Research Report

Automated detection of hidden failures in heating and photovoltaic systems of 'Nul-op-de-meter'-housing Projects

Wietse Harmsma

University of Twente, Enschede

BOUWGROEP DIJKSTRA DRAISMA

UNIVERSITY OF TWENTE.

<Blank page>

Research Report

Automated detection of hidden failures in heating and photovoltaic systems of 'Nul-op-de-meter'-housing Projects

Version 8

3-2-2022

Author
Wietse Harmen Harmsma BSc
s1570285
w.h.harmsma@student.utwente.nl

Supervisors University of Twente
dr. M.C. van der Heijden
dr. C.G.M. Groothuis-Oudshoorn

Supervisor bouwgroep Dijkstra-Draisma
W.M. de Vries MSc

University of Twente
Drienerlolaan 5, 7522 NB, Enschede
The Netherlands

Preface

Dokkum, January 2022

Dear reader,

I am very happy to present you my thesis “Automated detection of hidden failures in heating and photovoltaic systems of ‘Nul-op-de-meter’-housing projects”. This thesis reports on a study of energy usage data from the net zero emission housing projects of Bouwgroep Dijkstra-Draisma. The goal of this study was to advance the detection of hidden failures in heating and photovoltaic systems. This thesis is written as my graduation assignment for the master Industrial Engineering and Management (IEM) at the University of Twente.

This project absolutely has been a challenge, as I faced the limits of what was possible with the available data. I could not have completed this thesis without the support of my supervisors. First of all, I want to thank dr. Matthieu van der Heijden for his dedicated guidance throughout this project as well as his lectures over the course of the master program. Secondly, I want to thank dr. Karin Groothuis-Oudshoorn for contributing as a second reader.

I also gratefully thank my colleagues from Bouwgroep Dijkstra-Draisma. Foremost, my inhouse supervisor, Wietse de Vries, for his enthusiasm and guidance throughout this project and for providing me the opportunity to do my graduation assignment at BGDD. I furthermore want to thank Rutger and Coen for gladfully sharing their valuable insights and support, and for getting me up-and-running on this project, and Renze, for providing the emergency IT support even in evening hours and during the holidays.

With the prospect of completing my studies at the University of Twente, I want to thank everyone who made my time in Enschede so memorable. A special mention goes to Ilse, Magnus, het 52e, my housemates at Toko Chico and everyone at the D.R.V. Euros and the master corner.

I hope you find this report both enjoyable and informative, and invite you to keep me informed on any thoughts or suggestions.

Kind regards,

Wietse Harmsma

Table of Contents

Abbreviations

Bagging	Bootstrap aggregation
BGDD	“Bouwgroep Dijkstra-Draisma”
CART	Classification and regression tree
CM	Condition monitoring
CMFD	Condition monitoring and fault detection
COP	Coefficient of performance
CSR	Corporate Social Responsibility
EP	“Energieprestatie”
EPG	“Energieprestatiegarantie”; The guarantee that that a NOM-project meets the determined standards regarding energy use and isolation.
EPV	“Energieprestatievergoeding”; A fee landlords can charge tenants of social housing projects if EPG is given.
FDD	Fault detection & diagnosis
HDD	Heat Degree Day
HVAC	Heat, Ventilation & Air Conditioning
NOM	“Nul-op-de-meter”; Concept to describe that sustainable measures are taken in a social housing project in compliance to special Dutch regulations. (see “EPV”). Synonyms: “EPC-0”, “energienota-nul”, “energienotaloos”, “near zero energy buildings (nZEB)”.
OLS	Ordinary Least Squared
PCA	Principal component analysis
PCR	Principal component regression
PV	Photovoltaic
ROC	Receiver operating characteristic
S.D.	Standard deviation
SLA	Service-Level Agreement
SVM	Support vector machine

Management summary

Problem description

In this thesis, we aim to solve the problem of houses in 'Nul-op-de-meter' projects of Bouwgroep Dijkstra-Draisma to not meet the agreed energy requirements to be 'Nul'-op-de-meter' due to delayed detection of faults in the heating, ventilation and air conditioning (HVAC) and photovoltaic (PV) systems. We study 5 unique projects, which cover a total of 162 houses.

Current situation

First, we estimated the past consequences of delayed detection of faults in the PV and heating systems of the studied projects. To achieve this, we estimate the response of each system while not in fault. The 'normal' system responses are compared to the observed system responses during periods of failure, which are visually identified in readings from these systems' monitoring systems. Normal system responses are modelled using a nearest-neighbour heuristic, which uses the observed responses at locations that are physically in the neighbourhood of the candidate system as input, and piecewise ordinary least squares linear regression, which uses outdoor weather data as input.

We find that the largest consequences of detection delays relate to faults in the PV and heat pump systems. We find an estimated annual total consequence of delayed detection of faults to be about 2388 kWh, or 5% of the observed annual energy short on the service-level agreements. This estimate, however, considers only the detection delays that where visually identifiable in past system responses. Considering these results, we find that failures in the heat pump and PV systems are unlikely to be a major cause for not meeting the energy requirements in the SLAs.

Solution approach

We inquire the scientific literature for fault detection models in heat pump and PV systems. Recognizing the limited amount of available fault data, we also discuss techniques in which models are trained only on data from fault-free operation. For fault detection in PV systems' responses, we propose the implementation of the 'SolarClique' algorithm (Iyengar et al., 2018). For fault detection in heat pump systems, we propose the use of principal component regression (PCR) models.

We train SolarClique models to estimate the responses of fault-free operation of each of the studied PV systems. Likewise, we train PCR models to estimate the responses of fault-free operation of each of the studied heat pump systems.

By validating the developed models against independent validation data, and running statistical tests for heteroscedasticity, heterogeneity and normality, we find that most of the statistical test of a valid model are rejected. We propose that the statistical tests are rejected because of faulty system responses remaining in the validation data due to the limitations of our approach of isolating suspecting periods of failure in these data by visual approximation.

Although both the SolarClique and principal component models do not meet the requirements to be considered valid, we still apply the models to an independent test set, to see how well they perform as fault detection models. To move from regression to classification, we construct a decision interval centred on the regression estimates. For both model approaches, the width of the decision interval is determined by multiples of the sample standard deviation. The widths of the decision intervals are associated to some degree of likelihood that a future response, given the studied system is not in fault,

will fall within the decision interval. Therefore, if three consecutive observation of system energy use, or yield, fall outside of the decision interval, the model classifies the candidate system as being in fault.

Results

Both the SolarClique and the PCR models are ran for a range of widths of the decision intervals. The resulting models are scored on how well they reproduce the classifications assigned by visual approximation. We quantify the performance of both techniques using accuracy, precision and recall. Accuracy is the fraction of all observations correctly classified. Precision measures the fraction of positive prediction that is classified correctly. Recall measures the fraction of positive observations correctly identified by the model.

Both the dataset of visually marked faults of the PV systems and the dataset of visually marked faults of the heat pump system are greatly unbalanced. Of the labels assigned to the usage data of the PV system in the test set, 82% of the observations is visually marked as 'no fault' and 18% of the observations is visually marked as 'fault'. Of the labels assigned to the usage data of the heat pump system in the test set, 17% is visually marked as 'fault' and 83% is visually marked as 'non fault'.

The performance of both techniques is shown below;

Summary of the performance of the SolarClique models with 4, 3, 2 and 1 standard deviation decision intervals (S.D.) around the mean model estimate.

Interval	Accuracy	Precision	Recall
4 S.D.	0.82	0.04	0.01
3 S.D.	0.81	0.09	0.02
2 S.D.	0.78	0.13	0.06
1 S.D.	0.65	0.16	0.26

Summary of the performance of the PCR models with decision intervals around the mean model estimate with a 4, 3, 2 and 1 standard deviation decision intervals (S.D.) around the mean model estimate.

Interval	Accuracy	Precision	Recall
4 S.D.	0.83	0.00	0.00
3 S.D.	0.83	0.42	0.00
2 S.D.	0.83	0.24	0.01
1 S.D.	0.79	0.19	0.09

From these results, we make the follow conclusions:

- The models resulting from either approach do not perform notably better than a random classifier.
- For any width of the decision interval, the precision of the models' classifications is very low. This suggests that the 'fault' classifications given by these models are not a reliable indicator of a visually identifiable 'fault' occurring in the candidate system.
- For any width of the decision interval, also the recall of the models' classifications is very low. These models therefore also have no use for filtering which readings of the energy use of a heat pump system or energy yield of a PV system may be further studied in order to detect system faults that could also have visually been identified.

While we find little overlap between the energy yield or usage observations that are visually marked as being from a system in fault and the readings that are classified as anomalous by the models, we found several readings that are classified as 'fault' by either the SolarClique or PCR model and which, in hindsight, can be related to a sudden change in trend of the response. This indicates that the proposed visual approximation approach may also not be a good indicator of the ground-truth state of the studied systems. Without reliable classifications of the historic usage data, there is no way of knowing whether the models or the visual approximation approach cause these poor results.

Recommendations

The proposed SolarClique and PCR models, without further validation, should not be implemented for detection failures in an operational setting.

The primary challenge for developing and validating fault detection models in this case is the lack of reliable data on when the systems were historically in fault. Future efforts should be put in the collection of reliable data on the state of the monitored systems, including data on precisely when these systems where in fault.

Furthermore, we find that detection delays of failures in the heating and PV systems are unlikely to be a major cause for not meeting the SLA's. We recommend exploring other causes for the studied projects to not always meet the agreed energy requirements.

CHAPTER 1. **Introduction**

1.1 **Background information**

This report describes a graduation project performed at Bouwgroep Dijkstra-Draisma (BGDD). BGDD is a leading construction company in the north of the Netherlands. Its coverage is approximated by the area within a one-hour radius of Dokkum and Bolsward. Its operations cover most aspects of construction projects, including development, construction work, maintenance, renovation and reworking existing sites to be more sustainable. The nature of its projects is residential buildings, school-buildings, offices, factories, and gymnasia, among others.

BGDD is committed to its corporate social responsibility (CSR). It is awarded level 3 certification on the CSR performance ladder, thereby proving to adhere to the relevant standards of quality (International Organization for Standardization. (ISO), 2019d, 2019c), social responsibility (ISO., 2019a) and information security (ISO, 2019b), and is the proud holder of a variety of other certificates and leader in the Cobouw top 50 of 2020, proving its commitment to people, planet and profit (Bouwgroep Dijkstra-Draisma, n.d.).

The focus of this research is the net energy use of BGDD's "Nul-op-de-meter" (NOM) housing projects. A "Nul-op-de-meter" (NOM) project is a social housing project that meets the NOM-criteria. These criteria include that the buildings are well-isolated, have a low demand for heat and sustainably produce enough energy to fulfil demand for heating, cooling, ventilation and system monitoring (Rijksdienst voor Ondernemend Nederland, n.d.) . If a landlord can prove compliance with the NOM-criteria, thereby able to provide an 'energieprestatiegarantie' (EPG), the landlord is allowed to charge "energieprestatievergoeding" (EPV) as an incentive to invest in sustainable social housing. The concept has been introduced since 2013 and the number of NOM-houses in the Netherlands shows exponential growth (Stroomversnelling, 2019, 2020).

BGDD has embraced the concept of NOM-housing. It has delivered 423 houses for which EPG is provided and is working on another 524 NOM-houses. Recurring elements of these projects are the use of prefab isolated façades, installation of photovoltaic (PV) panels and, most notably, installation of gas free heating, ventilation and air condition (HVAC) systems.

With service-level agreements (SLAs) BGDD guarantees proprietors that its NOM-projects comply to the required criteria. This includes that, under normal circumstances, more energy is yielded from the PV panels than is required by the residents for space heating, domestic hot water (DHW) and household appliances. If the criteria in the SLA are not met, compensation is paid. In extend, BGDD is responsible for monitoring that its projects meet the NOM-criteria in practice, by collecting and analysing data on electric energy and heat usage.

The HVAC systems used in the NOM projects of BGDD are heat pump systems. A heat pump extracts heat from a source, moves and distributes this heat at a higher temperature (Hundy, 2016). Common sources of latent heat are outdoor air, ground, or a body of water. Despite generally having a high electric energy input to heat output ratio, heat pump systems use a large amount of electric energy to operate. To generate this energy in a sustainable way, and reduce net electric energy use, the NOM projects of BGDD are equipped with photovoltaic (PV) cells. PV systems collect solar irradiation energy and transform this energy into electric energy.

1.2 Research motivation

1.2.1 Problem description

When committing to a NOM project, BGDD initiates an SLA with the involved housing corporation. These SLAs state a guaranteed amount of electric yield from the PV installation and a maximum for the electric energy used by the heating system for every involved house. In practice, the SLAs on BGDD's NOM projects are not always fulfilled. This means that BGDD is accountable to pay compensations relative to the amount of electric energy short of the agreed service-level. Naturally, it is in the interest of BGDD to limit these occurrences.

The magnitude of the problem becomes apparent when we look at the amounts of energy short of the agreed service-level in past years. For the 5 projects studied in this thesis, BGDD was liable for an average annual total of 45,000 kWh which was yielded less from the PV systems or used more by the heating systems than stated in the SLAs. This corresponds to a value of about € 9.450,-. According to BGDD, past cases of not meeting the SLAs could predominantly be attributed to either one or a combination of failures and/or flaws in the designs or realization of the heating system.

Some failures greatly affect the net energy use of a house. Most notably, a PV system that is in fault is no longer able to yield the required amount of electric energy. On the other hand, a fault in the heating system generally causes the heating system to seize function, making it use less energy, not more. However, some faults in the heating system, e.g., evaporator or compressor fouling, manifest as a reduction of energy efficiency.

Besides the nature of the failure, the consequences of failure also depend on the duration of the failure. We distinguish two phases of the failure period the 'detection delay' and the 'logistic delay and repair time'. The 'detection delay' is defined as the time after a fault has occurred that is required for BGDD to become aware of the fault. The 'logistic delay and repair time' is the time required by BGDD to complete a service activity after they are notified of the fault.

Another commonly heard explanation within BGDD for not meeting the energy requirements stated in the SLAs is summarized as 'unjust assumptions in the systems' designs'. Design of the heating and PV system and the service levels in the SLAs and are founded on assumptions on the performance of the equipment, thermal properties of the building and usage of the residents. These assumptions are based on *a priori* inductive models, past experiences and information provided by material and system suppliers. Among other things, practice shows that model outcomes are sometimes misinterpreted, models may be too simplistic to capture the complex dynamics of residents' behaviors and information from suppliers tends to be overly optimistic. As a result, the realized systems' performances may differ from how they were designed, so that the systems are incapable to meet the expected performance immediately after deployment.

1.2.2 Problem statement

The goal of this project is to reduce the annual energy deficiency that BGDD is liable for. Following Heerkens & van Winden (2012), we define the problem as a deviation between a norm and a reality, in this format we define the problem as follows:

> *"The annual energy use of the heating systems exceeds and/or the annual energy yield of the photovoltaic systems is less than the amounts agreed in the service level agreements."*

We have summarized the causal structure of the problems in a "problem cluster" (Heerkens & van Winden, 2012). This problem cluster is shown as Figure 1. At the top level of the problem cluster we have the problem statement. Following the accounts given by BGDD, the main causes for not meeting

the energy requirements in the SLAs are unjust assumptions in the system design and failures.

It is unknown what the respective contributions of each of these causes is to not meetings the SLAs, however, there is a widespread belief within BGDD that failures in the system play a large role. In this thesis, we adopt this assumption and focus on solving the limited availability of the heating and PV system. In reliability engineering, the availability of a system is dictated by the frequency and duration of a failure. So, the identified causes for the low availability of the heating and PV systems in BGDD's NOM projects are a too long duration of failures and a too high frequency of failure.

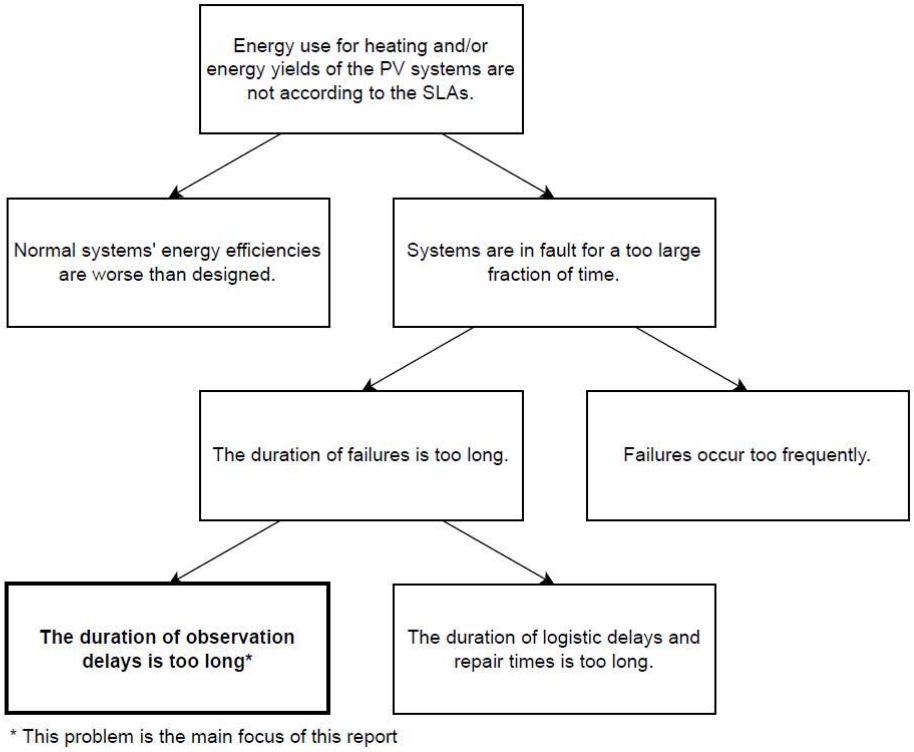

Figure 1 Problem cluster for this thesis, the core problem tackled in this thesis is marked in bold.

In this thesis we aim to reduce the duration of failures. We recognize that the duration of a failure is determined by two distinct periods of time. First, a failure in the heating or PV system needs to be reported to BGDD, we refer to this as the 'observation delay'. Once a failure is reported, there is some time required for BGDD to realize the repair.

BGDD is primarily reliant on the residents to detect and report failures. According to BGDD, however, some failures are not recognized by the residents, because the utility of a failed subsystem is taken over by another system. This is thought of as the main reason for long detection delays and the core problem that we aim to solve in this thesis.

One case of failures which are masked by some other system, described by BGDD, is that of PV failure which is masked by the energy supply of the electric grid. PV panels are installed in all BGDD's NOM projects, but the NOM-projects are also connected to the electric grid, which supplies the NOM-houses of electricity when demand exceeds power output from the PV panels, at night, in winter or during peak consumption. A failure in the PV installation may cause partial or complete loss of PV capacity, but this reduction in PV yield is compensated with energy from the electric grid and therefore

often remains unnoticed by residents.

A second case, described by BGDD, is heat pump failure which is masked by the heat supplied by the auxiliary heater. In most systems, the reduction of heating capacity is compensated by an increased use of the auxiliary heater, which is to assist the system when it is not able to produce the required heat or is in defrost mode. Since the energy efficiency of the auxiliary heater is much lower than that of the heat pump, a fault in the heat pump often translates to a reduction of the energy efficiency and an increase in system energy use. As this does not affect the capacity of the system to deliver heat, it is generally unnoticed by the residents, until the energy bill is received. This problem is not unique to BGDD for residential buildings, Lazzarin & Noro (2018a) observed a similar problem at a school building.

We do not explore the possibilities for reducing the frequency of failures, as, according to the accounts of BGDD, the dominant issue is that failures are not always reported as they occur. Furthermore, the systems are relatively new and an individual system is installed at each location. This means that common solutions for reducing the frequencies of failure, such as increasing the amount of preventive maintenance or redesigning the system, are likely costly whilst having minimal effect.

1.2.3 Case description

We look at 5 unique NOM-projects of BGDD, which contain a total of 162 homes. Every house is equipped with a PV system for power output, a connection to the electric grid and an electric heating installation. All houses in this project are equipped with an individual heat pump installation.

Of the studied heat pumps, 2 are air ventilation systems and 3 are geothermal systems. Furthermore, the systems are installed in a variety of configurations and from a variety of manufacturers. An overview of the considered projects in shown in Appendix A.

1.2.4 Research objective

The goal of this project is to reduce the duration of failures by reducing the time required to detect these failures. We propose to use a statistical failure detection model to achieve this. Each house in BGDD's NOM-housing projects is equipped with a monitoring system. The monitoring systems are designed to track the PV power output, electric energy used for space heating and DHW and the heat energy generated. The readings of the monitoring systems are used as input for the classification models.

If successful, the failure detection model is integrated into a dashboard, which can be accessed by BGDD's after-sales service department. They are then able to contact residents and proactively check for hidden faults or issues with the system design and repair the failure so that less energy is lost on inefficient operation of the system in fault.

1.3 Methodology

In this paragraph, we discuss how we aim to solve the problem of not meeting the energy requirement in the SLAs due to undetected failures. The chapters are provided in sequential order.

1.3.1 Chapter 2: Current situation

As shown in Figure 1, we have identified multiple causes that may lead to not meeting the energy requirements stated in the SLAs. The goal of the first part of this research is determine what amount of energy yield less or used more than agreed in the SLAs can be attributed to detection delays specifically.

From monitoring the systems, we collected data on the historic energy yield of the PV system and use

of the heating system. So, to determine the contribution of the detection delay, we compare the realized energy yield or use with an estimated yield or use for the systems if they would not have been in fault. We answer the following research questions:

1. What amount of energy yield less or used more than agreed in the SLAs can be attributed to detection delays?
   1.a. When was a system in fault and not reported to BGDD?
   1.b. How much energy would have been yielded or used or yielded if the systems would not have been in fault during these periods?

To answer these questions, we analyse historic service reports and monitoring data. This includes preparations like pre-processing, labelling, and pairing the monitoring data to the service reports. Furthermore, we develop a data-driven regression model for the normal energy yield or use of every monitored system.

1.3.2 Chapter 3: Literature Review

In the next phase, we inquire the scientific literature for failure detection techniques that can be used for this project. In this phase we answer the following research questions:

2. Literature review: Which failure detection techniques proposed in literature can we apply for detecting system failure in photovoltaic and heat pump systems?
   2.a. Which failure detection techniques are adopted to detect system failures in photovoltaic systems?
   2.b. Which failure detection techniques are adopted to detect failures in heat pump systems?
   2.c. Which failure detection models are suitable for failure detection in the studied case?

1.3.3 Chapter 4: Model specification and validation

Having studied the current situation and which failure detection techniques are used in literature, we select the techniques that seem most suitable to detect failures in the systems studied in this thesis. We construct models using these techniques and validate these models against historic observations. The result of this chapter is a collection of failure detection models for PV and heat pump systems.

1.3.4 Chapter 5: Test results

Next, we evaluate the performance of the failure detection models on the test set. We answer the following questions:

3. Can implementation of the proposed failure detection models improve failures detection in the studied systems' responses?

1.3.5 Chapter 6: Conclusion and recommendations

Finally, we conclude the report by stipulating the key findings. We also state future steps to be taken by BGDD to solve the problem of not meeting the energy requirements stated in the service level agreements. We answer the following questions:

4. What future steps are to be taken by BGDD to reduce annual energy short of amount stated in the SLAs?
   4.a. What conditions are required for implementation of the proposed models?
   4.b. What further research effort should be realized by BGDD?

CHAPTER 2. **Current situation**

2.1 **Introduction**

For the 5 projects studied in this thesis, BGDD was liable for an average annual total of 51,904 kWh in 2018, 41,332 kWh in 2019 and 40,227 kWh in 2020 which was yielded less from the PV systems or used more by the heating systems than stated in the SLAs. This corresponds to a value of about € 10,900. -, € 8,680. - and, € 8,448. – respectively. In this chapter, we aim to quantify which part of energy these shorts can be attributed to detection delays in particular.

To answer the research question, we answer each of the sub-questions related to this research question:

1.a. When was a system in fault and not reported to BGDD?
1.b. How much energy would have been yielded or used or yielded if the systems would not have been in fault during these periods?

We adopt a 3 part approach. To visualize this approach, the cumulative energy use of a single system under failure is shown in Figure 2.

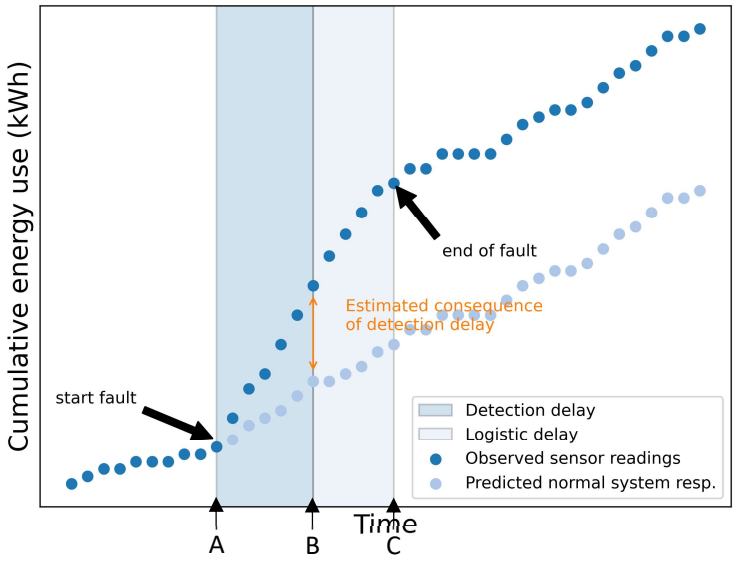

Figure 2 Energy loss principle due to fault masking, with A, the time the fault started, B, the time BGDD is notified of the fault and the order is created and C, the time the order is completed.

1. First, we estimate when detection delays have occurred for every past instance of failure. For the instance depicted in Figure 2 this means we estimate the location of label A.
2. Next, we make an estimate of the energy yield or use for the monitored systems when not in fault, which we refer to as the 'normal system response'. This is analogues to approximating the light blue dotted line in Figure 2. We also compute the cumulative difference between the observed energy use, or yield, and the normal system response, which is analogous to determining the length of the orange arrow in Figure 2.
3. Finally, we add all the estimated consequences of individual fault detection delays during a year to estimate the annual consequence of fault detection delay.

This chapter is organized as follows. In section 2.2, we will introduce which data is available for our analysis. Sections 2.3 and 2.4 address the first 2 steps of the 3 part approach and answer the sub-

questions related to the research question. The results are further discussed in section 2.5 and 2.6.

2.2 Available data

For the analysis, we make use of three separate data sources: (1) the service orders, (2) the monitoring data and (3) weather data. We discuss each of these sources and the respective transformations that we made to the data in these sources to improve useability.

2.2.1 Service orders

The service order data is extracted from BGDD's ERP system. Service orders are a log of most processes that occur at the after-sales service department of BGDD. A service order is created every time a resident or housing corporation notifies BGDD of an issue that cannot be immediately resolved or when a service mechanic is scheduled. These orders report most maintenance activities, whether executed by one of BGDD's service mechanics or by a third party. A service order contains the following relevant information:

1. The address at which the service was executed, each service order strictly serves a single address.
2. A problem description by the resident or housing corporation in case of a perceived problem, or a task description by the service department in case of preventive maintenance.
3. A description of the taken actions by the dedicated service mechanic.
4. The date and time the service order was created.
5. The date the service order was completed.

This dataset therefore offers a fairly complete view of the maintenance history of each system.

2.2.2 Usage data

The second available dataset consists of the usage data. The usage data are the data collected from readings of the monitoring systems. Features in this dataset include the total amount of heat supplied, electric energy used by various parts of the heating system and yield from the PV system. A complete enumeration of the monitored features is listed in Table 4 of Appendix B**.**

For every house, the applicable readings are stored in time series. Time stamps are given to a minute precision. The sensors are read at irregular intervals and the time between consecutive readings at a single sensor is about 50 minutes. Moreover, the dataset contains only the last reading of every day for each respective sensor. In practice, this means that most of the sensor readings in the monitoring data is timed between 11 and 12 PM of a given day. For simplicity, we map the sensor readings to evenly spaced intervals, at 12 PM of every day.

2.2.3 Weather data

The third available dataset contains historic weather data. The weather data are retrieved from the publicly available database of the Royal Dutch Meteorological Institute (KNMI). The weather data used for any specific house is that of the KNMI weather station nearest to that house. The weather data contain the following features for each weather station:

- the outdoor temperature in degrees Celsius.
- wind speed in 0.1 m per second
- solar irradiation in J/cm^2
- wind chill in degrees Celsius equivalent

These features are previously selected by BGDD as they are considered important for the forecast of the heat demand. Readings are collected as daily averages and on a daily interval and stored as time series.

