morrison, and J. Bursill, "Development and implementation of a thermostat learning algorithm," *Science and Technology for the Built Environment*, vol. 24, no. 1, pp. 43–56, 2018.

[24] B. Huchuk, W. O'Brien, and S. Sanner, "A longitudinal study of thermostat behaviors based on climate, seasonal, and energy price considerations using connected thermostat data," *Building and Environment*, vol. 139, pp. 199 – 210, 2018.

[25] B. Huchuk, W. O'brien, and S. Sanner, "Exploring smart thermostat users' schedule override behaviors and the energy consequences," *Science and Technology for the Built Environment*, vol. 27, no. 2, pp. 195–210, 2020.

[26] M. F. Touchie and J. A. Siegel, "Residential hvac runtime from smart thermostats: characterization, comparison, and impacts," *Indoor Air*, vol. 28, no. 6, pp. 905–915, 2018.

[27] W. Van der Ham, M. Klein, S. A. Tabatabaei, D. J. Thilakarathne, and J. Treur, "Methods for a smart thermostat to estimate the characteristics of a house based on sensor data," *Energy Procedia*, vol. 95, pp. 467–474, 2016.

[28] A. P. Rogers, F. Guo, and B. P. Rasmussen, "A change point detection algorithm with application to smart thermostat data.," *ASHRAE Transactions*, vol. 126, no. 1, 2020.

[29] P. C. Mahalanobis, "On the generalized distance in statistics," *Proceedings of the National Institute of Sciences (Calcutta)*, vol. 2, pp. 49–55, 1936.

[30] H. B. Mann, "Nonparametric tests against trend," *Econometrica: Journal of the econometric society*, pp. 245–259, 1945.

[31] M. G. Kendall, "Rank correlation methods. london: Charles griffin & co," 1975.

[32] A. Kulkarni and H. von Storch, "Monte carlo experiments on the effect of serial correlation on the mann-kendall test of trend," *Meteorologische Zeitschrift*, vol. 4, no. 2, pp. 82–85, 1995.

[33] C. Libiseller and A. Grimvall, "Performance of partial mann–kendall tests for trend detection in the presence of covariates," *Environmetrics: The official journal of the International Environmetrics Society*, vol. 13, no. 1, pp. 71–84, 2002.

[34] H. Theil, "A rank-invariant method of linear and polynominal regression analysis (parts 1-3)," in *Ned. Akad. Wetensch. Proc. Ser. A*, vol. 53, pp. 1397–1412, 1950.

[35] P. K. Sen, "Estimates of the regression coefficient based on kendall's tau," *Journal of the American statistical association*, vol. 63, no. 324, pp. 1379–1389, 1968.

[36] C. Neme, S. Nadel, and J. Proctor, "Energy savings potential from addressing residential air conditioner and heat pump installation problems," American Council for an Energy-Efficient Economy Washington, DC, 1999.

[37] H. Do and K. S. Cetin, "Data-driven evaluation of residential hvac system efficiency using energy and environmental data," *Energies*, vol. 12, no. 1, p. 188, 2019.

[38] J. H. Ward Jr, "Hierarchical grouping to optimize an objective function," *Journal of the American statistical association*, vol. 58, no. 301, pp. 236–244, 1963.

[39] P. J. Rousseeuw, "Multivariate estimation with high breakdown point," *Mathematical statistics and applications*, vol. 8, no. 283-297, p. 37, 1985.

[40] G. E. Box and D. R. Cox, "An analysis of transformations," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 26, no. 2, pp. 211–243, 1964.

[41] P. J. Rousseeuw and B. C. van Zomeren, "Unmasking multivariate outliers and leverage points," *Journal of the American Statistical Association*, vol. 85, no. 411, pp. 633–639, 1990.

[42] K. Vakili and E. Schmitt, "Finding multivariate outliers with fastpcs," *Computational Statistics & Data Analysis*, vol. 69, pp. 54 – 66, 2014.

[43] R. A. Maronna, R. D. Martin, V. J. Yohai, and M. Salibián-Barrera, *Robust statistics: theory and methods (with R)*. John Wiley & Sons, 2019.

[44] P. J. Rousseeuw and K. V. Driessen, "A fast algorithm for the minimum covariance determinant estimator," *Technometrics*, vol. 41, no. 3, pp. 212–223, 1999.

[45] W. A. Stahel, *Breakdown of covariance estimators*. Fachgruppe für Statistik, Eidgenössische Techn. Hochsch., 1981.

