Should you clean your solar panels now?

by

Xiang Gao

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Management Sciences

Waterloo, Ontario, Canada, 2014

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Photovoltaic (PV) applications worldwide are being increasingly deployed; however, the performance of PV systems is greatly affected by external factors, such as weather, environment, and terrain. These factors can form anomalies on solar panels, and lead to low performance. We present a data driven approach to identify anomalies on panels, based on the power output data and using our simplified solar irradiance model. The approach includes the in-plane solar irradiance model, disaggregation and detection of anomalies, and a decision tree to classify the types of the anomalies. The detection sensitivity is adjustable to suit different production environments. This methodology can be applied in multiple areas, such as anomaly alerts, energy loss analysis on different interval bases, and so on. The approach has been tested using real data collected in two cities in Ontario, Canada. The classification has a 85% precision rate for the detected anomalies.

Acknowledgements

I would like to thank Professor L. Golab for supervising this research, and Professor S. Keshav for providing technical guidance. I would like to thank Nicole Keshav for helping me to correct grammatical errors.

I would also like to thank Bo Hu, who implemented and helped me to maintain the data collection device. I would like to thank Toronto and Region Conservation Authority (TRCA), who provided data for this research.

I would like to thank Professor Stan Dimitrov and Professor Parmit Chilana for being the readers of this thesis.

Dedication

This thesis is dedicated to all the ones I love.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

Photovoltaic (PV) technology (i.e. solar panels), the most important application of solar energy, is being revisited as an important sustainable alternative to fossil fuel, and has received increasing attention recently. Installed capacity (rated capacity), is the maximum capacity at which a PV system is designed to run with optimal sun exposure. In fact, the annual installed capacity was about 29.6 Gigawatts (GW) in 2011 and approximately 31 GW in 2012 [4]. Especially, following the nuclear disaster in Japan, there has been a growing tendency to rely on more secure power, and there is the potential to install an additional 100 gigawatts (GW) of PV by 2015 [56]. Many countries now heavily depend on PV for a large proportion of their power supply. For example, by June 6, 2013, PV electricity production in Germany had reached 23.4 GW, which met 39% of national peak electricity needs [33]. Figure 1.1 depicts the PV installation trend over the last 10 years.

Meanwhile, with the development of related technology and the growth of the market, the price of PV panels has been decreasing for decades, and it is believed that the unit price will keep decreasing in the future. Figure 1.2 depicts the change of the price of PV panels in the last thirty years [1]. Above all, the pervasiveness, economics, and security make solar energy an important renewable energy resource. Based on the historical data and recent analysis, solar energy will play a dominant role in the new energy revolution.

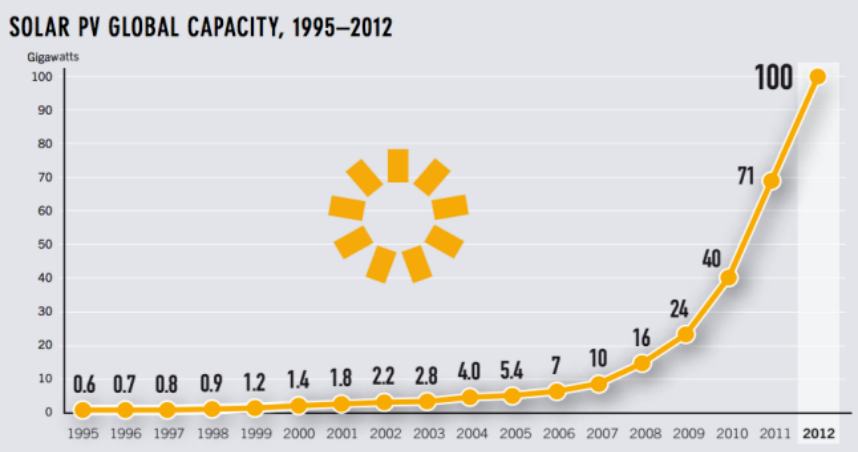


Figure 1.1: Global solar PV power capacity grew from about 2.2 GW in 2002 to 100 GW in 2012. Source: Renewables 2013 Global Status Report - http://www.ren21.net/REN21Activities/GlobalStatusReport.aspx

Currently, the PV conversion rate (efficiency) is still low, compared with other energy forms. So one essential direction to obtain more PV power is to develop materials with a higher photoelectric conversion rate. Another major direction is to make sure that the PV system is working to its maximum capability, that is, to maintain its efficiency.

However, unlike fossil fuel, solar energy is vulnerable to external factors, such as clouds, shadows, dust, and so on. For instance, studies by Salim, et al., [58] indicate that there is a 32% reduction in solar panel performance after eight months' accumulation of dust in Riyadh, Saudi Arabia; Wakim's [64] experiments conclude that the performance reduction is 17% after six continuous dry days in Kuwait city, Kuwait, respectively. Moreover, with increases in solar module efficiency, the area required for installation has decreased, thus underlying the increasing importance of detecting and removing anomalies that affect efficiency.

One solution to deal with the dust is to frequently clean it with water. Zorrilla-Casanova, et. al.,'s experiment in Malaga, Spain, indicates the following conclusions [67]: dirty panels can recover with even very light rain, for instance, below 1 mm. However, in many solar resource rich areas, such as the Sahara and Middle East, the climate is dry

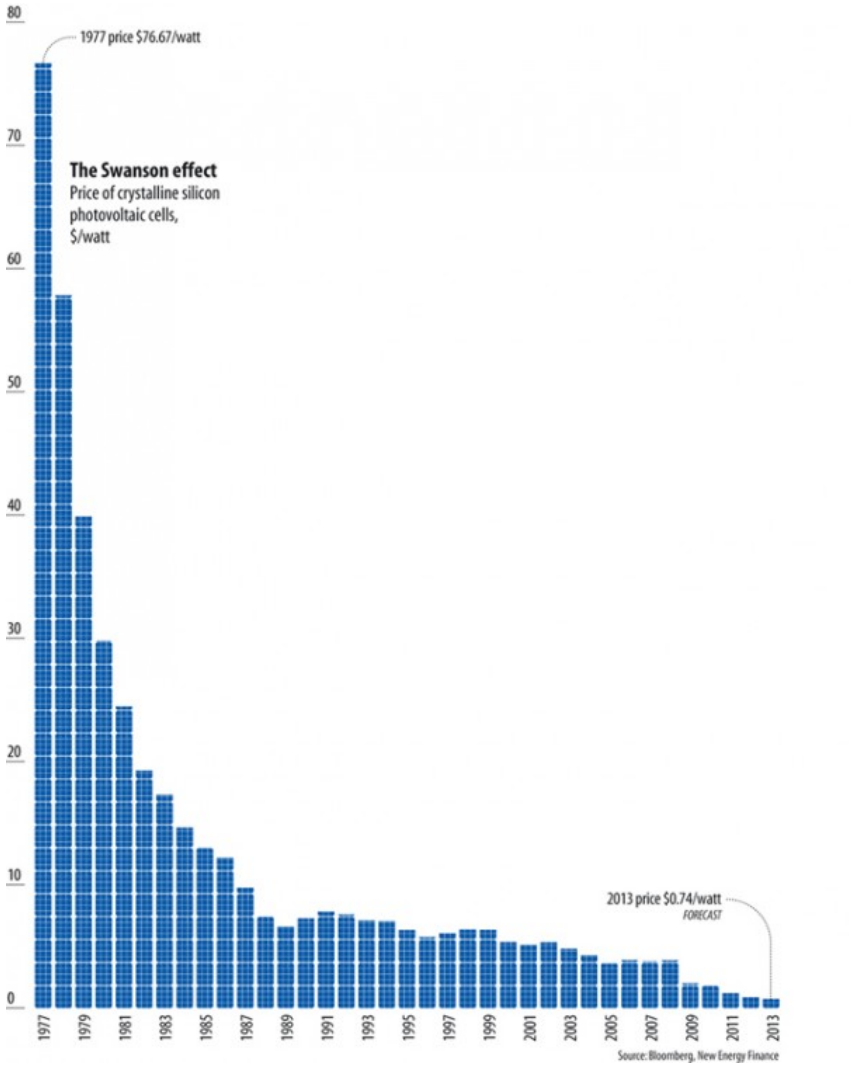


Figure 1.2: The price of solar PV panels dropped about 100 times over from 1977 to 2012. Source: Cost of Solar - http://costofsolar.com

and hot. Hence, water shortages are problematic in these locations, and it is unrealistic to clean the solar panels every day.

Even if we have enough water to clean the panels every day, this would not be feasible, not only because it is laborious, but also because many large scale deployed solar panels are very far away from cities or placed on the roofs of tall buildings. Hence, an automatic anomaly detection and identification method is needed. For instance, if the maintainers know that the low performance of the PV panels are due to dust accumulation, and there is no rainfall in a near future, they can address the particular anomaly with proper tools and approaches.

To achieve this goal, solely based on the power output data, time series analysis of the power output of PVs is adequate [32]. However, these anomalies detected by time series analysis include many of those caused by weather conditions, such as clouds and sudden overcast, which cannot be eliminated. Worse still, anomalies due to weather can be mixed with other types of anomalies, making identification of actionable anomalies, such as snow, dust, and shadows, more difficult.

Another intuitive way to automatically detect anomalies is to compute the expected power output, based on the expected solar intensity at a given time of a day, in a given location. If the observed power output is much less than the theoretical output, an anomaly is detected. Hence, the problem is how to obtain the theoretical power output, as normally a PV system only records real power output. When the anomalies are detected, we can extract the features and use trained classifers to identify the type of the detected anomaly. This data driven approach is what we focus on in this thesis.

1.2 Definitions

A *solar cell* (photovoltaic cell) is an electrical device that converts the energy of light directly into electricity by the photovoltaic effect. It is also the basic unit of a solar panel. A *solar module* is assembled with several solar cells using wires. A *solar panel* is a set of solar modules electrically connected and mounted on a supporting structure. A *solar array*

(solar string) is a linked collection of solar panels [8]. Figure 1.3 depicts these concepts in PV systems. Our work focuses on the solar panel level and the string level.

Solar irradiance is used to indicate the density of solar radiation, which is defined as the power of solar radiation incident on a unit area surface. A pyranometer is the most generally used device to measure solar irradiance. *Horizontal irradiance* is the irradiance arriving on a horizontal surface, and *In-plane irradiance* is the irradiance arriving on an inclined surface. *Efficiency* is a solar panel's maximum capability of converting solar energy to power [7]. *Theoretical power output* is the expected power output, based on the in-plane irradiance and the efficiency. *Real power output* is the measured value of power output. Typically, there are voltage meters and current meters installed for a PV system, and the real power output can be calculated by the power equation:

$$Power = Current * Voltage \tag{1.1}$$

Performance ratio (PR) is the ratio between the real power output and theoretical power output of a solar panel given the current weather conditions, which is used to evaluate the solar panel's performance.

Generally, *horizontal irradiance* or *global irradiance* refers to *horizontal global irradiance*, which includes three components: direct radiation, diffuse radiation, and albedo radiation. We will elaborate on these concepts in the next chapter.

1.3 Problem statement

The procedure of data driven solar panel anomaly detection and classification is depicted in Figure 1.4. The input includes real power output and irradiance data, which can be horizontal or in-plane; we need an extra step to compute the in-plane irradiance with our in-plane model, if the input irradiance is horizontal irradiance. Intuitively, a PV system achieves its maximum input energy yield when its solar panels face exactly towards the sun. Therefore, most of the solar panels are installed with a slope, and possibly an orientation angle. However, many available irradiance measurements, especially those with the public

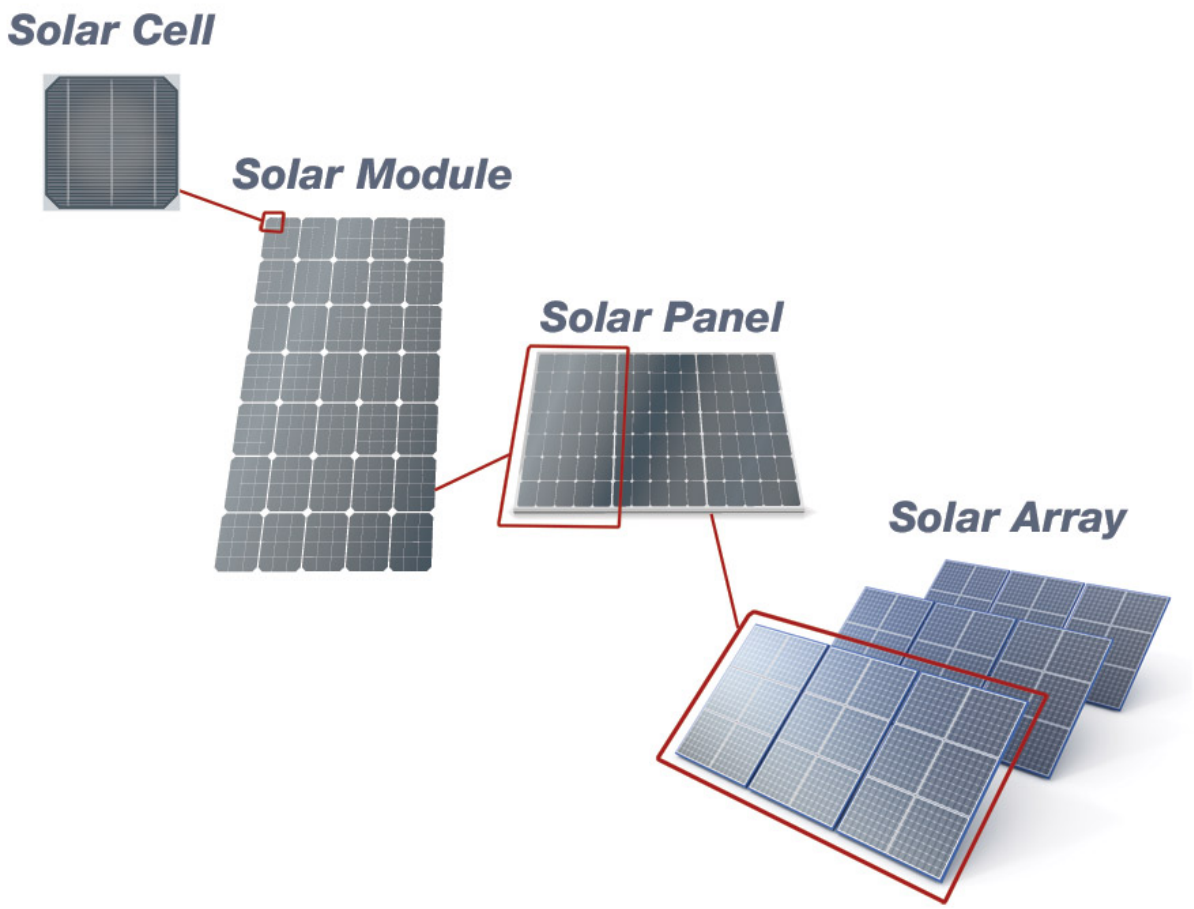


Figure 1.3: PV system disaggregation. Source: Photovoltaic Array Fundamentals http://etap.com/renewable-energy/photovoltaic-101.htm

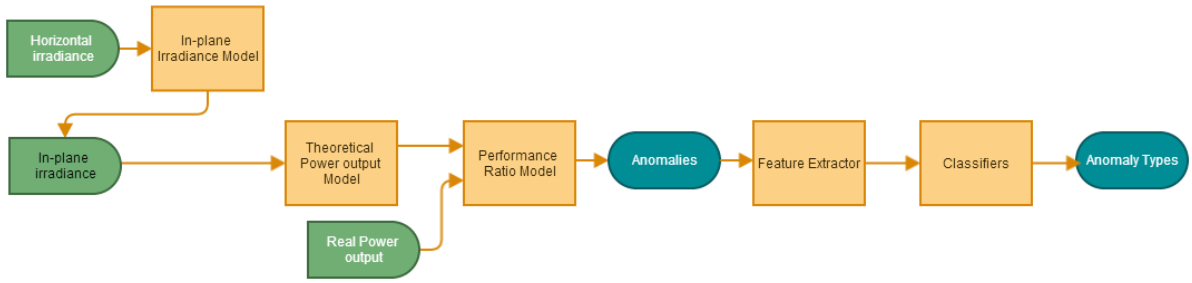


Figure 1.4: System diagram

data access, are just horizontal measurements. Hence, modelling in-plane irradiance is a crucial step in many use cases.

Based on the system diagram, there are four sub-problems which need to be addressed:

- Model in-plane irradiance
- Model theoretical power output given in-plane irradiance
- Determine which features from the detected anomalies is significant for classification
- Test the classification algorithms for anomaly classification on solar panels.

The output of this system is the detected anomalies, and the classified anomaly types.

We are the first to disaggregate the anomalies into two basic categories, *non-actionable*, which is mainly casued by weather, and *actionable*, which usually is caused by controllable factors. Our objective focuses on the identification of actionable anomalies. There is nothing we can do to improve solar panel performance on a cloudy day, however, we can take corresponding actions after identifying certain actionable anomalies. Other anomalies include panel failure, direct cover and indirect cover: the panel failure means that the panel is broken so that it is not able to function normally; the direct cover means that anomalies, such as leaves, snow, and dust, are in contact with the panels; the indirect cover means that

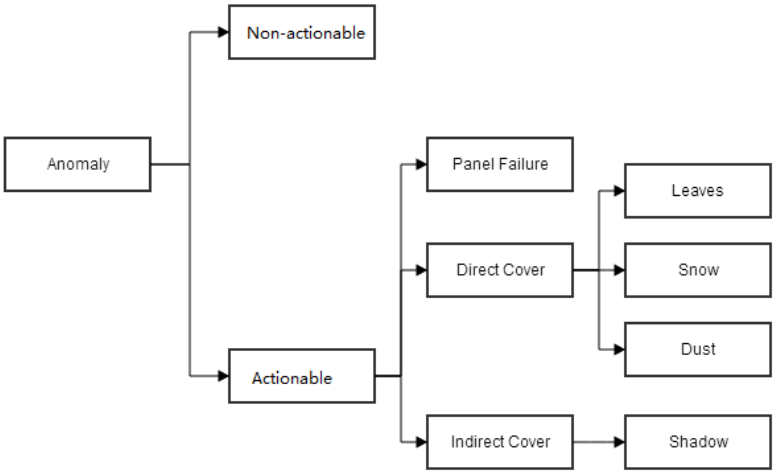


Figure 1.5: Objective anomaly identification

anomalies, such as shadows, are not in contact with the panels, however, they still block beams. Moreover, we developed metrics to describe the features of various anomalies, and applied these metrics to the classification of many types of anomalies. Figure 1.5 depicts our disaggregation of the anomaly types. Notice that more anomaly categories, high level or low level, can be added based on particular scenarios.

1.4 Solution Overview

Our approach contains two main stages: detecting anomalies and identifying which category an anomaly falls into.

1.4.1 Detection

To detect an actionable anomaly, we set a threshold for PR. Recall that the PR definition already takes weather anomalies into account, so if the PR drops below the threshold, an actionable anomaly is considered to exist. According to the definition of PR, the real

power output and the theoretical power output is necessary to compute PR. Therefore, we developed and refined two models: the first one addresses the input energy data, which is the in-plane irradiance; the second one, an efficiency model, converts the in-plane irradiance to theoretical power output. With the assumption that the real power output is available for most PV systems, we can compute the PR for each panel using its theoretical power output and real power output.

In-plane irradiance can be measured directly with a pyranometer. Irradiance measurement is regarded as anomaly-free, because pyranometers are small, and many material technologies are applied to isolate anomalies, such as electrostatic anti-dust glass and heaters for melting snow. However, in many solar panel deployment cases, especially small solar energy applications, installing pyranometers along with data collecting systems is economically unrealistic. For example, there is a PV system deployed on a building roof, on the campus of the University of Waterloo, along with a monitoring system which records only power output data. To integrate the measurements of pyranometers into this legacy system is uneconomical and infeasible given that it is not open and programmable. Therefore, we utilized physical models and campus weather station data to compute the in-plane irradiance.

We developed our efficiency model based on Skoplaki's efficiency equation [60], and specification from the solar panel manufacturers. The model was calibrated and validated with collected data.

1.4.2 Classification

We investigated real data collected in the natural environment, and extracted an important metric, Coefficient of Variance (CV), to first classify the actionable anomalies into direct covering, such as dust and snow, and indirect covering, such as shadows. We extracted more features to further classify the two direct covering anomalies.

We examined several classification algorithms by ten-fold cross validation, and discovered that a decision tree with a small number of levels can achieve an 85% accuracy in classifying the three types of actionable anomalies.