For the studied houses, the nearest weather station is either located in Leeuwarden (53.22°N, 5.75°E) or Eelde (53.12°N, 6.58°E). For some houses, the distance to the nearest weather station is considerable, which increases the probability that the weather conditions at the studied houses differs from that observed at the weather station.

2.3 Marking the fault period

In this subsection, we discuss how we determined when the studied systems where historically in fault, but BGDD was not yet notified of the fault. These periods are coined the 'fault detection delay'.

2.3.1 Method

To identify any period that constitutes a detection delay, we require both the moment that a system first showed a fault and the moment that the fault became known by BGDD. We assume that a service order is created immediately as BGDD is notified of a failure. The moments that faults became known by BGDD are therefore retrieved from the service order data, being the timestamps on which the orders where created.

On the other hand, we have no direct insight in the moments that a system first showed a failure. To approximate this moment, we use the fact that the faults relevant to in this project affects the energy use of the heating system or energy yield of the PV system. Therefore, to estimate the moments a fault starts, we attempt to visually mark changes in the trend of the system readings; For each service order, we plot the sensor readings about the service order's creation date. This is done for all sensors installed in the house at the address of the service order. In these plots, we mark any visual change in trend that appears to precede the order. A sample of such a plot, and the marked fault start are shown as Figure 3. The formal implementation of this heuristic is given as pseudocode in Appendix C.

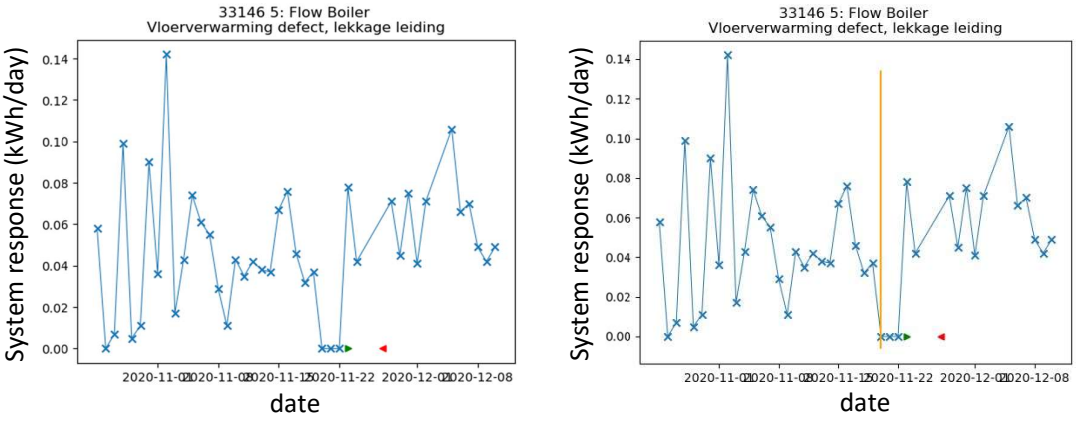

Figure 3 (left) a sample plot with a trend change preceding the order creation (green triangle) and the order completion date (red triangle) and (right) the fault start marked in this plot.

Sometimes, it is not feasible to visually identify a change in trend. One obvious reason is that the fault does not affect the process monitored by the sensor that is plotted. Another reason is that the fault is resolved within a single day, so that the change in trend cannot be seen on a daily precision. In either case, the consequence of that fault on the energy yield or use by the apparatus monitored by that sensor is expected to be minimal and therefore is expected to have a limited effect on the outcome of this analysis. Another option is that, regardless of the care that was put in the visual inspection, a clear change in trend was missed.

2.3.2 **Results**

We were able to visually distinguish a detection delay for 78 of the 224 service orders. A grid plot showing which part of the monitoring data is marked as representing a fault period is shown in Appendix H. Figure 4 shows for which service order categories we most commonly were able to distinguish a detection delay of more than one day. Having a detection delay of more than one day means that (1) such faults can be visually recognized from the sensor readings and (2) these faults are not immediately communicated, possibly because of a failure masking mechanism.

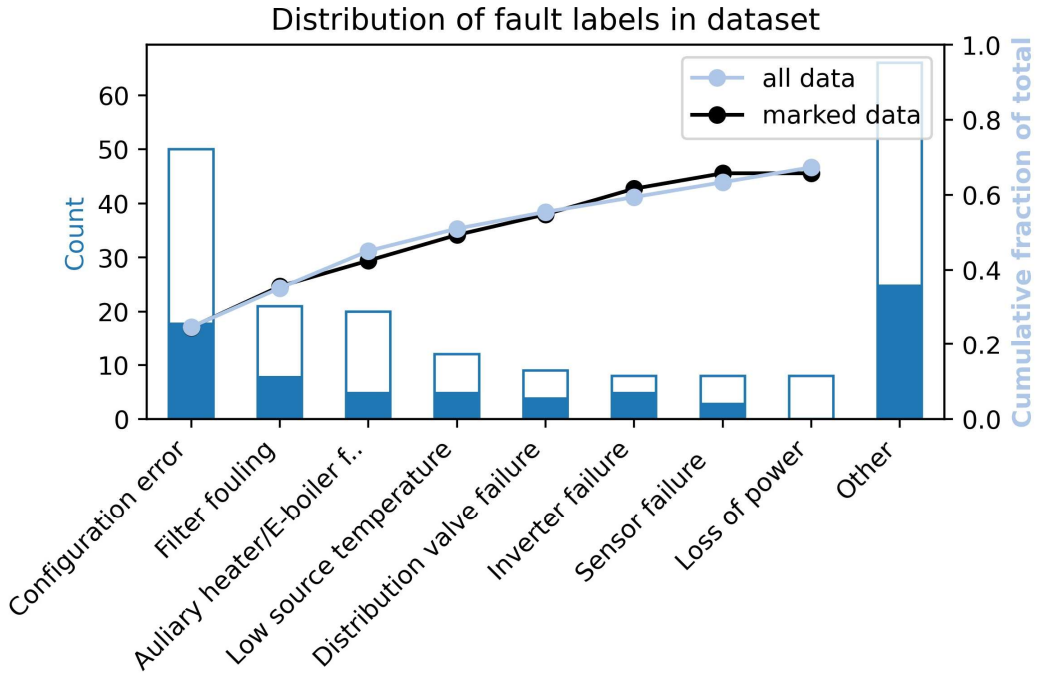

Figure 4 Distribution of the marked service order (filled) vs all service orders (opaque) into the fault categories of Appendix M. 'Other', here, denotes all service orders that occurred less than 5 times in the original service orders.

We compute the duration of the detection delay of the most labelled faults. The results are shown in Figure 5. We find that, of the visually recognized cases occurring more than 5 times during the monitored period, inverter failures, on average, take longest to be reported by the residents, followed by faults in the heat pump, such as configuration errors of the settings and filter fouling.

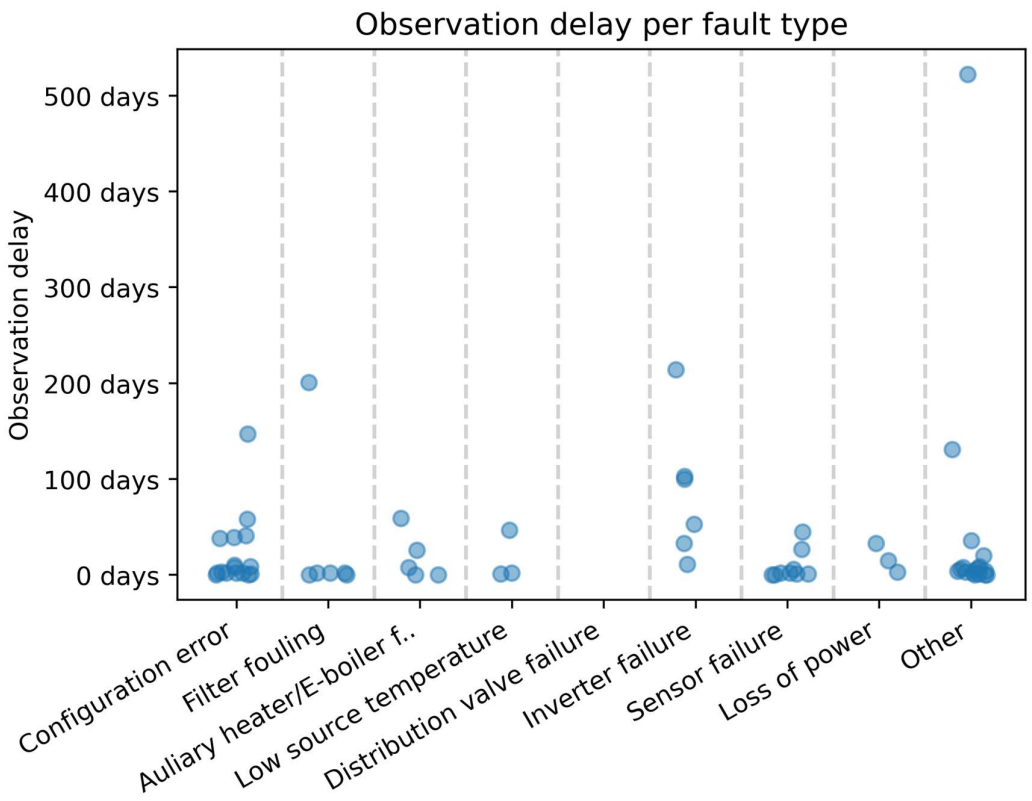

Figure 5 Scatterplot of the detection delay of the labelled faults. 'Other', here, denotes all service orders that occurred less than 5 times in the original service orders. Some jitter has been applied across the x-axis in order to visualize overlapping points.

A single large detection delay of 522 days was observed in the category 'Other'; this involved a source ventilator failure. The estimated aggregate annual number of readings during a detection delays is 806 readings/year. This means that this fault was neither visually identified in the system's usage data as it occurred, nor was it recognized during annual billing. This can only be the case if the impact of this fault to the energy use of the system is limited, which is indeed observed.

2.4 Estimating the normal system response

The second step in estimating the consequences of fault detection delay is to develop a model for the energy yield or use for the monitored systems when not in fault, which we refer to as the 'normal system response'.

2.4.1 Method

In this subsection, we introduce two relatively simple models to estimate the normal system response. This models are complementary, as they rely on different principles for estimating the normal system response of a candidate system.

Our first approach is A K-Nearest Neighbour (K-NN) implementation, which uses sensor readings from other houses with similar building characteristics as input. This approach exploits the assumption that the PV and heating systems may operate similarly to other houses with the same building characteristics and a similar geographical location. We propose a method that estimates the normal system response as the average of the concurrent sensor readings of the K systems that historically

are most similar to the specified sensor readings. Pseudocode for our implementation of the K-NN heuristic is shown as Figure 19 in Appendix D.

By visually estimating the periods that some apparatus was in fault and removing these periods from the usage data, we gain a dataset that hopefully contains mostly sensor readings from fault-free apparatus. On this 'fault-free' usage data, the correlation analysis with the contemporary weather data observations was performed to find how much the weather conditions affect the energy use of the heating system and yield of the PV system. Our second approach is backed by the results of a correlation analysis, which are found in Appendix E. The results show a strong negative correlation (r= -0.81 to -0.64) between the wind chill measured at the nearest weather station and the energy use of the heat pump as well as a strong positive correlation (r= 0.91 to 0.96) between the energy yield from the PV system and the solar irradiance measured at the nearest weather station. Our second approach is an ordinary least squares (OLS) simple linear regression approach, using weather data as an input.

2.4.1.1 **K-Nearest Neighbour implementation**

To determine which systems most closely relate to the candidate system, we compute the distance between the past responses of the candidate system and the past responses of each of the systems that are of the same project and type as the candidate system. Having already mapped all sensor readings to the same time intervals allow us to use a distance metric that determines distance by comparing only concurrent sensor readings of multiple systems. We adapt the intervector Euclidean distance metric as formulated in Chicco et al. (2006).

Equation 1 Euclidean distance between two vectors of past sensor readings, x and y, with H the length of either vector (Chicco et al., 2006).

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\frac{1}{H}\sum_{h=1}^{H}(x_h - y_h)^2}$$

Next, we decide how many neighbors to consider. With a too small value for K, the estimate is dependent on just a few signals and therefore not robust in the case a neighbor shows irregular sensor readings. If K is too large, however, it is unlikely that there is still sufficient similarity between the studied signal and all the selected neighbors. At last, it may occur that there are less than K neighbors available. If less than K neighbors are available, we use all the available neighbors. In this section, we set K to 3, which was found to be less than the available number of suitable neighbors in most of our cases, yet returned fairly accurate estimates for most cases.

Finally, for every individual apparatus that is monitored, a 3-Nearest Neighbour regression is calculated to predict the normal system response. As an input, this regression uses the concurrent energy yield or use of the 3 apparatus of the same type in other houses from the same project that show the response closest to the candidate apparatus. The distance between responses is calculated using Equation 1. The predicted response is equal to the pointwise unweighted average of the selected neighbors. If any of the selected responses has a missing value, the estimate for that respective day is returned empty. If three or less neighbors are available, the unweighted average of the responses of all the available neighbors is returned.

2.4.1.2 **(Piecewise) simple linear regression model**

2.4.1.2.1 PV system

For every monitored PV system, an ordinary linear regression is calculated to predict the electric energy yield from the respective system using only the solar irradiance, given in kJ/m^2, measured at the weather station that is geographically nearest to that system. The energy yield of the normal

system response is estimated by Equation 2. The values for β_0 and β_1 are found by ordinary least squares approximation.

Equation 2 Simple linear regression, where $\hat{y}$ is the estimated energy yield by the PV system in kWh, x is the solar irradiance measured at the geographically nearest weather station in kJ/m^2, β_0 is the intercept and β_1 is a coefficient.

$$\hat{y} = \beta_0 + \beta_1 x$$

2.4.1.2.2 Heating system

Similar to the energy yield of the PV system, the energy use of a heating apparatus cannot be negative. However, while the levels of observed solar irradiance are strictly positive, the levels that wind chill can take are unbounded. Moreover, in the case of the heat pump energy use, there is baselevel of energy used to heat domestic hot water (DHW). Above a certain wind chill, space heating is no longer required, so that the electric energy uses of the heating system flattens to the amount used for heating DHW, which is more or less constant across time and independent of wind chill. This implies that past a certain wind chill level, the response of the system is no longer dependent on the observed wind chill. An example of this response pattern is shown on the top of Figure 6.

Alternatively, there are several heat pump apparatuses that are equipped to function as a cooling installation by reversing the heating cycle. This means that once the wind chill levels exceed the point at which space heating is required, the function of the heat pump apparatus changes to cooling. As the wind chill increases further, the cooling load will increase, and the electric energy use of the heat pump increases accordingly. A sample of the electric energy use of such systems is shown at the bottom of Figure 6.

The level of wind chill at which space heating is no longer required is dependent on the internal heat load of a house and is therefore unique for every house. However, a general approximation for this temperature, can be referenced by the base temperature for a “Heat Degree Day” (HDD) as calculated by the KNMI, which is 14 degrees Celsius equivalent (Wever, 2008). We initially adopt this 14 degrees equivalent as the critical value.

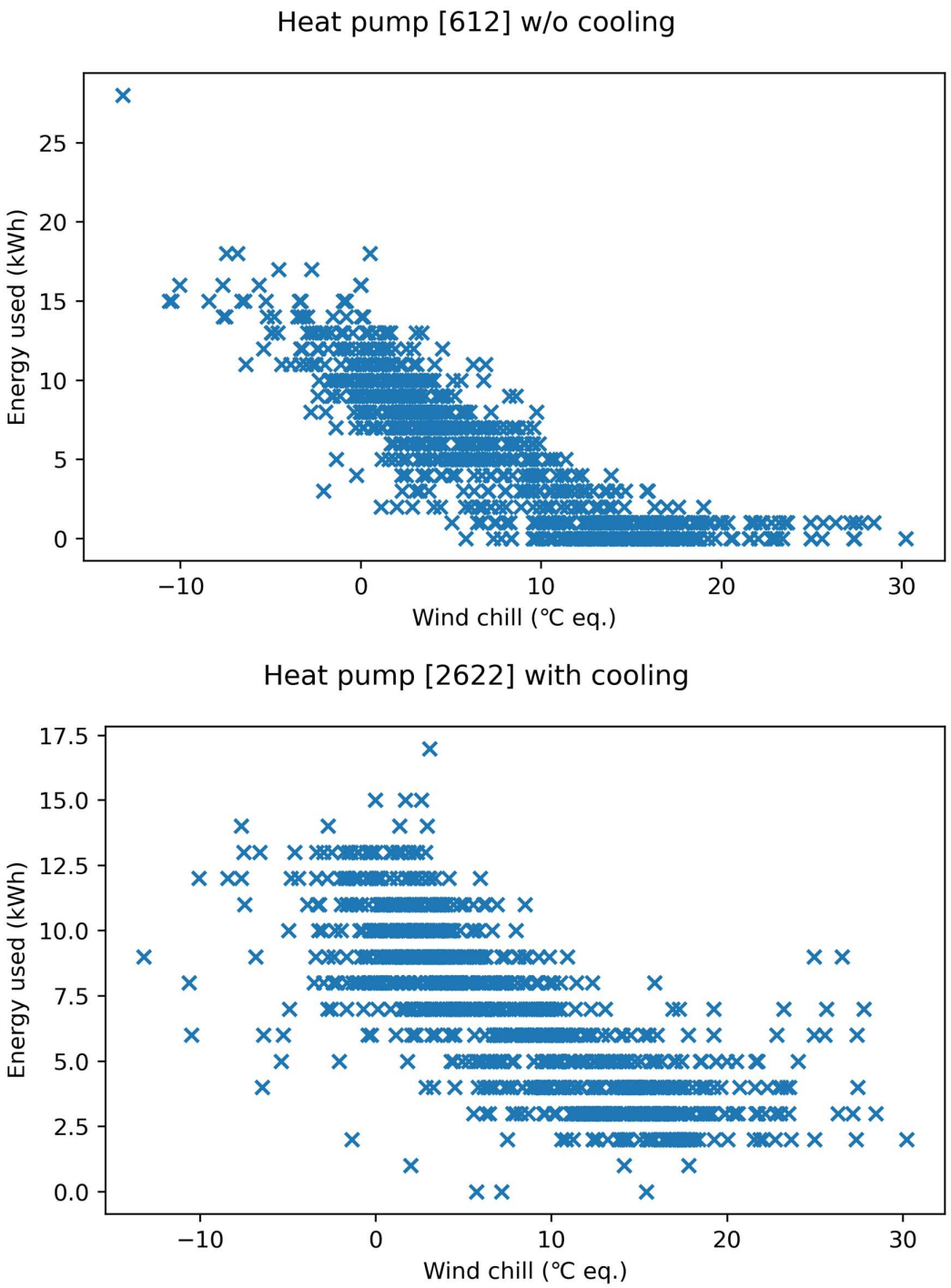

Figure 6 Energy use observations of (top) a heat pump without cooling system and (bottom) a heat pump with cooling system, plotted against wind chill observations.

We model the described dynamic using a piecewise linear regression. Initially, two separate linear regression models are defined; the first is defined for a wind chill of 14 degrees Celsius equivalent or higher and a second model is defined for a wind chill lower than 14 degrees Celsius equivalent. For every individual apparatus that is not a PV system, such a piecewise linear regression is calculated to predict its energy use, based on the wind chill measured at the weather station that is geographically

nearest to that apparatus and given in degrees Celsius equivalent. The initial prediction for energy use is calculated using Equation 3. The values for β_{00}, β_{01}, β_{01} and β_{11} are found by ordinary least squares approximation on the individual segments.

Equation 3 Discontinuous piecewise linear regression model where x is the wind chill in degrees Celsius equivalent, y is the estimated energy use of the candidate apparatus, β_{00} the intercept of the first segment at x=0, β_{10} the coefficient of the independent variable of the first segment. β_{01} the intercept of the second segment at x=14, β_{11} the coefficient of the independent variable of the second segment.

$$y = \begin{cases} \beta_{00} + \beta_{10}x & if\ x \leq 14 \\ \beta_{01} + \beta_{11}x & if\ x > 14 \end{cases}$$

After solving Equation 3, an adjustment is made to the model to improve performance when the true level of wind chill at which space heating is no longer required differs from the initial 14 degrees Celsius equivalent. Due to the nature of the processes, the slope of the first segment (β_{10}) is always negative, whereas the slope of the second segment (β_{11}) is either positive, in case of cooling, or, if not cooling, less steep than the slope of the first segment. If we interpolate both lines, these intersect approximately at the true level of wind chill at which space heating is no longer required. Therefore, instead of Equation 3, Equation 4 is used to estimate the energy use of the subject apparatus. In this way we gain a continuous function that better fits the true level of wind chill at which space heating is no longer required.

Equation 4 Discontinuous piecewise linear where x is the wind chill in degrees Celsius equivalent, y is the estimated energy use of the candidate apparatus, β_{00} the intercept of the first segment at x=0, β_{10} the coefficient of the independent variable of the first segment. β_{01} the intercept of the second segment at x=14, β_{11} the coefficient of the independent variable of the second segment.

$$y = max \begin{cases} \beta_{00} + \beta_{10}x \\ \beta_{01} + \beta_{11}x \end{cases}$$

As a result, some error is introduced in the OLS approximation, as some of the samples in the training data are used in the approximation of the coefficients for the wrong segment. We argue that this is not a problem, since the outcomes of these models will only be used as a rough estimate and the consequences of these errors should be minimal because the initial assumption of 14 degrees Celsius equivalent will be relatively close to the true estimate and either segment is expected to have sufficient training data to compensate for a few outliers.

2.4.1.3 **Model validation**

2.4.1.3.1 Validation approach

2.4.1.3.1.1. Measure of goodness-of-fit

Several methods exist to evaluate the goodness-of-fit of the models' estimates to the observed values. These include common metrics such as the Mean Squared Error (MSE), the Mean Absolute Percentage Error (MAPE) and the estimation bias.

With the MSE, the squared errors are averaged across estimates, which causes this measure to be sensitive to outlier sensor readings. As we expect to not have visually identified all fault periods in the usage data, it is likely that some outlier readings remain. To overcome this, we adopt a robust measure of spread that is less sensitive to outliers, the Median Absolute Deviation (MedAE), which uses Equation 5.

Equation 5 Absolute Error with y the vector of observed sensor readings, where $\hat{y}_i$ is the i-th element in the vector of sensor reading estimates, y_i is the i-th element in the vector of observed sensor readings and n is the number of elements in each vector.

$$MedAE(y, \hat{y}) = median(|y_1 - \widehat{y_1}|, |y_2 - \widehat{y_2}|, \dots, |y_n - \widehat{y_n}|)$$

To compare model accuracy across projects and different systems, we require a scaled measure, such as the MAPE. The MAPE, however, is sensitive when applied to small values and is undefined when the observed sensor reading is zero (Rob J Hyndman & Koehler, 2006). To overcome these limitations, we adopt the Mean Archtangent Absolute Percentage Error, which is given by Equation 6 (Kim & Kim, 2016).

Equation 6 Mean Archtangent Absolute Percentage Error (Kim & Kim, 2016), with A the vector of observed sensor readings and F, the vector of estimates. If the A_t denominator is 0, the fraction is approximated as infinity.

$$MAAPE = \frac{1}{N}\sum_{t=1}^{N} \arctan\left(\left|\frac{A_t - F_t}{A_t}\right|\right)$$

2.4.1.3.1.2. Validation heuristic

To validate the models, 5-fold cross validation is performed using the 'fault-free' usage data. That is the usage data without the visually identified fault periods. After computing a MedAE and MAAPE for every iteration in the 5-fold cross-validation, the measures are averaged across iterations so that a single MAAPE and MedAE are computed both for the estimates of the 3NN models and for the OLS models of every one of the 156 monitored PV and 152 monitored heat pump systems.

The 3-Nearest Neighbour models are trained on contemporary observations of other systems. The training data in the 5-fold cross validation approach are therefore essentially meaningless for this heuristic. Still, for comparability between the two estimation approaches, we use the same validation data for either estimation approach.

2.4.1.3.2 Validation conclusions

The results of running a cross-validation on each of the studied PV, heat pump and auxiliary heating systems are described in Appendix F. The distributions of the parameter-levels fitted to the (piecewise) linear OLS regression models to estimate the normal system response of the PV systems, heat pump systems and auxiliary heater systems are shown in Figure 32, Figure 33 and Figure 34 of Appendix G respectively. No such figures can be made for the nearest neighbour models, since these models are nonparametric.

In summary, we find that most nearest neighbour estimates have a lower MedAE and MAAPE than any OLS estimates when estimating the normal system response of the PV systems. In contrast, we do not find such differences between the nearest neighbour and OLS estimates for the normal system responses of the heat pump or auxiliary heater systems.

There are a few estimates with high MedAE. We find these to be the result of periods of no or little energy use in either the training or validation data. We explain these observations as visually unrecognized system failures, or otherwise deviating response that do not fit the normal system response. The quality of the estimates of these systems cannot be guaranteed and these models should therefore not be used to estimate the consequences of failure.

For the PV systems and the auxiliary heater systems, the systems for which the poor estimates are made do not occur in the service orders for which we wish to estimate the normal system response.

For the heat pump systems, the nearest neighbour models return poor estimates for 6 systems for which we also require a normal system response estimate. The OLS models return poor estimates for 5 systems for which we require a normal system response estimate.

From these results, we draw the following conclusions;

The 3-NN model is the preferred model for estimating the power output of the nominal response of the PV system, as it is more accurate than the OLS model. We find the 3-NN estimates to be sufficiently accurate to use for providing a rough estimate for the consequences of failure; the median absolute error is less than 3 kWh per day for most 3-NN models regardless of the project studied, which is about 17.1% of the average daily yield throughout a year. Moreover, the median of the median absolute error is about 1 kWh per day or 5.7% of the average daily yield throughout a year.

For the heat pump apparatuses, we cannot make a clear distinction of which model performs best. We therefore take the average of both models unless one of the models performs poorly, which potentially adds to the robustness of the estimates as individual errors are cancelled out. If we look at the median of the MedAE, the models for energy use of the heat pumps are slightly less accurate than that of the 3-NN models for energy yield of the PV system. However, the observed MedAE are still mostly less than 3 kWh per day, or about 36.3% of the average daily energy use throughout a year, and the median of the MedAE per project is about 1.2, or 14.5% of the average daily energy use throughout a year, which we deem sufficiently accurate for providing a rough estimate of the consequences of failure.

Based on these findings:

- The normal system response of the PV system is estimated by the 3-nearest neighbour heuristic.
- The normal system response of the heat pump systems is estimated by the pointwise average of the 3-nearest neighbour heuristic and piecewise simple linear regression estimates.

The service orders related to systems for which either a poor 3-NN or OLS estimate is made are shown in Table 5 and Table 6 of Appendix F. Since the estimates from these models are not accurate, these are not included in our results. Instead, we include a manual estimate of the consequences of these service orders:

- Some service orders describe faults that do not logically affect the energy use of the heating system, we estimate no consequences of these faults to the energy use of the heat pump.
- For the other service orders, it is either the 3-NN or the OLS model that has a poor performance, but never both. If the 3-NN model performance poorly, we therefore use only the OLS model and vice versa. The total additional annual consequence of these cases is 168 kWh, of which a 163 kWh contribution of a single failure in project 33146 is most notable.

2.4.1.3.3 Estimating a confidence interval of the annual consequences of fault detection delay

To model the uncertainty, we compute a confidence interval over the final estimates of annual energy short. We first make the assumption that the estimation errors are i.i.d. and follow a normal distribution, which is an implicit assumption of the OLS linear regression model, but not necessarily the case with the 3-nearest neighbour regression model. Then, we use the computed MedAE's and Equation 7 to estimate the sample standard deviation.

Equation 7 Sample standard deviation (σ_1) estimate using formula for the modified z-score, with MedAE is the median absolute deviation (Iglewicz & Hoaglin, 1993; Rousseeuw & Croux, 1993)

$$\hat{\sigma}_1 \approx \frac{1}{\Phi^{-1}(3/4)} * MedAE \approx 1.4826 * MedAE$$

We then estimate the standard deviation of the annual energy short by entering the sample standard deviation estimate of Equation 7 into Equation 8.

Equation 8 Estimated standard deviation of energy short, where σ_1 is the sample standard deviation of the estimation error of a single system and n is the average number of days of observation delays per year

$$\hat{\sigma}_{year} = \hat{\sigma}_1 * \sqrt{n}$$

2.4.2 **Results**

For each visually identified fault detection delay, we compute the difference between the estimated normal system response and the observed response. The K-NN and OLS models are used to estimate the normal system response for every day marked as detection delay. We then compare these estimates to the observed energy use of the heating system and yield from the PV system.

