# What's Wrong with my Solar Panels: a Data-Driven Approach

Xiang Gao, Lukasz Golab and S. Keshav
University of Waterloo
200 University Avenue West
Waterloo, Ontario, Canada N2L 3G1
{x39gao,lgolab,keshav}@uwaterloo.ca

## ABSTRACT

Solar panels have been improving in efficiency and dropping in price, and are therefore becoming more common and economically viable. However, the performance of solar panels depends not only on the weather, but also on other external factors such as shadow, dirt, dust, etc. In this paper, we describe a simple and practical data-driven method for classifying anomalies in the power output of solar panels. In particular, we propose and experimentally verify (using two solar panel arrays in Ontario, Canada) a simple classification rule based on physical properties of solar radiation that can distinguish between shadows and direct covering of the panel, e.g,. by dirt or snow.

## 1. INTRODUCTION

Photovoltaic (PV) technology, i.e., solar panels, has been rapidly dropping in price and increasing in popularity worldwide [7]. The monitoring and measuring capability of PV installations has also improved. While it used to be possible only to measure the total power output of an array of solar panels, micro-inverters (which are devices that covert Direct Current generated by an *individual* panel into Alternating Current) now make it possible to measure the power output of each individual panel at fine granularities (e.g., every minute or every five minutes). Thus, solar panel data analytics is becoming an important area of research and practice.

The power output of a PV system depends on solar intensity and the panels' efficiency of converting light into power (typically 15-20 percent). Additionally, even a perfectly-functioning panel on a sunny day will produce little power if it is shaded or covered by dust or dirt. For instance, many large-scale PV installations are located on farmlands and/or near country roads, which makes them vulnerable to dust, mud, pollen and other types of soiling. Furthermore, even if a farm site is chosen to be shadow-free, grass may eventually grow tall enough to cast shadows on the panels. Numerous studies have observed power drops of 40 or more percent due to shaded, dirty and snow-covered panels [1, 2, 3, 4, 6, 8, 11, 15, 18, 23, 25].

A simple solution is to frequently clean the panels. However, this is not feasible in desert locations that suffer from water shortages, or in remote large-scale installations where an automated sprinkler system is prohibitively expensive. Some PV installations include cameras that monitor the panels, but it may be difficult to tell from videos or still images whether the panels are dirty (see, e.g., Figure 4 in Section 5). Thus, in practice, PV systems often operate in less than ideal conditions.

The problem we address in this paper is how to determine, in a data-driven fashion, what is wrong with a solar panel, on a per-panel rather than per-array basis. Since most large-scale PV systems are equipped with sensors that measure solar intensity and power output at regular intervals, we propose a simple classification approach to explain *anomalies* (i.e., drops) in the produced power based on these time series. This is a challenging problem because it is not obvious how to distinguish between different types of anomalies, and therefore it is not obvious which *features* of the data to use for classification.

We take a first step towards data-driven classification of anomalies in PV power output based on fine-grained per-panel data. Our solution exploits the physical properties of solar radiation. We observe that obstructions which do not touch the panels, such as shading, affect the power output in a subtly different way than dirt or snow lying on the panels. Based on this observation, we derive simple features from the power output time series that distinguish between shadows and soiling. We tested the proposed idea using data obtained from two real PV installations in the province of Ontario, Canada, and obtained 85 percent accuracy.

An obvious limitation of the proposed solution is that it can only tell shadows apart from direct cover, but it cannot distinguish between different types of direct cover (such as dust, dirt, or leaves) or between direct cover and physical panel malfunctions. Nevertheless, this simple classification can already be helpful to PV owners as it can suggest when the panels are due for a cleaning and when unexpected shadows arise. Our preliminary results are promising, and we hope that this paper encourages further research in solar panel data mining.

The remainder of this paper is organized as follows. Section 2 presents the necessary background in solar panel monitoring and defines our problem; Section 3 discusses related work; Section 4 presents our solution; Section 5 describes our experimental results; and Section 6 concludes the paper with directions for future work.



.

## 2. PRELIMINARIES AND PROBLEM STATEMENT

We begin with a simple example of the factors affecting the power output of a solar panel with the help of Figure 1. The curve labeled "1" corresponds to the maximum solar intensity times the surface area of the panel throughout a hypothetical day, on which the sun rises at 6:00 and sets at 20:00. If the sun were shining all day, there were no clouds, and the panel was able to convert 100 percent of the solar radiation into power, curve 1 would be the maximum power output throughout the day. Chapter 20 of [14] describes how to estimate the maximum clear-sky solar intensity given the time of day, day of year, latitude and tilt angle of the panel, all of which determine the relative position of the panel with respect to the sun.