1.5 Challenges and Contributions

1.5.1 Challenges

- As a data driven approach, data availability is the greatest challenge. Particularly, we lacked evidence to label the detected anomaly samples for training. Hence, we had to manually observe anomalies on an installation site, where the access was limited. Moreover, during our experiment period, the environment in south Ontario was rainy and relatively clean, so many anomalies were very difficult to observe, such as dust. Therefore, we had to simulate them by manually sprinkling dust on the panels. Furthermore, our panels are on the roof and we physically could not create artificial shadows.

- The quality of the available data was also a challenge. We used two datasets for this thesis: the TRCA data collected by Toronto and Region Conservation Authority (TRCA), which includes on-site irradiance data, power output, and pictures recording evidence of anomalies on solar panels at an interval of 5 minutes; the UW (University of Waterloo) data, which contains the power output, recorded by our PV system, and the irradiance data, published by the UW weather station. For instance, the resolution of the pictures from TRCA data is 600*800, which is not fine enough to observe dust anomalies. Moreover, the granularity of the UW weather station data is 15 minutes, which is not fine enough to build and verify in-plane irradiance models.

1.5.2 Contributions

Our specific contributions are:

- We developed a series of models to compute theoretical power output; the input supports any sort of available irradiance measurements, in-plane or horizontal.

- We used these models to develop an approach for anomaly detection and classification. We are the first to disaggregate the anomalies into two basic categories,

actionable and non-actionable. We focused on the need for simple detection and the identification of actionable anomalies. We are also the first to further disaggregate the actionable anomalies, and to develop metrics for the critical features, which facilitate the following classification of anomalies greatly.

- We validated our approaches on two real data sets, and achieved an accuracy of 85% for classification of the three types of actionable anomalies.

1.6 Applications

Our methodology can be extended and applied in many solar energy scenarios. Here are some examples.

1.6.1 Anomaly warning

Anomaly warnings can be triggered and sent to maintainers when an anomaly is identified, and maintainers will handle the specific anomaly according to the warning. For long distance locations, administrators can customize the thresholds to determine the situations when warnings should be generated.

1.6.2 Energy loss estimation

Energy loss estimation will interest many PV system investors. Existing systems, such as Enphase (introduced in Section 2.5), can be used to estimate energy losses; however, the users never know whether the energy loss is avoidable, because they do not separate the actionable anomalies from the non-actionable anomalies. Based on our approach, the estimation of energy loss due to non-actionable factors (weather) and actionable factors can be computed, respectively. If the energy loss due to weather is great, a new location should be considered. In contrast, by analysing actionable anomalies leading to energy loss, the owners can adjust the deployment of panels, and install necessary maintaining equipments to increase power production.

1.7 Organization

The rest of this thesis is organized as follows: the next chapter introduces the related work. Chapter 3 describes the data sets we used for this thesis. Chapter 4 discusses how to build models of irradiance and theoretical power output, and offers the corresponding evaluation. Chapter 5 covers the methodology for detecting and classifying anomalies, and experimental evaluations as well. Chapter 6 presents the conclusion, and proposes the future work.

Chapter 2

Related Work

In this chapter, a list of related work is introduced as follows: Section 2.1 is about the physical model of solar irradiance arriving on horizontal surfaces. Section 2.2 describes the conventional way, which is complex, to compute in-plane irradiance based on horizontal irradiance. Section 2.3 presents the concepts of efficiency and the performance rate of solar panels. Section 2.4 describes experiments done by researchers to understand the effects of various anomalies. Section 2.5 is about the related work of anomaly detection on solar panels. Section 2.6 discusses the related work of solar panel anomaly classification.

2.1 Modelling Theoretical Solar Irradiance on Horizontal Surfaces

Luque, et al., [41] give basic physical concepts for computing theoretical solar irradiance: imagine putting a solar panel just outside of the atmosphere, which means there are no detrimental effects from air, then the resulting solar irradiation received by unit area perpendicular to the beam is called *solar constant*: $B_0 = 1367W/m^2$, where W is watt, m^2 is square meter.

Next, *air mass* (AM) is employed to describe the cleanliness of the atmosphere on clear days; the more clean the atmosphere is, the more irradiance will arrive on a horizontal

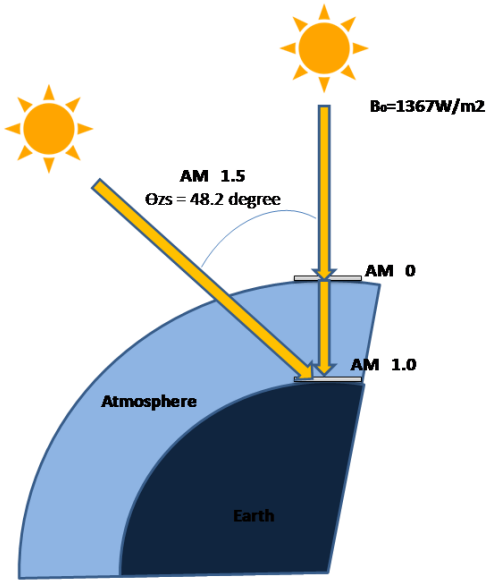


Figure 2.1: Solar irradiance incidence

surface. As shown in Figure 2.1, $AM = 0$ indicates the situation in which a beam lies perpendicular to a panel located above the atmosphere, while $AM = 1$ represents the situation in which a beam lies perpendicular to the surface of the earth. In general, an increasing air mass displaces the solar spectrum towards the red. The solar radiation w.r.t. $AM = 1$ is $1000W/m_2$, which is just the value used in standard tests for PV cells. The zenith is an imaginary point perpendicularly above a particular location; as shown in Figure 2.1, the zenith is the normal of the two panels pointing to the sun. In most situations, there is an angle θ_{zs} between the zenith and the beam incidence, so air mass can be represented by $cos\theta_{zs}$. Intuitively, $AM = 1$ when $\theta_{zs} = 0$, which indicates that beams incident perpendicularly; however, incident beams travel a longer distance and scatter more in the condition of larger $cos\theta_{zs}$. Figure 2.1 depicts these variables. Luque, et al., [41] give the following equation for computing AM:

$$AM = \frac{1}{cos\theta_{zs}} \tag{2.1}$$

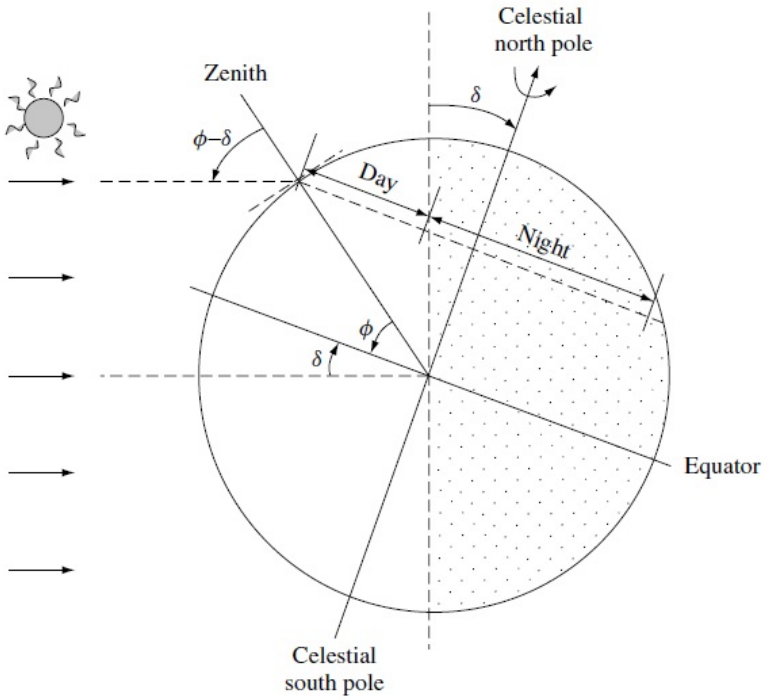


Figure 2.2: Relationships between the related angles and EarthSun position at solar noon

Intuitively, angle θ_{zs} is determined by three factors: where the sun beams lie perpendicularly on the Earth (determined by the motion of the Earth around the Sun, δ), where the panels are located (determined by the latitude, ϕ), and the time of the day (determined by the Earth's rotation, ω). The following describes a procedure to obtain these arguments.

Based on Figure 2.2, solar declination δ is employed to compute the angle θ_{zs}; the solar declination indicates the angle between the equatorial plane and the line connecting the center of the Earth and the center of the Sun[41]. The solar declination can also be regarded as the latitude at which the sun beams lie perpendicularly on particular days, for instance, $\delta = 0$ on spring equinox (20th/21st March) and autumn equinox (22nd/23rd September), and $\delta = 23.5$ on summer solstice. Luque, et al., [41] give the following equation for computing δ:

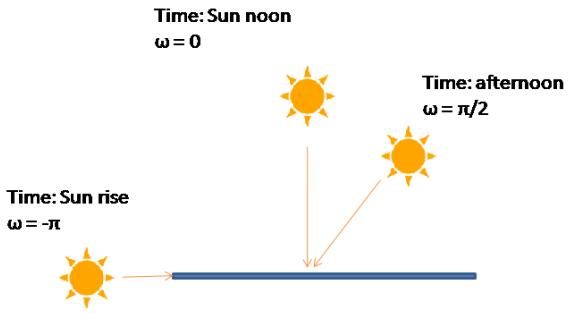


Figure 2.3: ω representation

$$\delta = 23.45^o \left[\frac{360(d_n + 284)}{365}\right] \tag{2.2}$$

where d_n is the sequence number of a particular day in one year, i.e. January 1st equals 1, January 2nd equals 2,... December 31st equals 365. Notice that the celestial pole is perpendicular to the equator, and angle ϕ (representing geographic latitude) has an inverse direction w.r.t. angle δ (representing solar declination). So the incidence angle θ_{zs} at solar noon is expressed by $\phi - \delta$.

The variable ω is used to compute the horizontal irradiance at an arbitrary time in a day. This represents true solar time, by the difference between solar noon and the moment of a day in terms of a 360 degree rotation. Figure 2.3 shows how to use the ω to represent a particular time of the day. In practice, ω is defined based on this principle, but can be customized to meet different granularity requirements. For instance, let T be the current solar time expressed in minutes (e.g., 9:00 represented by $540 = 9 * 60$ minutes), so the corresponding ω is expressed as:

$$\omega = \frac{T - 720}{720}\pi \tag{2.3}$$

Therefore, at any given moment, the angular coordinates (θ_{zs}) of the sun are calculated from the equation given by Luque, et al., [41]:

$$cos\theta_{zs} = sin\delta sin\phi + cos\delta cos\phi cos\omega \tag{2.4}$$

The track in which the Earth travels around the Sun is not a strictly circular orbit, which results in the variation in distance between the Sun and the panels throughout a year, such that the solar irradiance varies w.r.t. date. We use ε to correct the solar irradiance computation on a particular day. In practice, a useful expression for the so-called eccentricity correction factor given by Luque, et al., [41] is:

$$\varepsilon_0 = 1 + 0.33cos\left(\frac{360d_n}{365}\right) \tag{2.5}$$

Now, we have all the information needed to compute the maximum possible solar irradiance at a given time and location on the Earth. Meinel, et al., [42] give an empirical equation to compute the theoretical solar irradiance G on a horizontal surface, which is a regression of the irradiance data collected in a desert in California. Note that since ω is used in this equation, it can plot the theoretical irradiance for the entire daytime.

$$G = B_0\varepsilon_0 \times 0.7^{AM^{0.678}} \times cos\theta_{zs} \tag{2.6}$$

Equation (2.6) is an empirical physical model; the two constants, 0.7 as the base and 0.678 as the exponent, are mainly determined by the local air quality, and derived by fitting the equation to the data collected in California. We use this physical model to compute horizontal irradiance; as the data we used is collected in Ontario, we need to fit this equation to local data by changing the values of the two constants.

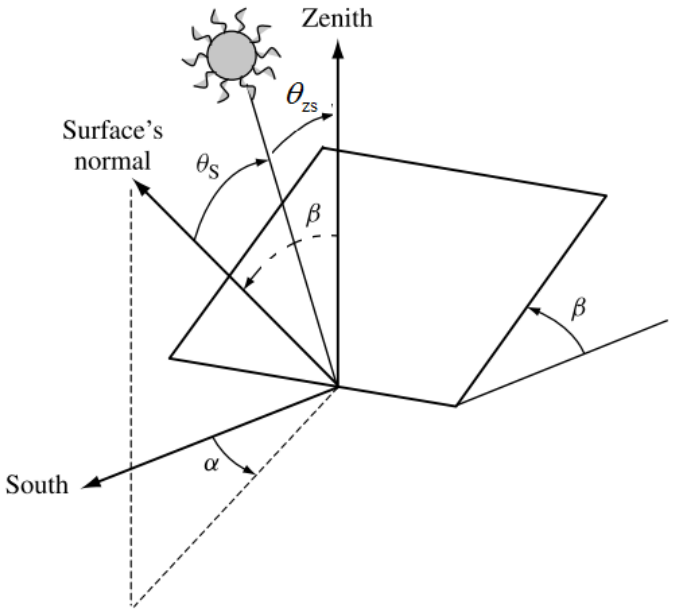


Figure 2.4: Receiver position (slope β, orientation angle α) and sun beam incidence angle θ_s

2.2 Modelling Theoretical Solar Irradiance on Inclined Surfaces (In-plane irradiance)

We did not use this conventional model to obtain in-plane irradiance, rather, we developed a much simpler and equivalent model to compute the in-plane irradiance. However, we implemented this conventional method and used the results to evaluate our model.

In practice, solar panels are usually mounted on a slope (β) to increase the equivalent receiving area, and to reduce rain or snow accumulation as well. Possibly, there is an orientation angle (α) due to construction or landform limitations. These two angles are depicted in Figure 2.4.

There are verified physical theories and approaches for modelling solar irradiance on inclined surfaces. When solar radiation passes through the atmosphere, it is modified

by the interaction with the components there. Some of them, such as clouds, reflect radiation. Others, such as ozone, oxygen, carbon dioxide and water vapour, have significant absorption at several specific spectral bands. Water droplets and suspended dust also cause scattering. All these processes mean that the solar radiation incident on the panel can be decomposed into three components: direct radiation, diffuse radiation, and albedo radiation. Direct radiation is made up of beams reaching the surface in a straight line from the Sun. Diffuse radiation is the radiation scattered towards the receiver. Albedo radiation is the radiation reflected from the ground. The sum of these three components is called global radiation, generally [41].

Figure 2.5 is from the Handbook of Photovoltaic Science and Engineering [41], which depicts the components of solar radiation.

The conventional procedure for calculating global irradiance on an inclined surface, G(β, α), is given by Luque, et al., [41] (in northern hemisphere, α is negative if it is facing east). The first phase was to derive the direct, diffuse, and albedo components from horizontal global irradiance, respectively. Next, they computed the irradiance of the three components on the inclined surface using corresponding equations. Finally, they summed up the three parts to obtain the global irradiance on the inclined surface. They also introduced the angle between solar incidence and normal to panels, θ_s (see Figure 2.4), and an equation to compute it:

$$\begin{aligned} cos\theta_s = sin\delta sin\phi cos\beta - sin\delta cos\phi sin\beta cos\alpha + cos\delta cos\phi cos\beta cos\omega \\ + cos\delta sin\phi sin\beta cos\alpha cos\omega + cos\delta sin\alpha sin\omega sin\beta \end{aligned} \tag{2.7}$$

In many cases, panels are oriented due south (in the northern hemisphere), that is, $\alpha = 0$. Correspondingly, the Equation (2.7) will be simplified to Equation (2.8):

$$\begin{aligned} cos\theta_s = sin\delta sin\phi cos\beta - sin\delta cos\phi sin\beta + cos\delta cos\phi cos\beta cos\omega \\ + cos\delta sin\phi sin\beta cos\omega \end{aligned} \tag{2.8}$$

Global irradiance on a horizontal surface is computed in last section, and there are models to compute the three components on an inclined surface [41], respectively. Khoo,

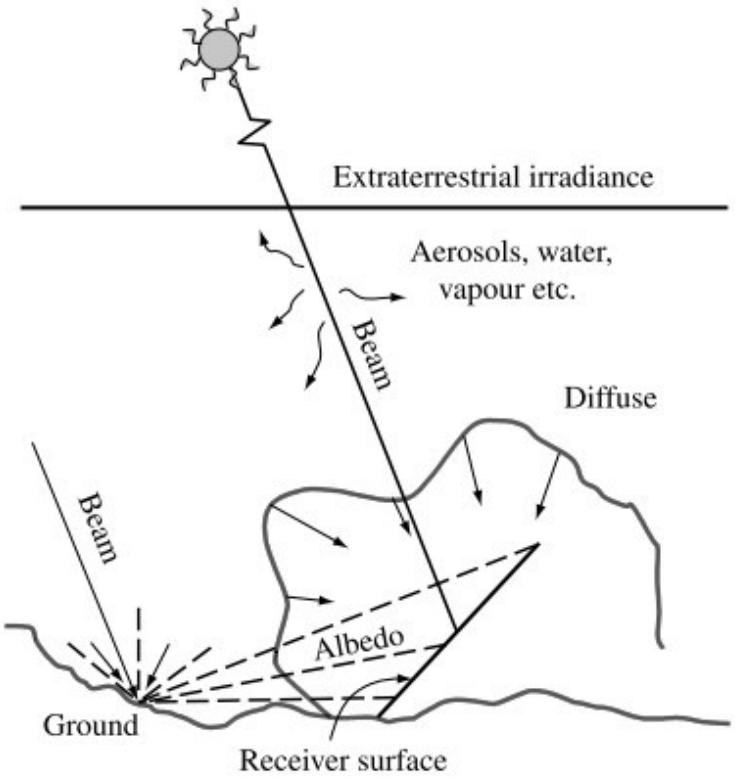


Figure 2.5: Different components of solar radiation. Source: Handbook of Photovoltaic Science and Engineering [41], page 113.

et al., [35], Yang, et al., [65] and Al-Rawahi, et al., [14] used this approach to compute local global irradiance on inclined surfaces. In practice, since the energy concentrates in the direct and diffuse components (above 99%), a general method is to extract the direct or diffuse part from global irradiance and omit the albedo part [41]. Therefore, either the diffuse irradiance fraction derived from regression of local irradiance data or the absolute diffuse irradiance value from a pyranometer must be available. For the fields where a pyranometer that supports measuring diffuse components is not available, alternatively, the regression of local irradiance data is the last option. However, the regression is heavily dependent on the local historical irradiance data but the data may be unavailable or in coarse granularity (e.g., monthly). Moreover, the irradiance data may vary in different years and may vary in the long run. Therefore, we developed a simpler model, which will be shown in Chapter 4, to compute in-plane irradiance. Nonetheless, we used this conventional complex model as evaluation. Due to the lack of pyranometers measuring the diffuse component, we reference local historical data from NASA [**?**].

Many other ways are also used to obtain in-plane irradiance, e.g., Veldhuis, et al., [63] employed a 3D modelling tool to simulate real-time irradiance with the shadow interference, but they ignored the weather factors. Pinker, et al., [52], Krotkov, et al., [37] and Mueller, et al., [46] researched how to employ satellite data to predict irradiance. However, these approaches highly depend on satellite data in fine spatial and time granularity, and the model is even more complicated. Due to the different objectives, we did not implement or evaluate these methods.

2.3 Efficiency

Solar panel efficiency η is the maximum capability for converting solar radiation energy to electricity energy. η is defined by following equation [7]:

$$\eta = \frac{P_{max}}{I * A} \tag{2.9}$$

where P_{max} is the expected electricity power output of a proper-functioning solar panel, I is the irradiance with unit W/m^2, and A is the solar panel surface area. In practice,

the efficiency of a single cell is called cell efficiency, which mainly depends on the material technology. The efficiency of a whole panel is called module efficiency. Given the loss from cell connections, the module efficiency is slightly lower than the cell efficiency. Module efficiency was used in this thesis since we detected anomalies on a panel basis.

Generally, the cell efficiency and module efficiency are available in the manufacturer's manual. Notice that the efficiency is measured in standard test conditions (STC: Air mass AM 1.5, irradiation 1000 W/m^2, Cell temperature $25C$) [7]. Usually PV manufacturers also provide another efficiency reference, which is measured in nominal operating cell temperature (NOCT: Cell temperature $45C$, irradiation 800 W/m^2, ambient temperature $20C$, Wind speed $1m/s$). Therefore, neither the precision nor the rangeability of the efficiency reference from manufacturers meets the requirement for detecting anomalies in any condition, so a process to verify the reference data and fit the changing conditions is necessary.