[46] C. Leys, O. Klein, Y. Dominicy, and C. Ley, "Detecting multivariate outliers: Use a robust variant of the mahalanobis distance," *Journal of Experimental Social Psychology*, vol. 74, pp. 150 – 156, 2018.

[47] Y. Hu, D. P. Yuill, A. Ebrahimifakhar, and A. Rooholghodos, "An experimental study of the behavior of a high efficiency residential heat pump in cooling mode with common installation faults imposed," *Applied Thermal Engineering*, vol. 184, p. 116116, 2021.

[48] M. Bayazit, "Nonstationarity of hydrological records and recent trends in trend analysis: a state-of-the-art review," *Environmental Processes*, vol. 2, no. 3, pp. 527–542, 2015.

[49] W. M. Alley, "A note on stagewise regression," *The American Statistician*, vol. 41, no. 2, pp. 132–134, 1987.

[50] W. M. Alley, "Using exogenous variables in testing for monotonic trends in hydrologic time series," *Water Resources Research*, vol. 24, no. 11, pp. 1955–1961, 1988.

[51] E. P. Smith and K. A. Rose, "Trend detection in the presence of covariates: stagewise versus multiple regression," *Environmetrics*, vol. 2, no. 2, pp. 153–168, 1991.

[52] S. Yue, P. Pilon, B. Phinney, and G. Cavadias, "The influence of autocorrelation on the ability to detect trend in hydrological series," *Hydrological processes*, vol. 16, no. 9, pp. 1807–1829, 2002.

[53] S. Yue and C. Wang, "The mann-kendall test modified by effective sample size to detect trend in serially correlated hydrological series," *Water resources management*, vol. 18, no. 3, pp. 201–218, 2004.

[54] K. H. Hamed and A. R. Rao, "A modified mann-kendall trend test for autocorrelated data," *Journal of hydrology*, vol. 204, no. 1-4, pp. 182–196, 1998.

[55] B. Önöz and M. Bayazit, "Block bootstrap for mann–kendall trend test of serially dependent data," *Hydrological Processes*, vol. 26, no. 23, pp. 3552–3560, 2012.

[56] A. El-Shaarawi, "Environmental monitoring, assessment and prediction of change," *Environmetrics*, vol. 4, no. 4, pp. 381–398, 1993.

[57] E. J. Dietz and T. J. Killeen, "A nonparametric multivariate test for monotone trend with pharmaceutical applications," *Journal of the American Statistical Association*, vol. 76, no. 373, pp. 169–174, 1981.

[58] H. R. Künsch, "The jackknife and the bootstrap for general stationary observations," *Annals of Statistics*, vol. 17, no. 3, pp. 1217–1241, 1989.

[59] K. Sheppard, S. Khrapov, G. Lipták, mikedeltalima, R. Capellini, esvhd, Hugle, JPN, X. RENE-CORAIL, M. E. Rose, MRabba, and jbrockmendel, "bashtage/arch: Release 4.11," Nov. 2019.

[60] M. Hussain and I. Mahmud, "pymannkendall: a python package for non parametric mann kendall family of trend tests.," *Journal of Open Source Software*, vol. 4, p. 1556, 7 2019.

[61] ASHRAE, "2017 ashrae handbook: Fundamentals," 2017.

[62] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[63] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1701–1708, 2014.

APPENDIX A

PERFORMANCE COMPARISON OF MODIFIED MANN-KENDALL STATISTICAL TESTS

A.1 Comparison of the Modified Mann-Kendall Tests

Paper D applied modified Mann-Kendall tests to identify faulty residential HVAC systems with degradation trends. This is because the system operational data is autocorrelated, but the original Mann-Kendall test is designed for independent data. There are four popular modified Mann-Kendall tests to account for the autocorrelation, namely trend-free pre-whitening [52], variance correction proposed by Hamed and Rao [54], variance correction proposed by Yue and Wang [53], and block bootstrapping [55]. Finally the author chose the moving block bootstrap method due to its higher effectiveness. For validation purpose, this section introduces the other three modified Mann-Kendall tests, and then compares their performance with the moving block bootstrap method by evaluating the Type I error.