As we are strictly interested in the adverse effect of faults to the ability to meet the SLAs, we only consider the observations for which the observed energy consumed by an apparatus exceeds the estimate or, in case of the PV system, the realized yield is less than estimated.

The total consequence of observation delay to not meeting the energy requirements of the SLAs, expressed in kWh of electric energy, is estimated by the sum of all the remaining differences between the observed energy use, or yield, and the estimated normal response of the respective system. The results, normalized to an annual average, are shown in Figure 7.

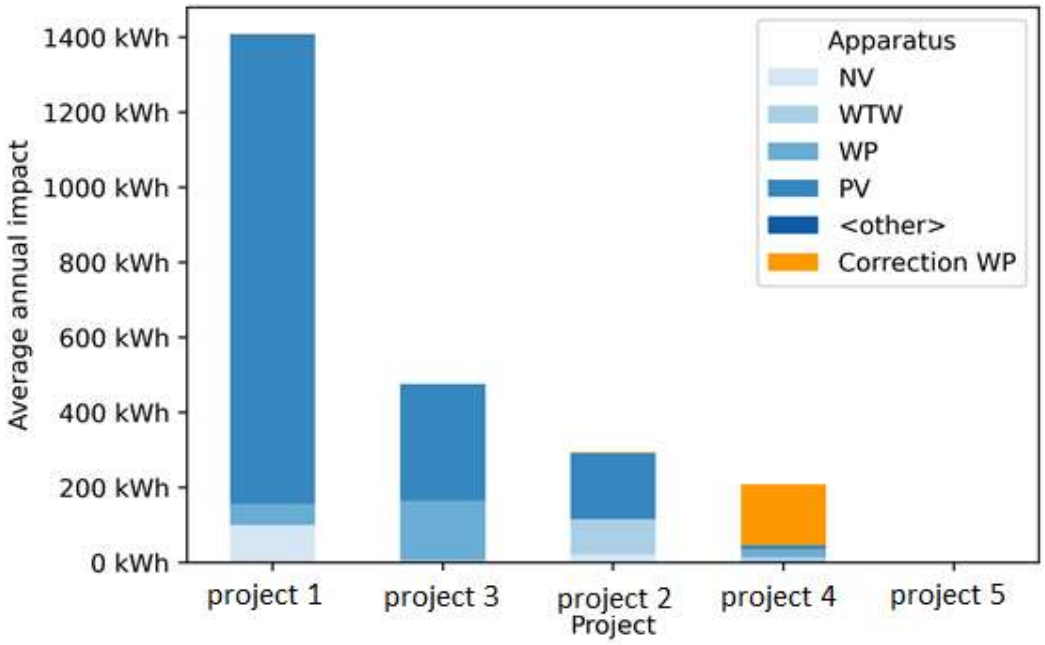

Figure 7 Distribution of annual estimated consequences of detection delay per project, with PV is photovoltaic system, WTW is the heat exchange unit, NV is the auxiliary heater, WP is the heat pump and '<other>' is the aggregate impact of all other apparatus. 'Correction WP' is the manual correction for poor model estimates.

Project 5 does not contribute to the annual impact. This is because there were no service orders of any failures in this system that relate to the PV or heating system. Looking at the individual subsystems contributing to the estimated energy loss during recognition delay, the estimated energy loss in the PV system is the largest, followed by the heat pump, which is forms a large portion of the estimated energy loss during recognition delay in project 3.

The estimated total annual energy loss over all 5 projects is 2,388 kWh, which approximately equals half the benchmark amount of total energy use at a single house, including the energy used by the HVAC and DHW system as well as the energy reserved for personal use. 2,388 kWh can be considered a lower limit for the consequences of detection delays, because we only considered the detection delays that we could identify in the sensor readings visually. 2,388 kWh is circa five percent of the observed annual energy short on the SLAs.

Some uncertainty is introduced by the estimation models. We compute the size of this uncertainty using the approach discussed in subsection 2.4.1.3.3 of this chapter. We observe a MedAE of about 3 kWh per estimate. Assuming that estimation errors are i.i.d. and normal distributed and following Equation 8 from subsection 2.4.1.3.3, the standard deviation per estimate approximately 4.44 per estimate. We have 2418 annual readings during detection delay. So, the distribution of the annual error is approximately normally distributed with mean 2,388 kWh and variance of 15,876. In this case, the probability of having at least the observed shorts, 51,904 kWh in 2018, 41,332 kWh in 2019 and 40,227 kWh in 2020, due to only detection delay is neglectable. This implies that the consequences of detection delay are not the only cause for not meeting the SLAs and other causes must contribute to this issue.

2.5 Discussion

According to our findings, the faults with the largest consequences occur in the PV systems. This is also the type of system where the most persistent detection delays are observed. Specifically, there are a few cases of inverter failure that both took a long time to be recognized and caused the PV system to not yield any energy at all. Besides the faults in the PV system, some of the consequences of faults can be attributed to faults in the heat pump apparatus. Faults related to the heating system for which a notable detection delay is observed are filter fouling and configuration errors.

The large difference between the estimated consequences of failure and the observed shortages on the SLAs, imply that most of the energy short is not caused only by the faults that we were able to visually identify. We suggest a few explanations for these results:

First, the visual identification of faults is difficult. Consequentially, it is likely that there are some faults that have not visually been identified but do affect the energy use of the heat pump system or yield of the PV system. For example, as shown in Figure 35 in Appendix H, there are many missing readings in the monitoring data. If a reading from the monitoring system is missing, we have no way to identify deviating responses and are unable to identify faults. Any faults that occurred during such a period are not included in our results.

Furthermore, it is very difficult to visually detect small changes in trends of the responses. Whereas sudden failures may be recognized, gradual faults may remain unrecognized.

On the other hand, we note that, because of the aforementioned remarks, we are particularly likely to miss faults that have a lesser effect on the response of the systems and, contrary, the faults that we were able to identify are likely those that have a more considerable effect. In our results, distinguish a detection delay for 78 service orders, whereas our service orders contain 224 orders that relate to either the heating system or the PV system.

We approximate the distribution of the energy short by a normal distribution. In this case the probability of having at least the observed shorts due to detection delays only is practically zero. Contrary to the leading view at BGDD, we belief it is unlikely that the detection delays have a major contribution to the observed shortages in the systems' energy yield or use. Instead, we propose that a major contribution to these shortages originates from inaccurate assumptions of the systems' performance during model design. These can result from e.g., the systems performing being worse than stated by the supplier, the residents using the system differently than intended in the systems' design or the weather conditions being different from how they are considered in the systems' design.

2.6 Conclusion

In this chapter we visually identified when the studied systems where historically in fault.

- We quantified the annual consequences of observation delay to not meeting the energy requirements of the SLAs to be about 2388 kWh, of which the largest contribution, about 1900 kWh, comes from unobserved failures in the PV systems.
- Based on these findings, we believe it is unlikely that the increased energy use or loss of energy yield during detection delays is the primary contributor to not meeting the SLAs. Instead, we hypothesize a major cause of the unmet SLA's to be that the systems operate with a poorer energy efficiency from how they are designed.

Solving issues related to the design of the heating and PV system is out of the scope of this project. Regardless, we have shown that there is some benefit for BGDD to identify faults before they are reported by residents or housing corporations. Moreover, we have identified the systems that are

most likely to contribute to useful energy loss during the recognition delay, being the PV systems, the auxiliary heater, and the heat pump.

CHAPTER 3. **Literature Review**

3.1 **Introduction**

In this chapter, we aim to answer research question 2 , which reads “Which classification techniques are proposed in literature and can we apply for detecting system failure in photovoltaic and heat pump systems in our case?”, and the related sub-questions. We answer this question by studying the scientific literature for fault detection techniques to apply to our case. We discuss the techniques most widely adopted in literature for fault detection in heating and PV systems, and which of these techniques are most suitable to our case.

First, we introduce the concept of *fault detection and diagnosis (FDD)*. This is the act of monitoring the normal operation of a system for signs that indicate failures (Isermann, 1984). The rise of maintenance 3.0 and smart systems, gave a boom to this discipline and to the related scientific literature accordingly. We distinct two approaches, ‘*analytical techniques’* and ‘*statistical techniques’*.

3.1.1 **Analytical techniques**

Analytical techniques, also referred to as ‘model-driven approaches’, ‘dynamic modeling techniques’ or ‘white box approaches’, are techniques wherein the behavior of a system is modeled using prior knowledge on the physical laws that govern the processes occurring in the system (Gertler, 1988; Katipamula & Brambley, 2005; Reddy, 2011). Faults are recognized as a deviation of the observed process from the modeled process.

Analytical models require extensive knowledge of the modeled system and the underlying physical processes, which, in our case, include the complex dynamics of the behavior of the residents. Prior attempts of BGDD to analytically explain some of the studied systems did not yield satisfactory results, because the generalizing assumptions about the behavior of the residents did not seem to hold. We therefore extend our inquiry to statistical techniques, which may implicitly capture some of the dynamics of the behavior of the residents.

3.1.2 **Statistical techniques**

Statistical techniques, also ‘data driven approaches’, ‘black box approaches’ or ‘process history-based techniques’, do not rely on *a priori* understanding of the system and its dynamics. Instead, these approaches build an implicit or explicit stochastic model, which is based on large amounts of historic process data (Venkatasubramanian et al., 2003). Faults materialize as changes in the parameters of these stochastic systems and so fault diagnosis is the problem of detecting these changes (Basseville & Nikiforov, 1993).

We distinct two types of statistical models. The first are classification models that require samples of the responses of systems during failures as training input. The second are outlier detection models that are trained only on the responses of systems that are not in failure and therefore require no sample data of responses during failures.

3.1.2.1 **Models trained on failure data**

The most prevalent type of study in our inquiry describes systems in a testing arrangement, so that the system is operated and monitored in a controlled setting. Faults are artificially imposed on the systems to collect fault data. The result of this approach is an abundance of system responses of a system under failure and datasets with each response labeled with the fault(s) imposed on that system and the external conditions at the time of reading.

Labeled failure data allow for classification models to be trained on the fault labels. However, the

systems studied in our project are not monitored in a controlled setting. We have no insight in the control system of the studied systems and can therefore only visually estimate when the studied systems were in fault. As a result we do not have such a sample of fault data, nor is it feasible to collect such data. Following (Zhao et al. (2013), we propose the use of outlier detection models to overcome these limitations.

3.1.2.2 **Outlier detection models**

No sample data of failure responses is required to train outlier detection models, to develop these models, just a sample of (mostly) fault-free system responses is required. Outlier detection models follow the same common structure; First, a regression model is trained. The goal of this regression model is to estimate the normal system response. Next, the distance between the normal system response estimate and the observed response is computed. If this distance exceeds some threshold, the observed response is considered unusual, i.e. different from the majority of the data, and therefore suspicious of resulting from an measuring error, data contamination or system failure (Zhao et al., 2013; Zimek & Schubert, 2017). Finally, to further filter the cases system failure from the cases of measuring error or random variation, we may consider only contiguous outliers, apply the outlier detection technique to the moving average of some number of observations or apply some other smoothing procedure to the observations.

Because the outlier detection approach does not rely on a sample of previously observed failures, outlier detection models can assist in the identification of failures that have not previously been observed. On the other hand, with outlier detection techniques, we identify unusual responses without considering the cause of the unusual response. In contrast to the previously discussed classification models, this means the outlier detection models may not exclusively identify failures, but may also indiscriminately identify the responses of other irregularities, including but not limited to, changing operating conditions, environment or usage.

In the next sections, we will discuss some outlier detection models that are proposed in literature for the detection of failures in PV and/or heating systems. In section 3.2, we discuss SolarClique; a group-based outlier detection model specifically tailored to fault detection in PV systems. In section 3.3, we discuss principal component regression, a linear regression approach that offers a solution for the collinearity between input features.

3.2 **SolarClique**

First, we discuss SolarClique, an outlier detection model specifically tailored to PV systems. The technique is developed by Iyengar et al., (2018). Unlike many fault detection models for PV systems, this algorithm does not rely on weather data or an analytical model of the PV system. With SolarClique, outliers are found by quantifying the distance between the observed response in a candidate system and the concurrent responses of systems that are similar to the candidate. The technique has gained some attention in literature, e.g. by (Park et al., 2020).

SolarClique is based on the assumption that there is some similarity between the energy yield of a candidate PV system and the yield of geographically nearby systems. This is backed by the fact that yield of a PV system is predominately the result of the solar irradiance reaching the system, which is generally similar over small geographic distances. We have made a similar assumption in Chapter 2, when implementing the 3-NN model. We have shown that, if accounted for the differences in capacity of various systems, fairly accurate estimates can already be made using just this assumption. Additionally, Iyengar et al., (2018) argue that there is also a component of the observed energy yield at a PV system that is dependent on the characteristics and state of that system and therefore not typically shared across geographically close systems. Examples of this include the effect of degradation or damages, the angle of the PV panels, fouling of the PV panels or shadows from nearby objects.

SolarClique is a three-part algorithm to detect anomalous energy yield in PV systems. The first part of the SolarClique algorithm involves making an estimation of the PV yield that is shared across geographically nearby systems. This is done by running a regression model with the PV yield of the candidate system as the dependent variable and the concurrent PV yield of a group of nearby systems as the independent variable. To make an estimate of the shared component of PV yield, Iyengar et al. (2018) propose to use of a bootstrapped regression model. A variety of regression models is tried. Each model uses a sample of the observations on the energy yield of geographically nearby systems to predict the energy yield of the candidate PV system. The final estimate of PV yield is the pointwise average of the estimates of each of these regression models.

In the second step, the estimate of the regression model is subtracted from the observed yield of the candidate system. What remains is considered the component of PV yield that is unique to the candidate system. The next step of the SolarClique algorithm is to identify seasonality in the component of the PV yield that is unique to the candidate system, as it is argued that this is likely due to shadow formation from nearby structures and therefore does relate to a failure. For this, Iyengar et al. (2018) propose the use of Seasonal Trend decomposition using Loess (STL), which is a popular and robust method for time series decomposition to isolate recurring seasonal patterns from a time series.

The final step of the algorithm is to detect statistical outliers in the remaining component of observed PV yield. These anomalies could point to faults or damages. As a threshold, Iyengar et al. (2018) propose to use a 4-standard deviations prediction interval about the estimated mean which is computed using bootstrapping. Specifically, the prediction interval is constructed from the residuals of the regression model estimates of the models in the bootstrap used to construct the final estimate of the PV yield that is shared across the group.

To filter the anomalies that are likely indicating faults from those resulting from random variation or measuring errors, Iyengar et al. (2018) only consider the cases where 3 contiguous anomalies occur.

3.3 Principal component regression

The second outlier detection technique that we discuss in principal component regression (PCR). This is technique is strictly a regression technique and can be combined with many distance measures and/or smoothing procedures for detecting outliers and faults.

Principal component regression combines principal component analysis (PCA) with OLS linear regression. PCA is a well-known statistical technique favored for its intuitive appeal. In this technique, the independent variables are transformed into a new group of uncorrelated and orthogonal variables called ‘principal components’, that capture the collinearity in the original variables (Reddy, 2011).

One of the uses of PCA is for dimensionality reduction. A few principal components may capture most of the variability of the response. Therefore, variables can be used than in case of the original variables, resulting in a sparse estimate of the original model.

Since there is no multicollinearity between the principal components, ordinary least squares linear regression can simply be performed on the principal components to gain out-of-sample estimates that are more robust to changing patterns in the independent variables. This is referred to as ‘Principal component regression’.

Once accounted for the effects of the selected features, outliers in the system response are detected using some distance measure to quantify the distance between the regression model’s estimates and the observed system responses. Various examples exist of the application of PCR in an HVAC outlier

detection context, e.g. Chen & Lan (2009) and Yu et al. (2017). Both use distance measures based of the squared prediction error and Hoteling's T-squared statistic. Another study combines PCA and SVM for outlier detection in an heating system's responses (Han et al., 2010).

Albeit fairly easy to implement, the downside to this technique is that the original independent variables are transformed into orthogonal variables, so that the physical meaning of the orthogonal variables is lost (Reddy, 2011). Furthermore, the accuracy of the model is dependent on the selecting the right features to be included in the model.

We propose to use PCR models for estimating the normal system response of the studied heat pump systems. The exogeneous variables for these models consist of the available weather data and sensor readings from the monitoring systems installed at each house in this project. The responses of heat pump apparatuses are more dependent on the behavior of the residents and the dynamics of the house, there is therefore a much higher heterogeneity between the responses of these apparatuses. We therefore find that a group-based outlier detection approach similar to SolarClique might not be suitable for these systems.

3.4 Conclusions

In our case we have no accurate information of when the studied systems where historically in fault. Our best approximations of when the systems where in fault follow from visual approximation as discussed in 2.3. We therefore opt for techniques that use only normal system responses for training, which means that we will use outlier detection techniques.

- For the PV systems, we have already shown that the nearest neighbor technique was fairly accurate. To further extend on this, we opt for the group-based SolarClique technique.
- For the heating systems, we propose to build PCR models for the estimate of the normal system response. On top of the regression estimate, a decision interval is constructed to identify outliers.

In the next chapters, we will explore the implementation of SolarClique and PCR for anomaly detection in the responses of the systems of our case. We discuss the choices and adjustments that we make to apply these techniques.

CHAPTER 4. **Model specification and validation**

4.1 Introduction

In this chapter we design and validate SolarClique and PCR models to estimate the energy yield of the PV system and energy use of the heat pump apparatus respectively. In section 4.2 we discuss the choices to be made for the implementation of the SolarClique algorithm as a fault detection model for the PV systems. In section 4.3 we discuss how to implement PCR for fault detection in the heat pump systems.

4.2 SolarClique

In this section we discuss each of the stages of the SolarClique algorithm; first, we discuss the regression model used to predict the yield of a candidate PV system. Then, we discuss why we, contrary to what is proposed by Iyengar et al., (2018), do not remove any seasonal components. Finally, we discuss how we approached the detection of anomalies in our case.

4.2.1 Constructing the regression models

When constructing the regression models, we need to decide on which features to use and how to fit a regression on these features. Both aspects are discussed in the following sub-sections.

4.2.1.1 Feature selection

As discussed when introducing the SolarClique algorithm in section 3.2 of 0, the training data used in the SolarClique algorithm consist of historic system responses of geographically nearby PV systems over a period of time.

To train their models, Iyengar et al. (2018) use the energy yield readings of 76 large PV sites located in Austin, Taxes. However, they show that high accuracy can be achieved even when few geographically nearby sites are available. Moreover, we reason that the PV systems that show a response that is most similar to the system for which the normal response estimate is to be made are probably of the same project as the candidate system, as these are geographically closest and share other characteristics with the candidate system, such as the orientation and manufacturer of the PV panels. Therefore, we limit the training data to be the historic yields of the PV systems that are of the same project as the candidate system. Restricting the number of systems to consider reduces the time required to compute a solution and is likely to improve the performance of the solution as the model is less likely to overfit.

The resulting training data can be described as a multidimensional time series, or matrix of energy yield, where the column indices describe the PV systems and the row indices describe the days at which the corresponding yields where observed; Let *S* be the number of systems in the project of the candidate system and *T* the number of days used for training, then the training data is given as matrix $\boldsymbol{M}$.

Equation 9 Matrix $\boldsymbol{M}$ with training data of a SolarClique model, with $x_{t,s}$ is the PV yield of system s at day t

$$\boldsymbol{M} \in \mathbb{R}^{T\times S} = \begin{array}{r} \\ day\ 1 \\ day\ 2 \\ \dots \\ day\ T \end{array} \begin{array}{c} \begin{array}{cccc} system\ 1 & system\ 2 & \dots & system\ S \end{array} \\ \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,S} \\ x_{2,1} & x_{2,2} & \dots & x_{2,S} \\ \vdots & \vdots & \ddots & \vdots \\ x_{T,1} & x_{T,2} & \dots & x_{T,S} \end{bmatrix} \end{array}$$

For training the SolarClique models, we use the PV yield data collected from 23-05-2017 to 12-10-2020, so that *T* is 1116. The number of systems differs per project, but every house in any project is equipped with a PV system. Consult Appendix A for the number of houses in each project.

4.2.1.2 **Regression technique**

In the SolarClique algorithm, PV system normal response estimates are made using a bootstrap of regression models. As discussed in Chapter 4, different regression techniques can be selected for use within the SolarClique algorithm. Iyander et al. (2018) ran a number of different regression techniques, of which random forest regression performed best. We adopt these findings and also use random forest to construct the sub-models.

Due to the bootstrap resampling in the SolarClique algorithm, each of the random forest sub-models is trained on a random sample, with replacement, of the training data. Following Iyengar et al. (), the size of each bootstrap sample is 80% of the size of the original training data, so that each sub-model is trained on the observed responses of randomly selected PV systems that are in the same project as the candidate system, allowing duplicate systems. This bootstrap is set to draw 100 samples.

Next, random forest regression models are fitted on each of the subsets of the training data in the bootstrap as follows:

1. Of the respective subsets of the training data, again, bootstrap random resamples are drawn, with replacement. Note that since a random forest regression requires a bootstrap of regression trees that this means the training data is bootstrapped twice.
2. A regression tree is fit on each of these resamples. The regression tree algorithm fitting procedure consists of the following steps, in recursion:
    a. Determine the PV system and threshold of PV yield of the respective system so that the average of the observed energy yield of the candidate system corresponding to either side of this threshold best approximate the PV yield observed in the candidate system. The measure of goodness-of-fit, following Iyengar et al. () is the Mean Squared Error (MSE).
    b. Split the dataset on this split.
    c. Continue to the next iteration on either branch of the split, until the number of observations in the respective branch is 1 or the maximum recursion dept is met.
3. After fitting a regression tree to each of the resamples in the bootstrap, each tree returns a day-to-day estimate of the PV yield of the candidate system given the observed PV yield at other PV systems in the respective project.
4. Finally, the random forest estimates consist of the average of the estimates of the regression trees on each respective day.

The random forest parameters are set to draw 100 bootstrap resamples and run with a recursion depth of 10 splits. The size of each sample is the same as the size of the subset of the training data, which was Empirically found to be a good default value (*1.11 Ensemble methods — scikit-learn 1.0.2 documentation*, n.d.). This corresponds to the default number of iterations and sample size used in the scikit-learn stack (Pedregosa et al., 2011). Illustratively, if the candidate PV system is installed in a house in project 1, there are 53 PV systems in this project other than the candidate PV system. In the first bootstrap, the energy yield of 42 systems, allowing duplicates, is drawn. Then, each of the regression trees in the random forest model is fitted to a sample of energy yield of 42 systems, are randomly drawn from the 42 systems in the first bootstrap, again allowing duplicates. Allowing duplicate PV systems to occur in a single bootstrap sample, causes variety between samples and improves the robustness of the resulting estimate.

Finally, a SolarClique regression estimate is made for each of the 162 studied PV systems. The SolarClique regression estimates are made by taking the average of each respective day of the 100

random forest estimates. So, the SolarClique regression estimates are time series with a length of 1116.

4.2.2 Removing seasonal components

In the original SolarClique algorithm Iyengar et al. (2018), propose to use seasonality and trend decomposition to identify any seasonal effects that affect only the candidate system, such as shading from a nearby structure. Such shading may occur only part of every day or at certain times a year and, in their case, does not require any service. They propose to use a short seasonal period, such as a week.

In contrast to the case discussed by Iyengar et al. (2018), in which sensor readings were collected at an hourly interval, the sensor readings in this project have been collected on a daily interval. This means that we cannot identify any seasonality within a day. Moreover, due to the nature of our case, we do not want to exclude local effects, as these have not been considered in the systems design. Instead, we would want our model to classify such cases as anomalous, so that it can be determined whether the source of this affect can be resolved. We do not include seasonal and trend decomposition in our algorithm.

4.2.3 Anomaly detection

The final step of the SolarClique algorithm is the identification of anomalies. Residuals are computed from the SolarClique regression estimate and the observed sensor readings. We compute a 1-, 2-, 3- and 4-standard deviation prediction interval about the estimated mean. This standard deviation is the computed standard deviation across the estimators returning from the different bootstrap iterations. An alert is raised in the case of more than *k* consecutive readings outside of the prediction interval. On the one hand, a large *k* decreases the number of false positives, on the other hand, every increment of *k* introduces additional detection delay to the system. We set *k* to 3 days.

4.2.4 Validation

4.2.4.1 Validation approach

To validate the SolarClique algorithm, we use the MAAPE and MedAE on the regression estimates. Using the same goodness-of-fit measures that we used to validate the models in Chapter 2, allows for direct comparison between the estimation accuracy of 3-NN and SolarClique.

Execution of the SolarClique algorithm is much more computationally intensive, due to the large number of repetitions of the bootstrap. To run this algorithm in 5-fold validation for all the monitored PV systems would be time-consuming, taking about 28 hours to complete for all locations. Instead of the 5-fold validation used in Chapter 2, we therefore select a separate independent validation dataset, to validate the trained model on. The validation dataset covers 3 months of sensor readings.

After building the validation model, we analyze the residuals by creating some diagnostics plots for every model. Moreover, the residuals are checked for heteroscedasticity, autocorrelation and normality using quantitative measures; The residuals are tested for heteroscedasticity using the White test (White, 1980), for temporal autocorrelation using the Stoffer-Toloi test (Stoffer & Toloi, 1992) and for normality using a Chi-squared test for normality. Each test is performed with an alpha of 0.05. More information on the White and Stoffer-Toloi statistical tests is found in Appendix J and Appendix K. While it is not a necessity that these characteristics hold for random forest regression models, running these tests is still informative for actions to take to improve the performance of the model.

4.2.4.2 Validation results

The dispersion of MAAPEs and MedAEs of the SolarClique regression estimates are shown in Figure 8. In general, the dispersion of the estimates from the SolarClique algorithm is smaller than the

dispersion of the estimates from the 3-NN algorithm. As a result, the number of outliers has increased, but the MAAPE and MedAE of the outlier-estimates is lower.

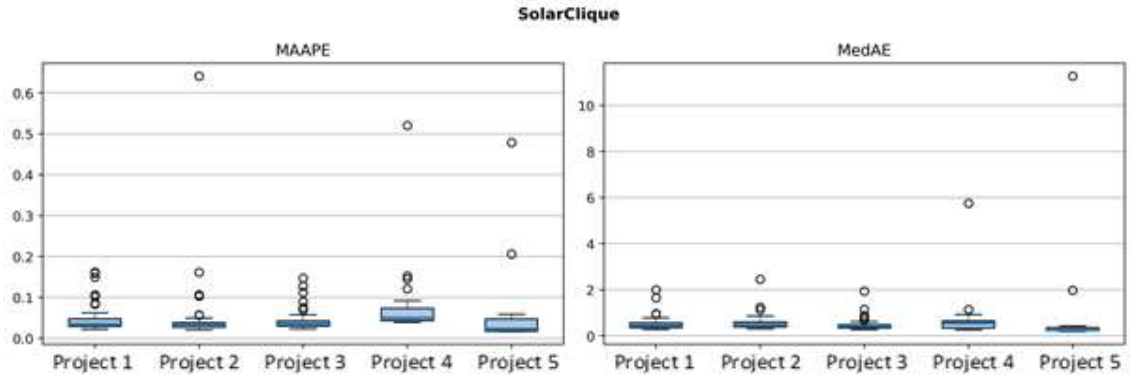

Figure 8 Validation results of the SolarClique models

The MedAEs of the SolarClique estimates are rarely greater than 1 kWh, or about 5.7% of the average daily energy yield of a PV system throughout a year. More, precisely, assuming the residuals are i.i.d. and normally distributed, this implies that approximately 96 percent of the estimates is less than two standard deviations off. In this case, by implementing Equation 7 introduced in Chapter 2, we find that approximately 96 percent of the estimates is less than 2.96 kWh off, or 16.9% of the average daily energy yield of a PV system throughout a year.

Our implementation of the White test for heteroscedasticity is reported in Appendix J. For 32 of 156 SolarClique models' estimates, the null hypothesis of homoscedasticity of the White test is rejected.

Our implementation of the Stoffer-Toloi test for autocorrelation is reported in Appendix K. For 73 of 156 SolarClique models' estimates, the null hypothesis of temporal independence between the residuals of the Stoffer-Toloi is rejected, which indicates that there exists some autocorrelation between the residuals.

We apply a chi-squared goodness-of-fit tests, assuming a normal distribution with the sample mean and standard deviation as its parameters. The p-values from these tests are shown in Figure 14. For 57 of the 156 SolarClique models' estimates, the null hypothesis of normality of the Chi-squared test is rejected.