Buday, et al., [19] proved that wind speed has little impact on the efficiency, it only impacts the ambient temperature. As we have the on-site ambient temperature data, we can ignore the wind speed impact to the empirical efficiency equation.

Colli, et al., [21], and Dolara, et al., [24] analyzed PR w.r.t. irradiance on a yearly basis. Since they used daily aggregation irradiance, it is not capable to detect the short-time anomalies, such as shadows.

2.4 Anomaly effects

The research below focuses on the anomaly effects of external factors, but they did not go further to utilize the results to detect and classify anomalies.

2.4.1 Effects of shadows

Shadow effects have been investigated by physical experiments and computer simulations. Ramabadran, et al., [55], Taha, et al., [62], Deline, et al., [23] investigated the effects of

shadows on solar panels by simulating shadows using models, and their aim is to find a versatile model that can simulate random shading scenarios. Basically, diodes are used to simulate the sheltered cells. However, their experiments are developed and validated using software; the results have not been validated in fields or applications, and have been excluded from this thesis. We had the shadow anomaly data from a productive solar panel system for this thesis.

In practice, shadows are not a critical problem for solar systems. As large scale solar plants are located in open fields, and for small application, panels will be deployed at high-altitude positions, such as roofs, to avoid shadows.

2.4.2 Effects of dust

Dust may diminish the performance of solar panels by blocking all the irradiance components, and thereby influencing the energy yield. Factors that determine to what extend dust affects solar panels performance include types of dust, density of dust, distribution of the dust on panels, rate of accumulation, humidity, panels tilting angles, rainfall interval, wind speed and direction, panel orientation angles, and so on. In production, dust or soiling is the biggest damaging factor, therefore, it is one of the most important objectives in PV system performance research.

Most of the dust-related work focuses on the correlation between dust property and energy loss. Sulaiman, et al., [61] conducted an indoor experiment and recorded that the energy loss due to two different types of dust, mud and talcum, with quantified dust thickness, is up to 20%. The advantage for indoor experiment is interference free, but spot lights used in such experiments cannot provide enough irradiance (< 500) in some cases, which degrades the efficiency slightly. We did not run an indoor experiment because solar panels in production are deployed in the field, so indoor experiments cannot reflect the real conditions.

Outdoor experiments are all based on natural dust accumulation. Zorrilla-Casanova, et al.,'s experiments in Malaga indicate the following conclusions [67]: the dirty panels can recover with even light rain, below 1 mm; daily energy loss is negatively related with the

sun height, which means dust affects performance more in the morning and afternoon than at noon; the maximum energy loss is 25% after one month's accumulation, and most loss is below 20%. In Bangladesh, Rahman, et al., observed that the panel performance drops in the morning by 35% and 20% at noon after one month's natural dust accumulation [54]. We referenced their results in our outdoor experiment, especially for the maximum energy loss w.r.t. dust layer thickness.

For other impacting factors, Goossens, et al., [31] and El-Shobokshy, et al., [25] found that smaller size particles will lead to more significant degradation in PV panels. Referencing these results, we ran the dust experiments with various granularity of dust and sand. Moon [45] discovered that dust has an only 2% impact on tilted panels but has a 50% impact on flat ones. Garg's [28] experiment also verified this conclusion. Elminir, et al., [26] quantified the deposition density on panels tilted at different angles and orientations after six months, and the results showed that the density drops nearly linearly with the increase of tilting degree. Moreover, they gave the regression of the transmittance loss against the deposition amount. The advantageous condition for their experiment was that there were only two rainfalls during the six months. The variation effects by various orientation angles are greatly related with the prevailing wind, which will bring dust particles. Due to many limitations, we did not reimplement their work. For instance, the slopes of panels cannot be adjusted by us. Moreover, we ran the experiments in summer, which is the rainy season in Canada, so a natural dust accumulation and the corresponding energy loss was difficult to observe.

To maintain performance of panels, Mohamed, et al.,'s experiment in Libya showed that weekly cleaning kept the PV performance losses between 2% and 2.5% [44]; however, the interval may need to change in other environments. As their work is more in the electromechanical direction, it is beyond the scope of this thesis.

2.4.3 Effects of snow

Snow accumulation on panels is complicated, and is mainly affected by ambient temperature, wind speeds, tilting angles, and surface properties [51]. Once a snow layer is formed, the intensity of light penetrating the snow layer and reaching panels has an exponentially

decreasing relationship with the depth of the snow layer. Approximately 20% of incident radiation will be available at 2cm snow depth, while only 4% is available at 10cm depth [30]. These correlations have been demonstrated empirically by O'Neill, et al., [49] and Curl JR, et al., [22].

Brench [17] performed a snow effect experiment with different tilting angles. The results showed that with a 30 degree tilting angle, average daily energy loss is 45% if the snow depth is greater than 1 inch, and 11% if the snow depth less than 1 inch; on panels with a 40 degree tilting angle, this number drops to 26% and 5%, respectively. Andrews, et al., [11] built a model to predict energy loss due to snow on panels by time series analysis. Andrews, et al., [12] and Becker, et al., [16] built models and predicted yearly energy loss percentage is up to 3.5% and 2.7% , respectively, based on experimental data and meteorology information. However, their conclusion is not fine enough to detect and identify snow anomalies.

Our dataset contains some snow anomaly cases, but there is no snow thickness data. So, we use other features to classify snow anomalies, such as PR.

2.4.4 Effects of physical failures

Kuitche, et al., [38] investigated effects due to various failure modes on PV panels in desert climate conditions, and used a decision tree to identify reasons for failure. They addressed the failure anomaly and developed a ranking metric for failure modes, which includes PV cell degradation, broken inverters, connection cut between modules, and so on. Different values mean the differences in terms of severity, occurrence, and detection possibility of a failure. However, failure anomaly is beyond the scope of this thesis due to data availability and limitation of our experiment condition. For instance, we could not break our panels on purpose given the high cost.

2.5 Anomaly detection

We compute the performance ratio (PR) to detect anomalies. As follows is related PR work:

Meydbray, et al., [43] summarized the main procedures to estimate PV performance by giving an abstract methodology: first use an irradiance model to obtain in-plane irradiance from horizontal irradiance, and then use an efficiency equation to predict theoretical power output based on the in-plane irradiance and other impacting factors. We implemented an instance of this methodology to compute theoretical power output; however, our work scope is much beyond that. For instance, we used the theoretical power output to obtain PR, and used the PR to detect anomalies.

Free solar energy design assistant software available online can be used to obtain PR, such as the toolkit from National Renewable Energy Laboratory (NREL http://en.openei.org/apps/SWERA/). It provides services covering investment assessment, historical data access, and theoretical power model. However, the estimation is on a daily or yearly aggregation basis, and solely based on historical data. Therefore, it is not suitable for anomaly detection.

There also are many business solar energy software tools, such as Velasolaris (http://www.velasolaris.com/english/home.html), PVresourse (http://www.pvresources.com/siteanalysis/software.aspx), and Enphase (https://enphase.com/enlighten/), which can be used to detect anomalies. These tools can compute theoretical power and visualize real produced power. However, Enphase's and PVresourse's theoretical prediction assumes a perfect sunny day always, so users have no idea why the real power output is deviating from theoretical data. Velasolaris' prediction can involve the impact of shadow anomalies with its 3D model, but it does not disaggregate the anomaly factors further, and is not able to identify directly cover anomalies as we did.

2.6 Anomaly classifications

Existing research in solar panel anomaly classification or identification is from different points of views, and on different levels.

Anderson, et al., [13] and many other astrophysicists investigated solar system anomalies in space, which are in different context from the applications on the Earth.

Gauthier, et al., [29] owned a patent for detecting cracks and other imperfections on solar panels with a beam scanning. Kumar, et al., [39], Rothwarf, et al., [57], and Kawano, et al., [34] investigated silicon cells or circuit in the panels which cause anomalies. In contrast to these more material-oriented studies, our research investigated anomalies on a solar system level.

On the solar panel level, Hu, et al., [32] detected anomalies by building a time series model, and used statistical characteristics for classification. However, they did not disaggregate the anomalies into actionable and non-actionable; rather, they took all ten defined anomalies equally. As most of the anomalies are the combinations of actionable and non-actionable anomalies, such as snow in overcast days, or shadows in cloudy days, this approach is not appropriate with the assumptions that all anomaly types are equal and only one anomaly occurs at one time. In contrast, we first disaggregated anomalies and obtained theoretical power output including the impacts of non-actionable factors. Moreover, the computing procedure of the time series anomaly detection algorithms is complex and the results are not physically explainable. Finally, due to the limitation of the data availability, they used synthetic data for training and testing while we used real data.

Chapter 3

Data Description

Two sets of data are used in this project; they are measured from a string of solar panels deployed in Northwest Toronto by the Toronto and Region Conservation Authority (TRCA), and five strings of solar panels deployed on the campus of the University of Waterloo (UW data), respectively. Toronto and Waterloo are two cities, an hour's drive apart, in Ontario, Canada.

Table 3.1 summarizes the two datasets used in this thesis.

DataSet	Location	Size	Interval	Horizontal Irradiance	In-plane Irradiance	Power output	Ambient Temperature	Wind Speed	Pictures
1	Toronto	1 year	1 min	yes	yes	yes	yes	yes	yes
2a	Waterloo	2 months	1 min	no	no	yes	no	no	no
2b	UW weather station	2 months	15 min	yes	no	no	yes	yes	n/a

Table 3.1: Summary of the two datasets

3.1 TRCA data

TRCA data is collected by the Toronto and Region Conservation Authority (TRCA) and stored in a database. A string of 15 solar panels is deployed facing due south with a 30 degree slope, and the panels are from five manufacturers. For each maker and model, a set of three sample solar panels is deployed. The current, voltage, and temperature of

each panel is measured independently, channels are used to differentiate panels. Ambient Temperature, Wind Speed, Wind Direction, Horizontal Irradiance, In-Plane Irradiance are measured and use dedicated channels. The data collection covers from December, 2011 to December, 2012. The power output data is in one minute granularity. Table 3.2 depicts the schema of the TRCA data.

Tables	schema
Current	MeasurementID: CHAR(8), Date: DATETIME, Reading: DOUBLE, Channel: CHAR(4)
Voltage	MeasurementID: CHAR(8), Date: DATETIME, Reading: DOUBLE, Channel: CHAR(4)
Temperature	MeasurementID: CHAR(8), Date: DATETIME, Reading: DOUBLE, Channel: CHAR(4)
Ambient Temperature	MeasurementID: CHAR(8), Date: DATETIME, Reading: DOUBLE, Channel: CHAR(4)
Wind Speed	MeasurementID: CHAR(8), Date: DATETIME, Reading: DOUBLE, Channel: CHAR(4)
Wind Direction	MeasurementID: CHAR(8), Date: DATETIME, Reading: DOUBLE, Channel: CHAR(4)
Horizontal Irradiance	MeasurementID: CHAR(8), Date: DATETIME, Reading: DOUBLE, Channel: CHAR(4)
In-Plane Irradiance	MeasurementID: CHAR(8), Date: DATETIME, Reading: DOUBLE, Channel: CHAR(4)

Table 3.2: Schema of TRCA data

TRCA data also contains pictures recording the evidence of anomalies on solar panels at an interval of 5 minutes. The pictures cover from Jul. 3, 2012 to Jan. 6, 2013. The resolution of pictures is 600*800 pixels. Figure 3.1 is a picture from the TRCA data, taken at 10:05 am, Aug. 4, 2012. Notice that we cannot observe the dust on panels due to the low resolution; however, we can confirm that there is no snow on panels because the panels are black.

3.2 University of Waterloo (UW) data

UW data is composed from two parts: the power output data collected by our experimental data collection system, and the weather data collected by the weather station of University of Waterloo.

Figure 3.1: An evidence picture from TRCA data

3.2.1 UW power output data

The power output data is collected from 5 strings of solar panels deployed on a building roof (EV3) on the campus of UW. All panels face 26.11 degree south by east, with a 15 degree slope. There are 15 sensors measuring 15 panels distributed in the 5 strings. UW power output data is stored in plain text files. Each sensor generates one text file daily, and the data schema is depicted by Table 3.3.

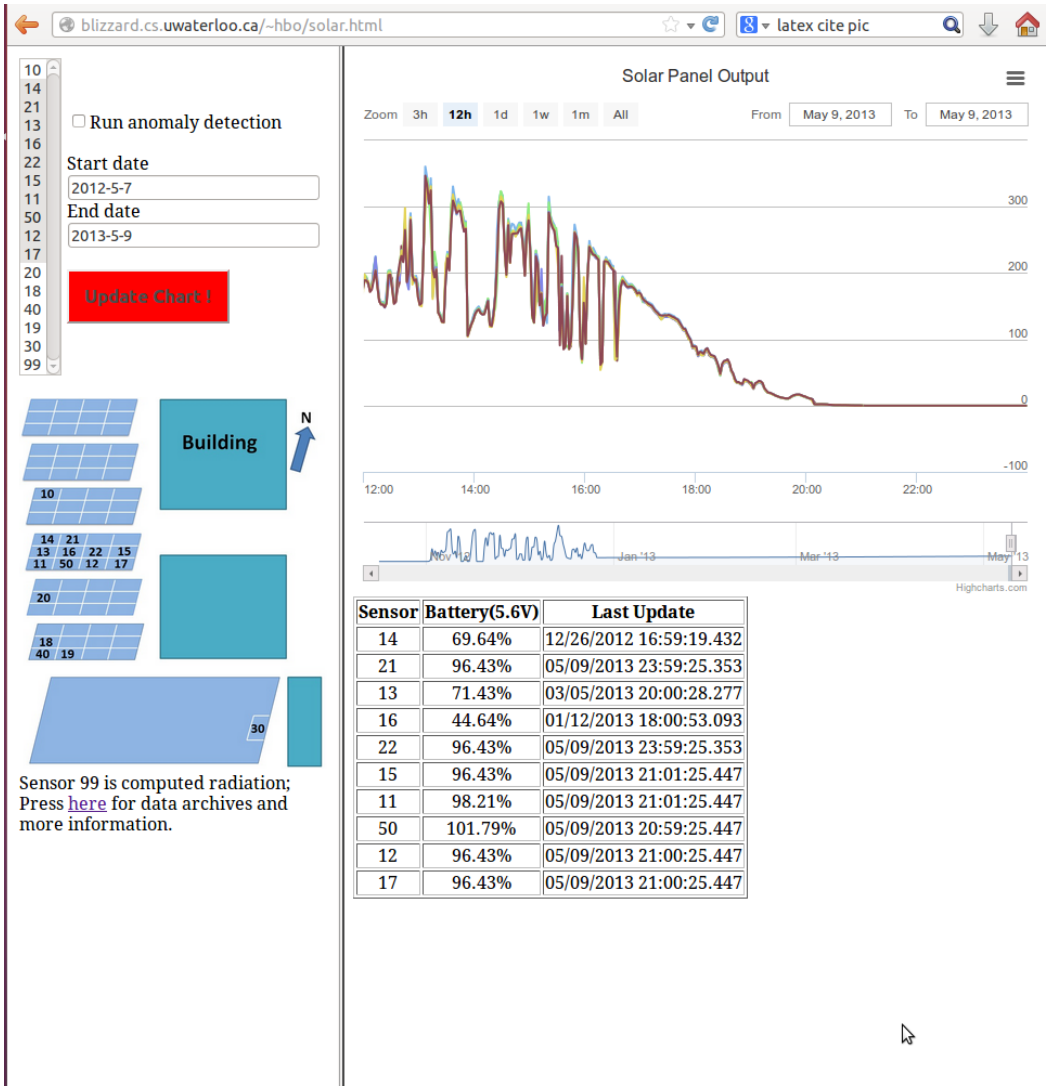
The data's archive and charting is available at http://blizzard.cs.uwaterloo.ca/~hbo/solar.html. Figure 3.2 is a snapshot of the web page developed by Bo Hu, where the left panel contains the sensors selection and a map, and the right panel contains the charts and a table indicating battery status (the data collection system is powered by batteries). 15 sensor IDs are marked on the map, and users can select specific sensors to check the power output. The light blue parallelograms with grids represent solar panels; there are six solar strings altogether, and the bottom one is the longest string. The numbers marked on some of the panels are the sensor IDs, which measured the power output of that panel. There are two big rectangles representing two skylight structures on the east of solar panels, and one smaller rectangle representing the entrance corridor structure to the roof. At sunrise, the three structures will generate shadows on some solar panels; however, we can observe only those on panel 30, because other panels with shadows are either mounted too high to observe or do not have sensors installed. The archive covers two months' data: November, 2012 and July, 2013. The power output data is in one minute granularity, too.

File	Schema
Data	sensor id, voltage (unit volt), battery value (0.1 volts), current value (unit ampere), year, month, day, hour, minute, second and millisecond

Table 3.3: Schema of UW data

3.2.2 UW weather data

The UW data includes meteorology data (UW weather data), which is obtained from the weather station of UW http://weather.uwaterloo.ca/data.html. The weather data is at an interval of 15 minutes, and the data includes Ambient Temperature, Wind Speed,



Sensor	Battery(5.6V)	Last Update
14	69.64%	12/26/2012 16:59:19.432
21	96.43%	05/09/2013 23:59:25.353
13	71.43%	03/05/2013 20:00:28.277
16	44.64%	01/12/2013 18:00:53.093
22	96.43%	05/09/2013 23:59:25.353
15	96.43%	05/09/2013 21:01:25.447
11	98.21%	05/09/2013 21:01:25.447
50	101.79%	05/09/2013 20:59:25.447
12	96.43%	05/09/2013 21:00:25.447
17	96.43%	05/09/2013 21:00:25.447

Figure 3.2: A snapshot of UW data access

Wind Direction, and Horizontal Irradiance. Notice that there is no in-plane irradiance data available, hence, we have to compute the in-plane irradiance using our empirical model in the next chapter.

In general, local weather stations have solar irradiance data, however, the granularity varies. Solar irradiance data can be obtained by installing pyranometers on the same rack of solar panels or nearby. Notice that shortwave radiation values include most of the solar irradiance (0.3 to 3 micrometers wavelength), which can be used as solar irradiance values.

Environmental factors may affect the precision of measurement. For instance, the pyranometers used by the weather station of UW are located by the Columbia Lake, and are mounted horizontally. There is a tower in the south, which will project a shadow on the pyranometers in some time of a year. This effect is seen from about Jan. 9 to Mar. 7, and then again from Oct. 5 to Nov. 21, accounting for the dip in solar radiation readings.

External effects should be taken into account when using the measured irradiance data. Figure 3.3 depicts the horizontal irradiance measurements from the UW weather station on Nov. 8, 2012. Based on the picture, the irradiance is zero before sunrise and after sunset. The irradiance begins to yield from sunrise, and the peak appears at about 11:00, which is the true solar noon of that day, it decreases as the Sun is moving to the west. All the observations are in accordance with common sense, and the fluctuation reflects the changeable weather conditions. For instance, it changed from clear to cloudy, then recovered, and became cloudy again after some time.

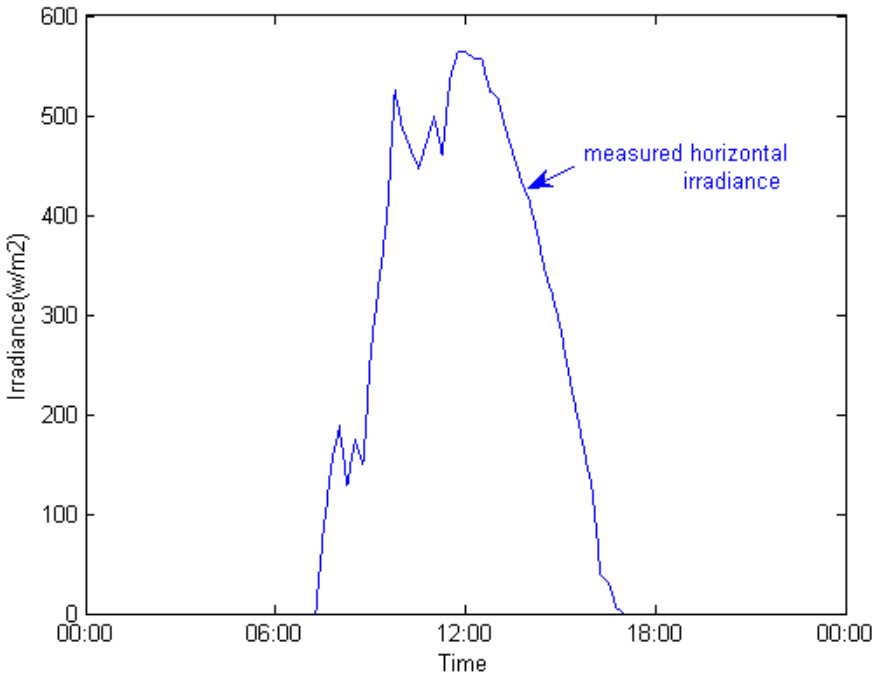


Figure 3.3: Measured horizontal solar irradiance, UW data

Chapter 4

Models and Validation

Theoretical power output as a reference is required to detect anomalies in our approach, since the detection procedure is simply to compare whether the PR is lower than a threshold. Hence, the best reference is the power output from an anomaly-free panel of the same model at the same location; however, this is usually unavailable in practice. Therefore, we introduce a method for computing the theoretical power output with our irradiance model and efficiency equation. In Section 4.1, we start by investigating how to build a local empirical horizontal irradiance model, and then we discuss the procedure to develop a simple model to compute the in-plane irradiance. Finally, we develop a method to find the empirical efficiency equation for a particular PV system. In Section 4.2, we validate our models by comparing the computed results based on our models with the measurements.