A.1.1 The Trend-Free Pre-Whitening Mann-Kendall Test

[52] proposed a four-step approach to remove the lag-one autocorrelation component as follows:

1. The slope b of the time series X_t is estimated by the Sen's slope, and the time series is detrended by

$$X'_t = X_t - bt \tag{A.1}$$

2. The lag-one autocorrelation coefficient ρ_1 is estimated from X'_t and removed by

$$Y'_t = X'_t - \rho_1 X'_{t-1} \tag{A.2}$$

3. The trend of the original time series is added back by

$$Y_t = Y_t' + bt \tag{A.3}$$

4. The original Mann-Kendall test is performed on the new series Y_t.

A.1.2 The Variance Correction Mann-Kendall Test Proposed by Hamed and Rao

[54] proposed to modify the variance of the S-statistic in order to account for the serial autocorrelation. The new variance is calculated by:

$$\sigma^2 = \frac{n(n-1)(2n+5)}{18} \cdot \frac{n}{n_s^*} \tag{A.4}$$

where

$$\frac{n}{n_s^*} = 1 + \frac{2}{n(n-1)(n-2)} \cdot \sum_{i=1}^{n-1}(n-1)(n-i-1)(n-i-2)\rho_i \tag{A.5}$$

In the equation, n is the sample size, n_s^* is termed the effective sample size, and ρ_i is the lag-i autocorrelation coefficient of the *ranks* of the *detrended* data. Note that using the ranks of data ensures the estimated autocorrelation is robust, and detrending the data by the Sen's slope guarantees the serial autocorrelation not being contaminated by the existence of trend [53]. Also, the author mentioned only significant ρ_i in the dataset should be used. The asymptotic normality of the Z-statistic does not change with the modification.

A.1.3 The Variance Correction Mann-Kendall Test Proposed by Yue and Wang

Similar to the the variance correction method proposed by Hamed and Rao, [53] proposed another formula to calculate the effective sample size n_s^*, where

$$\frac{n}{n_s^*} = 1 + 2\sum_{i=1}^{n-1}(1 - \frac{i}{n})\rho_i \tag{A.6}$$

In this equation, ρ_i is the lag-i autocorrelation coefficient of the *detrended* data.

A.1.4 Type I Error Comparison of the Modified Mann-Kendall Tests

The Type I error of the four modified Mann-Kendall tests is evaluated at 0.1% significance level, which means at most one out of a thousand fault-free systems will be mistakenly detected as faulty. Table A.1 shows the Monte Carlo simulation of 50,000 randomly generated time series without trend. The lag-one autocorrelation coefficient (ρ_1) ranges from 0 to 0.8, and sample size (n) ranges from 30 to 100. The autocorrelated time series is generated by autoregressive model of order one (i.e. AR(1) model):

$$X_t = \rho_1 X_{t-1} + \epsilon_t \tag{A.7}$$

where X_t is data at time t, X_{t-1} is data at time $t-1$, and ϵ_t is the random noise term. For the variation correction methods, the first ten lags are considered; for the moving block bootstrap method, 50,000 bootstrap samples (N) are taken for each time series, and the block length (L) is 7 when the sample size is smaller or equal to 50, and is 10 otherwise.

From the results, trend-free pre-whitening and variation correction methods all have problems maintaining the Type I error lower than 0.1%. The trend-free pre-whitening method has unacceptably high Type I error when high autocorrelation exists. This is because the trend-free pre-whitening method only considers the first lag component, but in a time series with high autocorrelation the followed lags can also have significant impacts on the test results. The two variance correction methods have the similar performance, but still have high Type I error when $\rho_1 > 0.4$. In comparison, the moving block bootstrap method is capable of keeping the Type I error below the 0.1% significance level when $\rho_1 < 0.6$ and generates much fewer false positives. Because false positives are almost intolerable in fault detection and diagnosis of residential HVAC systems, the moving block bootstrap method is selected to perform the trend test using smart thermostat data.

A.2 Comparison between the Partial Mann-Kendall Test (PMK) and the the Partial Mann-Kendall Test with Moving Block Bootstrap (PMK-MBBS)

Table A.2 shows the Monte Carlo simulation results of Type I error between the PMK test and the PMK-MBBS test.

Table A.1: Type I error under 0.1% significance level for four modified Mann-Kendall tests with different lag-one autocorrelation coefficient (ρ_1) and sample sizes (n). Each case is obtained by 50,000 Monte Carlo simulations.