PV systems usually include a *pyranometer* – a device that measures the solar intensity reaching the panels. The pyranometer is tilted at the same angle as the panels and is designed to stay clean and snow-free. The curve labeled "2" corresponds to the actual solar intensity times the surface area of the panel through the day. Drops in curve 2 compared to curve 1 indicate clouds, and in practice, curve 2 may be much more "noisy" than shown; see, e.g., Figure 2 and Figure 3.

Of course, a solar panel cannot convert all the radiation into power, i.e., its *efficiency* is not 100 percent. PV manufacturers typically specify efficiency as a function of temperature (solar panels tend to be more efficient at lower temperatures) [20]. Curves 3 and 4 in Figure 1 are derived by applying an efficiency formula to curves 1 and 2, respectively. That is, curve 3 is the expected power output given a perfectly sunny day, and curve 4 is the expected power output after taking clouds into account. Note that the area between curves 3 and 4 corresponds to power loss due to clouds, which is unavoidable.

There are two common ways to compute curve 4. One is to start with the solar intensity measured by a pyranometer, as described above, and adjust it according to the efficiency function. If there is no pyranometer onsite, another way is to select one panel as a reference panel and use its actual power output as the expected power output. Of course, this panel, to which we refer as a *reference panel*, must be clean and problem-free.

Finally, curve 5 shows the actual power output of the panel, as measured by a sensor connected to the micro-inverter. Ideally, curve 5 should be identical to curve 4. In Figure 1, the actual power output drops below the expected power output around 11:00, which could be due to external factors such as shadow or dirt. Note that the area between curves 4 and 5 corresponds to power loss due to such external factors, many of which are avoidable, e.g., by cleaning the panels.

We are now ready to state the problem we want to solve. We are monitoring a PV array consisting of multiple panels. We are given 1) enough enough information to compute the expected power output time series (curve 4), e.g., the corresponding solar intensity and temperature time series plus the performance ratio function, and 2) for *each* panel, we are given an actual power output time series (curve 5). Our goal is to identify and classify time intervals during which curve 5 significantly drops below curve 4, as we will formalize in Section 4. We assume that the input time series have a fine granularity (e.g., one measurement every 5 or 15 minutes). The frequency of identifying and classifying anomalies in the power output depends on the application; for concreteness, we assume that at the end of each day, we need to analyze the current day's data.

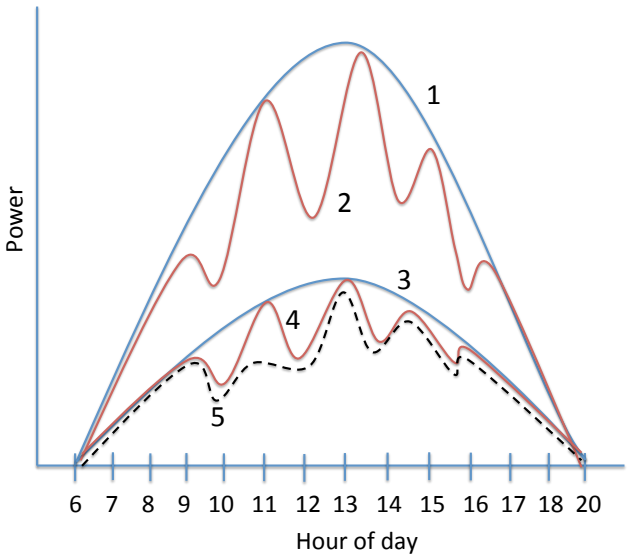


**Figure 1: Example of solar panel output assuming perfect efficiency and a sunny day (1), perfect efficiency and clouds (2), actual efficiency without (3) and with (4) clouds, and with other factors (5).**

## 3. RELATED WORK

There has been a great deal of research on understanding and attributing the power loss of a whole PV array due to weather and the external environment. Field trials and simulations were done to model and characterize the effects of cloud cover (see, e.g., [13]), air pollution (see, e.g., [11]), shadows (see, e.g., [4, 17, 23]), dust and dirt (see, e.g., [3, 6, 8, 15, 25]), and snow (see, e.g., [1, 2, 18]). The goal of this body of work was mainly to estimate the percentage power loss over an extended period of time, perhaps as a function of the type or thickness of snow or soiling. Rather than studying a particular factor in a controlled environment (e.g., using clean and dirty panels side-by-side), our work aims to infer the underlying factors based on (per-panel) power output and solar intensity data.