4.1 Models

4.1.1 An empirical model for local theoretical horizontal irradiance

Recall Equation (2.6), it appears again below for convenience:

$$G = B_0\varepsilon_0 \times 0.7^{AM^{0.678}} \times cos\theta_{zs} \tag{4.1}$$

In Chapter 2 this equation can be used to compute the horizontal irradiance for California. The computed theoretical solar irradiance fits perfectly with locally measured irradiance values, which form a cosine shape throughout a sunny day without any weather anomalies. Notice that Equation (2.6) is a curve fitting irradiance data collected in California, and that different places in the world will have different constants depending on air quality, and so on. Hence, it is necessary to fit different constants, rather than 0.7 and 0.678, to local data.

Mathematically, the alteration to the exponent from 0.1 to 0.9 (0.1 interval) maintains the shape of the curve, and shifts the curve up with a lower peak. Hence, the alternation of the exponent is mainly used to fine tune the equation. The effect of the change to an exponent on a selected day (Mar. 11, 2012) is depicted by Figure 4.1 (the base number is fixed to 0.8).

However, the alteration to the base from 0.1 to 0.9 leads to a more peaked shape change. The effect of alteration to the base from 0.1 to 0.9 (0.1 interval) is depicted by Figure 4.2 (the exponent is fixed to 0.41).

To determine the precise parameters fitting real irradiance in Toronto, we conducted the following steps:

First, we selected sunny days in the year. Daily *radiation* is the integration of irradiance in an entire day. Intuitively, high solar radiation is a necessary condition of sunny days. However, radiation on summer days is generally higher than in winter, due to higher incidence angles. Moreover, to assure the model be valid across a year, it is essential to select the representive days distributed evenly in a year. Therefore, we chose the top 10% of days with the highest solar radiation from each month as candidate days.

According to Equation (2.6), irradiance on a perfect sunny day froms a cosine like shape curve. However, clouds cover will deform the smooth cosine curve, as depicted by Figure 3.3. From the perspective of spectrum analysis, a deformed curve can be regarded as a lower amplitude cosine signal plus a series of high frequency signals. Correspondingly, in the frequency spectrum, a perfect cosine signal has only the main lobe (the lobe containing peak amplitude), that is, the amplitudes of its side lobes (the lobes containing local peak amplitude) are zero. However, there are many non-negligible side lobes for a

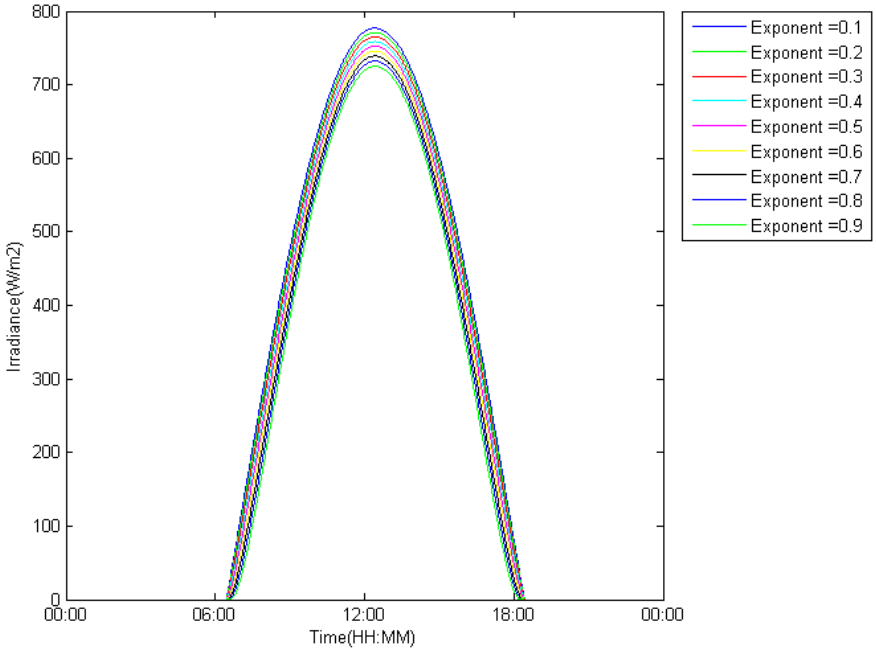


Figure 4.1: Alteration to the exponent from 0.1 to 0.9

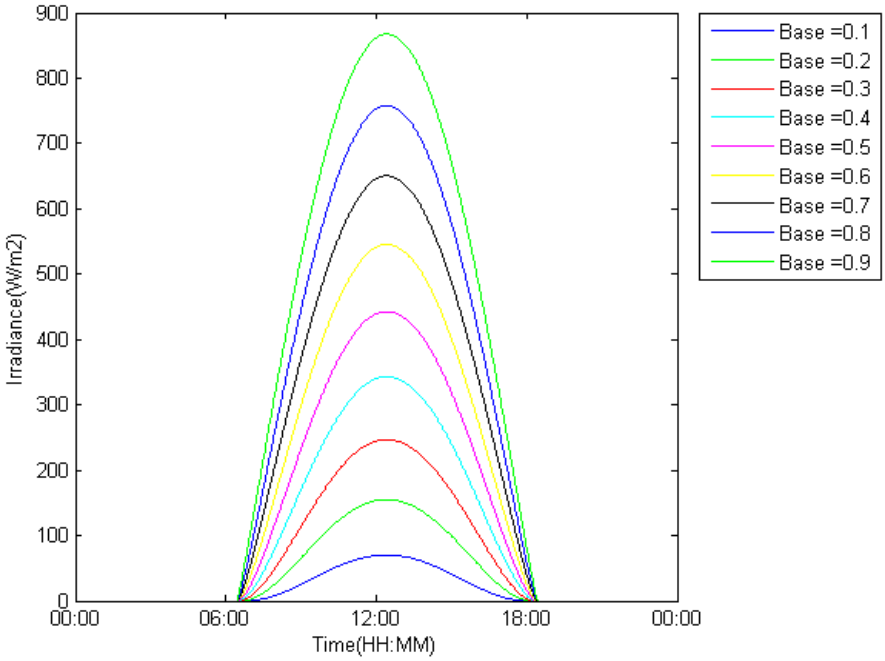


Figure 4.2: Alteration to the base from 0.1 to 0.9

deformed cosine signal. So we can use these two criteria to evaluate whether a signal has a smooth cosine shape [50]: Peak-to-Sidelobe Ratio (PSLR) and Integration-to-Sidelobe Ratio (ISLR).

The PSLR is computed by Equation (4.2), where $S_{firstlobemax}$ is the amplitude of the first lobe (the side lobe next to main lobe), and S_{max} is the amplitude of the main lobe:

$$PSLR = 20 * log_{10}(S_{firstlobemax}/S_{max}) \tag{4.2}$$

The unit of PSLR is dB; since the maxima of the first lobe is always smaller than the main lobe, the PSLR is always negative. The bigger the absolute value is, the less the interferences deform the shape. Figure 4.3 depicts the PSLR of the irradiance on the preliminarily selected sunny days:

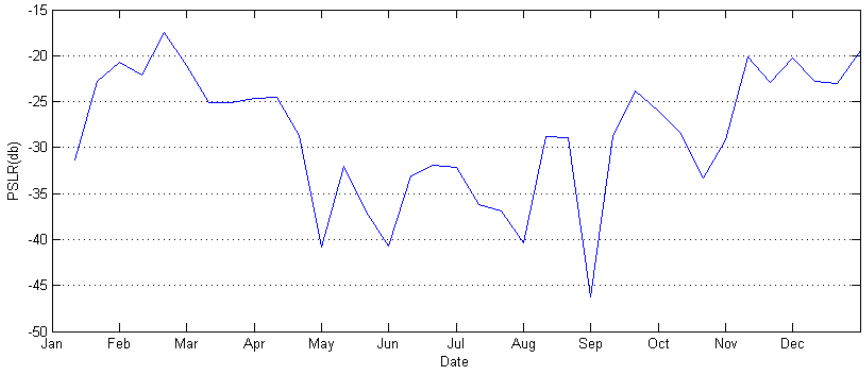


Figure 4.3: PSLR of the three days with the highest radiation in each month

Based on the PSLR, the irradiance on the selected days of winter is relatively not so smooth as those in the three other seasons. Hence, 21 days are selected, on which the absolute value of the PSLR is greater than 25dB, for further inspection.

Notice that small PSLR cannot guarantee that the curve is smooth. For instance, the PSLR on Jan. 15, 2012, is -31.29dB, a medium value of the PSLR set; however, the irradiance curve on that day is deformed, based on the measurements. The irradiance on Jan. 15, 2012 is depicted by Figure 4.4:

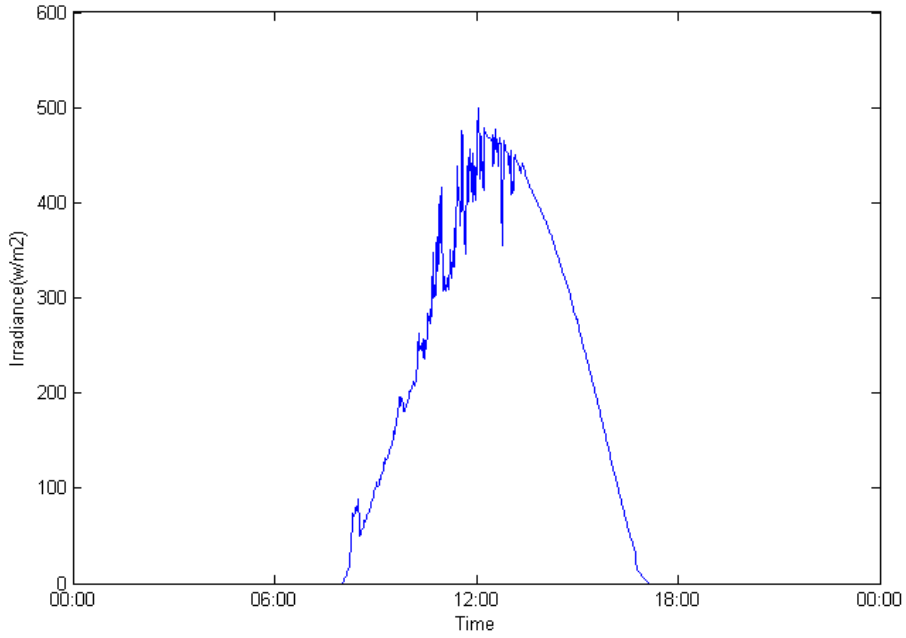


Figure 4.4: Measured horizontal irradiance on Jan. 15, 2012, Toronto

There are obvious cuts in the morning according to Figure 4.4. Sometimes, there are many comparable side lobes with close amplitudes. In these cases, only comparing the first lobe with the main lobe cannot reflect the severity of the interferences caused by side lobes. ISLR can reflect the total energy distributed in all the side lobes by computing the ratio between the integration of all side lobe energy with the integration of main lobe energy. Hence, we also employed ISLR to help to select clear days, which is depicted by Figure 4.5:

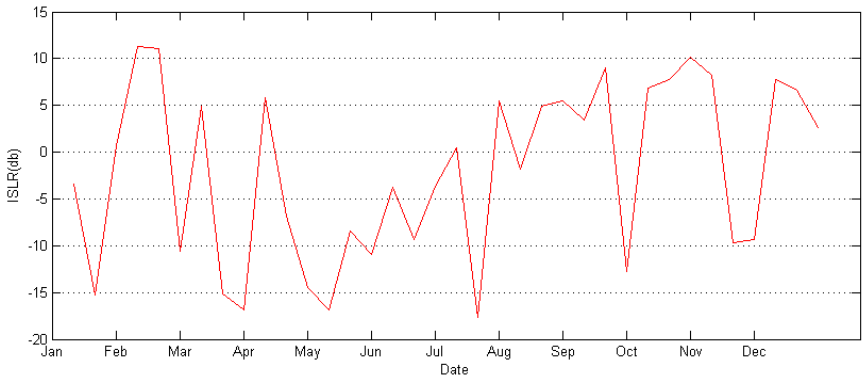


Figure 4.5: ISLR of the three days with the highest radiation in each month

The unit of ISLR is dB; however, compared with PSLR, which is always negative, ISLR can be positive, which means the integration of the energy of all side lobes is higher than the main lobe. Correspondingly, the signal is greatly deformed. Therefore, we selected the days with ISLR less than -5dB, and derived the intersection of the two sets of days according to ISLR and PSLR.

However, there is some inconsistency between the two charts. For instance, the ISLR on Nov. 16, 2012, is -9.6dB, a value indicating a relatively smooth curve, while that day is filtered according to the PSLR. Therefore, we manually inspected the measured solar irradiance of the selected days to exclude any types of anomaly that would deform the irradiance curve, and remove the days with broken data simultaneously. Finally, 11 completely sunny days with smooth irradiance curves are selected. These days are distributed evenly across three seasons: spring, summer, and fall.

We used irradiance data from the completely sunny days selected, for fitting and verification. Let S be the set of the sunny days, d be the the days in S, t be the minute in a day, Im be the measured irradiance, and Ic be the computed irradiance. The Ic can be computed by equation:

$$Ic = B_0\varepsilon_0 \times BASE^{AM^{EXP}} \times cos\theta_{zs} \tag{4.3}$$

where the ε_0 is a function of d, and both AM and θ_{zs} are functions of d and t.

Using minimum square error (MSE) as criteria, the objective function is:

$$min \sum_{d} \sum_{t} (Im_d(t) - Ic_d(t))^2 \qquad d \in S, t = 0, 1, ...1439 \tag{4.4}$$

According to the mathematical analysis for the two constants, the higher the base is, the higher peak value the curve achieves. Considering that the latitude of Toronto is higher than that of California, so the peak should not exceed the constant fit in California. Hence, we set the objective function subject to:

$$BASE, EXP \in [0, 1] \tag{4.5}$$

It is a nonlinear least-squares optimal problem, or equivalently, we want to find the values for $BASE$ and EXP, which can achieve the MSE for the objective function (4.4). We used the Nonlinear Least Squares (Curve Fitting) function "lsqnonlin" in the Matlab Optimization Toolbox to solve this problem. The result is depicted by Table 4.1:

BASE	0.84
EXP	0.51
R-square	0.91

Table 4.1: Results of the Nonlinear Least Squares function

Notice that the $BASE$ in this result is greater than that in Equation (2.6), that can be explained by the higher cleanliness in the city Toronto than in the desert. The high R-square value indicates that the equation with the two contants of this result fits the data

well. Above all, the optimal equation, based on TRCA irradiance data, for computing horizontal theoretical irradiance in Toronto is:

$$G = B_0\varepsilon_0 \times 0.84^{AM^{0.51}} \times cos\theta_{zs} \tag{4.6}$$

4.1.2 A simple model for theoretical in-plane irradiance

The conventional method must separate the three components of solar irradiance, by applying local historical data to the empirical equation. This step will generate a fair amount of errors. In fact, even though we can exactly obtain the three components, the following computing procedure is still complex and time consuming. Hence, we develop a simple model to calculate the global irradiance on inclined surfaces w.r.t. the known parameters, θ_s. Recall Equation (2.7), which is used to compute this angle and is listed below for convenience:

$$\begin{aligned} cos\theta_s = sin\delta sin\phi cos\beta - sin\delta cos\phi sin\beta cos\alpha + cos\delta cos\phi cos\beta cos\omega \\ + cos\delta sin\phi sin\beta cos\alpha cos\omega + cos\delta sin\alpha sin\omega sin\beta \end{aligned} \tag{4.7}$$

Suppose we project the horizontal surface to a plane, and the normal of that plane parallels with the incident solar beams, so the equivalent area for the horizontal surface S_h is $cos\theta_{zs}$ times of its original area S. Similarly, the equivalent area of an inclined surface after projection is $cos\theta_s$ times of S. Since the panels with the same size ought to receive the same amount of irradiance at the optimal angles, the irradiance on an inclined surface can be calculated by the following equation:

$$G_I = \frac{G cos\theta_s}{cos\theta_{zs}} \tag{4.8}$$

This equation simply uses the law of cosines in spatial geometrics and is able to compute in-plane irradiance in one step.

4.1.3 An equation to obtain measured in-plane irradiance

In-plane irradiance is the total irradiance that arrives on an inclined surface. Sometimes there is only local horizontal measured irradiance data available; hence, we develop an intuitive approach to derive in-plane irradiance with the aid of theoretical models.

First, we compute the ratio between theoretical in-plane irradiance and theoretical horizontal irradiance. Next, we then obtain equivalent in-plane irradiance measurements as products of the ratio in step one and the horizontal irradiance measurement. Equation (4.9) summarizes the procedures:

$$I_{In\_PlaneMeasurement} = \frac{I_{HorizontalMeasurement} * I_{TheoreticalIn\_PlaneMeasurement}}{I_{TheoreticalHorizontalMeasurement}} \tag{4.9}$$

4.1.4 An empirical efficiency equation

To obtain the efficiency equation w.r.t. the PV panels based on the TRCA data, we conducted following steps: first, we computed the PR of the five substrings of the solar panels during the daytime, w.r.t. the sunny days, respectively. Of the five substrings, the standard module efficiency can be derived from manuals. Next, we observe that efficiency is quite stable during a single day, whereas it varies slightly on different days. Figure 4.6 depicts the efficiency on the clear days selected in Section 4.1.1 on a panel.

Figure 4.7 depicts the correlation between temperature and efficiency on the same panel. As these days are distributed evenly across a year, the temperature has a wide coverage. From the Figure a negative relationship between temperature and efficiency can be observed.