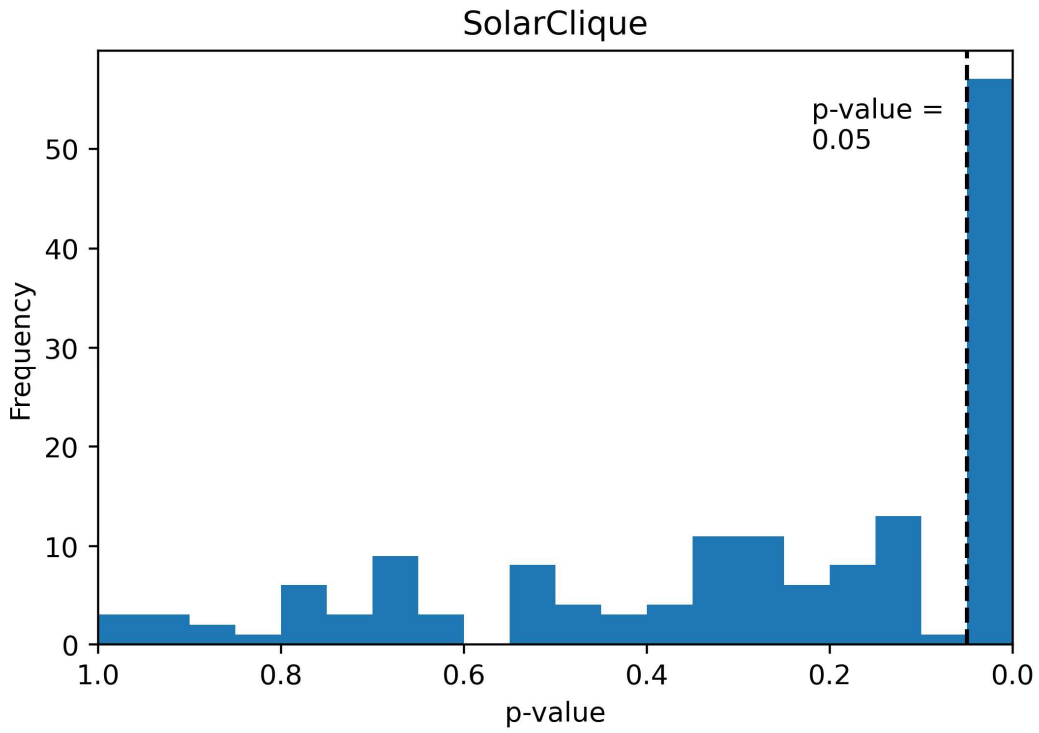

Figure 9 Distribution of the Chi-squared test p-values of the SolarClique models' residuals on the validation data, assuming a normal-distribution

4.2.4.3 **Validation conclusion**

4.2.4.3.1 Regression accuracy

The MedAE of the SolarClique regression model estimates, shown in the left of Figure 8, are generally lower than the MedAE of the model estimates of the 3-NN model introduced in Chapter 2. This suggests that the SolarClique regression model is, in general, more accurate than the 3-NN regression model previously proposed.

4.2.4.3.2 Heteroscedasticity

The null hypothesis of homoscedasticity is rejected for a larger than expected percentage of the residuals of the SolarClique validation models. There can be several explanations for the test pointing to rejection of the null hypothesis.

First, the White test pointing to rejection of the null hypothesis can be the result of a too small sample size (Greene, 2003). While we expect, under normal circumstances, that enough observations can be made during the 3-month validation period, there are 3 cases where there was a large number of missing observations in the validation data.

Secondly, the null hypothesis can correctly be rejected because there is actual heteroscedasticity in the residuals. Due to the nature of the SolarClique algorithm, we would not expect this to occur; Changes other than degradations and failures, are expected to affect also the neighbouring systems and therefore be incorporated in the SolarClique estimate.

Finally, the null hypothesis of the White test can sometimes be rejected due to model misspecification (Greene, 2002). We observe this particularly in the case of a previously unrecognized fault. We found that the heteroscedastic residuals are observed at geographically close locations. Looking at the

energy yield of these locations, we moreover find that there often is a single house in this group where the PV system shows irregular yield during some period, e.g., no yield during several days, indicating a previously unrecognized fault in that system. Since the SolarClique algorithm uses the energy yield at peer houses to estimate the energy yield at a candidate house, a dependency is created between the houses in a project. A fault in one house results in a less certain estimate for the houses that have a high weight assigned to that house. Such estimates are likely to explain the observed heteroscedasticity. Since the most dominant peers are often the houses closest to the candidate house, we observe that heteroscedasticity is observed mostly in groups of geographically close houses.

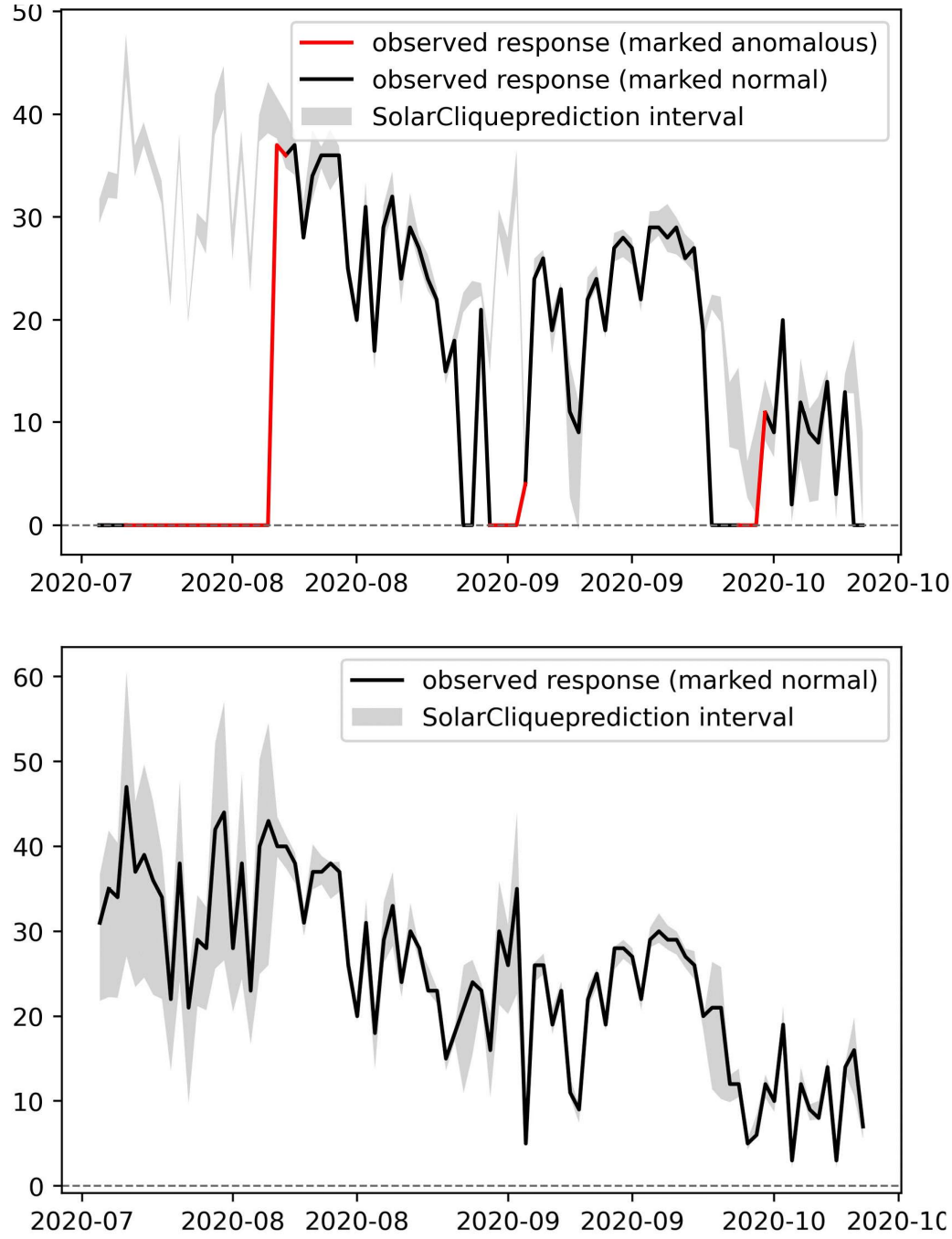

Figure 10 (Top) The energy yield of some PV system over time, multiple system failures occur and (bottom) the energy yield of a geographicaly nearby PV system over time. The prediction interval of the SolarClique estimate on the bottom system seems to be affected by the faults in the top system.

The listed causes likely explain most of the cases of heteroscedasticity in the residuals of the SolarClique validation models. However, without solving the missing observations and unrecognized faults in the validation dataset, we cannot assess if the null hypothesis of the White test would otherwise not be rejected in these cases. We cannot currently confirm nor deny whether there is heteroscedasticity in the residuals that is introduced due to misspecification in the SolarClique algorithm.

4.2.4.3.2.1. Autocorrelation

The SolarClique algorithm uses concurrent readings of a group of PV systems to estimate the energy yield at a candidate system. We would expect a great similarity between the autocorrelation of the estimation error within the daily readings of the candidate system and the members of the peer group. We would therefore expect this correlation to cancel out, resulting in independent residuals.

On the other hand, the outcomes of the Stoffer-Toloi tests performed on the residuals suggest that there are many cases where there is autocorrelation in the residuals of the SolarClique validation models. Looking into the specific cases, we find that 39 of the cases for which the null hypothesis of independence is rejected show significant negative autocorrelation at lag 1 and otherwise little autocorrelation.

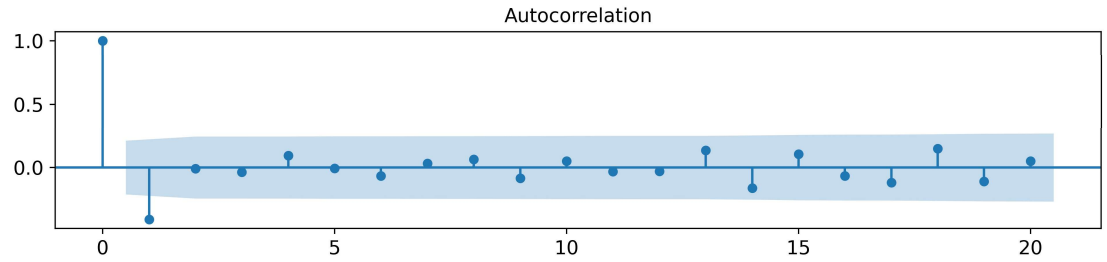

Figure 11 Autocorrelation function plot of a typical case for which the Stoffer-Toloi test was rejected.

This significant autocorrelation at lag 1 suggests that there is some alternating dependency in the data that is currently not modelled. One explanation for this is the fact that energy use is measured in integer kWh and the energy use per day is computed as the difference between observations. This introduces some negative autocorrelation at lag 1, because the decimal energy use is counted towards the next day. Another explanation for the regularity of autocorrelation at lag 1 may be the mapping of the readings to a daily interval. Because it is only the last interval at some date to be returned, it is likely that the energy use over a short interval is followed by the energy use over a long interval. As a result, a smaller difference is followed by a larger difference between observation. While this effect would be relatively small, at most the energy yield of 1 hour just before midnight, this effect may be significant at the scale of the residuals.

Alternatively, the responses of 26 of the remaining 34 cases for which the null hypothesis of independence is rejected show extended periods of little or no PV yield or can be related to a neighbouring system which shows such a response. These indicate a previously visually unrecognized fault in either the candidate system or one of its neighbours, as is also introduces as a reason for heteroscedasticity. These mechanisms are further discussed in section 4.2.4.3.2 but offers an alternative explanation of the rejection of the null hypothesis of independence.

If the errors are indeed autocorrelated, this would indicate that these models' estimates could be further improved by integrating some information about past observations into the model (Rob J Hyndman & Koehler, 2006). Due to the issues discussed above, however, we cannot safely state whether the residuals of the SolarClique validation models would show autocorrelation or not and such a pursuit would be productive.

4.2.4.3.3 Normality

From the Chi-squared test follows that 57 of the 156 estimates seem to have non-normally distributed estimation errors.

One of the characteristics of the response of a normally distributed random variable is homoscedasticity. For 18 of the 57 of the cases for which the null hypothesis of normality is rejected, the rejection can be explained by the heteroscedasticity because of previously unrecognized faults.

For most of the remaining cases for which the null hypothesis is rejected, the residuals of the SolarClique validation model contain one or a few outliers that affect the tails of the distribution. There can be several reasons for these outliers to exists, including inverter failures or maintenance and measuring error in the candidate system or one of its neighbours. Because these outliers may also indicate a fault, these cannot be ignored or removed from the data.

4.3 Principal component regression

In 0 we selected principal component regression (PCR) to detect anomalous behaviour in heat pumps. In this section we discuss how we constructed the regression models and introduce the anomaly detection approach that we used in combination with the PCR regression models.

4.3.1 Construction of the regression model

4.3.1.1.1 Feature selection

For the PCR models, a OLS linear regression is applied to the principal components from the independent variables.

The readily available data for the studied cases are weather data from the nearby weather stations and sensor readings from the monitoring systems installed at each house in this project, discussed in Chapter 2. If available for the candidate location, we include any of the following features:

1. Electric energy yield from the PV system in kWh
2. Thermal energy delivered to central heating in kWh
3. Thermal energy delivered to DHW in kWh
4. Electric energy consumed by the auxiliary heater in kWh
5. Electric energy consumed by the heat exchange unit or ventilation in kWh
6. Electric energy consumed by the electric radiator in kWh
7. Electric energy consumed by electric boiler in kWh
8. Outdoor temperature measured at the nearest weather station in degrees Celsius
9. Solar irradiance measured at the nearest weather station in J/cm^2
10. Wind chill measured at the nearest weather station in degrees Celsius equivalent
11. Wind speed measured at the nearest weather station in 0.1 m/s

As shown in Appendix H, there are many missing readings in the usage data. The weather data, on the other hand, are complete. The total number of missing readings counts up to 9.6% of the supposed training data. The distribution of missing readings across the features is shown in Figure 12.

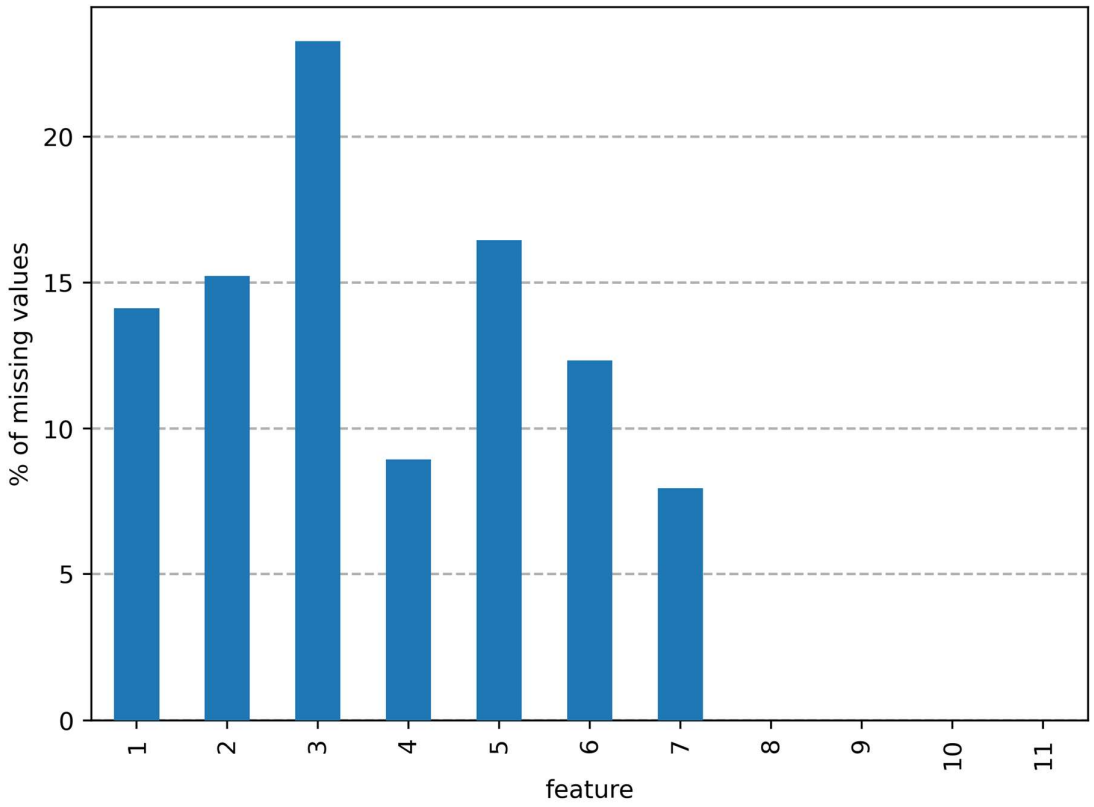

Figure 12 Average % of missing readings per feature

If a variable is included in the PCA-model, the training data may not include any missing readings for this feature. Moreover, if a reading of this feature is missing in the validation or test data, no estimate can be made for that day. To improve the applicability of the to be constructed models, we exclude any features for which more than 40% of the readings is missing in either the training or the test data of the candidate system. Which features are excluded differs per system. If any of the remaining features has a missing value, the estimate for that respective day is returned empty.

4.3.1.1.2 Finding the principal components

There exist multiple algorithms to compute the principal components from a matrix of observations. We adopt the Python-implementation from the Scikit-learn stack (Pedregosa et al., 2011).

4.3.1.1.3 Dimensionality reduction

One of the potential uses of the principal components is dimensionality reduction. It is often the case that a subset of the principal components can be used to describe most variability in the regression model. This implies that we could include only these components in the regression model, without affecting the outcome of the model much. However, by limiting the number of components, we introduce an, albeit small, bit of bias. In our case, there is no reason to apply dimensionality reduction, as the number of original independent variables is small, and the model is fast to evaluate.

Most notably, a potential benefit of dimensionality reduction is that it may be used as some sort of regularization procedure. This means that we may reduce the chance of overfitting at the cost of introducing a small bias to the model. Because a lower estimation uncertainty is a desirable trait when using the model to detect anomalies, we propose to exclude some of the principal components. To

determine how many principal components to exclude for each individual model, we set a threshold of 95% of cumulative explained variance. Starting with the most discriminatory components, we include components until the cumulative explained variance exceeds 95%. Any component past this point is excluded. In general, this means that 5 or 6 principal components are used in the regression model, each derived of the 11 or fewer features in the training data for which more than 60% of the readings is available.

4.3.1.1.4 Regression technique

After extracting the principal components from the independent variables in the training data, we run a linear regression on the principal components. We assume a linear relation between the principal components and the dependent variables. The fit the regression model, we estimate the coefficients β using ordinary least squared linear regression.

4.3.2 Outlier detection

Similarly to how outliers are detected in the SolarClique regression models, we define the decision intervals about the estimated energy use of the heat pump systems using multiples of the standard deviation.

Traditionally, the prediction intervals for principal component regression assume a normal distribution for data and residuals. Some authors remark that the normality assumption is often not correct, since the eigenvalue-transformation in PCA often introduces non-normal, non-symmetric statistical uncertainties (Babamoradi et al., 2013). In this thesis, however, we follow the traditional assumption of a normal distribution for data and residuals.

We project the expected inference of future observations in a decision interval. The novel readings are classified as anomalous if they fall outside of this interval. The lower ('LBP') and upper bound ('UBP') of the prediction interval are computed using Equation 10 and Equation 11. Similar to with the SolarClique algorithm, we run our models with z ranging from 1 to 4.

Equation 10 Lower bound (LBP) of the decision interval of a SolarClique classification model, where z is the number of standard deviations that define the decision interval, σ^2_{res} is the sample standard deviation of the estimation errors, $\boldsymbol{X}$ is the matrix of independent variables in the training data and $\tilde{\boldsymbol{X}}$ is the matrix of independent variables in the test set.

$$LBP = \hat{\boldsymbol{y}}_{\boldsymbol{t}} - z * \sqrt{\boldsymbol{\sigma}^{\boldsymbol{2}}_{\boldsymbol{res}}(\tilde{\boldsymbol{X}}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-\boldsymbol{1}}\tilde{\boldsymbol{X}}^{\boldsymbol{T}} + \boldsymbol{I})}$$

Equation 11 Upper bound (UBP) of the decision interval of a SolarClique classification model, where z is the number of standard deviations that define the decision interval, σ^2_{res} is the sample standard deviation of the estimation errors, $\boldsymbol{X}$ is the matrix of independent variables in the training data and $\tilde{\boldsymbol{X}}$ is the matrix of independent variables in the test set.

$$UBP = \hat{\boldsymbol{y}}_{\boldsymbol{t}} + z * \sqrt{\boldsymbol{\sigma}^{\boldsymbol{2}}_{\boldsymbol{res}}(\tilde{\boldsymbol{X}}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-\boldsymbol{1}}\tilde{\boldsymbol{X}}^{\boldsymbol{T}} + \boldsymbol{I})}$$

4.3.3 Validation

4.3.3.1 Validation approach

The principal component regression is first validated using the same measures as used for the previous models. Similar to how the SolarClique regression was validated, we use a separate validation data consisting of 3 months of sensor readings. The residuals of PCR anomaly detection are also analyzed visually and using the White, Stoffer-Toloi and Chi-squared test.

4.3.3.2 **Validation results**

The dispersion of principal component regression MAAPEs and MedAEs are shown in Figure 13. The dispersions of MAAPEs are mostly comparable to the 3-NN and OLS estimates of Chapter 2. More importantly, the MedAEs of the principal component regressions are generally much lower than those of the 3-NN and OLS estimates in Chapter 2.

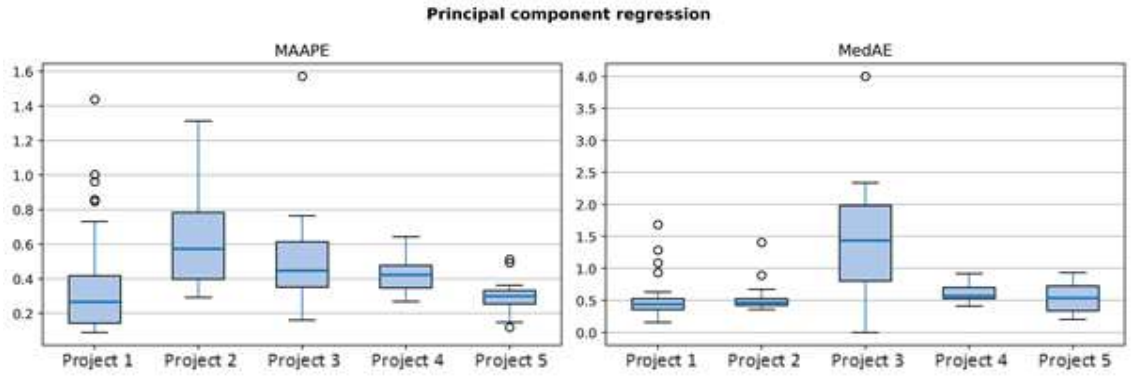

Figure 13 Validation results of the principal component regression model

Our implementation of the White test for heteroscedasticity is reported in Appendix J. For 90 of the 152 PCR models' estimates, the null hypothesis of homoscedasticity of the White test is rejected. This indicates that the variance of these estimates may be dependent on the time that the estimate was made.

Our implementation of the Stoffer-Toloi test for autocorrelation is reported in Appendix K. For 128 of the 152 PCR models' estimates, the null hypothesis of temporal independence between residuals from the Stoffer-Toloi test is rejected. This suggests these residuals exhibit autocorrelation.

We apply a chi-squared goodness-of-fit tests, assuming a normal distribution with the sample mean and standard deviation as its parameters. The p-values from these tests are shown in Figure 14. For 101 of the 152 PCR models' estimates, the null hypothesis of normality of the Chi-squared test is rejected.

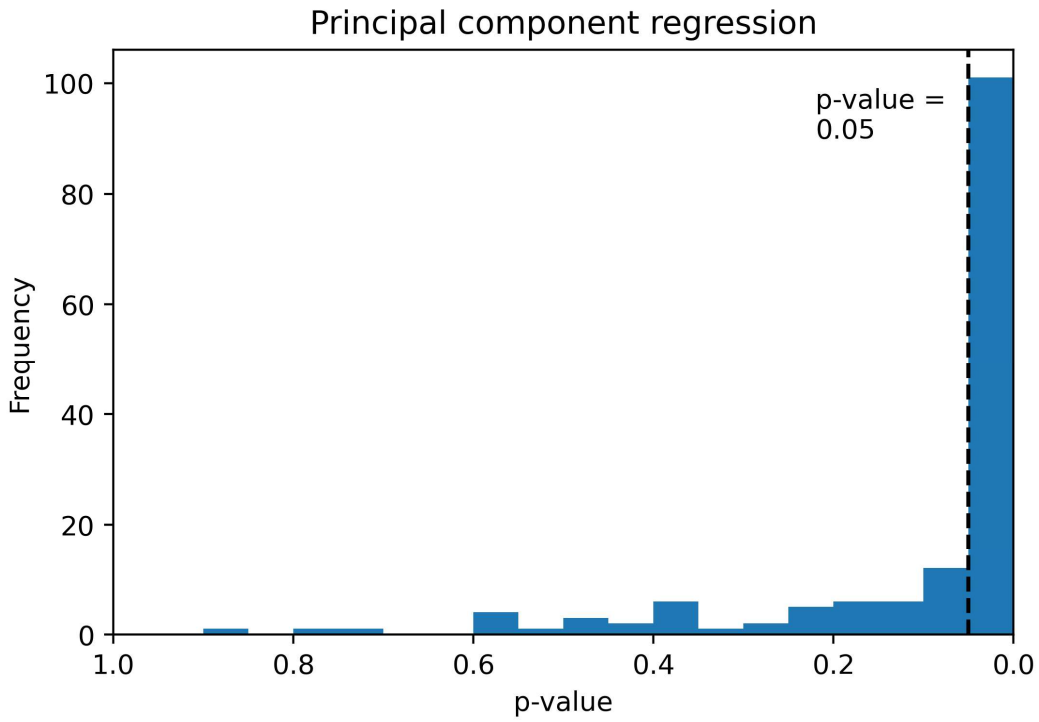

Figure 14 Distribution of the Chi-squared test p-values of the PCR models' residuals on the validation data, assuming a normal-distribution

4.3.3.3 **Validation conclusion**

4.3.3.3.1 Regression accuracy

The MedAE of the PCR model estimates, shown in the left of Figure 13, are generally much lower than the MedAE of the 3-NN and OLS model estimates of Chapter 2. This suggests that the SolarClique regression model is, in general, more accurate than the models previously proposed.

The MedAEs of project 3 are generally higher than those of the other projects. Looking into the validation data, we observe that there are some unusual levels of energy use in the original responses of these systems, this may be unrecognized failures. We can however not make this assumption without having a service order relating to these supposed faults, as this error could also be explained by an insufficient model; Another explanation is that there is another feature effecting the response, which is currently not included in the model, e.g., a different use of the system, or an unmonitored environmental effect.

4.3.3.3.2 Heteroscedasticity

By running the White test on the residuals, we find that the hypothesis of homoscedasticity is rejected for 90 of the 152 PCR validation models.

Looking further into the cases for which the null hypothesis of constant variance is rejected, we find some indications of heteroscedasticity that may be relevant when accessing the quality of the model estimates. We find three patterns in the residuals that occur similarly and about the same time across multiple systems. This indicates that there is some dynamic that is not captured in the current model, such as the fact that the weather conditions measured at the nearest weather station may be different from the local weather conditions affecting the candidate system or fact that heat is stored in the structure of the house so that more heating energy is required when there is a sudden decrease in

outdoor temperature compared to a gradual decrease in outdoor temperature. In total, 48 of the cases for which the null hypothesis of the White statistical test is rejected show these distinct patterns.

On the other hand, we also find some cases of which the outcome of the statistical test can be related to an issue regarding the input data.

First, we observe that most estimates for the heat pump energy use of the systems in project 3 and many estimates in project 4 are missing. This is due to a project-wide faults in the monitoring system, resulting in missing observations for both the dependent variable and some of the independent variables. This can explain 22 of the cases for which the null hypothesis of the White statistical test is rejected, as this test is sensitive to small sample sizes (Greene, 2003).

Second, we recognize that the null hypothesis of heteroscedasticity is rejected for all models that estimate the heat pump energy use of a system in project 5. Looking at the data used to train this model, we find that the first observations in the training data are from April 2019 and that, due to a project-wide fault in the monitoring system, there are no observations in the training data for the period between January 15, 2020, and April 15 2020. Finally, these models are validated on a test set, which contains observations over the period from 12 November 2020 to 16 June 2021. So, the validation set contains observations of the system usage and weather conditions in late January, February, March and the first half of April, but such observations do not occur in the training data. Therefore, heteroscedasticity is introduced because the estimates for this period are worse than those for the other periods.