In terms of anomaly detection, there are at least three related approaches, which we summarize below.

The first approach, mentioned in [16, 19, 24], is to periodically compute linear regressions of power output vs. solar radiation and power output vs. panel temperature to detect changes in the behaviour of panels. However, this approach is not meant to distinguish between different types of changes, and therefore anomaly classification was not discussed.

In [5, 21], the solution is to collect statistics about anomalies such as the magnitude of the power drop and the duration of the anomaly. The idea behind our solution is similar, but we show that a single feature is already sufficient to distinguish between shadow and direct covering of a panel. Furthermore, our solution does not rely on the magnitude

of the power drop since the same type of anomaly (e.g., dirt/snow) may cause a different amount of power drop in different circumstances (e.g., different thickness and density of snow or different types of dirt).

The third approach is based on machine learning. In [12], a decision tree classifier was constructed to predict the severity of a physical problem with a solar panel based on features such as discolouration or panel warping. While we also aim to classify anomalies in power output, we focus on external problems rather than hardware faults, and therefore our framework and features are different. In [10], several classifiers were tested on their ability to classify anomalies in PV power output based on statistical properties of the output time series. While our solution also classifies anomalies in power output, it is different from [10] in several ways. First, we assume that we are also given solar intensity data as input, which allows us to separate power drop due to cloud cover from other factors. Second, as we will show, we use simple and interpretable features of the output time series rather than complex statistical properties.

There is also a variety of commercial software tools for estimating and tracking the power produced by solar panels, and estimating power loss due to weather and other factors; examples include Enphase Energy's Enlighten[1], Locus Energy's PVIQ[2], PVSyst[3] and Tigo[4]. Some systems use rough estimates for shading and soiling losses based on historical data, while others include more sophisticated analytics. For example, PVIQ estimates loss due to shading by identifying seasonal patterns, e.g., a drop in power every morning throughout the summer may correspond to a morning shadow. Our solution does not require a year of training data. In general, our solution is complementary to, and may be incorporated in, the above systems to improve the accuracy of power loss estimation and attribution.

# 4. OUR SOLUTION

Recall that we are given an expected power output time series, computed using pyranometer measurements or using the power output of a clean reference panel, and an actual power output time series. Our goal is to explain anomalies in the actual power output. Also, recall that the expected power output already accounts for clouds, so any further drop in produced power is likely due to other factors such as dirt or shadow. The crux of our solution is the observation that dirt or snow, which physically cover a panel, affect the power output in a different way than shadows. We illustrate this observation with an example and then we explain it in terms of the physical properties of solar radiation.

## 4.1 Intuition and Physical Explanation

Figure 2 plots the expected ("theoretical") and actual ("real") power outputs (in Watts) of the solar panel circled in red in Figure 5; we will describe the PV array this panel comes from in Section 5.1. The measurements were taken on February 11, 2012, and, as can be seen, this panel is covered by snow. In general, this panel is producing roughly one third of the expected power. Notice that the real power output follows the fluctuations of the expected power output; that is, if clouds come out, the power output drops correspondingly.

Next, in Figure 3 we show another pair of theoretical and real power time series for another panel covered by a morning shadow (from about 9:00 till 11:00) on July 10, 2013. Notice that at that time the power output drops to roughly 20 Watts and generally *does not* follow the fluctuations of the expected power output. That is, whether it is sunny or cloudy, this shaded panel is producing (roughly) uniformly low power.

In order to explain these observations, we need to understand the physical properties of solar radiation [14]. It has two main components: *direct* and *diffuse*. Direct radiation reaches the surface of the Earth in a straight line from the sun without any reflection or scatter by the atmosphere. Diffuse radiation is scattered by the atmosphere and arrives at the surface of the Earth from all directions. There is also a third component, *albedo* radiation, which is the radiation reflected from the ground, but its effect on solar panels is negligible compared to the other two. On a clear sunny day, most of the radiation is direct. On a cloudy winter day even half the radiation may be diffuse depending on location.