Temperature accounts for these small changes in efficiency. Notice that temperature varies by as much as 40°C, such that its impact should not be ignored. From regression analysis results, we obtained a theoretical efficiency equation w.r.t. current temperature, which is a revision of Skoplaki's engineering equation [60]:

$$\eta_c = \eta_{T_{ref}} \left[1 - \beta_{ref} \left(T_c - T_{ref}\right)\right] - 0.06 \tag{4.10}$$

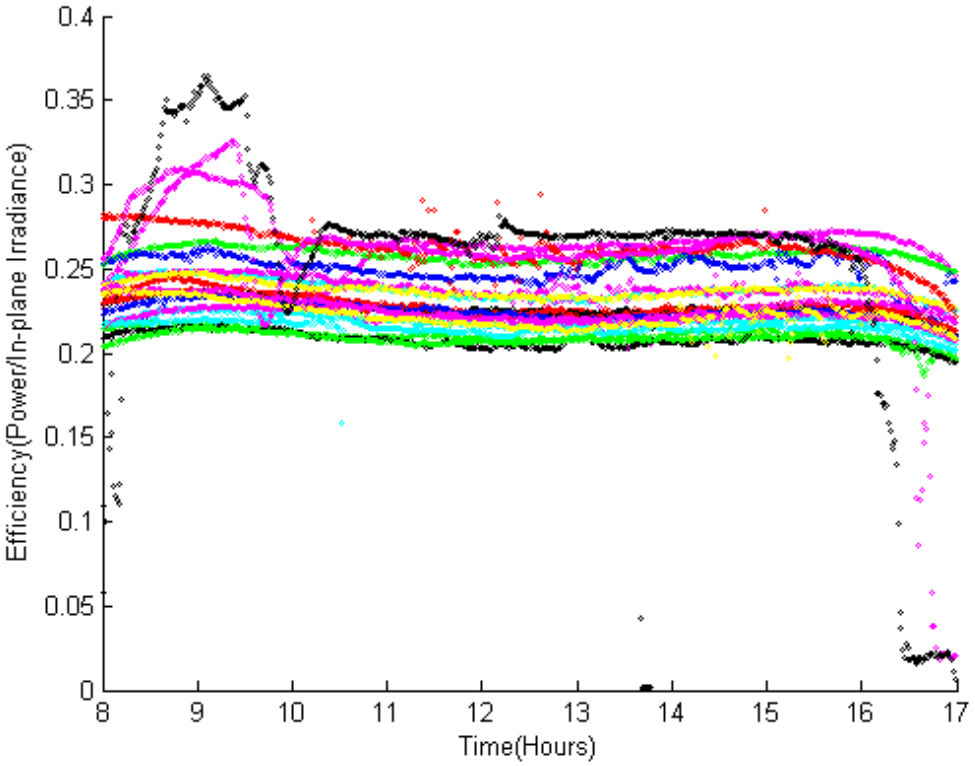


Figure 4.6: Efficiency on 21 sunny days of 2012 on panel 3, based on TRCA data

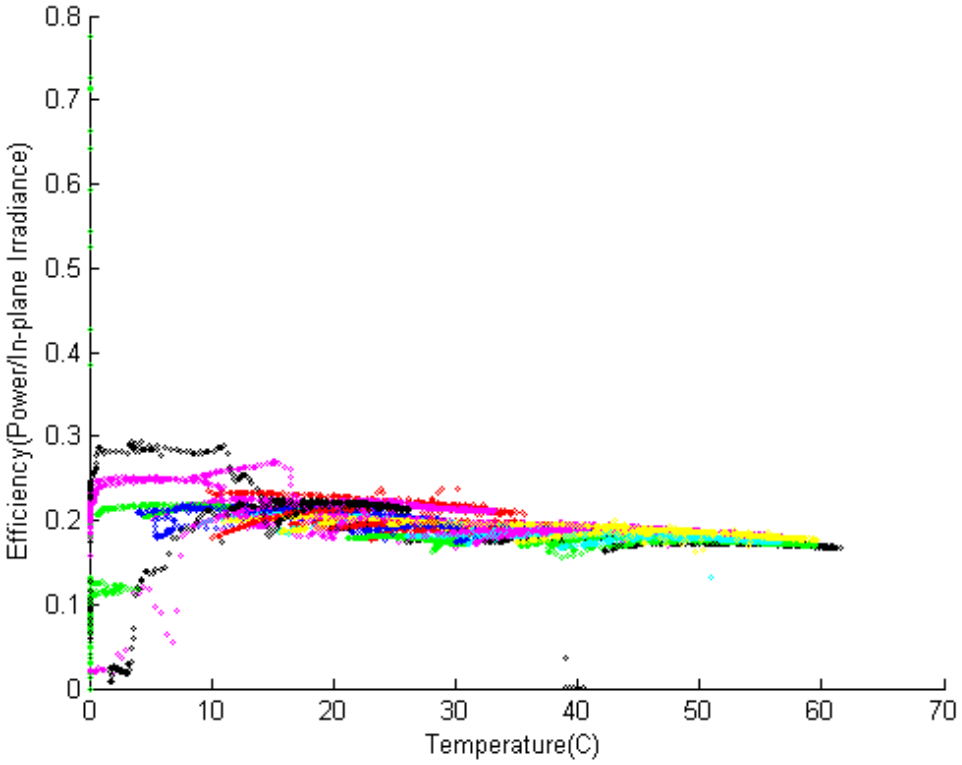


Figure 4.7: Correlation between temperature and power efficiency on panel 3, based on TRCA data

where T_{ref} is the reference temperature, 25°C; $\eta_{T_{ref}}$ is the efficiency at the reference temperature; β_{ref} is the coefficient determined by material properties, and is given by the manuals. Usually, β_{ref} varies from 0.0044 to 0.0047 per °C. Note that where T_c is the current temperature, η_c is the current efficiency. The constant -0.06 is obtained from the regression of our efficiency data on the clear days selected in Section 4.1.1. Evaluation results show that this equation fits up to two decimal places with measurement, which is precise enough for anomaly detections.

In addition, the PR could also be affected by irradiance; for some panels, it decreases linearly by 5% when the irradiance decreases from 500 to 200 W/m^2 [5]. Hence, we corrected the efficiency equation when irradiance drops into this range to diminish inaccurate detection in the morning and afternoon. The dust related research [67] shows that the dust has a greater effect when the incident angle is large (sunrise and sunset); however, the low irradiance at that time could also account for this.

Research is seldom conducted under conditions in which irradiance is below $200W/m^2$, due to high uncertainty and very low profits. Finally, we can assume the panel efficiency will not degrade due to materials for many years [5].

4.2 Validation

Since the model establishment is based on MMSE, the model assures a minimum MSE in fitting in-plane irradiance data from anomaly-free days. Considering the short distance between the two cities, we apply the horizontal irradiance model built with TRCA data (Toronto) on UW data (Waterloo). It has a 4.1% error rate on the anomaly-free sample days.

4.2.1 Theoretical in-plane irradiance model

We validated our theoretical in-plane models in two ways. For UW data, as the in-plane irradiance measurement was not available, we computed the in-plane irradiance using the conventional complex model, and compared the results with the results generated by our

simple model. Generally, the theoretical result is higher than the measurement. Notice that in-plane irradiance is usually higher than horizontal irradiance because the tilted panels often face the sun more. Figures 4.8 and Figure 4.9 depict the theoretical in-plane irradiance on Mar. 11, 2012 and on Sep. 16, 2012, respectively (panel slope $\beta = 15$ degree, orientation angle $\alpha = 23.11$ degree) on the UW campus. From bottom to top, the four curves are measured horizontal irradiance, which is obtained from the UW weather station and interpolated; theoretical horizontal irradiance, which is computed with our empirical equation; theoretical in-plane irradiance computed by our simple model, and theoretical in-plane irradiance computed by the conventional complex model. The external data used in the conventional model is retrieved from [**?**]. Notice that the solar noon is at 13:32 on that day, so the peak irradiance appears in the afternoon. There is a fair gap between the measurement and the horizontal irradiance for UW data, for two possible reasons. First, the empirical equation fits using Toronto data. Second, there is no evidence to prove that this measurement is taken on an anomaly-free day. However, the gap is still less than 5%, so it is sensitive enough to detect an anomaly with proper thresholds.

For TRCA data, we do have the in-plane irradiance measurements, so the measured in-plane irradiance is plotted as a reference. Figure 4.10 and Figure 4.11 depict the theoretical in-plane irradiance on Mar. 11, 2012 and Sep. 16, 2012, respectively (panel slope $\beta = 30^o$, orientation angle $\alpha = 0^o$), in Toronto, using both our simple model and the conventional model as well. The external data used in the conventional model is retrieved from NASA (https://eosweb.larc.nasa.gov/sse/). From bottom to top, the five curves are: measured horizontal irradiance; theoretical horizontal irradiance, which is computed with our empirical equation; theoretical in-plane irradiance computed by the conventional model; theoretical in-plane irradiance computed by our simple model, and measured in-plane irradiance. There is a negligible gap between the measurement and horizontal irradiance. Notice that our simple model actually performs better in fitting the measured in-plane irradiance.

To summarize the results of our validation experiments: our model fits the real horizontal solar irradiance curve well on arbitrary days and different locations, and for the in-plane irradiance, our simple model achieves comparable results to the conventional complex model.

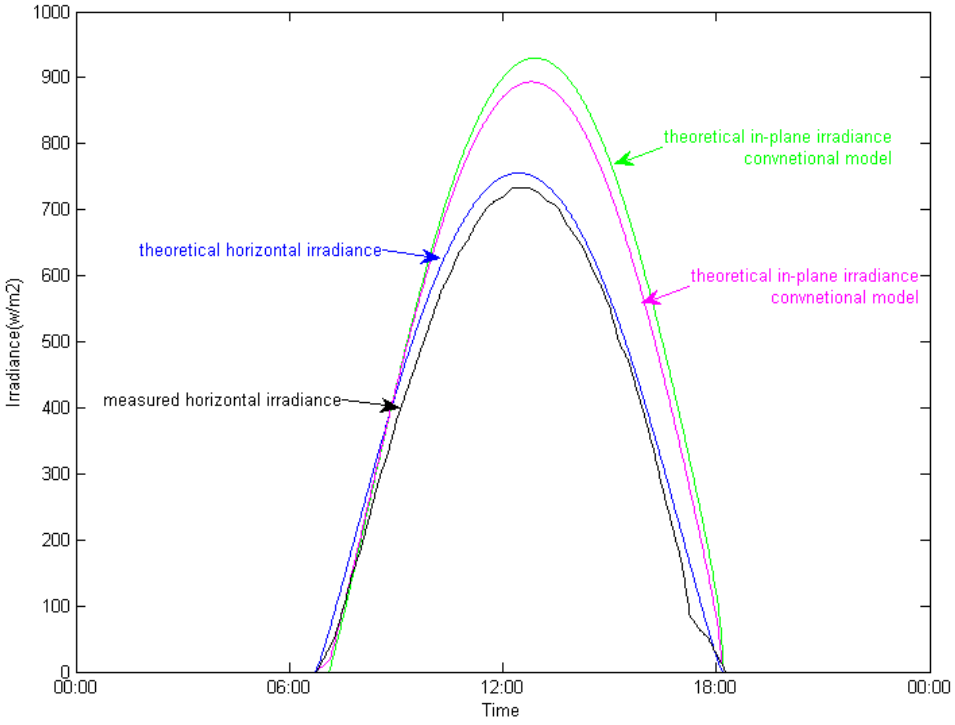


Figure 4.8: Theoretical in-plane irradiance on Mar. 11, 2012, UW

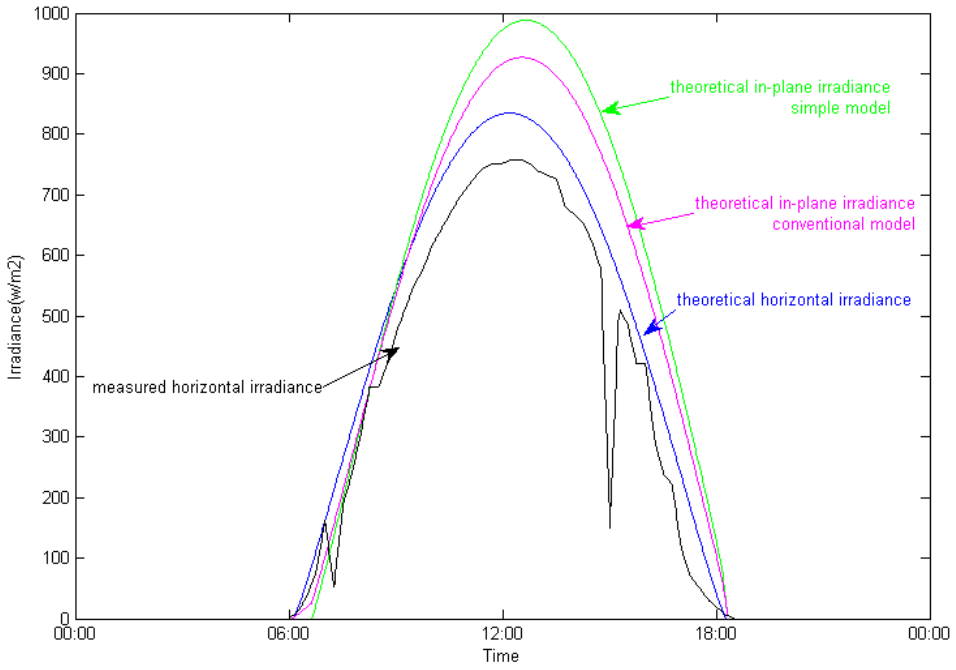


Figure 4.9: Theoretical in-plane irradiance on Sep. 16, 2012, UW

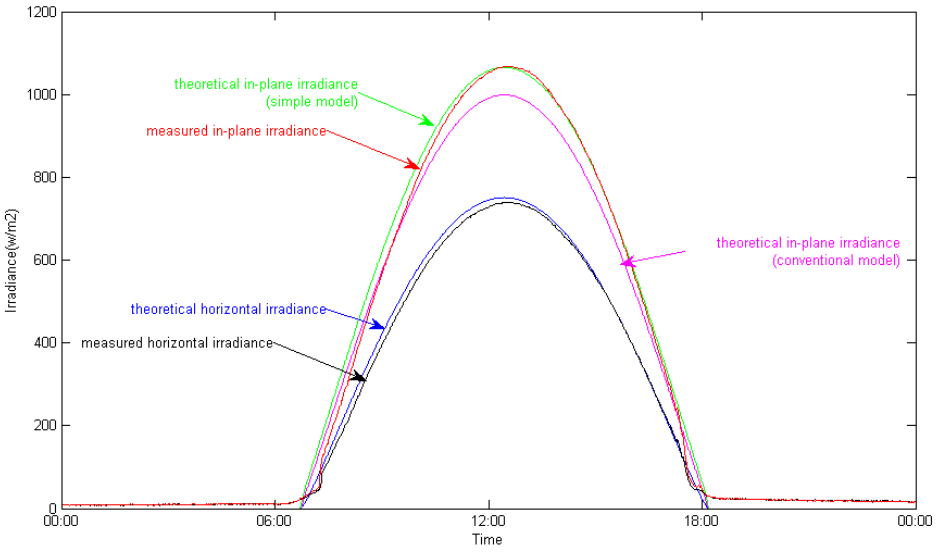


Figure 4.10: Theoretical in-plane irradiance on Mar. 11, 2012, Toronto

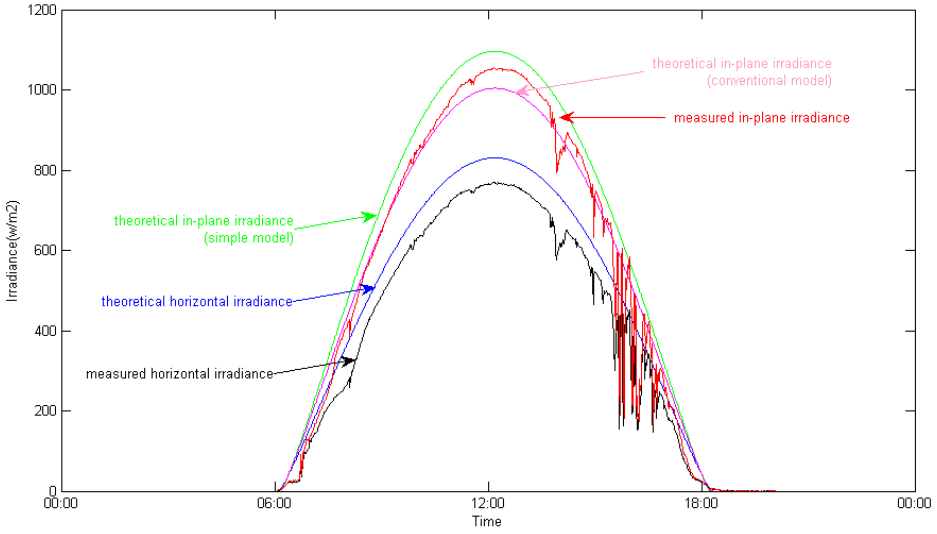


Figure 4.11: Theoretical in-plane irradiance on Sep. 16, 2012, Toronto

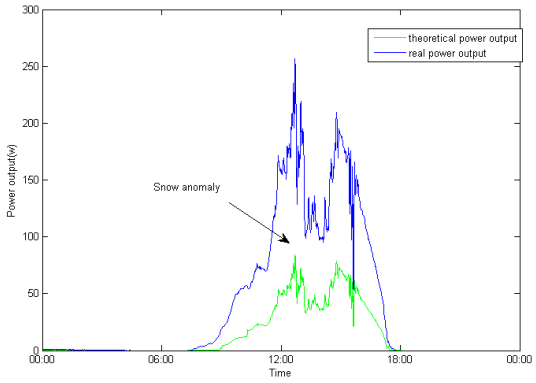


Figure 4.12: Power comparison, Toronto, Feb. 11, 2012

Figure 4.13: Corresponding Picture

4.3 Calculation of theoretical power output

The in-plane irradiance already contains the weather impacts, regardless of whether it is measured directly or derived from horizontal measured irradiance. Any gap between the theoretical power output and the real power output over a threshold should be regarded as an anomaly. We use PR to describe the magnitude of the gaps in the next chapter.

Figure 4.12 depicts the theoretical power output in Toronto on Feb. 11, 2012. Based on the calculated in-plane irradiance, we also plot the real power output (measured) for reference. The anomaly across the entire day is snow, which drops the power by 60% for most of the daytime. Figure 4.13 is the picture taken on that day, where the examined panel is circled in red.

Figure 4.14 and Figure 4.15 depict the theoretical power output in Toronto on Sep. 16, 2012 and Mar. 11, 2012, respectively, based on the measured in-plane irradiance. No actionable anomaly has been observed.

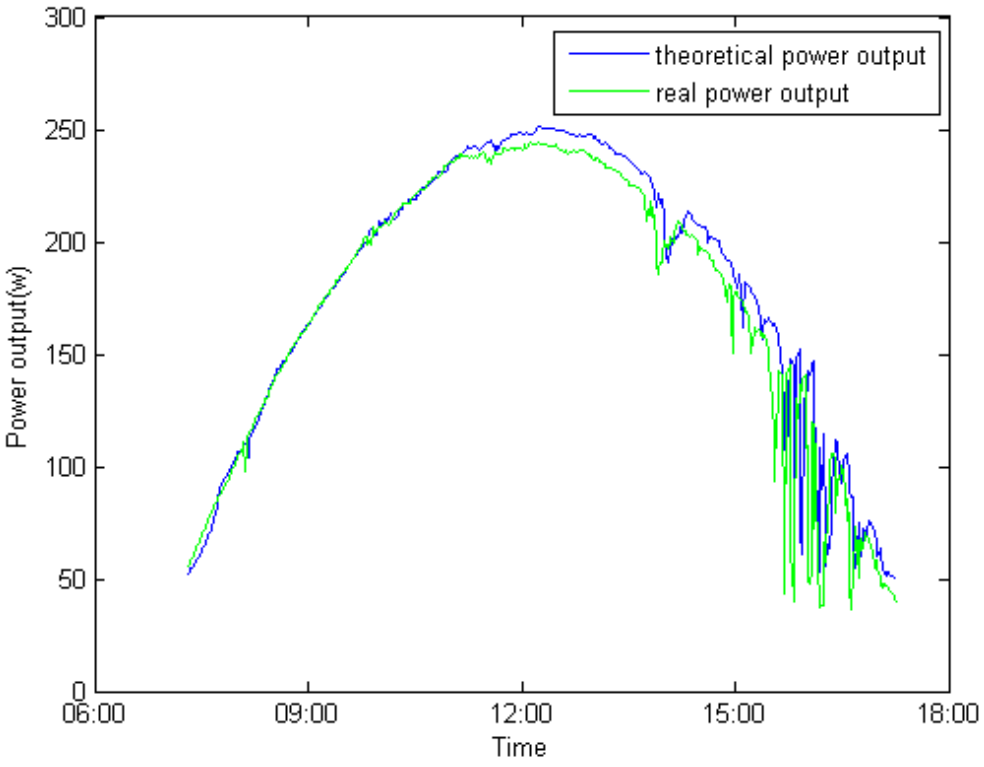


Figure 4.14: Comparison of theoretical power output with real output, Toronto, Nov. 29, 2012

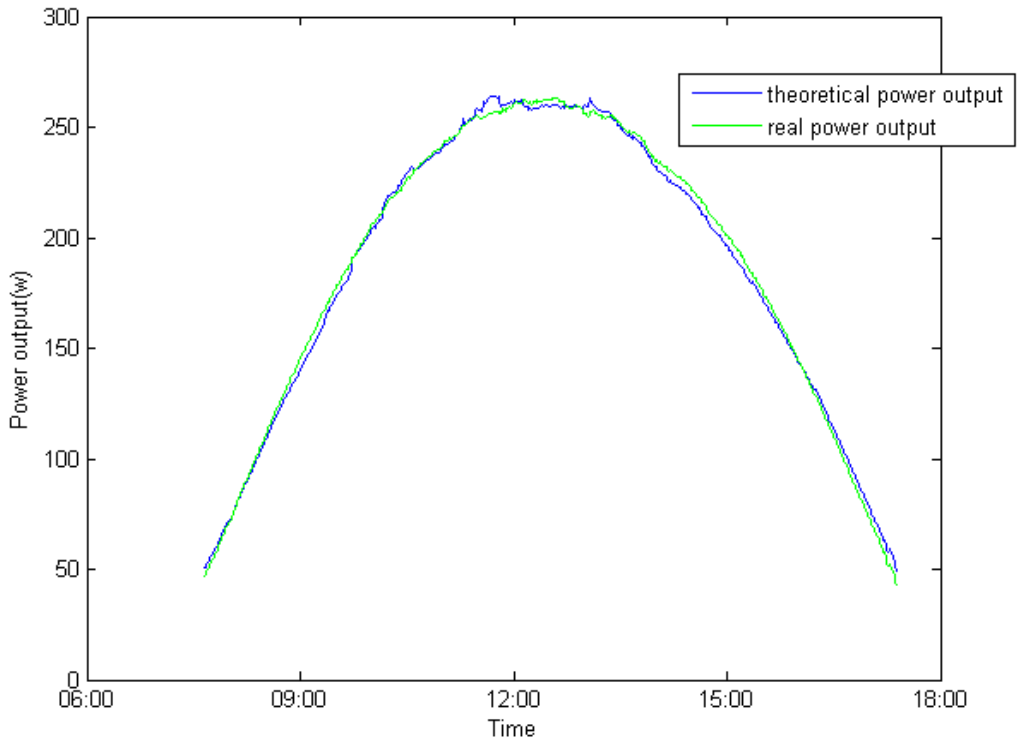


Figure 4.15: Comparison of theoretical power output with real output, Toronto, Mar. 11, 2012

Chapter 5

Anomaly Detection and Classification

Using the models in Chapter 4, we can obtain the PR of each panel, leading to detection and further classification. This chapter is organized as follows: Section 5.1 presents our methodology of detection and classification, and Section 5.2 discusses the experimental evaluation of the effects of anomalies, including why we need to collect anomaly samples, how we did that, and what we found. Section 5.3 is the evaluation, which includes the data quality, and discussion of the classification results.