4.3.3.3.3 Autocorrelation

The issues regarding the input data that introduce the heteroscadesticity in the residuals also are likely to explain a large portion of the cases for which the null hypothesis of the Stoffer-Toloi statistical test is rejected. In fact, there are only 3 cases where the White statistical test points to rejection of its null hypothesis, but the Stoffer-Toloi test does not.

Furthermore, like the PV yield observations, the heat pump energy use readings are given in integer kWh and are from the last hour of the respective date. Therefore, the residuals of the PCR validation models have some of the same issues that introduce a negative autocorrelation at lag 1 in the residuals of the estimates of the SolarClique validation models.

4.3.3.3.4 Normality

The null hypothesis that the residuals are normally distributed is rejected for 101 of the 152 sets of residuals of the PCR validation models estimates. This is, however, not surprising as the characteristics of the normal distributions are homoscedasticity and independence between observations.

On the other hand, we assumed normality when we formulated the lower- and upper bound of the prediction interval. Consequently, in practice, the probabilities associated to a type-I error may be very different from the assumed α. In other words, the probability of some observation to be classified as an anomaly may be larger than designed, so that the classification of the model become unreliable.

4.4 Conclusion

In this chapter we designed and validated SolarClique models to estimate the energy yield of the PV system and PCR models to estimate the energy yield of the PV system. Looking at the goodness-of-fit of these regression models, they outperform the 3-NN and OLS models' estimates discussed in in Chapter 2.

However, we have some remarks that make the validation outcomes unreliable:

- When inspecting a sample of models for which the statistical validation tests are rejected, we find indications of faulty system responses still being present in the training data due to the limitations of our approach for visually identifying such faults. This causes the models to be trained on faulty data.
- Furthermore, we find indications of faulty system responses being present in the validation data, causing the model estimates to be compared to faulty data. This may cause the statistical tests to falsely reject their null hypothesis.
- We found long periods of failure of the monitoring system, causing long periods of missing data in the training set. This affects the validation model accuracy and introduces imbalance in the training data.

CHAPTER 5. **Test results**

5.1 **Introduction**

In the previous chapter, we discussed our implementations of the SolarClique algorithm and PCR for predicting, respectively, the normal responses of PV and heat pump systems. We also validated the models' estimates against the observed responses to assess the models' fit. Although we faced some issues with the training data that complicate reliable validation of the models, we momentarily assume in this chapter that the SolarClique and PCR models are valid models for estimating the normal system response. This allows us to run the models on a test set to see whether the models have any utility for the detection of faults in the studied systems.

In this chapter we will answer the third research question of this thesis: 'Can implementation of the proposed failure detection models improve failures detection in the studied systems' responses?' Our approach to evaluate the performance is discussed in section 5.2. Results are shown in section 5.3 and discussion and conclusions can be found is section 5.4 and 5.5 respectively.

5.2 **Approach**

To test the ability of the models to detect possible failures in the PV and heating systems, we apply the models to an independent test set. Faults in this test set have been labelled using the visual approximation approach discussed in Chapter 2: Current situation. For each sensor reading on PV yield or heat pump energy use in the test set, we determine whether it is classified as a fault by the SolarClique and PCR model respectively. The classifications returned by the models are compared to the classifications assigned by visual approximation and confusion matrices are constructed from these results.

As briefly mentioned in Chapter 4, we run both the SolarClique and the PCR models over a range of widths of the decision interval. For the SolarClique models, we compute a 4-, 3-, 2- and 1-standard deviation interval about the estimated mean. For the PCR models, we run our models with $\alpha = 0.01$, $\alpha = 0.02$, $\alpha = 0.05$ and $\alpha = 0.10$, where α is the probability of a false positive. Generally, as the width of the decision interval decreases, the models are more likely to correctly detect faults, at the cost of a higher amount of false positive.

The performance of the classification models is evaluated using simple binary performance metrics; accuracy, precision and recall.

Accuracy is the fraction of all observations correctly classified. Accuracy is calculated using Equation 12. One thing to consider when using the accuracy metric is that both the dataset of visually marked faults of the PV systems and the dataset of visually marked faults of the heat pump system are greatly unbalanced. Of the labels assigned to the usage data of the PV system in the test set, 82% of the observations is visually marked as 'no fault' and 18% of the observations is visually marked as 'fault'. Of the labels assigned to the usage data of the heat pump system in the test set, 17% is visually marked as 'fault' and 83% is visually marked as 'non fault'. This means that the heuristic of classifying every observation as 'no fault', would yield an accuracy of 0.83 on the PV system fault labels and 0.82 on the heat pump, although such a heuristic has no practical use.

Equation 12 Accuracy, with the number of true positives (TP), the number of true negatives (TN), the number of false positives (FP), and the number of false negatives (FN).

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision measures the fraction of positive prediction that is classified correctly. It is therefore a measure of how trustworthy the positive outputs of the models are. It does not provide any information on the number of false negatives; positive observations that have been classified as negative. Precision is calculated using Equation 13.

Equation 13 Precision, with the number of true positives (TP), and the number of false positives (FP).

$$Precision = \frac{Number\ of\ correct\ positive\ predictions}{Total\ number\ of\ positive\ predictions} = \frac{TP}{TP + FP}$$

The last metric that we use to evaluate the classification models is recall. Recall, contrary to precision, measures the fraction of positive observations correctly identified by the model. It is therefore a measure of the generalizability of a model; A model with high precision but low recall is likely a good model for identifying the positive observations with some specific cause, but not for others. Recall does not give any information on the number of false positives. Recall is calculated using Equation 14.

Equation 14 Recall, with the number of true positives (TP), and the number of false negatives (FN).

$$Recall = \frac{Number\ of\ correct\ positive\ predictions}{Total\ number\ of\ positive\ observations} = \frac{TP}{TP + FN}$$

Furthermore, we compare the performance of the proposed classification techniques with a random classifier in a 'receiver operating characteristic' (ROC) curve. In the ROC curves, the recall is plotted against the false positive rate. Similar to a type-I error rate, the false positive rate is the probability of falsely classifying an observation as fault. In contrast to a type-I error rate, the false positive rate is computed of the post-prior results, e.g. after the model have ran. The false positive rate is computed using .

Equation 15 False positive rate, with the number of false positives (FP) and the number of true negatives (TN).

$$false\ positive\ rate = \frac{Number\ of\ false\ positive\ predictions}{Total\ number\ of\ negative\ observations} = \frac{FP}{FP + TN}$$

5.3 Results

5.3.1 SolarClique results

616 of the 34945 PV system readings in the test set where classified as faults by the SolarClique algorithm with a decision interval of 4 standard deviations above and below the regression model estimate. Table 7, Table 8, Table 9 and Table 10 in Appendix N show the confusion matrices of outlier detection with the SolarClique models, with each table the interval around the estimated value is decreased in size. Table 1 summarizes the performances of the SolarClique models.

Table 1 Summary of the performance of the SolarClique models for PV systems with 4, 3, 2 and 1 standard deviation decision intervals around the mean model estimate.

Interval	Accuracy	Precision	Recall	Total N.A.
4 S.D.	0.82	0.04	0.01	3743
3 S.D.	0.81	0.09	0.02	3743
2 S.D.	0.78	0.13	0.06	3743
1 S.D.	0.65	0.16	0.26	3743

We observe that accuracy ranges between 0.82 and 0.65 and declines with the width of the decision interval. Furthermore, the precision ranges between 0.04 and 0.16 and increases as the width of the decision interval decreases. Finally, the recall ranges between 0.01 and 0.26 and increases as the width of the decision interval decreases. These results relate to an increase in true positives and even greater increase in the false positives as the width of the decision interval decreases.

The most relevant result may be the precision. With a high precision, the 'fault' labels returned by the SolarClique classifiers are reliable leads for contacting the residents in order to find what caused this fault. The highest precision found is 0.26, which indicates that a reading classified as 'fault' by the SolarClique classifier, on average, has a probability of 0.26 of also being classified as 'fault' by visual classification. If we consider the visually classified faults to be the only faults that have occurred in the PV system during the period in which the test data has been collected, this indicates that the 'fault' labels of the SolarClique classifiers are fairly unreliable indicators of faults.

The recall is an estimate of the probability that a given reading that was visually classified as 'fault' will also be classified as a 'fault' by the SolarClique classification models. We are interested in the recall as a model with a high recall can be used effectively to filter readings in the usage data that are unlikely to indicate a fault. The recall-scores in Table 1, are not promising for using the SolarClique model in such a way.

The highest accuracy achieved by the SolarClique classification models is only 82%. Comparing this accuracy to the distribution of the visually assigned labels, we observe that classifying every observation as 'no fault' would yield about the same accuracy as is achieved using the SolarClique model.

The ROC curve in Figure 15 show the recall against the false positive rate. Similar to the observation on the accuracy, we note that the SolarClique classifiers perform only slightly better than a random classifier across the entire range of false positive rates.

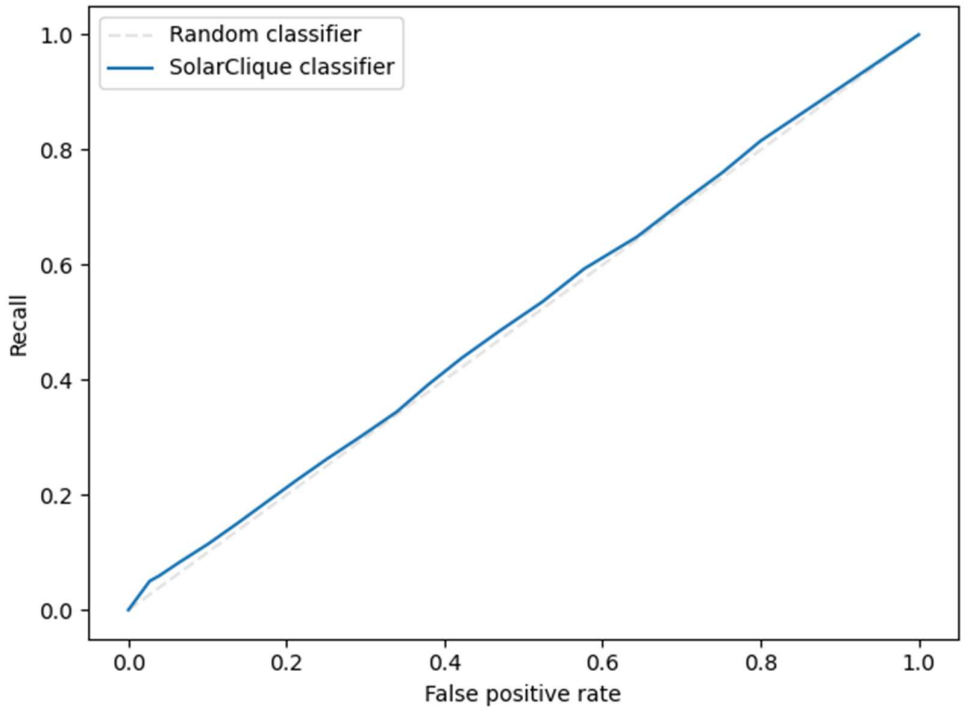

Figure 15 Receiver operating curve of the SolarClique models on the test set.

5.3.2 **PCR results**

189 of the 31388 readings in the usage data test set of the heat pump system are labelled as 'fault' by the PCR classification model with a decision interval of approximately 2 standard deviations above and below the regression model estimate. Table 11, Table 12, Table 13 and Table 14 in Appendix N show the confusion matrices of outlier detection with the PCR models, with each table the interval around the estimated value is decreased in size. Table 2 summarizes the performance of the principal component regression models.

Table 2 Summary of the performance of the PCR models for heat pomp readings with 4, 3, 2 and 1 standard deviation decision intervals around the mean model estimate.

Interval	Accuracy	Precision	Recall	Total N.A.
4 S.D.	0.83	0.00	0.00	5549
3 S.D.	0.83	0.42	0.00	5549
2 S.D.	0.83	0.24	0.01	5549
1 S.D.	0.79	0.19	0.09	5549

The results of the PCR classification models for fault detection in heat pump systems show a mostly

constant accuracy, precision and recall for all values of alpha. Particularly notable is the low recall of the models. This is a problem, as this means the vast majority of failures will remain unrecognized if these models are used. This only slightly improves when we reduce the width of the decision intervals. Of the sparse number of readings that are classified as 'fault' by the models, about 28% overlaps with readings that are visually marked 'fault'.

The highest precision observed in the results of the PCR classifiers is 0.42, corresponding to a decision interval of 3 standard deviations at either side of the estimated response. This is the highest precision observed for any decision interval of any of the models. However this level of precision is paired with a very low recall, rounded to 0. Furthermore, this precision corresponds to an accuracy of 0.83. Of the heat pump readings in the test set, 17% is visually marked as 'fault' and 83% is visually marked as 'non fault'. So, these models compare in performance to classifying each reading as 'non fault'. From these results, we deduce that there are very few observations that are classified as 'fault' by the PCR classification models, so that even if the precision is somewhat high, there is no use of these model estimates.

In extend, we observe that with a decision interval of 4 standard deviation above and below the regression model estimate, both the precision and recall are 0. It follows that, at this definition of the decision interval, all readings are classified as 'no fault'.

In line of the above observations, the ROC curve of the PCR classifier models, depicted in Figure 15, closely follows the non-discriminant line of the random classifier. This figure shows that the PCR classification models is outperformed by a random classifier in terms of recall, regardless of the width of the decision interval.

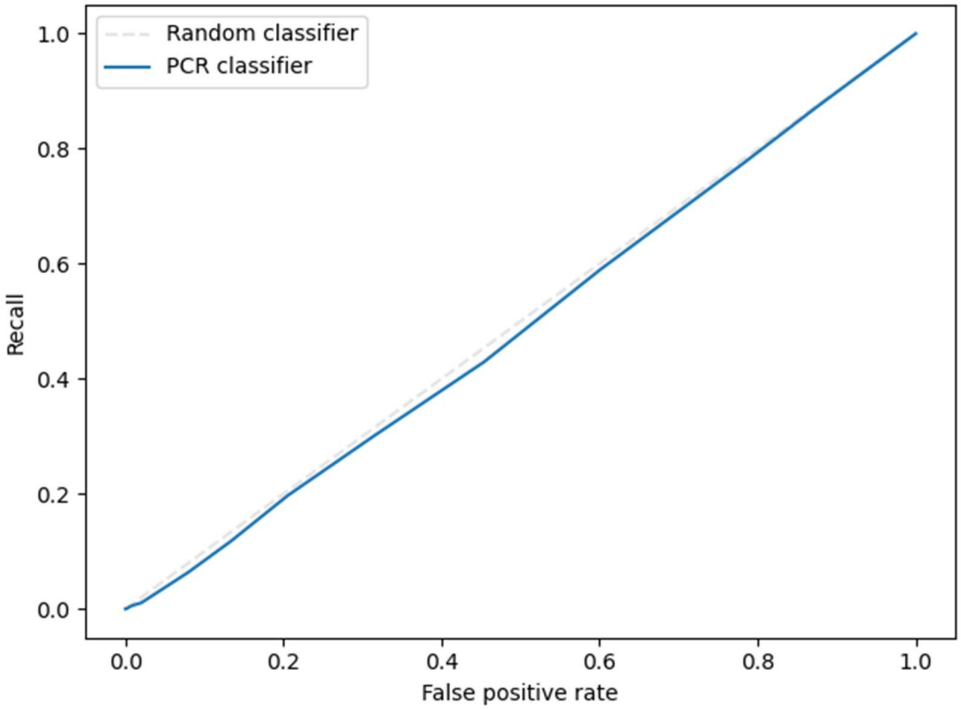

Figure 16 Receiver operating curve of the PCR classifier models on the test set.

5.4 Discussion

Notable in the classification results of both the SolarClique and PCR classification results is the very low recall for any width of the decision interval. This shows us that the current models are unable to successfully separate the visually identified 'fault'-readings from the visually identified 'non-fault' readings. This can either be explained by issues in the classification models, or, in line with our findings in the previous chapter, by possible unmarked failures during the initial visual classification.

The latter explanation is supported by our findings if we have a closer look at the observations that are marked as 'fault' by either the SolarClique or PCA classification models. We find that the labels returned by the outlier detection models sometimes correlate to what, in hindsight, could be explained as a sudden change in trend. Such a sudden change in trend may indicate that there was a fault in the candidate system, but that this was not recognized during visual approximation.

One example of such a case is shown in Figure 17. In this figure, both periods that are classified as containing outlier data fall within a period that is recognized visually as a change in trend and level. Moreover, this period is closed with the completion of a service order, which quite possibly marks the maintenance activity on the heating systems that has solved the fault. In hindsight, we state that this fault seems to have started at 2021-03-10. Due to the requirement for at least 3 consecutive readings outside of the decision interval, the PCR model first detected a fault on 2021-03-19, which is earlier than when BGDD was notified of the fault and created a service order at 2021-04-13. In this case, therefore, it would have been useful to act on the results of the PCR classification model.

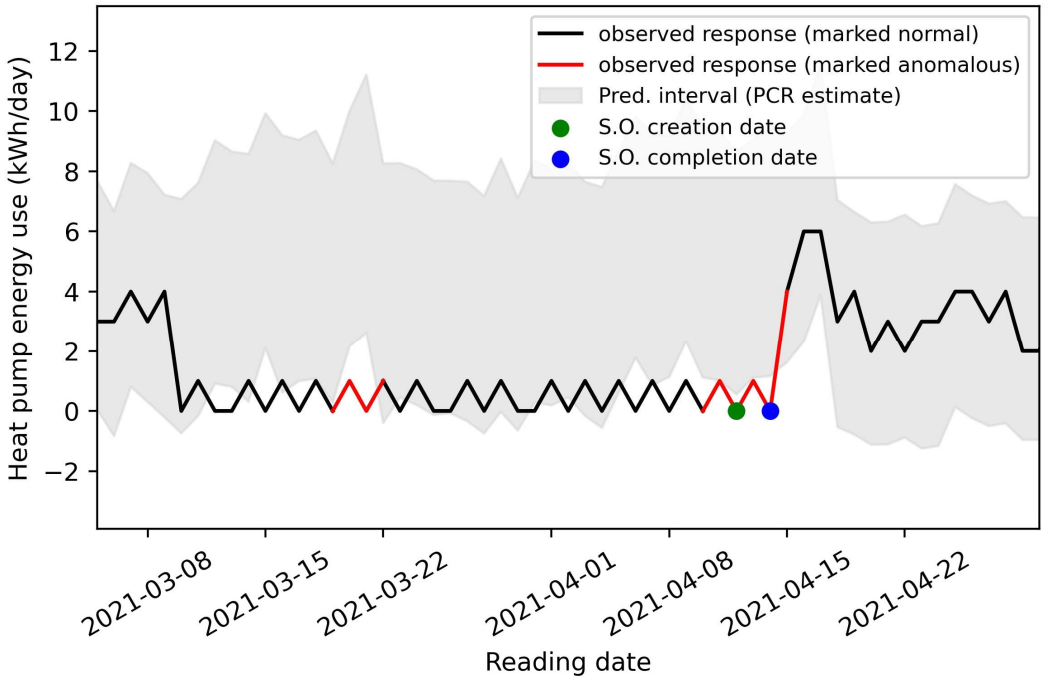

Figure 17 Energy use of a heating system with what seems to be an unrecognized, failure during March and the start of April.

A second potential fault which was not recognized visually, but is recognized by the SolarClique algorithm is depicted in Figure 18. In this case, the observed readings possibly indicate a fault where the capacity of the system is only partially reduced. Contrary, such results can also be the case of a poor estimate of the SolarClique algorithm. The difficulty with these cases is that we cannot be certain

that the outlier readings relate to a fault if there is no service order following the anomalous readings or if there is a service order, but its descriptions seem not to relate to a failure in the respective system. As a result, these readings may alternatively indicate something different than a system failure.

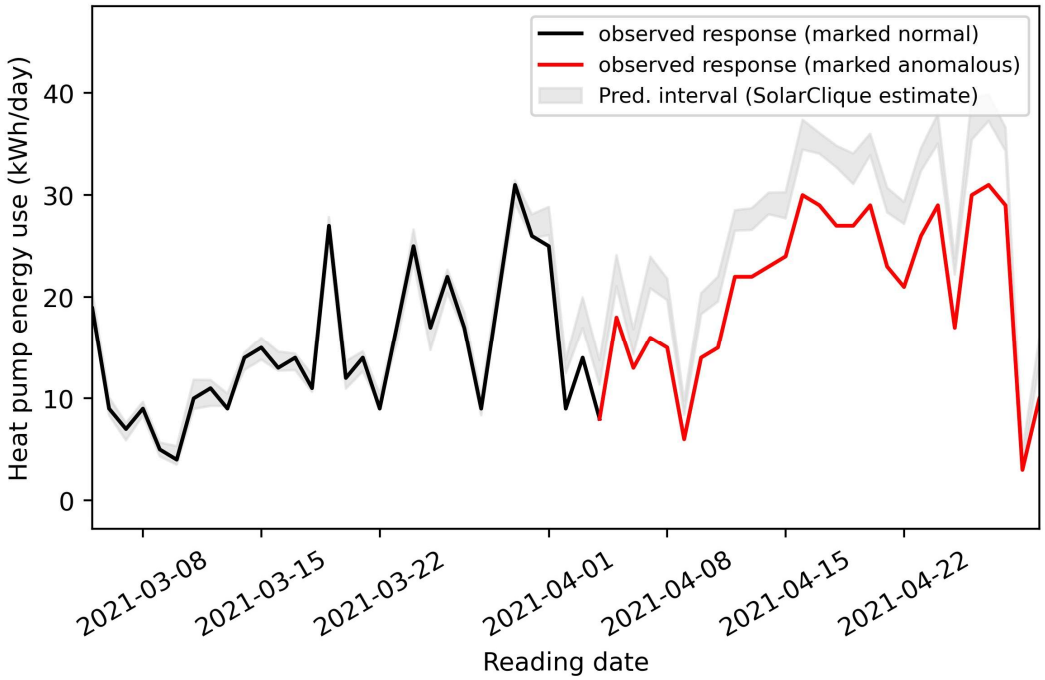

Figure 18 Energy yield of a PV system, a partial reduction of the PV capacity may be observed in the first quarter of April.

Upon visual reinspection of the readings labelled by the outlier detection models, we find that 259 of the 616 readings labelled as outliers by the SolarClique models with a 4-standard deviation decision interval and 19 of the 189 readings classified as outliers by the PCR models with a 2-standard deviation interval may, in hindsight, relate to failure. If these readings are indeed correctly classified by the classification models, this would mean that the models have performed much better than the previously discussed results indicate.

Unfortunately, there is no reliable way of knowing whether these past readings, and the other past readings labelled as failures by the classification models, are from periods of system failure. The lack of a reliable classification of the historic usage data remains an issue throughout this project and is a primary limitation for future endeavours in this direction. To solve this, we recommend to BGDD to explore the possibilities of acquiring data on the state of the studied systems from the control systems installed in these systems.

5.5 Conclusions

- We find little overlap between the readings visually marked as being from a system in fault and the readings that are classified as anomalous by the models. This shows that the ability of the SolarClique and PCR models to reproduce the visually identified classification is poor.
- We found several readings that are classified as ‘fault’ by either the SolarClique or PCR model and which, in hindsight, can be related to a sudden change in trend of the response. This indicates there are issues with the proposed visual approximation approach.
- Without reliable classifications of the historic usage data, there is no way of knowing whether the models or the visual approximation approach cause these poor results.

- Given this level of performance, it is unadvisable to use the proposed models for detection of failures in order to reduce detection delay. We recommend to explore the possibilities of acquiring data on the state of the PV and heat pump systems from the control systems of these systems.

CHAPTER 6. **Conclusion and recommendations**

This thesis discusses the use of outlier detection models to recognize failures in PV and heating systems. In this chapter, we summarize the key conclusions of this project and formulate some recommendations for further research efforts by BGDD.

6.1 **Conclusions**

Based on our findings, we arrive at the following conclusions:

1. The largest consequences of delayed detection of failures to not meeting the SLAs primarily relate to hidden failures in the PV system, followed by failures in the heat pump system. However, detection delays of failures in the heating and PV systems are unlikely to be a major cause for not meeting the SLA's. Further research is required to find definitive causes for not meeting the SLA's. One potential cause is that the performance of the heat pump systems, while not in fault, is worse than reported by its suppliers.
2. Both the proposed SolarClique and PCR models to classify faults cannot be validated, primarily due to a lack of reliable data on when these systems where historically in fault. These models also perform poorly on separating the visually identified 'fault' readings from 'no fault' readings.
3. The primary challenge for developing and validating fault detection models in this case is the lack of reliable data on when the systems were historically in fault. Without these data, the models to be used are limited to approaches that do not require a sample of fault data to be trained on. Even then, the lack of reliable labels is a recurring issue, which complicates both the training, validation and performance assessment of the proposed models.

6.2 **Recommendations**

Based on these conclusions, we recommend the following to be pursued in further research efforts of BGDD to improve performance on meeting the criteria in the SLA:

1. Future efforts should be put in quantifying the contribution of issues other than detection delay towards not meeting the SLAs. Possible issues include the heating systems performing not according to the specification given by its suppliers or false assumptions on the weather conditions or behavior of the residents.
2. The proposed SolarClique and PCR models, without further validation, should not be implemented for detection failures in an operational setting.
3. Efforts should be put in the collection of reliable data on the state of the monitored systems, including data on precisely when these systems where in fault. A possible solution to this is to gain access to the log of the control systems of the monitored installations. The scope of this data would have to cover a large period, e.g. 3 years, so that multiple samples of the response of the same system during some period in a year are available.

References

1.11 Ensemble methods — scikit-learn 1.0.2 documentation. (n.d.). Retrieved January 6, 2022, from https://scikit-learn.org/stable/modules/ensemble.html#forest

Alexander, R. A. (1990). A note on averaging correlations. *Bulletin of the Psychonomic Society*, *1990*(4), 335–336.

Babamoradi, H., Van den Berg, F., & Rinnan, Å. (2013). Comparison of bootstrap and asymptotic confidence limits for control charts in batch MSPC strategies. *Chemometrics and Intelligent Laboratory Systems*, *127*, 102–111. https://doi.org/10.1016/j.chemolab.2013.06.005

Basseville, M., & Nikiforov, I. (1993). *Detection of Abrupt Change Theory and Application* (Vol. 15).

Bouwgroep Dijkstra-Draisma. (n.d.). *Certificering*. Retrieved May 31, 2021, from https://bgdd.nl/wie-we-zijn/certificering/

Brockwell, P. J., & Davis, R. A. (2016). *Introduction to time series and forecasting*. https://doi.org/10.1007/978-3-319-29854-2

Chen, Y., & Lan, L. (2009). A fault detection technique for air-source heat pump water chiller/heaters. *Energy and Buildings*, *41*(8), 881–887. https://doi.org/10.1016/j.enbuild.2009.03.007

Chicco, G., Napoli, R., & Piglione, F. (2006). Comparisons among clustering techniques for electricity customer classification. *IEEE Transactions on Power Systems*, *21*(2), 933–940. https://doi.org/10.1109/TPWRS.2006.873122

Collenteur, R. A., Bakker, M., Caljé, R., Klop, S. A., & Schaars, F. (2019). *Pastas: Open Source Software for the Analysis of Groundwater Time Series*. https://doi.org/10.1111/gwat.12925

Dunsmuir, W., & Robinson, P. M. (1981). Estimation of time series models in the presence of missing data. *Journal of the American Statistical Association*, *76*(375), 560–568. https://doi.org/10.1080/01621459.1981.10477687

Gertler, J. J. (1988). Survey of Model-Based Failure Detection and Isolation in Complex Plants. *IEEE Control Systems Magazine*, *8*(6), 3–11. https://doi.org/10.1109/37.9163

Greene, W. H. (2003). *Econometric analysis*. Prentice Hall, Pearson Education International,.

Han, H., Cao, Z., Gu, B., & Ren, N. (2010). Pca-svm-based automated fault detection and diagnosis (afdd) for vapor-compression refrigeration systems. *HVAC and R Research*, *16*(3), 295–313. https://doi.org/10.1080/10789669.2010.10390906

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning : data mining, inference, and prediction*. Springer,.

Heerkens, J. M. G., & van Winden, A. (2012). *Geen probleem, een aanpak voor alle bedrijfskundige vragen en mysteries.* Business School Nederland.