Now, it is important to understand that shadow only blocks direct radiation, which would normally reach a solar panel in a straight line from the sun; diffuse radiation is not affected since it arrives from all directions. This is why the power output in Figure 3 drops and remains roughly constant. The only radiation getting through is diffuse, and this does not fluctuate when clouds come out. The peaks in theoretical power output are due to more direct radiation hitting the panel when the sky is clear. On the other hand, covering the panel with dirt or snow blocks both direct and diffuse radiation. This is why the power output in Figure 5 is roughly a constant fraction of the expected power output at all times: depending on the thickness and density of the snow, some fraction of all the radiation is blocked.

This simple property of solar radiation has been mentioned by prior work on PV performance analysis [5, 13, 17, 25]. Our contribution in this paper is to turn this observation into a classification feature, as we explain below, and experimentally verify its accuracy on real data.

## 4.2 Anomaly Classification

We now translate the above observations into features that may be used in classification. At any point in time, we define the *Performance Ratio* (PR) of a solar panel as the ratio of actual power produced to the expected (theoretical) power. That is, PR is the ratio of curves 5 and 4 from Figure 1, or the ratio of the two curves shown in each of Figures 2 and 3. For example, if the expected power is 100 Watts but the produced power is 40 Watts, the PR at that point in time is 0.4.

Let $S$ be a set of data points. The *Coefficient of Variation* (CV) of $S$ is a standard statistical metric, defined as the ratio of the standard deviation of the data points to their mean. Now, note that the Coefficient of Variation of the Performance Ratio (CVPR) is low in Figure 2 but higher in Figure 3. This is the main idea of the proposed solution.

The input to our problem consists of the expected and actual power output time series for a given solar panel, as discussed earlier. In the first step, we identify time intervals in which the PR is below some threshold $\tau_{PR}$. In the second step, we compute the CVPR for each such time interval. If

[1] http://enphase.com/enlighten/
[2] http://locusenergy.com/solutions/pviq-analytics/
[3] http://www.pvsyst.com
[4] http://www.tigoenergy.com/

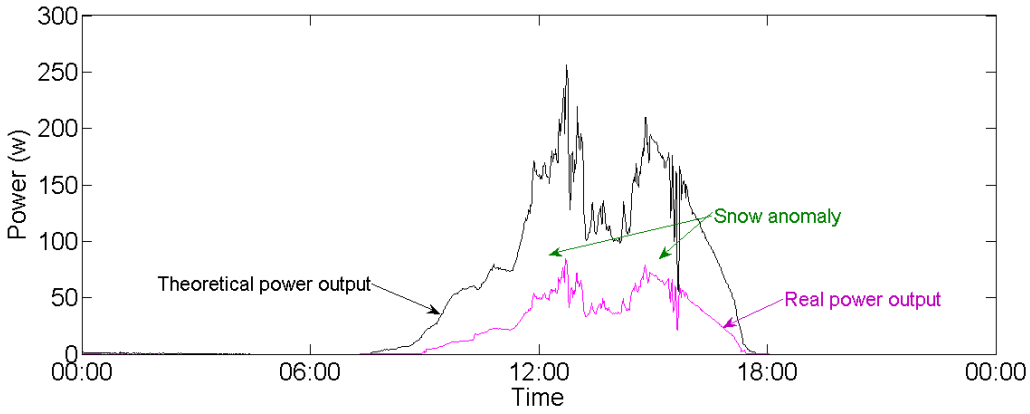


**Figure 2: Expected and actual power output of a panel covered by snow.**

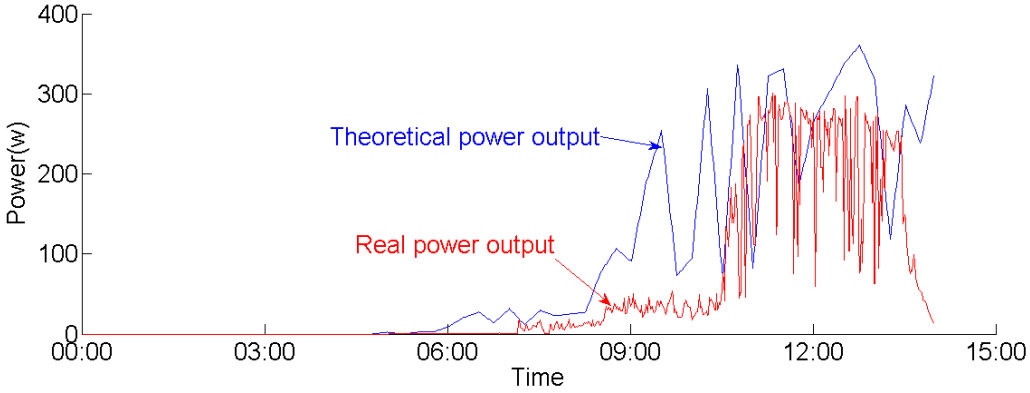


**Figure 3: Expected and actual power output of a shaded panel.**

the CVPR is below some threshold $\tau_{CVPR}$, we classify the anomaly as direct cover. Otherwise, we classify the anomaly as a shadow. We reiterate that per-panel data are required for this method. Otherwise, if, say, only one panel is shaded, then the whole array's PR may still be very close to one and no anomaly will be detected.