5.1 Methodology

5.1.1 Detection

Anomaly types can fall into two categories: non-actionable and actionable. It is not possible to eliminate the former one manually, however, it can be detected by comparing theoretical in-plane irradiance with measured in-plane irradiance. Figure 5.1 depicts this approach using TRCA data for Feb. 11, 2012, when there was snow on the panels. The weather anomalies can be detected by the comparison between the theoretical irradiance and the measured irradiance, regardless of whether it is in-plane or horizontal. Irradiance measurement is regarded as anomaly-free, because pyranometers are small, and many

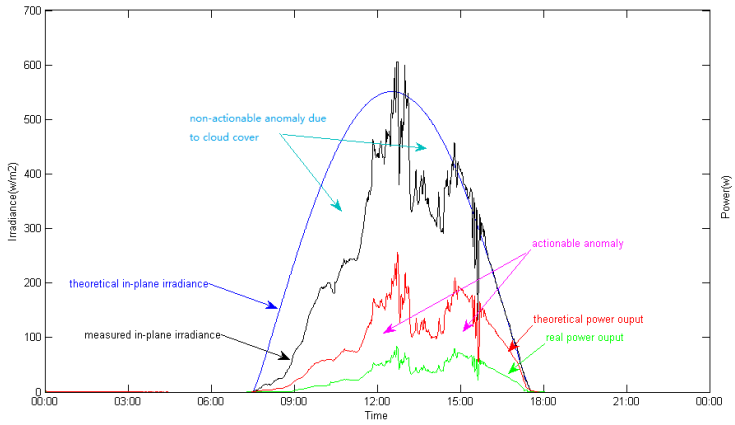


Figure 5.1: Anomaly disaggregation

material technologies are applied to isolate anomalies, such as electrostatic anti-dust glass and heaters for melting snow. Then, the theoretical power output can be computed based on in-plane irradiance and the efficiency equation. The large gap between theoretical power output and real power output indicates anomalies on panels. Notice that the measured irradiance exceeds the theoretical irradiance at noon, which can be explained by the over-irradiance caused by overcast in high latitude areas [66].

However, anomalies worth focusing on are those that can be removed. With in-plane irradiance and efficiency equation 4.10, the theoretical power output can be computed. Any significant decrease in real power output is due to certain anomalies. Multiple PR thresholds ranging from 0.5 to 0.95 have been tested in this project.

We check only the irradiance greater than 200 W/m^2, since efficiency with irradiance below 200 is not given by the manufacturer, and it is trivial due to the high uncertainty and low power profits. This is common in solar energy analysis. For example, Zorrilla-Casanova, et al., [67] analyzed the dust anomaly by checking only the time when the irradiance was greater than 200 W/m^2.

However, this threshold may occasionally delay anomaly detection. For instance, if it continues to be overcast after snowing, the snow anomaly will not be detected since the irradiance is very low on overcast days in winter. For example, according to the TRCA

data, it began to snow at 16:00, on Jan. 26th, 2012 in Toronto, but this anomaly could not be detected due to the very low irradiance, until 11:00, on Jan. 27th, when the irradiance was high enough.

In this project, we detected the samples only in daytime, from 8:00 to 18:00, during which panels yield 90% of the daily solar energy. Usually, only one anomaly occurs with a duration longer than 20 minutes per day. For real-time detection, the aggregations are maintained for each panel, and a profiling and classification process is conducted every 30 minutes, which will be discussed in the next section.

5.1.2 Classification

To classify different types of actionable anomalies, features need to be extracted and associated with certain types of anomalies. Anomalies on each panel are detected independently. However, in classification anomalies are the union of panel anomalies with overlapping existing time. Since an anomaly may affect several panels, and each affected panel has its own PR, the average PR (APR) of an anomaly is weighted by the duration of the anomaly on each affected panel.

This is the crucial observation that we make in this thesis: according to the UW data collected, shadows lead to low, but relatively stable, power output, which can be explained by the relatively stable diffuse component of irradiance. The diffuse component accounts for around 50% of global irradiance, and will not be blocked by anomalies, such as shadows, because it arrives on panels by an indirect path. However, direct cover anomalies, such as dust, will block all components of irradiance, including the direct component and the diffuse component, because there is no distance between these components and the panels. The extent to which irradiance is blocked by direct cover is mainly determined by the thickness and density of the anomalies. Therefore, the existing physical conditions lead to different effects of these two different types of categories: indirect cover leads to consistently low power output, regardless of the change of other components of irradiance; in contrast, the direct cover will lead to a low, but changing, power output, tracking the change of the irradiance. Figure 5.2 depicts the power output of panel 21 w.r.t. the theoretical power output on Jul. 10, 2013, which reflects the changing of irradiance. From the Figure it is

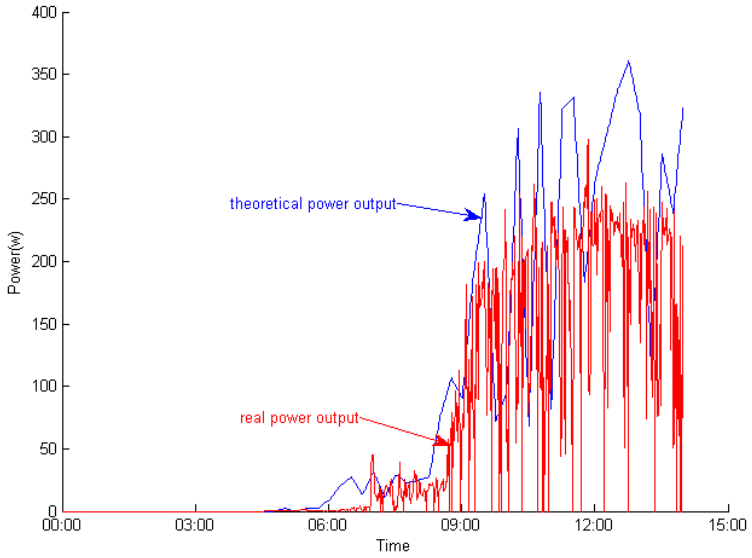


Figure 5.2: Panel 21 without shadows tracks irradiance, UW data

clear that the power output of this panel almost completely tracked the irradiance in the morning; recall that in Figure 3.2 in Chapter 3, panel 21 is at a high and open position and there are no shadows on it. In contrast, Figure 5.3 depicts the power output of panel 30 w.r.t. the theoretical power output. From the figure, the power output of panel 30 does not track the irradiance before 11:30 (it recovers before noon, as its orientation is south-east), rather, it maintains a consistent and low power output. Recall that in Figure 3.2 in Chapter 3, the corridor structure is beside the panel 30 and projects shadows on it.

The coefficient of variation (CV) is defined as the ratio of the standard deviation σ to the mean μ [18] of a sequence of values. CV reflects the deviation degree, so we used CV to describe how the power output tracks the variation of irradiance. Particularly, a panel has a shadow on it in the morning, but due to the relatively constant diffuse component in the irradiance, the absolute power output does not fluctuate much. However, power generated by panels that do not suffer from indirect cover anomalies are very sensitive to irradiance, which is impacted by changeable weather conditions. Recall that PR actually reflects the relation between the irradiance and power output, so the PR of indirect cover

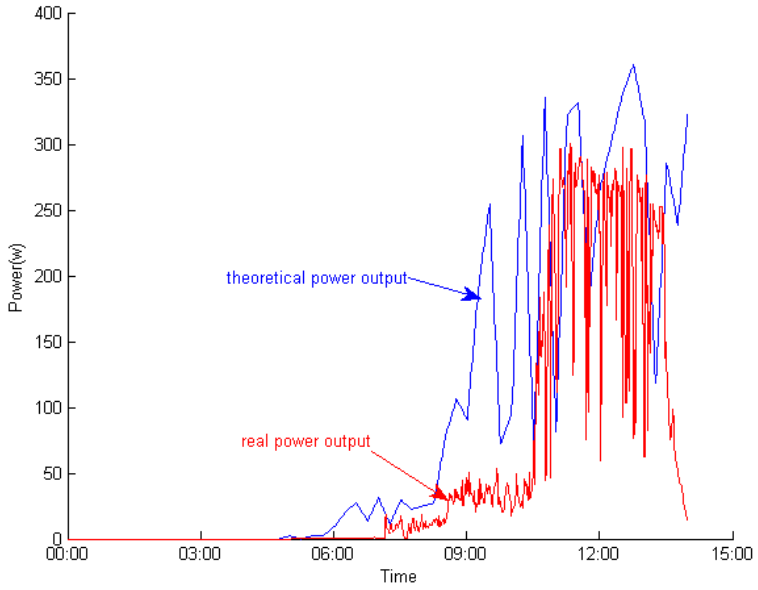


Figure 5.3: Panel 30 with shadows does not track irradiance, UW data

anomalies fluctuates greater than that of direct cover anomalies. Consequently, for indirect cover anomalies, such as shadows, their CV of PR is high, usually greater than 1.

In addition to CV of PR, several attributes are also useful to classify these anomalies :

Duration : According to the data with pictures from TRCA data, snow typically exists from one to three days. Based on manually observed samples, duration for dust is highly dependent on the precipitation intervals; shadows last for 6 hours at most, unless the panel is mounted facing north in the northern hemisphere, which is unrealistic. Finally, transient anomalies, such as dew or frost, disappear in 20 minutes based on the TRCA data.

Time : In practice, solar panel deployment will avoid facing a high block due south, so shadows often occur in the morning (for a structure located in the east) in our experiment, or at sunset (for a structure located in the west). The discovery time for dust anomalies can be arbitrary, when the dust accumulation is thick enough to be detected. Snow occurs at night in most cases, so it is discovered in the morning, but it may also snow in the morning and in the afternoon.

Number of affected panels : For dust and snow, all panels usually are affected. However, for shadows, several or all panels can be affected. In the samples we collected, shadows always affected several panels of the whole string.

Average PR : The PR is usually high for dust, due to its long accumulation time. However, it is low (lower than 0.4) for snow; this attribute should be customized based on local climate.

Season : It is mainly employed for snow identification, and can be replaced by month if a finer category is needed. In Canada, winter usually lasts from November to March.

5.2 Experimental evaluation of the effects of anomalies

From TRCA data, we set that panels whose PR below the threshold 0.9 are considered to have anomalies on them. For the detected 127 anomaly samples, we manually checked the corresponding pictures to label the samples. Finally, we collected 24 snow anomaly samples. Most of the other detected anomalies occured in the morning, and could be explained as frost or dew. However, other anomalies, such as dust, leaves, and shadows, had not been found.

Hence, we needed to conduct experiments to collect more samples of various types of anomalies. We had access to the roof of the building EV3 for one month, from Jun. 20 to Jul. 20, 2013, when we were able to run a series of experiments and collect a small number of anomaly samples. This data is the major component of the UW data, and it includes real power output collected by the system and manually recorded anomaly evidence.

5.2.1 Methodology

Recall that there is a tall structure to the east of panel 30 according to Figure 3.2, which will project shadows on panel 30 and its neighbour panels in the morning on clear days. Since this panel is mounted at a low position, we can observe shadows on it in the mornings

and record the starting and ending time for each anomaly. We retrieved the corresponding power output from archives based on these time records.

Southern Ontario had 140mm of precipitation distributed in 15 days of the experiment period; therefore, we observed few dust anomalies. We collected the dust anomaly samples by manually putting some dust on the panels. We first used papers to simulate dust, and then found that there were obvious differences between the two kinds of materials, which will mislead the experiment results. So we collected some fine sand and mixed it with dried soil, gradually put it on panels, and made sure that the effects of the dust anomalies fell into a reasonable range by referencing natural dust observations [67] [54]. The dust anomaly disappears after rainfall (even a light rain below 1 mm is enough to clean the panels [67]) or with a strong wind (depending on the humidity, wind direction and panel orientation). To prevent it from being blown away by wind, we humidified it by spraying water on it. Since it rained often during that time, we increased the forming speed of the dust anomalies by adding dust at an interval of 10 minutes. Otherwise, it is impossible to observe natural dust anomalies in summer in the Waterloo area, given the rainy climate and the clean air.

5.2.2 Results

We used real data collected in this project. Of the data collected in Toronto across 2012, we had 24 detected and labelled snow anomalies. For the other two types of anomalies, we performed experiments on the solar panels deployed on the UW campus. From the experiment, we derived 18 natural shadow anomalies by a structure on the roof, and 18 dust anomalies by gradually putting sand on panels manually.

The following describe our experiment conclusions:

For snow anomalies, the average PR ranges from 23% to 80%. The power output ranges from 4.9 to 129 watts, which corresponds to a melting process on a sunny day. Recalling the related work of the effects of snow anomalies, the two features of our snow samples are consistent with the results of previous work [30] [17] [11]. However, only four out of the 24 samples have an average power output greater than 30 watts, due to the low sun height in winter.

For shadow anomalies, the beginning time is always at sunrise, and the average power output ranges from 12.1 to 69 watts. The PR ranges from 15% to 50%, while most samples fall into the range of 30% to 45%. Of all the 18 samples, 16 have a CV of PR greater than 1, as mentioned in the previous section. Most durations are 6 - 8 hours, but this attribute varies greatly w.r.t. the block position and panel installation. Recall the related work of the effect of shadow anomalies, and note that the two features of our shadow samples are consistent with the simulation [62] [23].

For dust anomalies, irradiance drops, while the PR of the dirty panels rises when rainfall begins. Usually it takes less than 10 minutes to recover the PR with a medium rainfall. The energy loss percentage due to natural dust varies from 4% to 30% w.r.t. the density and thickness of the dust. Although dust is a sort of direct cover anomaly, the irradiance can penetrate through thin and sparse dust layers. Of all the 18 dust anomaly samples, 17 have a CV of PR less than 1. Since these samples are obtained from our experiments, we controlled the experimental conditions to ensure that our samples are consistent with the effect of naturally accumulated dust anomalies by referencing the previous work [31] [25] [67] [54].

Hence, we obtained a small set of real anomaly samples containing 60 labelled anomalies; these samples fell into three categories representing normal anomalies in Southern Ontario.

5.2.3 Other experiment discoveries

Other results of anomaly effects

- An A4 size paper drops the power output by 20%, which is in accordance with the experimental conclusions of [32]. The effect drops linearly with the increase of papers before it reaches zero.

- Each panel consists of 72 cells: covering half (36 cells) stops power generation; covering 16 cells drops performance by 86% - 88%, 5 cells drops by 75% - 80%. An A4 size paper is equal to about 3 cells in terms of size. We also tried different materials,

such as a piece of package paper, which is much thicker than a A4 paper. However, the effect has negligible differences compared with regular paper.

- Sometimes the real power exceeds the theoretical power (maximum 10%); this phenomenon can be explained by the over-irradiance caused by sudden overcast in high latitude areas [66].

Is the central panel temperature higher?

Intuitively, middle panels could have higher temperatures than the edge panels, because they are surrounded by the edge panels. However, based on the TRCA data, there is no obvious temperature difference between the central panels and the edge panels. In fact, the weighted average differences are greater than -0.5°C and less than 1°C during daytime on 70% of the days in 2012, and the maximum difference is 1.5°C. Figure 5.4 depicts the distribution of the difference between the panels in the center and the panels on the edge of the same string.

5.3 Experimental evaluation of anomaly classification

5.3.1 Data quality

Data quality is essential in our approach, since several datasets are referenced. There are four problems:

availability : The best real-time PR reference is an anomaly-free solar panel in the same model, which is mounted at the same site. However, this is usually unrealistic unless we can guarantee that we always have one anomaly-free and perfectly-functioning panel. Alternatively, the irradiance data, horizontal or in-plane, should be available on site. Since pyranometers are small, many material technologies can be applied to isolate an anomaly, for example, electrostatic anti-dust glass and a heater for melting snow. Otherwise, data can be found at local weather station, but the data precision may diminish based on the distance. The in-plane temperature was often unavailable; however, the ambient temperature

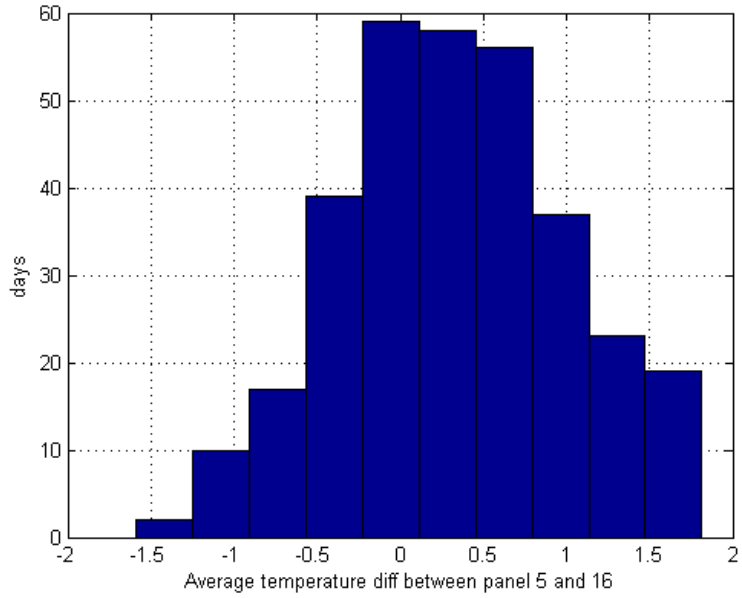


Figure 5.4: Distribution of the temperature difference due to position

collected by a local weather station can also be used. However, this may bring inaccuracy in the efficiency computation, since in-plane temperature may differ largely from in-plane temperature, due to insulation [12] and wind effects.

granularity : Notice that the irradiance data must have fine enough granularity (e.g. 1 minute); otherwise, some anomalies may be missed. For instance, the irradiance data from the UW weather station is recorded at an interval of 15 minutes, while the interval of power output data is 1 minute. Therefore, transient anomalies such as sudden clouds will be missed by the weather station, but will be captured by the power meter, and it is not possible to explain this gap correctly.

alignment : Discontinuous data is very common due to problems of sensors or collection systems. Therefore, aligning the data in each step, based on its time stamp, is necessary. For the missing data, it is better to discard than try to interpolate it.

extreme and nonsense values : Extreme values should also be detected and processed before applying. In this project, we have extreme values, such as -850°C in-plane tempera-

ture, and negative current readings. Notice that the negative irradiance values mean that a surface emits more radiation than it receives, so such values should not be considered broken. However, negative current values usually mean that the current meter is failed. We discarded these extreme and nonsense values to keep the detection accuracy.

5.3.2 Classification

This section is organized as follows: first we introduce the data used to build classifiers; next, we discuss the critical features as our main contributions to the classification; then we list six classification algorithms that are used to classify the anomalies; finally, we state the evaluation results and discussions.

Data

The data set used for classification contains 24 snow samples, which are collected from TRCA data with pictures as evidence, 18 dust samples, and 18 shadow samples. The dust samples and shadow samples are collected from UW data by the experiments described in Section 5.2.