Hundy, G. F. (2016). Refrigeration, Air Conditioning and Heat Pumps. In *Refrigeration, Air Conditioning and Heat Pumps*. Elsevier Science. https://doi.org/10.1016/c2014-0-03725-2

Hyndman, Rob J, & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International*

Journal of Forecasting, *22*(4), 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001

Hyndman, Robin John, & Athanasopoulos, G. (2018). *Forecasting: Principles and Practice* (2nd ed.). OTexts.

Iglewicz, B., & Hoaglin, D. C. (David C. (1993). *How to detect and handle outliers*. ASQC Quality Press.

International Organization for Standardization. (2019a). *Guidance on social responsibility (ISO Standard No. 26000:2010)*. https://www.iso.org/standard/42546.html

International Organization for Standardization. (2019b). *Information security management (ISO/IEC Standard No. 27000:2018)*. https://www.iso.org/isoiec-27001-information-security.html

International Organization for Standardization. (2019c). *Quality management systems — Fundamentals and vocabulary (ISO Standard No. 9000:2015)*. https://www.iso.org/standard/45481.html

International Organization for Standardization. (2019d). *Quality management systems — Requirements (ISO Standard No. 9001:2015)*. https://www.iso.org/standard/62085.html

Isermann, R. (1984). Process fault detection based on modeling and estimation methods-A survey. *Automatica*, *20*(4), 387–404. https://doi.org/10.1016/0005-1098(84)90098-0

Iyengar, S., Lee, S., Sheldon, D., & Shenoy, P. (2018). *So-larClique: Detecting Anomalies in Residential Solar Arrays*. https://doi.org/10.1145/3209811.3209860

Katipamula, S., & Brambley, M. R. (2005). Review article: Methods for fault detection, diagnostics, and prognostics for building systems—A review, part I. *HVAC and R Research*, *11*(1), 3–25. https://doi.org/10.1080/10789669.2005.10391123

Kim, S., & Kim, H. (2016). A new metric of absolute percentage error for intermittent demand forecasts. *International Journal of Forecasting*, *32*(3), 669–679. https://doi.org/10.1016/j.ijforecast.2015.12.003

Lazzarin, R., & Noro, M. (2018). Lessons learned from long term monitoring of a multisource heat pump system. *Energy and Buildings*, *174*, 335–346. https://doi.org/10.1016/j.enbuild.2018.06.051

Park, S., Park, S., Kim, M., & Hwang, E. (2020). Clustering-Based Self-Imputation of Unlabeled Fault Data in a Fleet of Photovoltaic Generation Systems. *Energies 2020, Vol. 13, Page 737*, *13*(3), 737. https://doi.org/10.3390/EN13030737

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. http://scikit-learn.org.

Reddy, T. A. (2011). Applied Data Analysis and Modeling for Energy Engineers and Scientists. In *Applied Data Analysis and Modeling for Energy Engineers and Scientists*. Springer US. https://doi.org/10.1007/978-1-4419-9613-8

Rijksdienst voor Ondernemend Nederland. (n.d.). *Energieprestatievergoeding (EPV)*. Retrieved May 31, 2021, from https://www.rvo.nl/onderwerpen/duurzaam-ondernemen/gebouwen/wetten-

en-regels/bestaande-bouw/energieprestatievergoeding

Rousseeuw, P. J., & Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, *88*(424), 1273–1283. https://doi.org/10.1080/01621459.1993.10476408

Seabold, S., & Perktold, J. (2010). statsmodels: Econometric and statistical modeling with python. *9th Python in Science Conference*.

Silver, N. C., & Dunlap, W. P. (1987). Averaging Correlation Coefficients: Should Fisher's z Transformation Be Used? *Journal of Applied Psychology*, *72*(1), 146–148. https://doi.org/10.1037/0021-9010.72.1.146

Stoffer, D. S., & Toloi, C. M. C. (1992). A note on the Ljung—Box—Pierce portmanteau statistic with missing data. *Statistics & Probability Letters*, *13*(5), 391–396. https://doi.org/10.1016/0167-7152(92)90112-I

Stroomversnelling. (2019). *Marktmonitor nul-op-de-meter, april 2019*. https://stroomversnelling.nl/wp-content/uploads/2019/04/Stroomversnelling-Marktmonitor-NOM.pdf

Stroomversnelling. (2020). *Marktmonitor nul-op-de-meter, juni 2020*. https://pages.stroomversnelling.nl/nom-marktmonitor-2020

Venkatasubramanian, V., Rengaswamy, R., Kavuri, S. N., & Yin, K. (2003). A review of process fault detection and diagnosis part III: Process history based methods. *Computers and Chemical Engineering*, *27*(3), 327–346. https://doi.org/10.1016/S0098-1354(02)00162-X

Wever, N. (2008). *Effectieve temperatuur en graaddagen : klimatologie en klimaatscenario's - Publicatiedatabank IenW*. http://publicaties.minienm.nl/documenten/effectieve-temperatuur-en-graaddagen-klimatologie-en-klimaatscen

White, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica*, *48*(4), 817. https://doi.org/10.2307/1912934

Yu, D., Yu, J., Sun, F., Deng, Y., Wu, Q., & Cong, G. (2017). Research on the PCA-based Intelligent Fault Detection Methodology for Sewage Source Heat Pump System. *Procedia Engineering*, *205*, 1064–1071. https://doi.org/10.1016/j.proeng.2017.10.171

Zhao, Y., Lehman, B., Ball, R., Mosesian, J., & De Palma, J. F. (2013). Outlier detection rules for fault detection in solar photovoltaic arrays. *Conference Proceedings - IEEE Applied Power Electronics Conference and Exposition - APEC*, 2913–2920. https://doi.org/10.1109/APEC.2013.6520712

Zimek, A., & Schubert, E. (2017). Outlier Detection. *Encyclopedia of Database Systems*, 1–5. https://doi.org/10.1007/978-1-4899-7993-3_80719-1

Appendix A: Overview of studied projects

Table 3 Project characteristics

Project number	Type of heat systems	Number of houses	Monitored period
1	Ventilation-air	54	02-2017 to present
2	Split geothermal system	33	04-2017 to present
3	Ventilation-air	40	04-2017 to present
4	Geothermal (ground-coupled heat exchange)	21	04-2018 to present
5	Geothermal (brine-water)	14	03-2019 to present

Appendix B: List of features returned from monitoring system

Table 4 Description of data collected by current monitoring system

	Field	Description	Unit
1	PV	Electric energy generated by solar panels	kWh
2a	Levering	Electric energy delivered by grid	kWh
2b	Ontvangst	Electric energy received by grid	kWh
3	CV (GJ)	Thermal energy consumption of central heating	GJ
-	CV (m^3)	Flow of water in central heating	m^3
4	WT (GJ)	Thermal energy consumption of domestic hot water (DHW)	GJ
-	WT ($m^{3)}$	Flow of water in domestic hot water (DHW)	m^3
5	WP	Electric energy consumed by heat pump	kWh
6	NV	Electric energy consumed by post heater	kWh
7	WTW	Electric energy consumption by heat exchange unit of ventilation	kWh
8	E-rad	Electric energy consumed by electric radiator	kWh
9	E-boil	Electric energy consumed by electric boiler	kWh
	Schaduwm.	Net energy delivered to grid (levering +/- ontvangst)	kWh
-	Temp	Temperature	kWh
-	Ventilatie	Energy consumed by ventilation	kWh
-	Bron (GJ)	Thermal energy consumed by ground-coupled heat exchanger	GJ
-	Bron (m^3)	Flow of water in ground-coupled heat exchanger	m^3

Appendix C: Pseudocode of the start of fault state marking heuristic

```
1  Procedure FindFaultStart:
2  Input: record    //Record instance
3  Output: Start date and time of the apparent fault state the record refers to. If it exists, None otherwise.
4     Set location <- location of record                                  //Retrieve record attributes
5     Set startdate <- initiation date of record
6     Set enddate <- completion date of record
7     Set fault_array to empty list                                       //Init empty list
8     For each sensor in the monitoring system of location do
9        Set t_fault <- None                                              //Reset t_fault from previous loop
10       Plot sensor readings over period (startdate – 30 days, enddate + 14 days)
11       If an abrupt change of trend is observed between startdate and enddate do
12                                                                        //We assume this constitutes a system
13                                                                        that has recovered from a fault state.
14          Set t_fault, the first point in time before startdate were the readings show an abrupt change in trend, if it exists.
15                                                                        //We assume this constitutes the start
16                                                                        of the fault state.
17          While t_fault seems to lie outside of plot do                 //The start of the fault state lies outside
18                                                                        of the plot, extend the plot.
19             Extend plotted range with xmin := xmin – 30 days and xmax := xmax + 14 days
20             Set t_fault, the first point in time before startdate were the readings show an abrupt change in trend, if it exists.
21          End while
22       End if
23       If not t_fault = None do
24          Set fault_array <- fault_array ∪ {t_fault}
25       End if
26    Next for
27    If not fault_array is empty:
28       Set output <- average(fault_array)
29    Else:
30       Set output <- None
31    Return output
32 End procedure
```

Appendix D: Pseudo-code for the K-Nearest neighbour estimation approach

Procedure K-NearestNeighbour
Input:
meterID :
s_date:
e_date:
K;
Output:
Estimate for sensor **meterID** between **s_date** and **e_date** if the sensor would not have been in fault.

Procedure:
1. Let **N** be the set of sensors with the same building type as sensor **meterID**
2. Let **r** be a list of size **K** filled with infinitely large numbers
3. Let **nn** be an empty list of size **K**.
4. For every sensor in **N**:
5. Calculate the distance between the fault free readings of sensor **meterID** and sensor **N**
6. If this distance is smaller than the largest value in **r**:
7. Replace the corresponding value of **r** with this distance
8. Replace the value of the corresponding index in **nn** with **N**
9. Let **result** be the daily arithmetic mean of the sensor readings of the sensors in **nn** over the period between **s_date** and **e_date**. If a sensor does not have a reading at that day, take the average of the other sensors
10. Return **result**

Figure 19 Pseudocode for the K-Nearest Neighbour heuristic

Appendix E: Linear correlation analysis

Introduction

In this exploratory analysis we study the (linear) relations between the monitoring data and the weather data.

Method

a. Pre-processing

We are only interested in the correlation between weather observations and the normal system response. Therefore, prior to running the correlation analysis, we remove from the data any readings that are made during which the studied system was suspecting to be in fault. These periods have been identified as discussed in section 2.3.

b. Compute correlations

We first compute the correlation coefficient between all features stored in either the monitoring data or the weather data for every individual house. The correlation coefficient between the responses of two features are computed using the Pearson correlation coefficient provided as **Equation 16**. This is a measure strictly for linear correlation.

Equation 16 *Pearson product-moment correlation coefficient between two variables, where X is the first series, Y is the second series and σ_X and σ_Y the standard deviation of X and Y respectively.*

$$r = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

c. Group results

To study the general tendencies within the entire dataset, we average the correlation coefficients of the individual houses into a single correlation coefficient for every feature at project level. We cannot, however, simply take the average of the individual correlation coefficient estimates, because the sampling distribution of the correlation coefficient is skewed (Silver & Dunlap, 1987). Instead, as proposed by Alexander (1990), we first transform the sampling distributions of every correlation coefficient to become normally distributed using Fisher z transformation denoted in Equation 17. Then, we take the average of the z-scores. Finally, we back-transform this average using the hyperbolic tangent to return the average correlation coefficient between two features at a project level.

Equation 17 Fisher z-transformation, with

$$z = \operatorname{arctanh}(r)$$

Results

Project 1

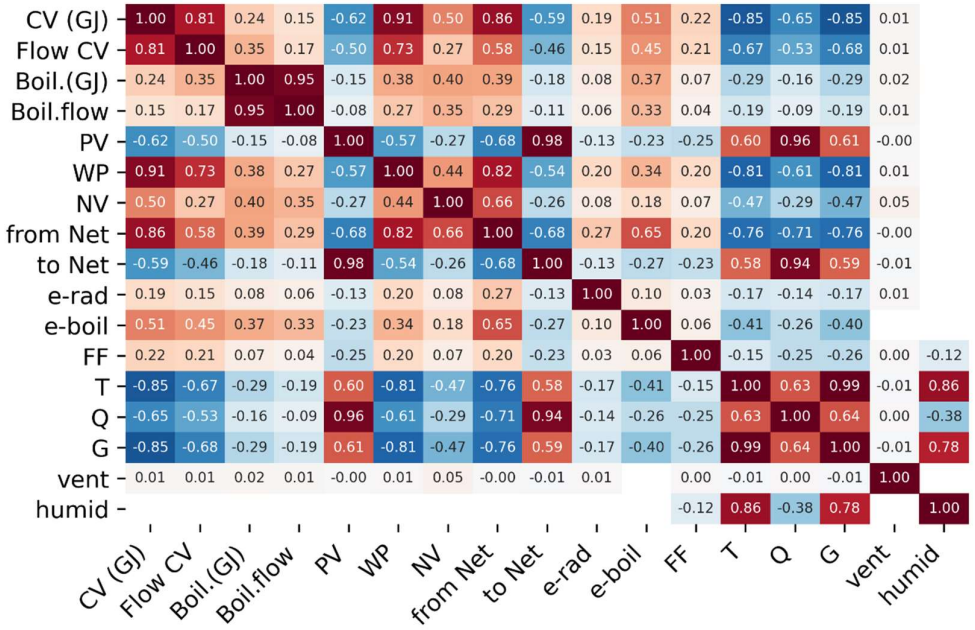

Figure 20 Correlation matrix for project 1, values denote the sample correlation coefficients of all readings within the project

Project 2

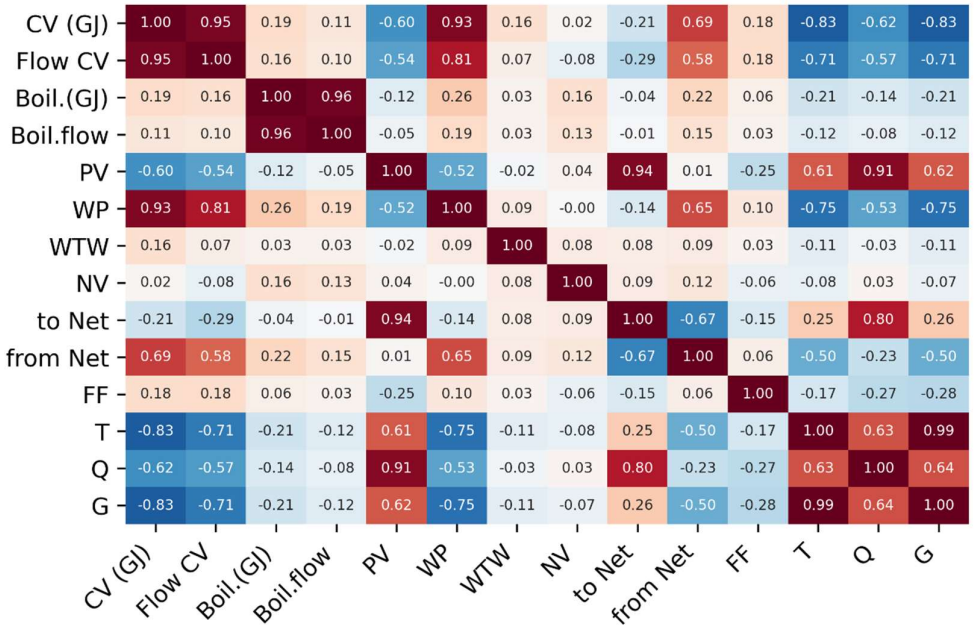

Figure 21 Correlation matrix for project 2, values denote the sample correlation coefficients of all readings within the project

Project 3

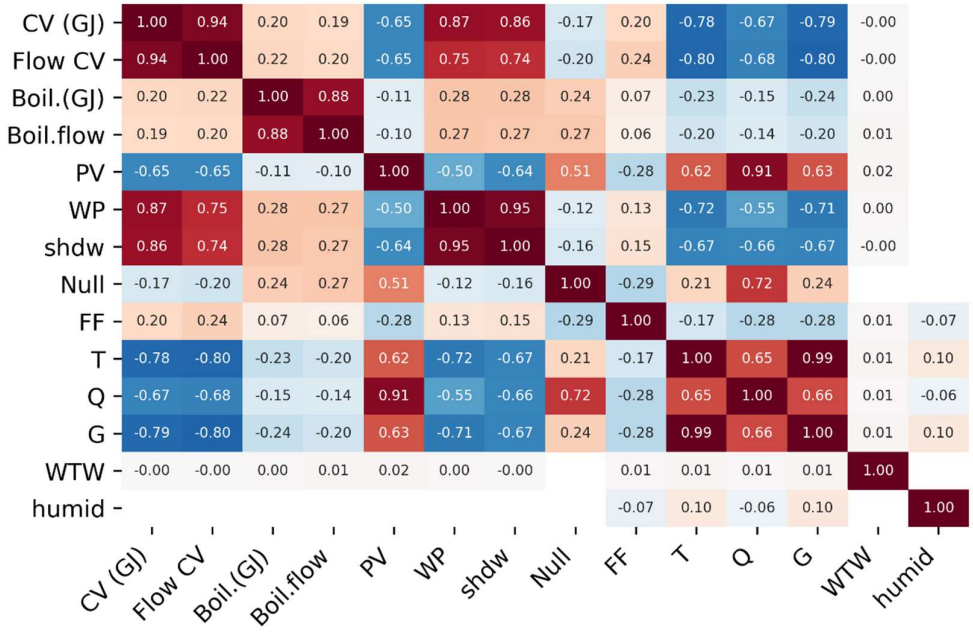

Figure 22 Correlation matrix for project 3, values denote the sample correlation coefficients of all readings within the project

Project 4

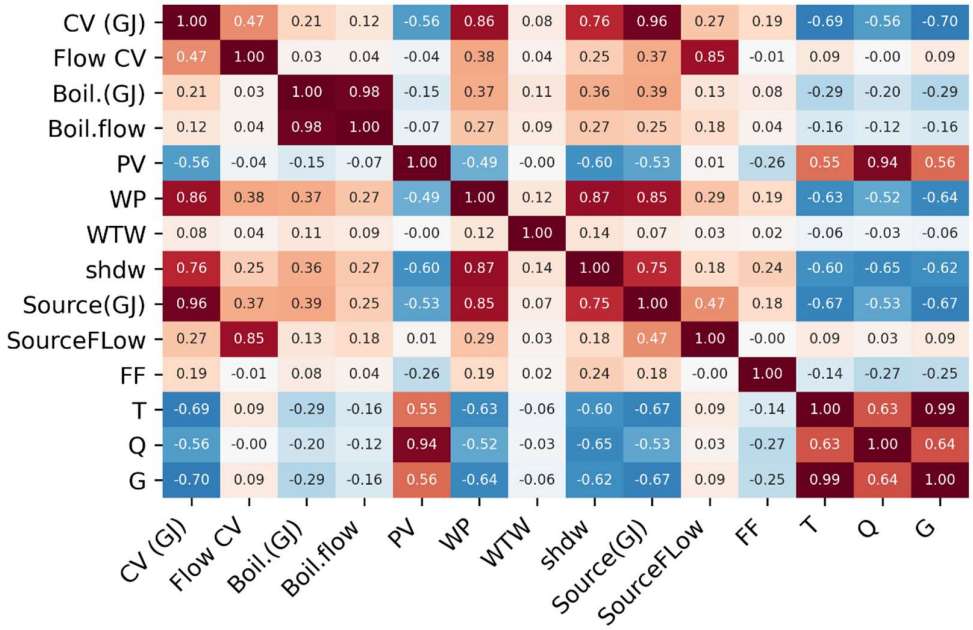

Figure 23 Correlation matrix for project 4, values denote the sample correlation coefficients of all readings within the project

Project 5

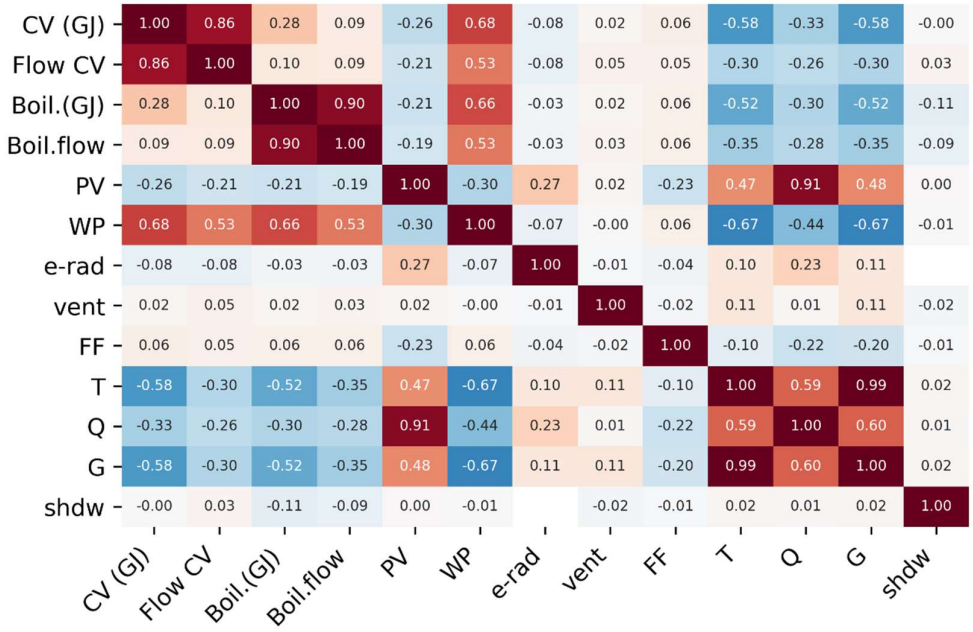

Figure 24 Correlation matrix for project 5, values denote the sample correlation coefficients of all readings within the project

Distribution of results

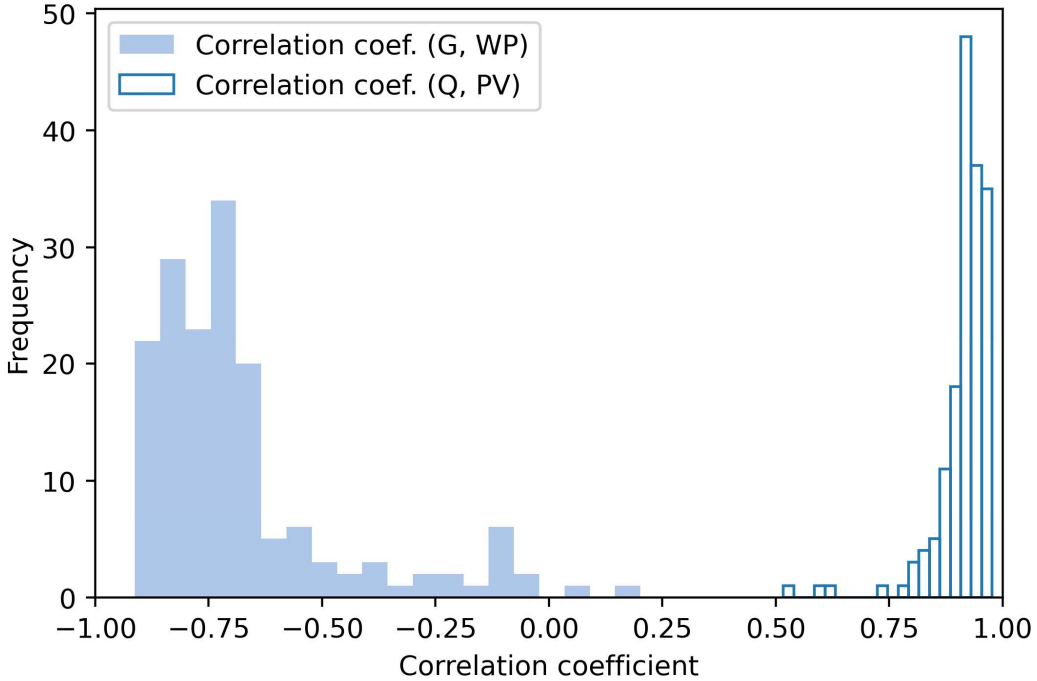

Figure 25 Distribution of correlation coefficients between wind chill (G) and heat pump energy use (WP) and solar irradiance (Q) and PV energy yield (PV)

Conclusions

In general, we found a strong negative correlation between wind chill and energy used by the heat pump apparatus; r = -0.81 for project 1, r = -0.75 for project 2, r = -0.71 for project 3, r = -0.64 for project 4 and r = -0.67 for project 5.

Furthermore, we found a similarly strong negative correlation between outdoor temperature and energy used by the heat pump apparatus; r = -0.81 for project 1, r = -0.75 for project 2, r = -0.72 for project 3, r = -0.63 for project 4 and r = -0.67 for project 5.

Finally, we found a similarly strong positive correlation between solar irradiance and energy yield by the PV system; r = 0.96 for project 1, r = 0.91 for project 2, r = 0.91 for project 3, r = 0.94 for project 4 and r = 0.91 for project 5.

Looking into the heat pump energy use responses that show the lowest correlation coefficient, we find that most of these can be explained. Many of these series have many missing readings, either because of sensor failure or because the system was identified of being in fault and thus a period of readings was removed. We find that there are certain periods during which the energy use of the heat pump system does not correlate to the changes in the wind chill. This is the case if the heating system has already reached the limits of its capacity, e.g., because space heating is not required when the difference between the wind chill and the target temperature is sufficiently small, so that the loss of heating energy due to inference can be corrected by the internal load of the house. If the remaining readings are mostly during spring or summer, seasons during which use of the heat pump for space heating is low, then the energy use of this system will seem uncorrelated to the observed wind chill.

Appendix F: Current situation models validation results

Introduction

In this appendix, we report the results of the validation of the 3-Nearest Neighbour (3-NN) and Ordinary Least Squares (OLS) simple linear regression models for estimating the normal system responses of the studied PV and heating system. We first discuss the results of the PV system response estimates. Second, we discuss the heating system response estimates. Particularly, we discuss the heat pump systems and, if applicable, the auxiliary heater, which are the primary components of the heating system.

PV systems

The MAAPEs and MedAEs of the 3NN and OLS models for normal system response estimates of the studied PV and heat pump systems are visualised in and Figure 27 respectively.

Figure 26 shows that most nearest neighbour estimates have lower MAAPE than any OLS estimates. In fact, the highest observed MAAPE for the nearest neighbour estimate, except for a few outliers, is lower than the lowest observed MAAPE for the OLS estimate in all the studied projects.

The 3-KNN returns poor estimates for 11 PV systems. These estimates are recognized as the outliers in the higher quartiles of Figure 26 and Figure 27. Looking further into these cases, these results can be explained by differences between the responses in the training and the validation data. Particularly, most of these outlier cases show an extended period of little or no PV yield in either the training or the validation data. The models that have been generated for these systems cannot be used to accurately estimate the normal response of these systems. Fortunately, most of the houses that relate to these systems do not appear in the service records. The records that do relate to a house for which the 3-NN approach does not return an accurate model are shown in Table 5. We exclude these records from our results, as none of the faults discussed in these records relate to the PV system. Therefore, it is expected that there are no consequences of these faults that affect the yield of the PV system, regardless of what is estimated by the respective models.

Table 5 Corrected estimates for the energy yield of the PV system, where 3-NN is '3-Nearest Neighbour'

Project	Estimation method	Service order nr.	Fault type	Estimated impact	Consequence according to model	#days Obs. Delay
1	3-NN	1	Loss of signal (Hardware defect)	0 kWh	3,385 kWh	131
1	3-NN	2	Filter fouling (in heating system)	0 kWh	65 kWh	201
4	3-NN	3	Unknown failure (in heating system)	0 kWh	7 kWh	2

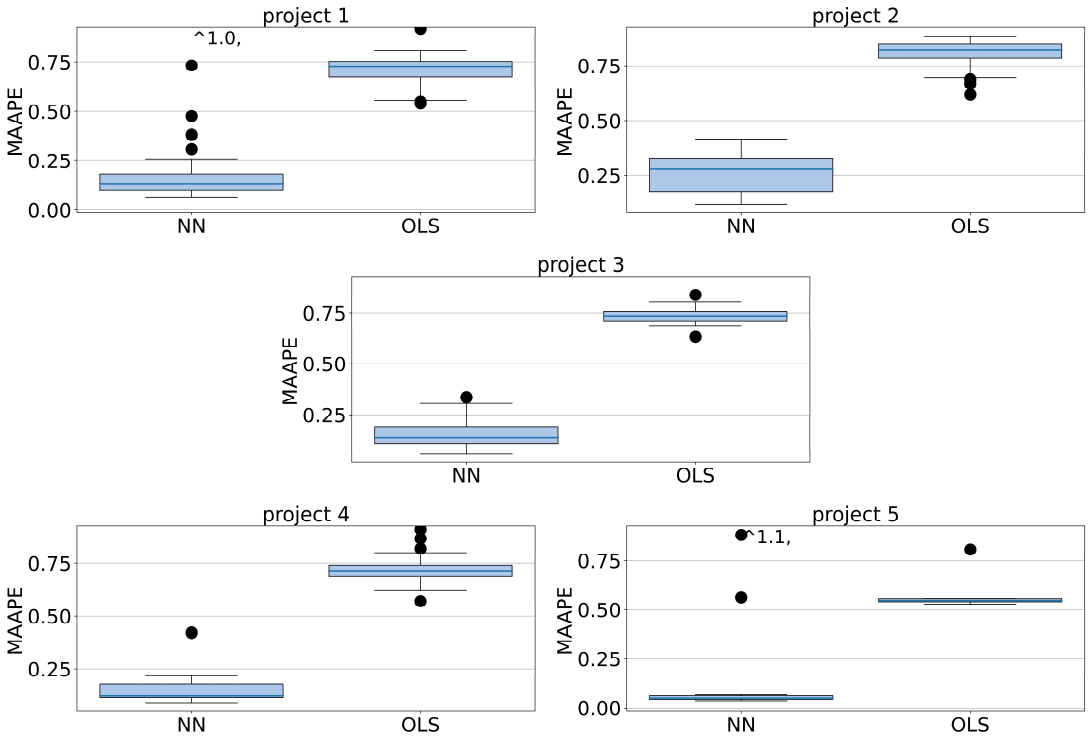

Figure 26 Boxplots of MAAPE of PV power output sensor reading estimates from the 'Nearest Neighbour' (NN) and 'Ordinary Least Squares regression' (OLS), far outliers are noted at the top of the respective boxplot after '^'.