Method	ρ_1	Type I Error $n = 30$	$n = 50$	$n = 100$
Trend-free Pre-whitening	0	0.09%	0.12%	0.12%
	0.2	0.44%	0.49%	0.55%
	0.4	1.78%	2.13%	2.63%
	0.6	6.18%	7.77%	9.00%
	0.8	16.9%	20.7%	23.8%
Variance Correction by Hamed and Rao	0	0.10%	0.12%	0.11%
	0.2	0.35%	0.47%	0.52%
	0.4	1.30%	1.49%	1.12%
	0.6	3.71%	3.02%	1.09%
	0.8	7.72%	4.11%	1.96%
Variance Correction by Yue and Wang	0	0.07%	0.11%	0.10%
	0.2	0.34%	0.48%	0.52%
	0.4	1.23%	1.60%	0.92%
	0.6	3.55%	2.85%	0.98%
	0.8	7.22%	3.75%	1.68%
Moving Block Bootstrap	0	0.02%	0.01%	0.01%
	0.2	0.03%	0.02%	0.03%
	0.4	0.04%	0.03%	0.05%
	0.6	0.09%	0.06%	0.06%
	0.8	0.21%	0.31%	0.25%

A.3 Performance of a Simplified Mann-Kendall Test with Moving Block Bootstrap (MK-MBBS)

In Paper D, when the Mann-Kendall test with moving block bootstrap (MK-MBBS) is applied to monitor the system operating conditions in real time, the author proposes to use only 5,000 bootstrap samples to reduce the computational cost. Meanwhile, the confidence level is lowered

Table A.2: Type I error of bivariate time series at multiple levels of cross-correlation (ϕ) and lag-one autocorrelation coefficients ($\rho_{1,1}$ and $\rho_{1,2}$). The results from PMK-MBBS and PMK are compared at 5% and 1% significance level. Note that some combinations of the coefficients are removed because Equation (6.21) should always be positive-semidefinite.

Sample size	Cross-correlation	Autoregressive coefficients		Type I error			
				Nominal level 5%		Nominal level 1%	
n	ϕ	$\rho_{1,1}$	$\rho_{1,2}$	PMK-MBBS	PMK	PMK-MBBS	PMK
30 (days)	0	0	0	5.4%	5.3%	0.7%	0.8%
	0	0	0.4	5.7%	4.3%	0.8%	0.7%
	0	0	0.8	5.7%	3.6%	0.9%	0.7%
	0	0.4	0	8.1%	13.2%	1.7%	4.7%
	0	0.4	0.4	7.9%	12.4%	1.7%	4.7%
	0	0.4	0.8	7.8%	10.4%	1.7%	3.4%
	0	0.8	0	12.4%	26.1%	4.0%	17.0%
	0	0.8	0.4	12.7%	24.9%	4.2%	16.6%
	0	0.8	0.8	11.9%	22.6%	3.8%	14.1%
	0.8	0	0	5.5%	5.0%	0.7%	0.9%
	0.8	0	0.4	5.4%	4.2%	0.7%	0.6%
	0.8	0.4	0	7.7%	11.9%	1.4%	4.5%
	0.8	0.4	0.4	8.3%	11.5%	1.7%	3.7%
	0.8	0.4	0.8	6.1%	10.8%	0.9%	2.7%
	0.8	0.8	0.4	11.8%	25.8%	3.8%	17.4%
	0.8	0.8	0.8	12.0%	23.4%	3.3%	14.1%
90 (days)	0	0	0	5.2%	4.5%	0.8%	1.1%
	0	0	0.4	5.2%	4.9%	1.1%	0.9%
	0	0	0.8	5.4%	4.4%	0.6%	0.8%
	0	0.4	0	6.3%	13.7%	1.2%	6.2%
	0	0.4	0.4	6.9%	13.6%	1.3%	6.1%
	0	0.4	0.8	6.3%	12.3%	1.2%	5.4%
	0	0.8	0	11.2%	27.7%	4.3%	20.0%
	0	0.8	0.4	12.5%	27.6%	4.3%	19.8%
	0	0.8	0.8	11.4%	27.0%	3.9%	19.7%
	0.8	0	0	5.1%	4.9%	0.7%	0.7%
	0.8	0	0.4	5.0%	4.9%	0.6%	0.9%
	0.8	0.4	0	7.3%	12.8%	1.8%	5.3%
	0.8	0.4	0.4	7.0%	12.7%	1.2%	5.1%
	0.8	0.4	0.8	5.0%	12.5%	0.8%	4.5%
	0.8	0.8	0.4	11.7%	28.2%	3.8%	20.4%
	0.8	0.8	0.8	11.5%	26.8%	4.1%	18.6%

from 99.9% to 99.8%. In this section, the performance of this simplified version of the MK-MBBS test is discussed.