Features

Based on the methodology description and experiment conclusion, we utilized the CV of PR and the average PR of anomalies as the two critical features for classification. Recall that a great fluctuation of PR reflects that the real power output does not track the irradiance, and indirect cover anomalies, such as shadows, are typically not tracking the change of irradiance. Therefore, we used the CV of PR to differentiate the indirect cover from the direct cover anomalies.

Further, based on the datasets, experiment conclusion, and other anomaly effect research [67] [54], we concluded that snow anomalies usually have lower PR compared with dust anomalies. Additionally, season, occurring time, number of affected panels, and duration are some intuitive features; however, they are not necessary reliable in all cases. For

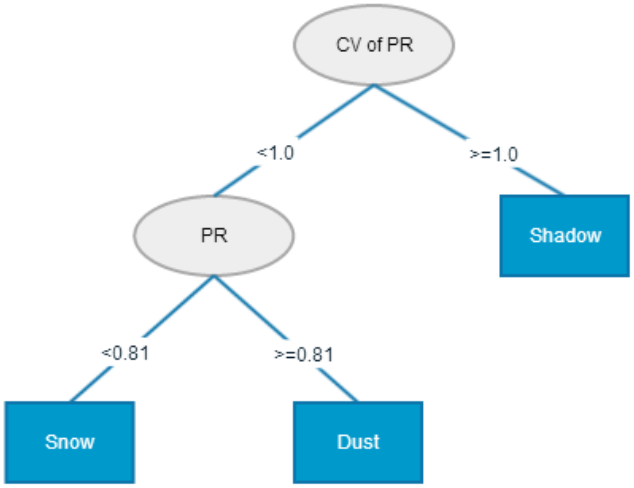


Figure 5.5: Manually constructed decision tree

instance, based on our samples, the time when shadows occur is always between 5:00am and 6:00am; however, this may vary due to different conditions, for instance, if the block structures are in the west, the shadows will occur in the afternoon. Hence, the two critical features are considered to be the most meaningful features, and all features are defined by the collection of the two critical features plus the four additional features. We evaluated the classification using both critical features and all features.

Algorithms

First, we constructed a decision tree manually using the two critical features. The classification decision tree is depicted by Figure 5.5, where the nodes in rectangles are classified anomaly types, and the nodes in ovals are the attributes used for classification, along with the corresponding conditions.

We also ran various classification algorithms on the data for comparison, which are introduced briefly as follows:

Quinlan's C4.5 decision tree [53] algorithm builds decision trees from a set of training data, using the concept of information entropy. Specifically, for each attribute, C4.5 finds the normalized information gain ratio from splitting this attribute. Then, it selects the attribute with highest information gain to create the decision node. Finally, C4.5 recurses on the left attributes and creates children decision nodes. Best-first decision tree (BF Tree) is an extension of the C4.5 algorithm, it uses Gini index [59] as the criteria to select the best splitting node. A Naive Bayes decision tree (NB Tree) generates a decision tree with naive Bayes classifiers at the leaves [36]. Functional tree (FT) is a classification tree that has logistic regression functions at the inner nodes and leaves [48]. SimpleCart implements minimal cost-complexity pruning, which can effectively avoid over-spanning of the leaves [40].

Support vector machine (SVM) is a supervised learning algorithm, which constructs a hyperplane in a high-dimensional space for separation. We used LIBSVM to run SVM evaluation [20].

K-Nearest Neighbors algorithm (kNN) classifies an object by a majority vote of its neighbors, and the object will be assigned to the class that is most common among its k nearest neighbors [10].

Evaluation

We applied ten-fold cross validation for the evaluation. The ten-fold validation means that the data set is divided into ten parts, and ten rounds of tests are conducted. For each round, nine parts are used for training and the remaining part is used for testing; the remaining part is selected alternately for the ten rounds. The final precision rate is the average of the result of the ten rounds.

The following is the definition of evaluation criteria. The TP rate is the true positive rate, which means the accuracy rate of prediction. The FP rate is the false positive rate, which means the classifier predicts an anomaly to be one type, but, actually it is not. The precision is also called positive predictive value (PPV), which indicates the percentage between correctly hit positive cases and all positive predictions. The recall is also called sensitivity, which indicates the percentage between correctly hit positive cases and all real

positive cases. F-Measure is the harmonic mean of precision and recall values, which is computed by equation 5.1:

$$F - Measure = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{5.1}$$

F-Measure is generally used to evaluate a classifier.

Results

Our manually constructed decision tree achieves an 85% accuracy rate, and it identifies 51 out of the total 60 samples correctly. Table 5.1 is the detailed accuracy by class, which lists all the criteria introduced in the previous section. Based on the detailed accuracy, all of the three classifiers are of high evaluation marks.

Class	TP rate	FP Rate	Precision	Recall	F-Measure	ROC Area
SNOW	0.791	0.055	0.905	0.791	0.84	0.770
SHADOW	0.944	0.110	0.773	0.944	0.85	0.888
DUST	0.833	0.071	0.833	0.833	0.833	0.803
Weighted Avg.	0.85	0.079	0.844	0.85	0.843	0.815

Table 5.1: Detailed accuracy by class

Table 5.2 is the confusion matrix of the classification; the confusion matrix indicates the incorrect classifications. According to the confusion matrix, three dust samples are incorrectly identified as snow. Given that the decision tree uses PR as the critical attribute to identify snow anomalies, snow anomalies with comparable PR to dust anomalies will be incorrectly recognized as dust anomalies. There is also some confusion between shadows and two other direct cover anomalies; these can be explained by the trade-off of the CV of PR. It will be improved with more attributes, such as time of occurrence, duration, and so on. The classification overall is simple since there are clear characteristics among these anomalies.

We also ran the WEKA data mining toolkit [9] to compare our classifier with more classifiers. The accuracy of tested classifiers is listed in Table 5.3:

a	b	c	<- classified as
19	2	3	a = snow
1	17	0	b = shadow
3	0	15	c = dust

Table 5.2: Confusion matrix

Classifier	Accuracy
C4.5	0.96
BF Tree	0.96
NB Tree	0.96
FT	0.91
Simple Cart	0.91
SVM (Linear)	0.90
SVM (deg-4 polynomial)	0.90
kNN (k = 1)	0.88
kNN (k = 3)	0.86
kNN (k = 5)	0.88

Table 5.3: Accuracy of more classifiers

The C4.5 decision tree achieves a precision rate of 96%, and the rules are depicted in Figure 5.6:

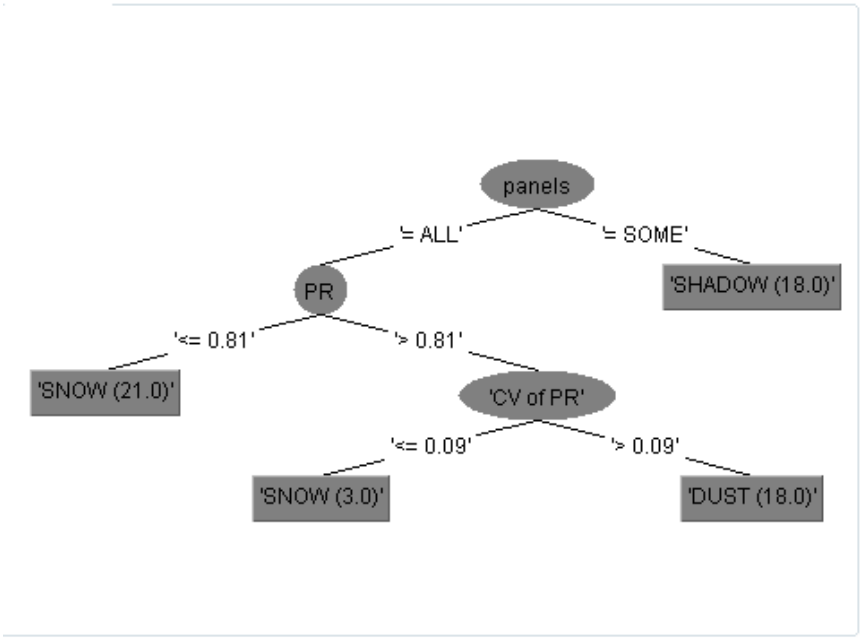


Figure 5.6: Classification tree of C4.5

Table 5.4 shows the detailed accuracy of C4.5 decision tree classifier. Notice that the shadow classifier is a perfect classifier because its F-Measure is 1. This can be explained by the limited representativeness of samples, because in many practical cases, shadow anomalies can also cover the entire strings. Because the shadow classifier is perfect, so the snow classifier, works as a childnode in the decision tree, is able to correctly identify all the snow anomalies.

Class	TP rate	FP Rate	Precision	Recall	F-Measure	ROC Area
SNOW	1	0.056	0.923	1	0.96	0.972
SHADOW	1	0	1	1	1	1
DUST	0.889	0	1	0.899	0.941	0.944
Weighted Avg.	0.961	0.022	0.963	0.961	0.96	0.972

Table 5.4: Detailed accuracy by class (C4.5 decision tree)

Table 5.5 is the confusion matrix of this classification. According to the confusion ma-

trix, two dust samples are incorrectly identified as snow. Given that the decision tree uses PR as the critical attribute to identify snow anomalies, snow anomalies with comparable PR to dust anomalies will be incorrectly recognized as dust anomalies.

a	b	c	<- classified as
24	0	0	a = snow
0	18	0	b = shadow
2	0	16	c = dust

Table 5.5: Confusion matrix of C4.5

The BF tree generates exactly the same rules, has the same accuracy result, and the same confusion matrix as the C4.5 decision tree, so we will not repeat the result and discussion.

The NB tree also generates the same accuracy result and the the same confusion matrix as the C4.5 decision tree. However, it generates two leaves, which are two naive Bayes models, and the corresponding rules of the NB tree are:

```
PR <= 0.815: NB 1
PR > 0.815: NB 2
```

Attributes	Panels		Time		Duration		PR		CV of PR		Total
	All	Some	<=7:30am	>7:30am	<=7.5hrs	>7.5hrs	<=0.53	>0.53	<=0.754	>0.754	
SNOW(0.52)	22	1	1	22	22	1	11	12	20	3	23
SHADOWS(0.45)	1	19	18	2	3	17	19	1	1	19	20
DUST(0.02)	1	1	1	1	1	1	1	1	1	1	2

Table 5.6: Naive Bayes model: NB 1

The two NB models, NB1 and NB2, are depicted in Table 5.6 and Table 5.7, respectively. The numbers in the cells are the number of corresponding samples, and the NB models then compute the conditional probability to predict the class.

Based on the rules, the classification first splits on PR = 0.815, then uses the corresponding naive Bayes model to further predict the types of anomalies. Notice that the

Attributes	Panels		CV of PR		Total
	All	Some	<=0.09	>0.09	
SNOW(0.17)	4	1	4	1	23
SHADOWS(0.04)	1	1	1	1	2
DUST(0.79)	19	1	1	19	20

Table 5.7: Naive Bayes model: NB 2

first leaf mainly processes the snow and shadow anomalies, while the second one mainly processes the dust anomalies.

The FT achieves an accuracy rate of 93%, and generates only one node. Correspondingly, it has a rule to classify the three types of anomalies. The three regression functions are depicted by Table 5.8:

Class	Function
SNOW	16.72 + [time] * -0.77 +[duration] * -0.02 +[PR] * -5.24 +[CV of PR] * -7.87 +[season=WINTER] * 3.55
SHADOW	-10.5 + [panels] * 22.19
DUST	-58.55 + [time] * 0.16 +[duration] * 0.01 +[PR] * 63.82 +[CV of PR] * 7.9

Table 5.8: Functions of FT

For each tested sample, every function is used to compute a result. This sample will be labelled as that class, if the function of that class generates the highest value for this sample. Since FT is based on logistic regression, it takes all the attributes as the same and is difficult to interpret. Hence, the incorrect classification covers all the three types of anomalies in the confusion matrix Table 5.9:

a	b	c	<- classified as
23	1	0	a = snow
0	17	1	b = shadow
2	0	16	c = dust

Table 5.9: Confusion matrix of FT

The SimpleCart achieves an accuracy of 91%, and because this algorithm is designed for generating compact trees, the generated rules are also simple:

```
panels=(SOME): SHADOW(18.0/0.0)
panels!=(SOME)
|  PR < 0.815: SNOW(21.0/0.0)
|  PR >= 0.815: DUST(18.0/3.0)
```

Notice that except for the number of panels, the rules only use PR as the predicate condition, so some of the snow and dust anomalies are incorrectly classified. For instance, a very thin layer of snow is incorrectly classified as dust. Table 5.10 depicts the confusion matrix for SimpleCart:

a	b	c	<- classified as
21	0	3	a = snow
0	18	0	b = shadow
2	0	16	c = dust

Table 5.10: Confusion matrix for SimpleCart

We tested two types of cores for SVM, linear and 3 degree polynomial, and tuned them to achieve its maximum accuaracy, respectively. We first determined the cost parameter, C, to be 1.0 by some iterations of test. For the polynomial core, the values of *gamma* and *degree* also need to be tuned. These two approaches achieve comparable results as the C4.5 decision tree; however, the efficiency of SVM, especially the one with the polynomial core, is much lower than C4.5.

Table 5.11 depicts the confusion matrix for SVM (linear):

a	b	c	<- classified as
24	0	0	a = snow
0	17	1	b = shadow
5	0	13	c = dust

Table 5.11: Confusion matrix for SVM (linear)

Table 5.12 depicts the confusion matrix for SVM (deg-3 polynomial):

a	b	c	<- classified as
22	0	2	a = snow
0	17	1	b = shadow
2	1	15	c = dust

Table 5.12: Confusion matrix for SVM (deg-3 polynomial)

Based on the confusion matrix, linear core SVM correctly identified all the snow samples; however, it made wrong predictions on shadow samples. The polynomial core did not achieve a better accuracy, which means this classification does not need a highly twisted hyperplane. However, based on the confusion matrix, we can see that the hyperplane cuts the space more evenly than the linear core, as the incorrect classification appears in every class.

The kNN achieves an accuracy rate of 88%, when k equals to 1 or 3, respectively. When k equals to 1 or 3, the algorithm generates the same confusion matrix. Table 5.13 depicts the confusion matrix for kNN (k = 1):

a	b	c	<- classified as
22	0	2	a = snow
0	18	0	b = shadow
5	0	13	c = dust

Table 5.13: Confusion matrix for kNN (k = 1)

As kNN uses all the attributes for classification, this result demonstrates that the selected features support the classifiers well. Again, the incorrect classification is concentrated in snow anomalies and dust anomalies.

We notice that all the tree based classifiers utilize the attribute of the number of panels; however, this attribute has not been proved to be reliable, due to some limitation of our experiments. For instance, the size of the data set is small.

Similarly, all the collected shadow samples occur between 5:00am and 6:00 am; however, in practice, shadows could occur in the morning or in the afternoon. Therefore, we reran

the evaluation with only two critical attributes: the average PR and the CV of PR. The result is depicted by Table 5.14 :

Classifier	Accuracy
C4.5	0.92
BF Tree	0.92
NB Tree	0.90
FT	0.86
Simple Cart	0.86
SVM (Linear)	0.86
SVM (deg-4 polynomial)	0.90
kNN (k = 1)	0.88
kNN (k = 3)	0.86
kNN (k = 5)	0.85

Table 5.14: Accuracy of more classifiers (using only two features)

The C4.5 decision tree still achieves a 92% accuracy rate, however, it is over-fit based on the generated rules shown in Figure 5.7:

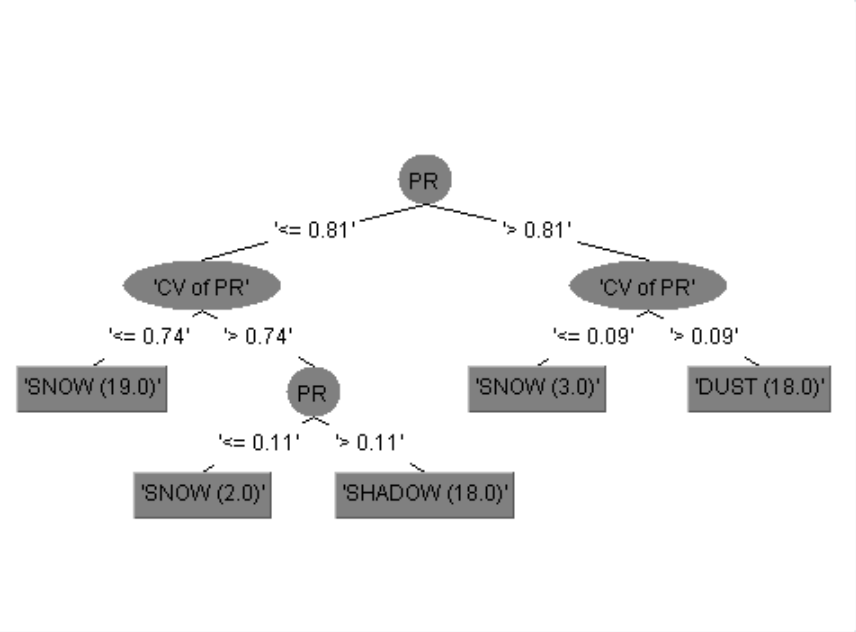


Figure 5.7: Classification tree of C4.5 (using only two features)

It is obvious that the rules reuse these two attributes as prediction conditions and use very fine values to split them several times. Therefore, this classifier is not able to fit more general cases, so we will not list its confusion matrix here.

The BF tree has the same result and also generates over-fitting rules:

```
PR < 0.815
|  CV of PR < 0.745: SNOW(19.0/0.0)
|  CV of PR >= 0.745
|  |  PR < 0.13: SNOW(2.0/0.0)
|  |  PR >= 0.13: SHADOW(18.0/0.0)
PR >= 0.815
|  CV of PR < 0.09: SNOW(3.0/0.0)
|  CV of PR >= 0.09: DUST(18.0/0.0)
```

Notice that the generated decision trees by the two different algorithms, C4.5 and BF tree, have the same structure, and the only differences are the condition values. This is because

they use different criteria to select the split value: C4.5 uses information gain, while the BF tree uses the Gini index. Correspondingly, the confusion matrices of the two classification are different; although their accuracy rates are of the same value.

The NB tree achieves a 90% accuracy rate and generates three naive Bayes models as the three leaves, which are depicted by Table 5.15, Table 5.16, and Table 5.17, respectively:

```
PR <= 0.815
|   CV of PR <= 0.745: NB 2
|   CV of PR > 0.745: NB 3
PR > 0.815: NB 4
```

Attributes	PR	CV of PR	Total
	All	All	
SNOW(0.91)	20	20	5
SHADOWS(0.05)	1	1	2
DUST(0.79)	1	1	20

Table 5.15: NB Model 2 (using only two features)

Attributes	PR		CV of PR	Total
	<=0.13	>0.13	All	
SNOW(0.13)	3	1	3	4
SHADOWS(0.83)	1	19	19	20
DUST(0.04)	1	1	1	20

Table 5.16: NB Model 3 (using only two features)

Attributes	PR	CV of PR		Total
	<=0.13	<=0.09	>0.09	
SNOW(0.17)	4	4	1	5
SHADOWS(0.04)	1	1	1	2
DUST(0.79)	19	1	19	20

Table 5.17: NB Model 4 (using only two features)

Since naive Bayes does not select important attributes, it actually benefits from being forced to use the two critical features. We observe that the decision tree achieves a more reasonable structure, in which each leaf, representing a Bayes model, a particular type of anomaly is processed. Again, because it builds trees simply based on the conditional probability, the incorrect classification covers all the types. Table 5.18 depicts the confusion matrix:

a	b	c	<- classified as
21	3	0	a = snow
1	17	0	b = shadow
2	0	16	c = dust

Table 5.18: Confusion matrix of NB Tree (using only two features)

The FT achieves an 86% accuracy rate, again, it generates a tree that contains only one node; however, the functions become simpler. Table 5.19 depicts the functions:

Class	Function
SNOW	5.91 + [PR] * -4.26 +[CV of PR] * -4.07
SHADOW	-2.67 + [PR] * 2.18 +[CV of PR] * 1.65
DUST	-26.43 + [PR] * 31.34 +[CV of PR] * 4.1

Table 5.19: Functions of FT (using only two features)

Similar to the NB tree, the incorrect classification of FT covers all the types in the confusion matrix 5.20:

a	b	c	<- classified as
22	4	0	a = snow
4	14	0	b = shadow
1	1	16	c = dust

Table 5.20: Confusion matrix of FT (using only two features)

The SimpleCart achieves an 86% accuracy rate, and a compact decision tree:

```
PR < 0.815
|  CV of PR < 0.745: SNOW(19.0/0.0)
|  CV of PR >= 0.745: SHADOW(18.0/2.0)
PR >= 0.815: DUST(18.0/3.0)
```

Table 5.21 is the confusion matrix:

a	b	c	<- classified as
18	3	3	a = snow
1	17	0	b = shadow
1	0	17	c = dust

Table 5.21: Confusion matrix of SimpleCart (using only two features)

Based on the confusion matrix and the generated rules, we conclude that when it uses PR to first identify the dust, some thin layers of snow are incorrectly classified; for some snow anomalies, the CV of PR is high enough to be incorrectly classified as shadows.