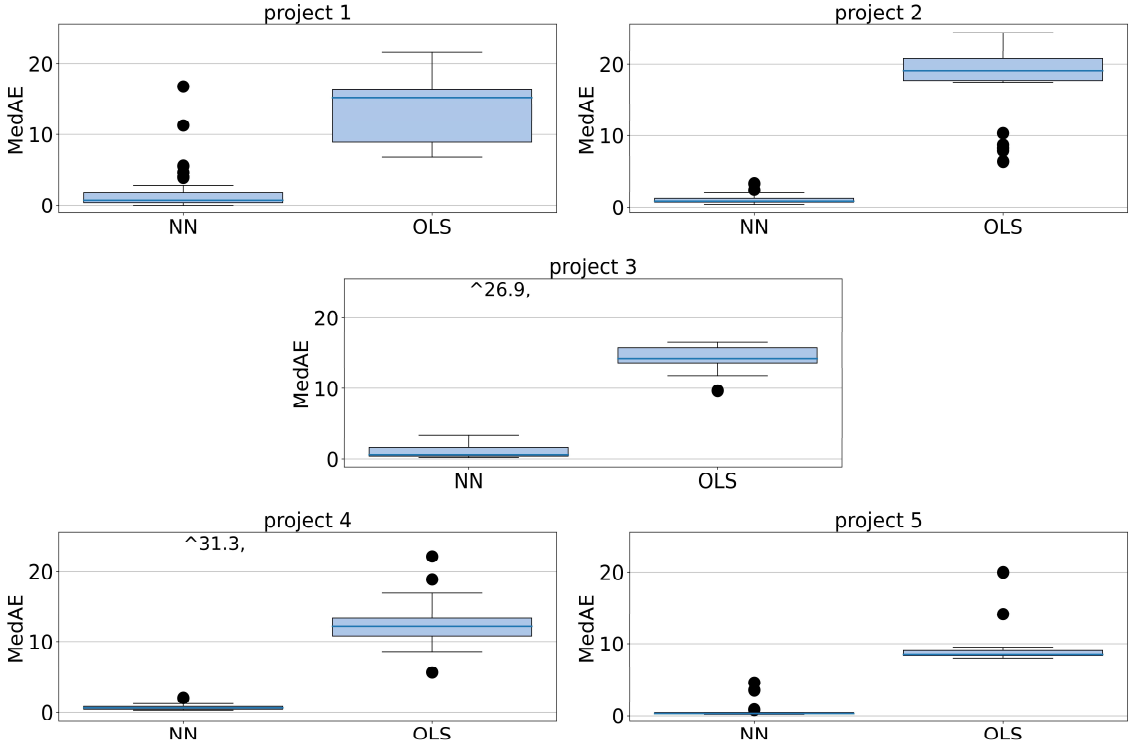

Figure 27 Boxplots of MedAE of PV sensor reading estimates from the 'Nearest Neighbour' (NN) and 'Ordinary Least Squares regression' (OLS) , far outliers are noted at the top of the respective boxplot after '^'.

Heating system

Second, we discuss the results of the validation of the models for the normal response of the heating system, of which the most important subsystem is the heat pump and possibly an auxiliary heater. In 3 of the 5 studied projects, the electric auxiliary heater is integrated in the heat pump, in which case energy use of the auxiliary heater is not monitored separately but part of the energy use by the heat pump.

Heat pump

The dispersion of average MAAPEs and MedAEs of the estimates for energy use of the heat pump per house are visualised in Figure 28 and Figure 29 respectively. In contrast to what we observe in the PV systems, we do not find the dispersion and medians the be strikingly different between approaches. This can be explained by the fact that the response of the PV systems is for a large part unrelated to the behaviour of the residents of a house, whereas the response of heat pump systems is, for a large part, dependent on the settings of the system and the internal heat load, which are dependent on the behaviour of the residents. Therefore, a lower variance is expected between the responses of the PV systems of different houses then what is expected between the responses of heating systems. As a result, the variance of the estimation error of the nearest neighbour regression model is higher for the heating systems, parring it variance of the OLS regression estimation errors. A similar result is found looking at the dispersions and medians of the MedAEs of the heat pump energy use estimates in Figure 29.

If we study the differences in dispersion between individual projects, we find that the dispersion of average MAAPEs for the estimates for projects that are equipped with a separate auxiliary heater, 1 and 2, are notably greater than the dispersions of average MAAPEs in the other projects.

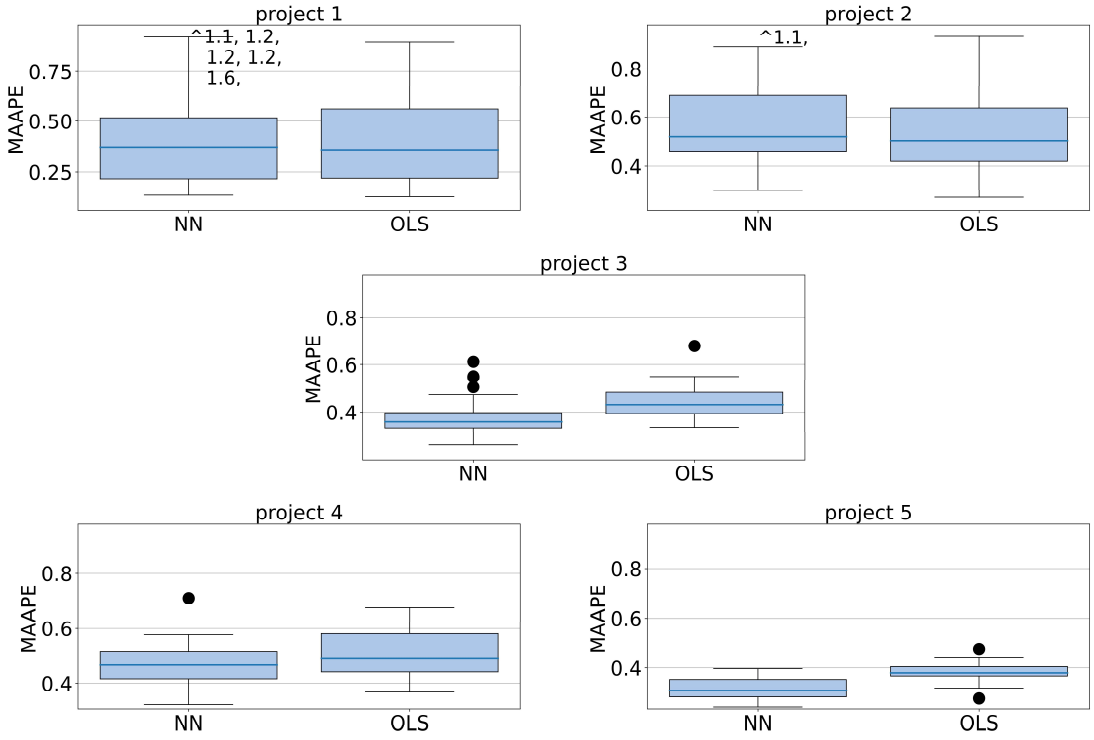

Figure 28 Boxplots of MAAPE of the heat pump electric use sensor reading estimates from the 'Nearest Neighbour' (NN) and 'Ordinary Least Squares regression' (OLS) , far outliers are noted at the top of the respective boxplot after '^'.

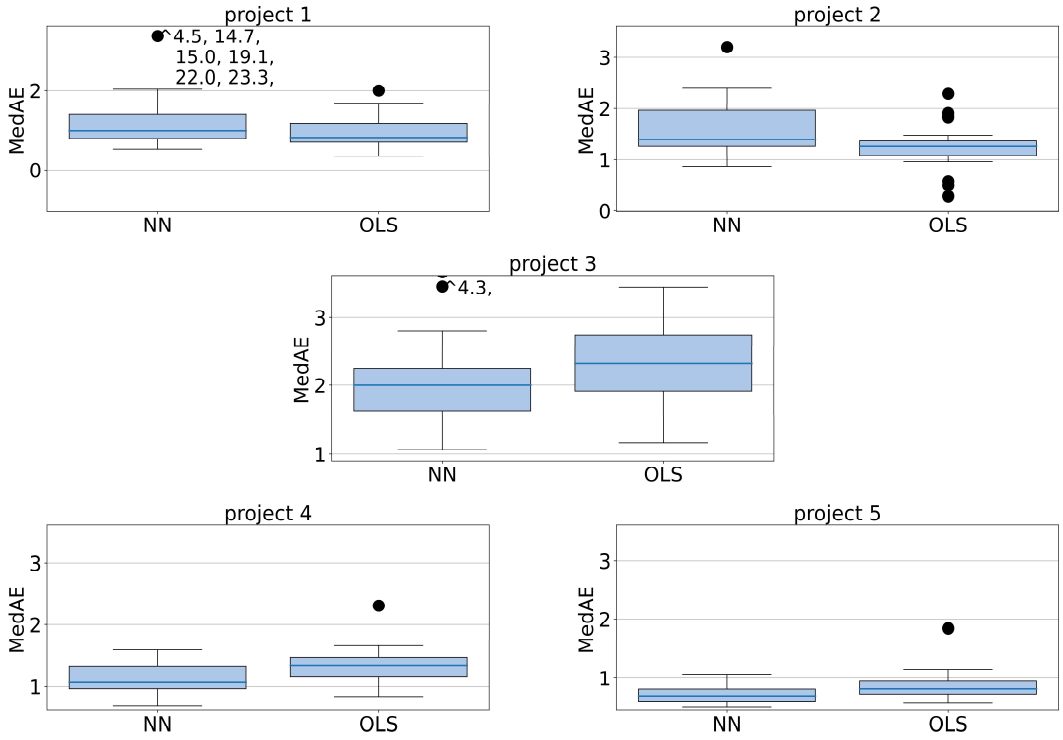

Figure 29 Boxplots of MedAE of the heat pump electric use sensor reading estimates from the 'Nearest Neighbour' (NN) and 'Ordinary Least Squares regression' (OLS) , far outliers are noted at the top of the respective boxplot after '^'.

Particularly notable are that both the nearest neighbour and the OLS estimates have a few estimates for project 1 that have a very high MedAE, which are recognized as the outliers in Figure 29. Similar to the outliers observed in the estimates of the PV system, these are explained by distinct periods of no or little energy use in either the training or validation data. This suggest there are unmarked periods of system failure. The quality of the estimates of these systems cannot be guaranteed and these models should therefore not be used to estimate the consequences of failure. Of the systems for which these poor 3-NN and OLS models are constructed, 6 and 5 occur in the service orders respectively. These service orders are shown in Table 6. Since the estimates from these models are not accurate, these are not included in our results. Instead, we include a manual estimate of the consequences of these service orders:

- Some service orders describe faults that do not affect the energy use of the heating system, we estimate no consequences of these faults to the energy use of the heat pump.
- For the other service orders, it is either the 3-NN or the OLS model that has a poor performance, but never both. If the 3-NN model performance poorly, we therefore use only the OLS model and vice versa. The total additional annual consequence of these cases is 168 kWh, of which a 163 kWh contribution of a single failure in project 4 is most notable.

Table 6 Corrected estimates for the energy use of the heat pump, where 3-NN is '3-Nearest Neighbour' and OLS is 'Ordinary least squares linear regression'

Project	Estimation method	Service order nr.	Fault type	Estimated impact	Consequence according to model	#days Obs. Delay
1	3-NN	4	Unknown failure	0 kWh (OLS)	2 kWh	2
1	3-NN	5	Loss of sensor accuracy	0 kWh	21 kWh	20
2	3-NN	6	Configuration error	0 kWh	56 kWh	9
2	3-NN	7	Wrongly installed	12 kWh (OLS)	40 kWh	4
2	3-NN	8	Inverter failure (PV system)	0 kWh	256 kWh	36
2	3-NN	9	Inverter failure (PV system)	0 kWh	0 kWh	11
1	OLS	10	Filter fouling	5 kWh (3-NN)	1 kWh	2
1	OLS	11	Low source temperature	3 kWh (3-NN)	2 kWh	7
1	OLS	12	Loss of sensor accuracy	0 kWh	21 kWh	20
3	OLS	13	Inverter failure (PV system)	0 kWh	16 kWh	33
4	OLS	14	Distribution valve failure	418 kWh (3-NN)	423 kWh	9

Auxiliary Heater

The dispersion of average MAAPE and MedAE of the energy use of the auxiliary heater per house is visualised by the boxplots in Figure 30 and Figure 31 respectively. Like the heat pump energy use estimates, we detect no large difference between the nearest neighbour and OLS regarding the dispersion and medians of MAAPEs or MedAEs for the estimates of the energy use of the auxiliary heaters.

Comparing the dispersion of the average MAAPEs, there is a notable difference between the two projects. The dispersion of the average MAAPEs in project 1 is greater than that of project 2. This does, however, not translate to a large difference in the dispersion or median of the average MedAEs. We also observe that the MAAPE of the estimates for project 1 is much higher than the MAAPE of the estimates for project 2.

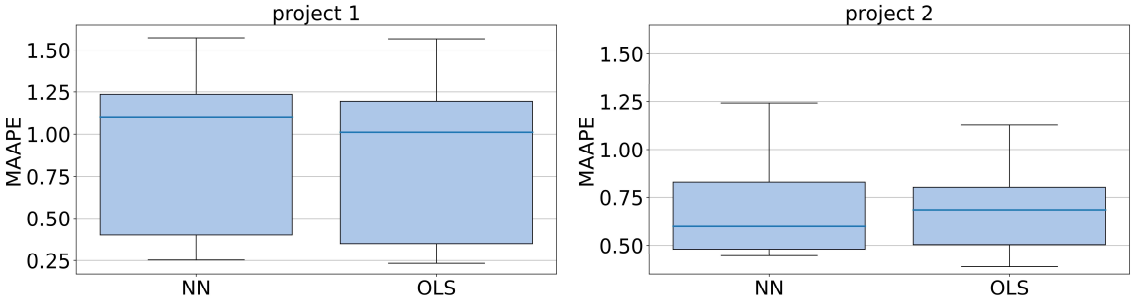

Figure 30 Boxplots of MAAPE of the auxilary heater electric use sensor reading estimates from the 'Nearest Neighbour' (NN) and 'Ordinary Least Squares regression' (OLS).

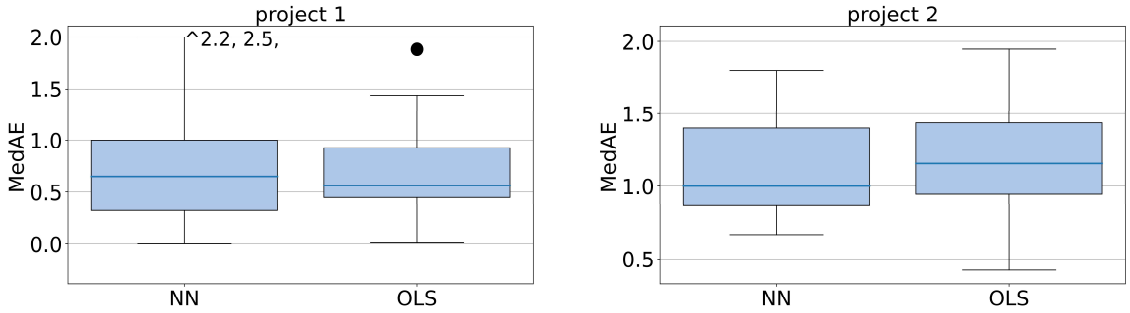

Figure 31 Boxplots of MedAE of the auxilary heater electric use sensor reading estimates from the 'Nearest Neighbour' (NN) and 'Ordinary Least Squares regression' (OLS).

None of the auxiliary heating systems for which a poor estimation model was constructed occurs in the service orders. The outliers in Figure 31 are therefore not an issue to our results in the following section.

Appendix G: Distributions of the parameter-levels fitted to the linear OLS regression models to estimate the normal system response

PV systems

The normal energy yields of the PV systems are estimated using equation 2.The distribution of the best found parameter levels for these models is shown in Figure 32.

Equation 2 Simple linear regression, where ŷ is the estimated energy yield by the PV system in kWh, x is the solar irradiance measured at the geographically nearest weather station in kJ/m2, β0¬ is the intercept and β1 is a coefficient.

$$\hat{y} = \beta_0 + \beta_1 x$$

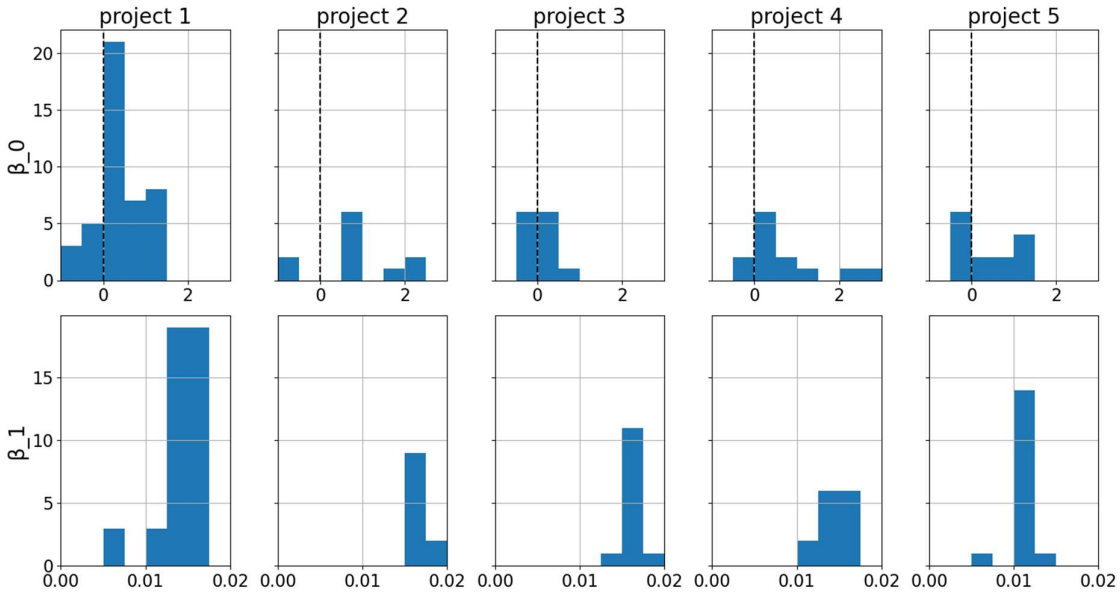

Figure 32 Distribution of the best found parameter levels for the ordinary least squares regression models to predict the normal system response of the studied PV systems.

Heat pump systems

The normal energy usage of the heat pump systems is estimated using equation 4.The distribution of the best found parameter levels for these models is shown in Figure 33 .

Equation 4 Discontinuous piecewise linear where x is the wind chill in degrees Celsius equivalent, y is the estimated energy use of the candidate apparatus, β_{00} the intercept of the first segment at x=0, β_{10} the coefficient of the independent variable of the first segment. β_{01} the intercept of the second segment at x=14, β_{11} the coefficient of the independent variable of the second segment.

$$y = max \begin{cases} \beta_{00} + \beta_{10} x \\ \beta_{01} + \beta_{11} x \end{cases}$$

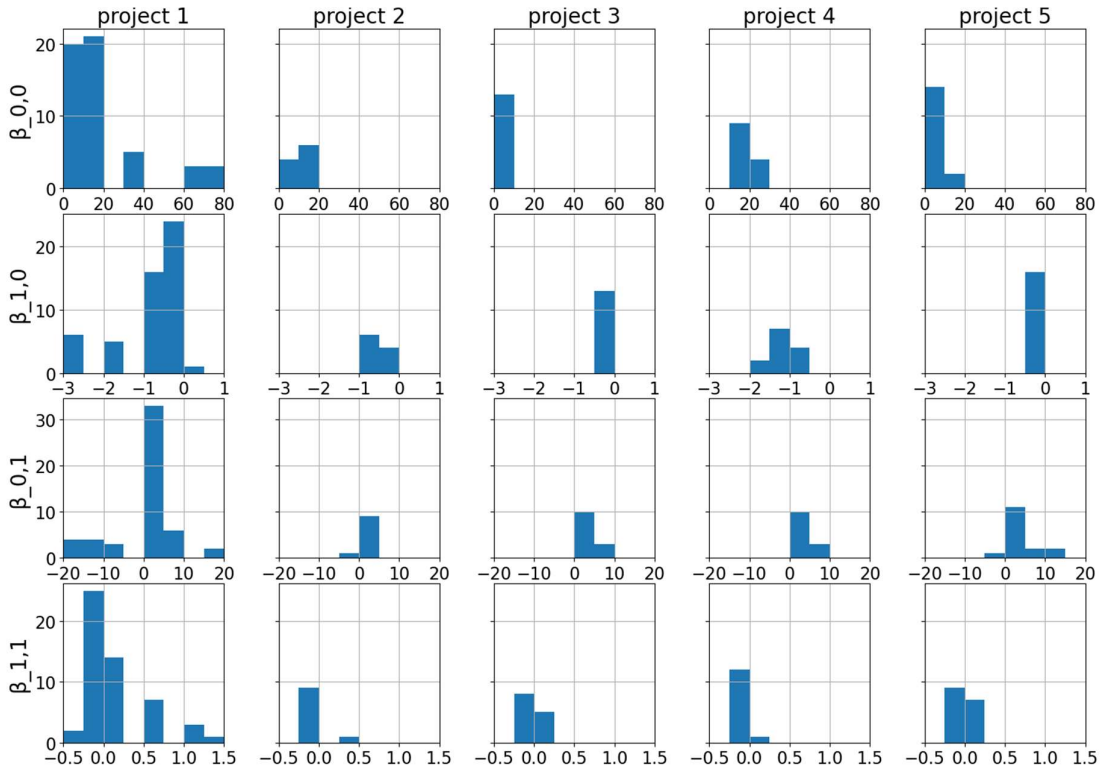

Figure 33 Distribution of the best found parameter levels for the piecewise ordinary least squares regression models to predict the normal system response of the studied heat pump systems.

Auxiliary heater systems

The normal energy use of the auxiliary heater systems are estimated similar to how the heat pump energy use is estimated, using equation 4. The distribution of the best found parameter levels for these models is shown in Figure 34.

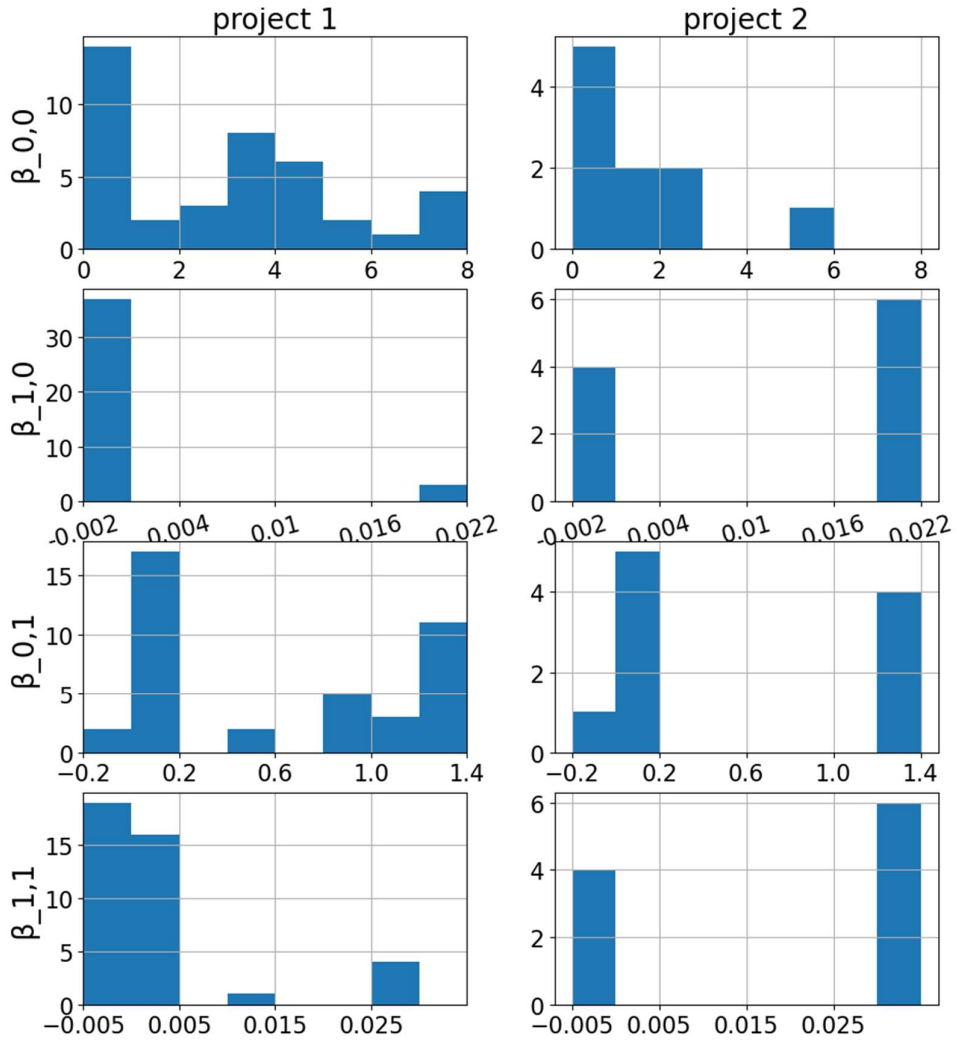

Figure 34 Distribution of the best found parameter levels for the piecewise ordinary least squares regression models to predict the normal system response of the studied heat pump systems.

Appendix H: Exploratory data plot of monitoring data after visual marking of fault data

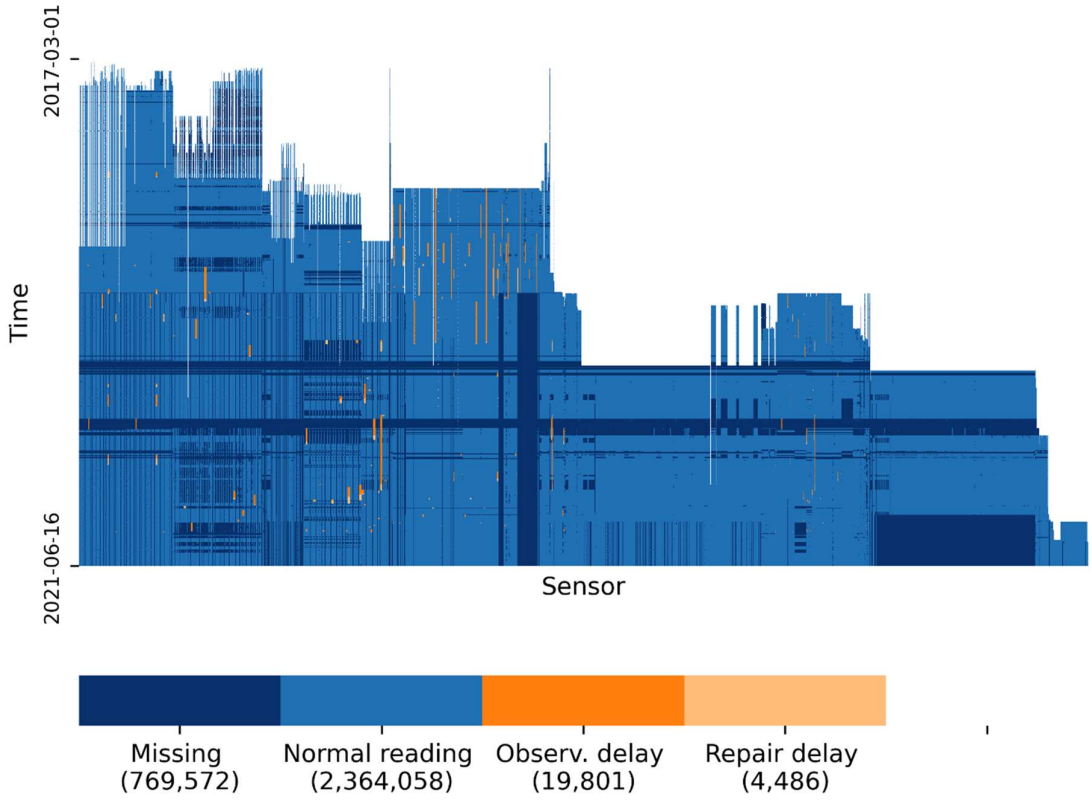

Figure 35 Exploratory data plot, with missing data marked dark blue, unmarked sensor readings in light blue, readings from the system while marked as being in detection delay marked in dark orange and readings form the system while marked as being in repair delay marked in light orange. Between brackets the frequency of the respective state.