Using 500 cooling effort time-series from residential systems, Figure A.1a shows the confidence level calculated by the simplified version of the MK-MBBS test for the time-series having a trend with above 99.9% confidence level. Most of these time-series has a confidence level of above 99.8%, and there are only a few false negative cases (shown in light blue). The result shows that 99.8% confidence level of the simplified version of the MK-MBBS test almost achieves a similar level of the Type II error.

On the other side, for the time-series labeled by the simplified version of the test as having an increasing trend with above 99.8% confidence level, Figure A.1b shows their real confidence level of having a trend. Most of these time-series has a real confidence level of above 99.4%, a little bit lower compared with 99.9%. However, the trivial loss of Type I error is worthy compared to the significant decline of computational cost.

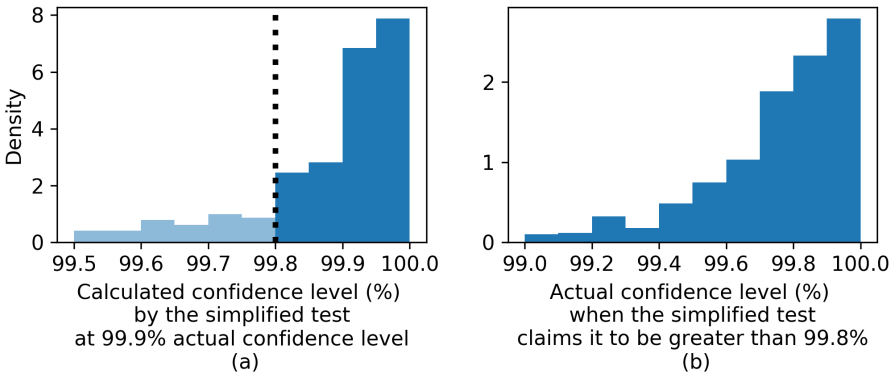


Figure A.1: Performance of the simplified version of the MK-MBBS test.

A.4 Performance of a Simplified Partial Mann-Kendall Test with Moving Block Bootstrap (PMK-MBBS)

Similar to the above Section, when the partial Mann-Kendall test with moving block bootstrap (PMK-MBBS) is applied to monitor the system operating conditions in real time, the author also proposes to use only 5,000 bootstrap samples to reduce the computational cost. Meanwhile, the confidence level is lowered from 99.9% to 99.8%. In this section, the performance of this simplified version of the PMK-MBBS test is discussed as well.

Using real thermostat data from 200 residential systems, Figure A.2a shows the confidence level calculated by the simplified version of the PMK-MBBS test for the time-series having a trend with above 99.9% confidence level. Most of these time-series has a confidence level of above 99.8%. Therefore, 99.8% confidence level of the simplified version of the PMK-MBBS test almost achieves a similar level of the Type II error.

On the other side, for the time-series labeled by the simplified version of the test as having an increasing trend with above 99.8% confidence level, Figure A.2b shows their real confidence level of having a trend. Most of these time-series has a real confidence level of above 99.6%.

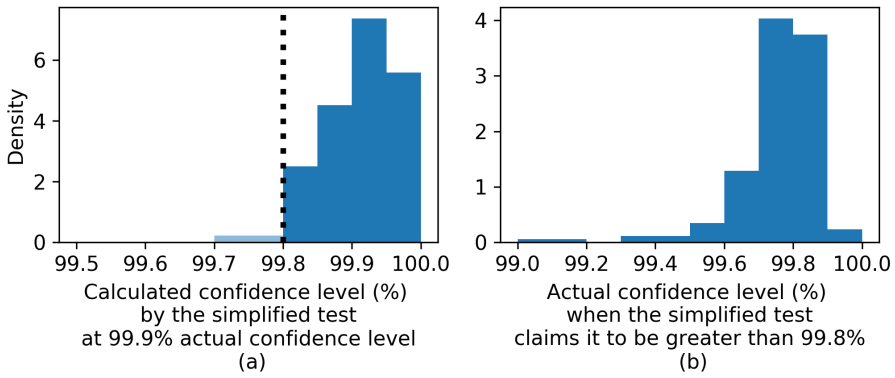


Figure A.2: Performance of the simplified version of the PMK-MBBS test.

APPENDIX B

PYTHON CODE

This appendix includes Python code used to analyze data throughout the dissertation.