The kNN achieves a decreasing accuracy with the increasing of k. However, the 1NN maintains its accuracy, which demonstrates that the two critical features are significant enough to differentiate the three classes.

With only the two critical features, SVM with linear core has a lower accuracy. However, the SVM with polynomial core maintains the accuracy, and the hyperplane is changed, based on its confusion matrix 5.22:

a	b	c	<- classified as
22	2	0	a = snow
3	15	0	b = shadow
1	0	17	c = dust

Table 5.22: Confusion matrix of SVM (deg-3 polynomial, using only two features)

When the classifiers are forced to use only the two critical features, all of tree based algorithms try to use PR to first identify the dust. This may not work well for other more general datasets, because compared with the PR range of different anomalies, the boundary between direct cover anomalies and indirect cover anomalies is more reliable. We believe that with larger sized and more representative samples, the generated rules will be more similar to our manually constructed decision tree.

In terms of the classification results, since the generated trees for C4.5 and for the BF tree are over-fitting, their pruned trees will have structures similar to SimpleCart, and have similar accuracy rate as well. The NB tree uses three naive Bayes models to classify the three types of anomalies, respectively, and it achieves a high accuracy rate. The only drawback is that the important features should be evaluated before applying this algorithm. The FT achieves the same accuracy rate as SimpleCart; however, the functions of FT are difficult to interpret. The SimpleCart algorithm generates the most close structure compared with our decision tree, with a very close accuracy rate. KNN (k = 1) achieves a medium accuracy and very high efficiency. Ignoring the over-fitting decision trees, the SVM with polynomial core achieves the highest accuracy. Therefore, given the data we have, SVM is the most reliable classifer, its drawbacks are the low efficiency and requirements of manually tuned parameters.

Above all, all the classifiers, regardless of whether they are simple or complicated, information gain based or regression based, work well. This is because we developed sufficient and reasonable features that can capture the physical essence. Therefore, these features are able to leverage many different classifiers to achieve high accuracy. In conclusion, the actionable anomaly classification, based on the data we have, is demonstrated to be a simple classification problem, given the appropriate features developed by us.

Chapter 6

Conclusion

We developed a comprehensible data driven approach to detect and identify the types of anomalies on solar panels, based on power output and weather data. In our dataset, some anomalies, unfortunately, are difficult to observe, especially performance degradation due to panel failures.

The following conclusion can be made, regarding our anomaly detection and classification framework:

1. We used comprehensible methodology to detect anomalies whose performance ratio (PR) drops below a threshold.

2. To obtain the PR, we need the in-plane irradiance data and a PR model to compute the theoretical power output, based on in-plane irradiance. In many cases, there are no pyranometers beside the panels, so the in-plane irradiance cannot be obtained directly. Hence, we developed a simple model to compute the in-plane irradiance based on the horizontal irradiance data, which usually is available on the website of the local weather station. We verified this simple model by comparing the result with the result of the conventional model, which not only requires more complicated computing, but also depends on more input.

3. With the support of in-plane irradiance, the theoretical power output can be derived using the PR model and anomalies can be detected by comparing the PR. Our methodology of building the in-plane irradiance model could be used in a variety of locations. For example, local horizontal irradiance can be obtained by modifying equation 2.6 to fit local irradiance data; in-plane irradiance on inclined surfaces can be obtained with our simple model. By adjusting the efficiency equation based on the solar panel attributes, it can be used for general theoretical power output predictions as well.

4. We disaggregated the anomalies into hierarchical categories. Moreover, we used CV as the essential feature to differentiate the two major types of actionable anomalies: direct cover and indirect cover. Multiple classifiers have been tested and our decision tree can precisely classify the anomalies with real data.

This framework, including detection, profiling, and classification methods, can be referenced based on local conditions. Abundant and representative data is beneficial to the precision of the classification.

6.1 Limitations

There are several limitations of this framework:

- At least one type of irradiance data is required. Moreover, the effectiveness of the framework is highly determined by the data quality. The more accurate the data is, the more precise the classification will be. Generally, using on-site irradiance data achieves more precise results than using weather station data.

- The granularity of the critical data, such as irradiance and power output, should be the same and sufficiently fine, because if one of them has a coarse granularity, anomalies during the intervals will be missed.

- There are some types of anomalies that cannot be classified due to the limited scope of our data, such as panel failures, leaves.

6.2 Future Work

6.2.1 Data collection

A major limitation of this project is lack of enough supporting data, especially evidence, which is necessary for supervised learning. To address this problem, we plan to deploy a camera on the EV3 building to collect pictures of anomalies. The other limitation is that the data collection system is not reliable, because the sensors are powered by batteries. Since sensors use wireless protocol to send data, the batteries need to be replaced every two weeks. Hence, we plan to replace the battery power supply with stable AC power.

Sufficient data is the foundation of all future work. For instance, with reliable data access, we can build a real-time anomaly detection and classification demo platform.

6.2.2 Streaming database

The UW data is currently stored in a format of plain text, and the TRCA data is stored in a conventional relational database. A streaming database can be used to enhance the data quality, such as data integrity and standardization. For instance, the streaming database is able to reject bad data insert operations caused by sensor or system problems. Moreover, for large scale deployment, the streaming database is able to improve the responsiveness and provide real-time identification services by keeping aggregation results in memory.

6.2.3 Capability of additional features and additional types of anomalies

We plan to add more features with the support of more data, for instance, anomaly accumulation time is defined by how long an anomaly takes to be detected. Intuitively, the accumulation time of dust is very long, while it is much shorter for shadows, which is usually less than one hour.

With sufficient real data, we can take more anomaly categories into scope, such as panel failures, leaves, and so on. We can also disaggregate current objective anomalies further.

Take failure as an example: there are many reasons accounting for a panel's failure, such as the degradation of PV cells, broken inverters, or disconnection between modules [3].

Additionally, a methodology for identifying anomaly combinations should be taken into scope. In this project we only collected single anomaly data, but, in practice, it is possible that several types of anomalies may occur simultaneously. The first step would be to identify the dominant anomaly, which needs to exclude the "noise" caused by other anomalies. The next step is to identify all the anomalies, if necessary. This problem can also be addressed by the approach introduced in the next section, the association rules.

6.2.4 Association rules

We have done some experiments to discover association rules, which occur when one panel has an anomaly on it, it is assumed that, after a certain time, there must be similar anomalies occurring on its neighbouring panels. Such rules are useful for capturing many features, such as when an affected area is enlarging, shifting, or stable. For instance, based on the TRCA data, we found that on the sunny days after snow, panels on the top of the string were the first to recover, and panels below it recovered sequentially (positive slope of the PR curve). Corresponding to the curves, we verified this phenomenon with pictures: snow began to melt on the top panels, then the current helped the panels below to melt.

An intuitive proposal is to use association rules to probe the change of affected areas: both in size and in direction. For instance, shadows caused by structures will enlarge in size and follow the same direction every day, while a failed panel will probably only affect a fixed area.

This approach first needs to find primitive rules (rule 1):

```
"If panel (A) has an anomaly on it
 Then its neighbour panel (I) will have an anomaly in T time"
```

Based on rule 1, the affected area's size (Enum SizeChange: Stable, Enlarging, Shrinking) and direction of movement (Enum Direction: Upward, Downward, Side) values can be probed, so we have the association rules in the format of :

```
"If a SizeChange with a Direction and Speed occur
 Then in a confidence (C) that a particular type of anomaly is running"
```

Other types of association rules could be:

- Association between external factors and anomalies, for instance, a long period of dry days with little wind will lead to a dust anomaly.

  ```
  "If a certain weather condition lasts for time (T)
   Then in a confidence (C) that a particular type of anomaly will occur"
  ```

- Association between different types of anomalies, for instance, a long period of coverage on single panel may lead to an entire string failure with certain solar system designs. [47]

  ```
  "If a type of anomaly lasts for time (T) and affects area (A)
   Then in a confidence (C) another type of anomaly will occur"
  ```

With more experimental data in the future, we plan to verify the existing rules and test the proposed rules.

6.2.5 Computer vision

Computer vision is able to acquire, process, and analyze images. Many researchers have proposed using this technology to monitor sustainable energy systems, such as wind turbines and solar panels [27]. We plan to utilize this technology to recognize the objects on solar panels from field images, or cross-verify the anomalies detected by our approaches in this project. By combining object recognition and weather forecasting, we can also develop a precise solar energy output prediction to optimize the power generation structure.

References

[1] 7 impressive solar energy facts. http://www.abb-conversations.com/2013/12/7-impressive-solar-energy-facts-charts/ as of 2014-05-05.

[2] Archive and charting of power data of solar panels on EV3, University of Waterloo. http://blizzard.cs.uwaterloo.ca/~hbo/solar.html

[3] Degradation and Failure Modes. http://www.pveducation.org/pvcdrom/modules/degradation-and-failure-modes as of 2014-05-05

[4] Global Solar PV Installations Increase to 31 GW in 2012. http://www.azom.com/news.aspx?newsID=36010 as of 2014-05-05.

[5] Manual:LDK-220D-225D-230D-235D-240D-245D-250D-255D-260D-20Monocrystalline Solar Module, http://www:osmsolar:com/LDKspec220:pdf

[6] Meteorology data from weather station, University of Waterloo. http://weather.uwaterloo.ca/data.html

[7] "Photovoltaic Cell Conversion Efficiency". U.S. Department of Energy. Retrieved 19 May 2012.

[8] "Small Photovoltaic Arrays". Research Institute for Sustainable Energy (RISE), Murdoch University. Retrieved 5 February 2010.

[9] Weka 3: Data Mining Software in Java. http://www.cs.waikato.ac.nz/ml/weka/ as of 2014-05-05.

[10] Aha, David W., Dennis Kibler, and Marc K. Albert. "Instance-based learning algorithms." Machine learning 6.1 (1991): 37-66.

[11] Andrews, Rob W., and Joshua M. Pearce. "Prediction of energy effects on photovoltaic systems due to snowfall events." Photovoltaic Specialists Conference (PVSC), 2012 38th IEEE. IEEE, 2012.

[12] Andrews, Rob W., Andrew Pollard, and Joshua M. Pearce. "The effects of snowfall on solar photovoltaic performance." Solar Energy 92 (2013): 84-97.

[13] Anderson, John D., and Michael Martin Nieto. "Astrometric solar-system anomalies." Proceedings of the International Astronomical Union 5.S261 (2009): 189-197.

[14] Al-Rawahi, N. Z., Y. H. Zurigat, and N. A. Al-Azri. "Prediction of hourly solar radiation on horizontal and inclined surfaces for Muscat/Oman." The Journal of Engineering Research 8.2 (2011): 19-31.

[15] Arthur, David, and Sergei Vassilvitskii. "k-means++: The advantages of careful seeding." Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics, 2007.

[16] Becker, Gerd, et al. "An approach to the impact of snow on the yield of grid-connected PV systems." Proc. European PVSEC (2006).

[17] Brench, Bronwyn L. "Snow-covering effects on the power output of solar photovoltaic arrays." NASA STI/Recon Technical Report N 81 (1979): 11551.

[18] Broverman, Samuel A. Actex study manual. Winsted, CT: Actex Publications .Page 104. 2001.

[19] Buday, Michael S. "Measuring irradiance, temperature and angle of incidence effects on photovoltaic modules in Auburn Hills." Michigan. Diss. University of Michigan, 2011.

[20] Chang, Chih-Chung, and C.-J. Lin. "LIBSVM : a library for support vector machines." ACM Transactions on Intelligent Systems and Technology, 2:27:1–27:27, 2011.

[21] Colli, Alessandra, and Willem J. Zaaiman. "Maximum-Power-Based PV Performance Validation Method: Application to Single-Axis Tracking and Fixed-Tilt c-Si Systems in the Italian Alpine Region." Photovoltaics, IEEE Journal of 2.4 (2012): 555-563.

[22] Curl JR, Herbert, John T. Hardy, and Ronald Ellermeier. "Spectral absorption of solar radiation in alpine snowfields." Ecology (1972): 1189-1194.

[23] Deline, Chris, et al. "A simplified model of uniform shading in large photovoltaic arrays." Solar Energy 96 (2013): 274-282.

[24] Dolara, Alberto, et al. "Performance analysis of a single-axis tracking PV system." Photovoltaics, IEEE Journal of 2.4 (2012): 524-531.

[25] El-Shobokshy, Mohammad S., and Fahmy M. Hussein. "Degradation of photovoltaic cell performance due to dust deposition on to its surface." Renewable Energy 3.6 (1993): 585-590.

[26] Elminir, Hamdy K., et al. "Effect of dust on the transparent cover of solar collectors." Energy conversion and management 47.18 (2006): 3192-3203.

[27] Erol-Kantarci, Melike, and Hussein T. Mouftah. "Wireless multimedia sensor and actor networks for the next generation power grid." Ad Hoc Networks 9.4 (2011): 542-551.

[28] Garg, H. P. "Effect of dirt on transparent covers in flat-plate solar energy collectors." Solar Energy 15.4 (1974): 299-302.

[29] Gauthier, Michael K., Emmett L. Miller, and Alex Shumka. "Solar cell anomaly detection method and apparatus." U.S. Patent No. 4,301,409. 17 Nov. 1981.

[30] Giddings, J. C., and E. LaChapelle. "Diffusion theory applied to radiant energy distribution and albedo of snow." Journal of geophysical research 66.1 (1961): 181-189.

[31] Goossens, Dirk, and Emmanuel Van Kerschaever. "Aeolian dust deposition on photovoltaic solar cells: the effects of wind velocity and airborne dust concentration on cell performance." Solar Energy 66.4 (1999): 277-289.

[32] Hu, Bo. "Solar Panel Anomaly Detection and Classification." (2012).

[33] Kamm, Raimund, "Germany reaches 23.4 GW of electricity from PV on June 6th", http://www.solarserver.com/solar-magazine/solar-news/current/2013/kw24/germany-reaches-234-gw-of-electricity-from-pv-on-june-6th.html as of 2014-05-05.

[34] Kawano, Ryuji, and Masayoshi Watanabe. "Anomaly of charge transport of an iodide/tri-iodide redox couple in an ionic liquid and its importance in dye-sensitized solar cells." Chemical communications 16 (2005): 2107-2109.

[35] Khoo, Yong Sheng, et al. "Optimal Orientation and Tilt Angle for Maximizing in-Plane Solar Irradiation for PV Applications in Singapore." 1-7.

[36] Kohavi R. "Scaling up the accuracy of naive-Bayes classifiers: A decision-tree hybrid" KDD. 1996: 202-207.

[37] Krotkov, N. A., et al. "Satellite estimation of spectral surface UV irradiance in the presence of tropospheric aerosols: 1. Cloudfree case." Journal of Geophysical Research: Atmospheres (19842012) 103.D8 (1998): 8779-8793.

[38] Kuitche, Joseph M., Rong Pan, and G. TamizhMani. "Investigation of Dominant Failure Mode (s) for Field-Aged Crystalline Silicon PV Modules Under Desert Climatic Conditions." (2014): 1-13.

[39] Kumar, Ankit, Srinivas Sista, and Yang Yang. "Dipole induced anomalous S-shape IV curves in polymer solar cells." Journal of Applied Physics 105.9 (2009): 094512.

[40] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, Charles J. Stone. "Classification and Regression Trees. Wadsworth International Group, Belmont, California. 1984.

[41] Luque, Antonio, and Steven Hegedus, Eds., Chapter 20. Handbook of photovoltaic science and engineering. John Wiley & Sons, 2011.

[42] Meinel, Aden B., and Marjorie Pettit Meinel. Page 132-134, "Applied solar energy." Addison-Wesley, Reading, MA(1976).

[43] Meydbray, Jenya, et al. "Pyranometers and Reference Cells: Part 2: What Makes the Most Sense for PV Power Plants?." (2012).

[44] Mohamed, Ali Omar, and Abdulazez Hasan. "Effect of Dust Ac-cumulation on Performance of Photovoltaic Solar Modules in Sahara Environment." J Basic Appl Sci Res 2.11 (2012): 11030-11036.

[45] Moon, M., "Google Studies How Diet Affects Solar Panel Efficiency". PC Magazine: Good Clean Tech. www:goodcleantech.com/2009/08/googlestudieshowdirtaffect.php as of 2014-05-05.

[46] Mueller, R. W., et al. "Rethinking satellite-based solar irradiance modelling: The SOLIS clear-sky module." Remote sensing of Environment 91.2 (2004): 160-174.

[47] Muenster, Ralf J. "Shade Happens". National Semiconductor. February 02, 2009. http://www.renewableenergyworld.com/rea/news/article/2009/02/shade-happens-54551 as of 2014-05-05.

[48] Landwehr, Niels, and Mark Hall, Eibe Frank (2005). Logistic Model Trees.

[49] O'Neill, A. D. J., and Don M. Gray. "Solar radiation penetration through snow." The Role of Snow and Ice in Hydrology, Proceedings of the Banff Symposium, International Association of Hydrological Sciences, Publ. No. 107. 1972.

[50] Oppenheim, Alan V., Alan S. Willsky, and Syed Hamid Nawab. Signals and systems. Vol. 2. Englewood Cliffs, NJ: Prentice-Hall, (1983):314-332

[51] Pfister, R., and M. Schneebeli. "Snow accumulation on boards of different sizes and shapes." Hydrological processes 13.1415 (1999): 2345-2355.

[52] Pinker, R. T., and I. Laszlo. "Modeling surface solar irradiance for satellite applications on a global scale." Journal of Applied Meteorology 31.2 (1992): 194-211.

[53] Quinlan J R. "C4. 5: programs for machine learning". Morgan kaufmann, 1993.

[54] Rahman, Mizanur, et al. "Effects of Natural Dust on the Performance of PV Panels in Bangladesh." International Journal of Modern Education & Computer Science 4.10 (2012).

[55] Ramaprabha, R., and B. L. Mathur. "A comprehensive review and analysis of solar photovoltaic array configurations under partial shaded conditions." International Journal of Photoenergy 2012 (2012).

[56] Ristau, Oliver. "Fukushima fallout could create 100 GW of newly installed PV". http://www.pv-magazine.com/news/details/beitrag/fukushima-fallout-could-create-100-gw-of-newly-installed-pv_100002499 as of 2014-05-05.

[57] Rothwarf, Allen. "The CdS/Cu_2S solar cell: Basic operation and anomalous effects." Solar Cells 2.2 (1980): 115-140.

[58] Salim, A. A., F. S. Huraib, and N. N. Eugenio. "PV power-study of system options and optimization." EC photovoltaic solar conference. 8. 1988.

[59] Shi, Haijian. "Best-first decision tree learning. "Diss. The University of Waikato, 2007.

[60] Skoplaki, E., and J. A. Palyvos. "On the temperature dependence of photovoltaic module electrical performance: A review of efficiency/power correlations." Solar energy 83.5 (2009): 614-624.

[61] Sulaiman, Shaharin A., et al. "Effects of Dust on the Performance of PV Panels." World Academy of Science, Engineering and Technology 58 (2011): 588-593.

[62] Taha, Mohammed Qasim, and Salih Mohammed Salih. "Performance Analysis of Photovoltaic Modules under Shading Effect." (2012).

[63] Veldhuis, A. J., et al. "Real-Time Irradiance Simulation for PV Products and Building Integrated PV in a Virtual Reality Environment." Photovoltaics, IEEE Journal of 2.3 (2012): 352-358.

[64] Wakim, F. "Introduction of PV power generation to Kuwait." Kuwait Institute for Scientific Researchers, Kuwait City (1981).

[65] Yang, Jin-huan, Jia-jun Mao, and Zhong-hua Chen. "Calculation of solar radiation on variously oriented tilted surface and optimum tilt angle." Journal-Shanghai Jiaotong University-Chinese Edition- 36.7 (2002): 1032-1036.

[66] Yordanov, Georgi Hristov, et al. "Overirradiance (cloud enhancement) events at high latitudes." Photovoltaics, IEEE Journal of 3.1 (2013): 271-277.

[67] Zorrilla-Zorrilla-Casanova, J., et al. "Analysis of dust losses in photovoltaic modules." World Renewable Energy CongressSweden. 2011.