Appendix I: SolarClique Algorithm

The algorithms and notation in this appendix are adopted from Hastie et al. (2009), Iyengar et al. (2018) the scikit-learn documentation (*1.11 Ensemble methods — scikit-learn 1.0.2 documentation*, n.d.; Pedregosa et al., 2011)

I. For j = 1 to J;

1. For b= 1 to B;
    a. Draw a bootstrap sample Q_0 of size N from the training data, with

$Q_0 = \left(\{X_{0,i} \middle| \, i = 1, \dots, N\}, Y_0\right)$

$X_{0,i} \in R^n$; the *i*th training vector of the bootstrap sample at the top node and

$Y_0 \in R^N$; the label vector of the top node

    b. Grow a tree T_b to the bootstrapped sample Q_0;

Let Q_m be the data at node m with N_m samples

$Q_m = \left(\{X_{m,i} \middle| \, i = 1, \dots N_m\}, Y_m\right)$

$X_{m,i} \in R^{N_m}, i = 1, \dots, N;$; the *i*th training vector of the bootstrap sample at the node m and

$Y_m \in R^{N_m}$; the label vector of the top node

and let

$$H(Q_m) = \frac{1}{N_m} \sum_{y \in Y_m} (y - \bar{y})$$

be the (MSE) impurity function, with $\bar{y}$ the mean of all values in Y_m.

For each candidate split θ = (j, t_m) consisting of feature j and the threshold-value for that feature t_m, partition the data into $Q_m^{left}(\theta)$ and $Q_m^{right}(\theta)$ subsets. Such that:

$$Q_m^{left}(\theta) = \left\{\left(X_{m,j}, Y_m\right) | X_{m,j} \leq t_m\right\}$$
$$Q_m^{right}(\theta) = Q_m \setminus Q_m^{left}(\theta)$$

Let the quality of the candidate split of node m be determined by the impurity function H;

$$G(Q_m, \theta) = \frac{N_m^{left}}{N_m} H\left(Q_m^{left}(\theta)\right) + \frac{N_m^{right}}{N_m} H\left(Q_m^{right}(\theta)\right)$$

Select the split that minimizes the impurity;

$$\theta^* = argmin_\theta \, G(Q_m, \theta)$$

Recurse for subsets $Q_m^{left}(\theta^*)$ and $Q_m^{right}(\theta^*)$ until m = the maximum allowable depth or N_m = 1

2. Given a vector of power generation values from nearby PV systems, W, predict the yield of the candidate system using the *j*th random forest submodel;

$$E[Y|W]_j = \frac{1}{B} \sum_{b=1}^{B} T_b(W)$$

II. Combine the ensemble of submodel estimates to make the final initial estimate;

$$E[Y|W] = \frac{1}{J} \sum_{j=1}^{J} E[Y|W]_j$$

and the vector of estimation errors;

$$\hat{L} = Y - E[Y|W]$$

And the vector of standard deviations of the estimation errors over the ensemble of models for each day.

III. Remove seasonal component of the estimation error;

Let *F* be a time series decomposition function to remove seasonal components and *A* be a vector of deseasonalized estimation errors.

$$\hat{A} = F(\hat{L})$$

IV. Detect anomalies;

Let *T* be the set of observed days and A_t the estimation error at day t, and σ_t the standard deviation of the estimation error at day t.

To be indicative of a failure, the fault should occur for at least *k* contiguous days. Therefore, an anomaly is defined as;

$$anomaly = \left(\hat{A}_t < -4\sigma_t\right) \bigwedge \ldots \bigwedge \left(\hat{A}_t < -4\sigma_{t+k}\right) \quad \forall t \in T$$

Appendix J: White statistical test on validation models' residuals

Introduction

In this analysis we perform the White statistical test on the residuals of all SolarClique and PCR regression models' estimates of the validation data. The White statistical test is a test of whether the variance of the residuals is dependent on the values of the independent variables.

The null and alternative hypothesis can be given as follows (Greene, 2003):

H_0: The variance of the residuals is equal.

H_1: The variance of the residual is not equal.

Method

The test statistic of the White statistical test is a Lagrange multiplier and given by Equation 18. Under H_0 the test statistic follows a Chi-squared distribution with P – 1 degrees of freedom, where P is the number of independent variables used in the model, including the constant (Greene, 2003; White, 1980).

Equation 18 Test statistic of the White statistical test, with n, the sample size and R^2 the squared multiple correlation coefficient from the regression (White, 1980).

$$LM = nR^2$$

The White test is very general, we need not make any assumptions about the nature of the heteroscedasticity. More specifically, this means that this test also returns a significant result if the relation between the change in variance of the residuals and the independent variables used as input to the regression model is non-linear (Greene, 2003). Moreover, because of this, H_0 is also sometimes rejected by this test if there in fact exists no heteroscedasticity, but there is another error in model specification in the model (Greene, 2003).

To compute the White test statistics of each set of residuals in the validation, the het_white function of the Statsmodels package for Python 3 was used (Seabold & Perktold, 2010). The returned value is the *p*-value.

Results

The distribution of *p*-values of the SolarClique models is shown in Figure 36. H_0 is rejected for 32 of the 156 models' residuals. H_0 is rejected for all cases where the *p*-value is smaller than 0.05.

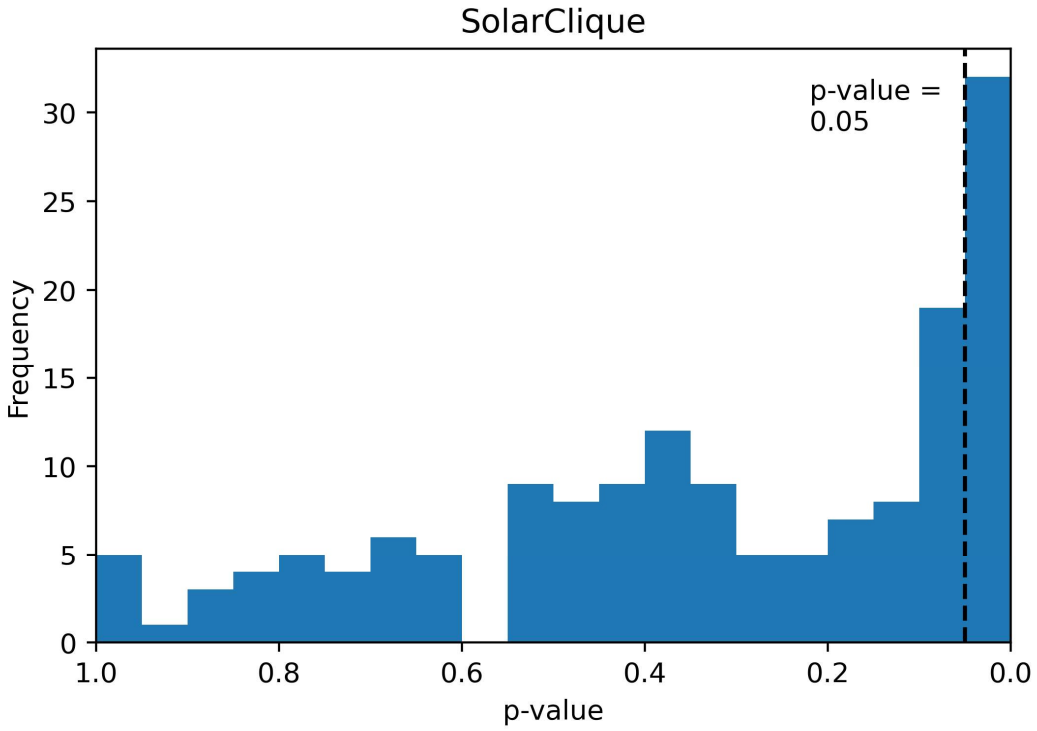

Figure 36 Distribution of the White test p-values of the SolarClique models' residuals on the validation data

The distribution of *p*-values of the principal component regression models is shown in Figure 37. H_0 is rejected for 64 of the 152 models' residuals. H_0 is rejected for all cases where the *p*-value is smaller than 0.05.

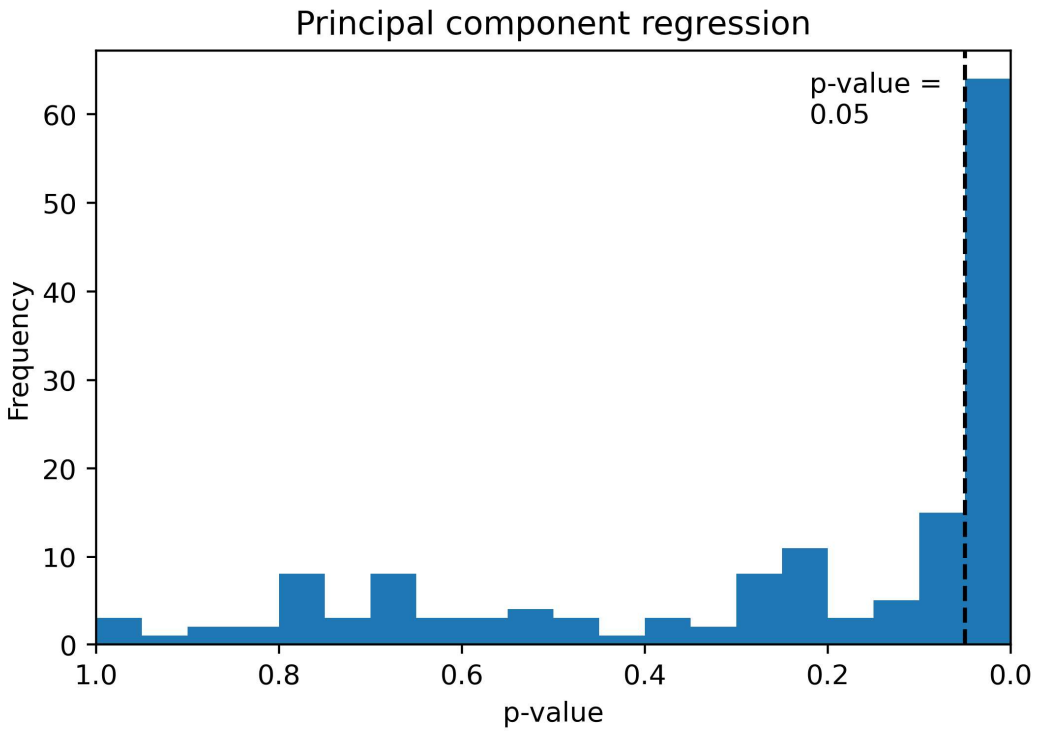

Figure 37 Distribution of the White test p-values of the principal component regression models' residuals on the validation data

Appendix K: Stoffer-Toloi statistical test on validation models' residuals

Introduction

In this analysis we perform the Stoffer-Toloi statistical test on the residuals of all SolarClique and PCR regression models' estimates of the validation data. The Stoffer-Toloi test is a statistical test of whether any group of sample autocorrelations is significantly different from zero. The null and alternative hypothesis can be defined as follows:

H_0: The autocorrelation of the residuals is independently distributed.

H_1: The autocorrelation of the residuals is not independently distributed, there exists autocorrelation.

Method

For sufficiently large samples, the sample autocorrelations of a sequence of independents and identically distributed observations with a finite variance approximately follow a normal distribution with mean 0 and a standard deviation of 1/n (Brockwell & Davis, 2016; Robin John Hyndman & Athanasopoulos, 2018). In the traditional *sample autocorrelation function,* this property is leveraged to identify significant autocorrelation in a sample by checking for significant outliers of autocorrelation up to many lags.

Instead of looking at individual levels of lag, the Stoffer-Toloi test considers a single statistic for the entire range of lags (Brockwell & Davis, 2016; Stoffer & Toloi, 1992). The Stoffer-Toloi statistical test is a correction on the Ljung-Box statistical test to support missing observation. The test statistic of the Stoffer-Toloi statistical test is shown in Equation 20. For a complete derivation see Stoffer & Toloi (1992).

Equation 19 Weight correction to k-th observation in the sample, with n, the sample size, and a(t) = 0 if the observation was missed at time t and 1 otherwise (Dunsmuir & Robinson, 1981).

$$C_a(k) = \sum_{t=k+1}^{n} \frac{a(t)a(t-k)}{n-k}$$

Equation 20 Test statistic of the Stoffer-Toloi statistical test, with n, the sample size, $C_a(k)$ is the weight computed by Equation 19, $r_e(k)$ an estimate of the sample autocorrelation at lag k and h the number of lags being tested (Stoffer & Toloi, 1992).

$$Q = n^2 \sum_{k=1}^{h} \frac{C_a(k) r_e(k)}{n-k}$$

The Stoffer-Toloi, similar to the sample autocorrelation function, assumes the autocorrelation at each lag is normally distributed, and continues that this implies that test statistic Q follows a Chi-squared distribution with *h* degrees of freedom, where *h* is the level of lag up to which the test is performed (Stoffer & Toloi, 1992).Therefore, we have that, for a significance level α, H_0 is rejected equation 14 holds.

Equation 21 Critical region for the rejection of H_0 of the Stoffer-Toloi test, with Q the test statistic of the Stoffer-Toloi test, $\chi^2_{1-\alpha,h}$, the 1-α-quantile of the chi-squared distribution with h degrees of freedom and h the level of lag up to which the test is performed.

$$Q > \chi^2_{1-\alpha,h}$$

To compute the Stoffer-Toloi test statistics of each set of residuals in the validation, the 'stoffer-toloi' function of the Pastas package for Python 3 was used (Collenteur et al., 2019). The returned value is the *p*-value. There is some debate on the number of lag orders to consider, we follow the general rule-of-thump, which is to use the square root of the sample size. With 3 months of validation data, and one reading per day, the square root of the sample size is approximately 9 orders of lag.

Results

The distribution of *p*-values of the SolarClique regression models is shown in Figure 38. H_0 is rejected for 73 of the 156 models' residuals. H_0 is rejected for all cases where the *p*-value is smaller than 0.05.

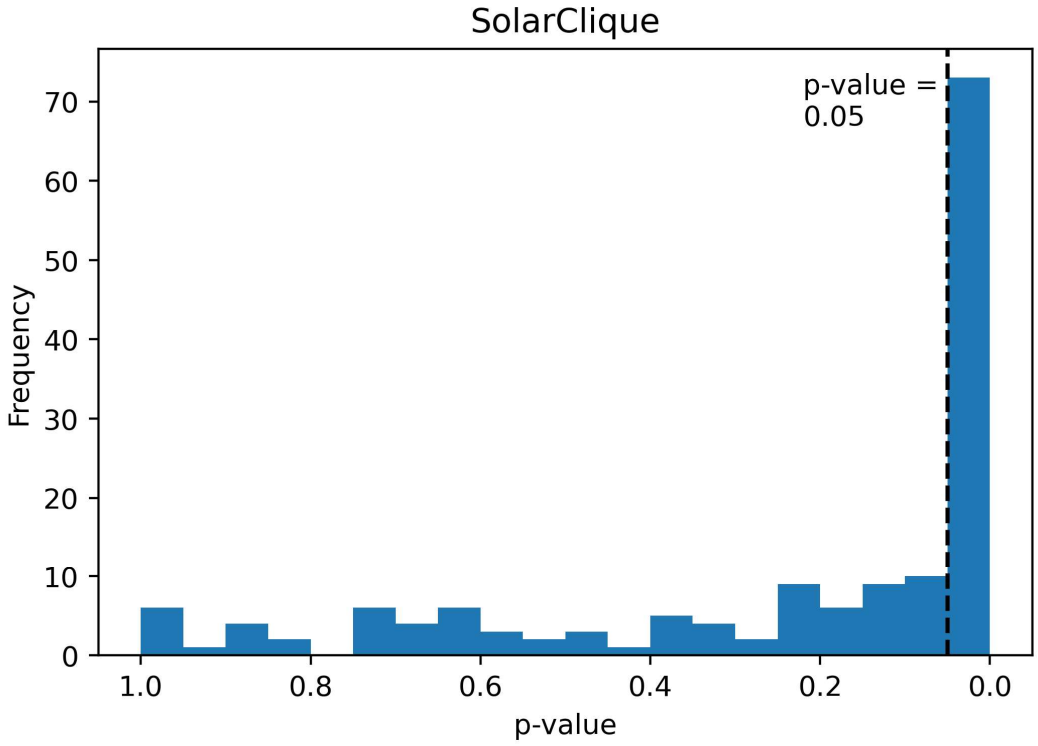

Figure 38 Distribution of the Stoffer-Toloi test p-values of the SolarClique models' residuals on the validation data

The distribution of *p*-values of the principal component regression models is shown in Figure 39. H_0 is rejected for 128 of the 152 models' residuals. H_0 is rejected for all cases where the *p*-value is smaller than 0.05.

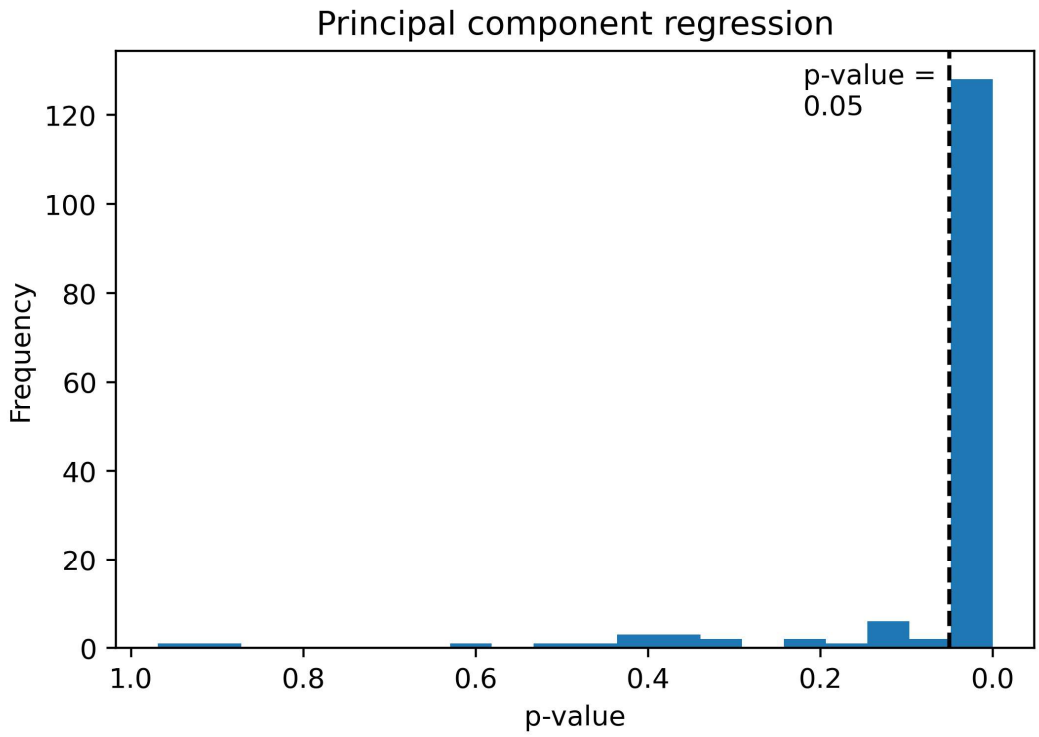

Figure 39 Distribution of the Stoffer-Toloi test p-values of the principal component regression models' residuals on the validation data

Appendix L: Principal component regression

The principal component regression model is an ordinary least squared simple linear regression of the dependent variable over some principal components. The principal components are found by transforming a matrix of the, normalized, original independent variables into an orthogonal matrix. The principal components are then in the columns of the new matrix.

We have a matrix ***X*** of data on the independent variables, containing *n* readings of *p* sensors and a vector $\mathbf{y}_{of}$ size *n*, with readings on the dependent variable. That is:

Equation 22 Definition of matrix of independent variables, with x_x being a vector with the x-th reading of every independent variable.

$$\boldsymbol{X} = [\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_n}]^T$$

To perform principal components regression, we transform the columns of ***X*** to principal components, such that:

Equation 23 *Definition of the matrix of principal components **Z**, with **X** the matrix of independent variables, **I** the identity matrix, **P** a matrix of eigenvectors of **X** and **D** a diagonal matrix of eigenvalues.*

$$\boldsymbol{X'X_t} = \boldsymbol{PDP'} = \boldsymbol{Z'Z}$$
$$\boldsymbol{P'P} = \boldsymbol{I}$$
$$\boldsymbol{D} = \boldsymbol{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$$

The values of ***Z*** are weighted averages of the original values in ***X***, and computed such that each column is uncorrelated to the others. Each of the columns of **Z** accounts for some variability in the training data. Excluding some of the components introduces some bias but allows for regularization and dimensionality reduction.

Appendix M: Fault labels

Index	Name
1	Preventive
1.1	Inspection
1.2	Preventive maintenance following MJOP
1.3	Other preventive maintenance
2	PV
2.1	Inverter failure
2.2	Short-circuit between strings
2.3	Open circuit
2.4	Bypass diode failure
2.5	Hightened cable resistence (corrossion, contact damage)
2.6	Not connected
2.7	DC-AC Switch failure
2.8	Optimizer failure
3	Monitoring system
3.1	Internal
3.1.1	Sensor failure
3.2	External
3.2.1	Loss of sensor signal (Hardware defect)
3.2.2	Loss of sensor signal (Link failure)
3.2.3	Loss of signal accuracy
3.2.4	Loss of signal precision
4	Settings and control
4.1	Configuration error
4.2	Software failure

4.3	RMU/Display/Thermostat failure
5	Electric circuit failure
5.1	Loss of power
5.2	Back-up battery failure
5.3	Short in installation
5.4	Installation/compontent not connected (not PV)
6	Source cycle
6.1	Low source temperature
6.2	Loss of flow in source
6.2.1	Source circulator fault
6.2.2	Source ventilator fault
6.2.3	Filter fouling
6.2.4	Retricted inflow source
6.2.5	Restricted return flow source
6.2.6	Source leakage
6.3	Reverse valve failure
6.4	Brine undercharge
7	Heat distribution circuit failure
7.1	Distribution valve failure
7.1.1	Distribution valve locked in open position (Communual systems)
7.1.2	Distribution valve locked in closed position (Communual systems)
7.2	Mixing valve failure
7.3	Loss of flow in distribution circuit
7.3.1	Air in central heating system
7.3.2	Filter fouling
7.3.3	Distribution circulator fault

7.3.4	Shut-off valves closed
7.3.5	Distribution circuit leakage
7.5	Heightened flow in distribution circuit
7.6	Auliary heater/E-boiler failure
7.7	Close
8	Refrigerent cycle failure
8.1	Expansion valve fault in closed position
8.2	Expansion valve fault in open position
8.3	Loss of pressure in expension tank
8.4	Filterdryer fouling
8.5	Compressor fault
8.6	Non-condenseable gas
8.7	Compressor inverter fault
8.8	Condensor fouling
8.9	Evaporator fouling
8.10	Checkvalve failure
9	Auxilary systems
9.1	Electric radiator failure
9.2	Ventilation failure
9.3	Heat exchanger failure
10	Residents' expectations exceed system capacity **(Excluded from analyis)**
11	Heat pump lacks start-up capacity
11	Wrongly installed
12	Unknown failure (Excluded from analyis)
13	Cancelled (Excluded from analyis)
14	Other:

Appendix N: Confusion matrices of fault detection models

SolarClique

Table 7 Confusion matrix of SolarClique fault detection for PV failures, with a 4-standard deviation prediction interval, where N.A. means 'not available' e.g. due to failure of the monitoring system. Last column and row denote subtotals.

	Visual class. **FAULT**	Visual class. **NO FAULT**	Visual class. **N.A.**	SolarClique, K=4
Model class. **FAULT**	32	765	0	797
Model class. **NO FAULT**	5667	28481	0	34148
Model class. **N.A.**	525	268	2950	3743
	6224	29514	2950	

Table 8 Confusion matrix of SolarClique fault detection for PV failures, with a 3-standard deviation prediction interval, where N.A. means 'not available' e.g. due to failure of the monitoring system. Last column and row denote subtotals.

	Visual class. **FAULT**	Visual class. **NO FAULT**	Visual class. **N.A.**	SolarClique, K=3
Model class. **FAULT**	108	1133	0	1241
Model class. **NO FAULT**	5591	28113	0	33704
Model class. **N.A.**	525	268	2950	3743
	6224	29514	2950	

Table 9 Confusion matrix of SolarClique fault detection for PV failures, with a 2-standard deviation prediction interval, where N.A. means 'not available' e.g. due to failure of the monitoring system. Last column and row denote subtotals.

	Visual class. **FAULT**	Visual class. **NO FAULT**	Visual class. **N.A.**	SolarClique, K=2
Model class. **FAULT**	349	2280	0	2629
Model class. **NO FAULT**	5350	26966	0	32316
Model class. **N.A.**	525	268	2950	3743
	6224	29514	2950	

Table 10 Confusion matrix of SolarClique fault detection for PV failures, with a 1-standard deviation prediction interval, where N.A. means 'not available' e.g. due to failure of the monitoring system. Last column and row denote subtotals.

	Visual class. **FAULT**	Visual class. **NO FAULT**	Visual class. **N.A.**	SolarClique, K=1
Model class. **FAULT**	1505	8136	0	9641
Model class. **NO FAULT**	4194	21110	0	25304
Model class. **N.A.**	525	268	2950	3743
	6224	29514	2950	

PCR

Table 11 Confusion matrix of PCR fault detection for heat pump failures, with alpha= 0.01, where N.A. means 'not available' e.g. due to failure of the monitoring system. Last column and row denote subtotals.

Visual class.	Visual class.	Visual class.

	FAULT	**NO FAULT**	**N.A.**	PCA,alpha= 0.01
Model class. **FAULT**	15	35	0	50
Model class. **NO FAULT**	5245	26093	1007	32345
Model class. **N.A.**	964	1626	2959	5549
	6224	27754	3966	

Table 12 Confusion matrix of PCR fault detection for heat pump failures, with alpha= 0.02, where N.A. means 'not available' e.g. due to failure of the monitoring system. Last column and row denote subtotals.

	Visual class. **FAULT**	Visual class. **NO FAULT**	Visual class. **N.A.**	PCA,alpha= 0.02
Model class. **FAULT**	24	62	4	90
Model class. **NO FAULT**	5236	26066	1003	32305
Model class. **N.A.**	964	1626	2959	5549
	6224	27754	3966	

Table 13 Confusion matrix of PCR fault detection for heat pump failures, with alpha= 0.05, where N.A. means 'not available' e.g. due to failure of the monitoring system. Last column and row denote subtotals.

	Visual class. **FAULT**	Visual class. **NO FAULT**	Visual class. **N.A.**	PCA,alpha= 0.05
Model class. **FAULT**	56	142	13	211
Model class. **NO FAULT**	5204	25986	994	32184

Model class. **N.A.**	964	1626	2959	5549
	6224	27754	3966	

Table 14 Confusion matrix of PCR fault detection for heat pump failures, with alpha= 0.10, where N.A. means 'not available' e.g. due to failure of the monitoring system. Last column and row denote subtotals.

	Visual class. **FAULT**	Visual class. **NO FAULT**	Visual class. **N.A.**	PCA,alpha= 0.10
Model class. **FAULT**	107	281	23	411
Model class. **NO FAULT**	5153	25847	984	31984
Model class. **N.A.**	964	1626	2959	5549
	6224	27754	3966	