The threshold $\tau_{PR}$ controls the aggressiveness of the above classification rule. A high value may lead to false positives, but a low value can miss some anomalies such as small shadows or delay the identification of anomalies such as dirt. The other threshold, $\tau_{CVPR}$, can be learned from labeled data. We will discuss threshold selection further in Section 5.

We point out two simple optimizations of the above classification rule. First, after we find a time interval with low PR, rather than computing CVPR from all the points within this interval, we can remove outliers (highest and lowest PR values in the interval) and compute the CVPR from the remaining points. This will help guard against data errors. The second optimization is to only consider anomalies occurring when the solar intensity is sufficiently high. During periods of low intensity (e.g., dusk or dawn), there is little power being generated and the PR can be noisy.

Note that our solution can easily be extended. For example, in the context of a decision tree, we may test the value of CVPR in the root node of the tree, and then add further tests on other attributes of the data to further specify the cause of a power drop (e.g., dust vs. leaves on the panel vs. bird droppings). That said, we believe that classifying anomalies into shadow vs. direct-cover is already very useful as it can determine when the panels are dirty, for whatever reason, and need cleaning.

# 5. EXPERIMENTS

This section describes our experimental results regarding the accuracy of the proposed classification rule and the accuracy of other classification algorithms that may be applied to our problem, starting with a description of our two data sets, followed by our findings.

## 5.1 Data

In order to test an anomaly classifier, we need examples of shading and soiling along with the corresponding (expected and actual) power output time series. We obtained these from the following two PV installations.

**TRCA:** an array of 15 panels, three each from five different manufacturers, located in Toronto, Ontario. The panels are facing due south with a 30 degree tilt and are managed by the Toronto and Region Conservation Authority (TRCA). This data set contains power output, solar intensity (from an on-site pyranometer), temperature and wind-


**Figure 4: Example of an image in the TRCA data set.**


**Figure 5: Example of an image showing snow-covered TRCA panels.**

speed measurements every minute for one year, from December 2011 till December 2012. We calculated expected power output (curve 4 in Figure 1) from the solar intensity time series and the efficiency formulas provided by the PV manufacturers. Additionally, we obtained an image data feed containing 600x800 photos of the panels taken every 5 minutes. Due to low resolution, we could not identify dust or dirt; see, e.g., Figure 4 taken at noon on August 3, 2012. However, we found 24 days with snow; see, e.g., Figure 5 taken at noon on February 11, 2012.

**UW:** an array of 15 panels installed on the roof of one of the University of Waterloo buildings, facing 26.11 degrees southeast with a 15 degree tilt. We obtained access to the array for one month, from June 20 till July 20, 2013. There is no pyranometer onsite, so we selected one panel as a reference panel and ensured it is always clean and anomaly-free. The power output of this panel was used as the expected power output (i.e., curve 4 in Figure 1). Furthermore, there is no camera on-site, so we manually inspected the panels several times a day and recorded the times and locations of shadows. We also manually covered the panels with varying amounts of dirt (consisting of fine sand mixed with dried soil) and measured the corresponding power drop.

**Table 1: PR and CVPR values of all 18 shadow anomalies**

| PR | CVPR |
|---|---|
| 0.44 | 1.82 |
| 0.5 | 0.75 |
| 0.35 | 0.91 |
| 0.51 | 1.87 |
| 0.34 | 1.99 |
| 0.45 | 0.93 |
| 0.41 | 1.42 |
| 0.46 | 1.82 |
| 0.45 | 1.17 |
| 0.35 | 1.17 |
| 0.19 | 3.45 |
| 0.33 | 1.53 |
| 0.31 | 2.18 |
| 0.3 | 1.87 |
| 0.31 | 2.42 |
| 0.15 | 6.3 |
| 0.3 | 1.91 |
| 0.47 | 2.13 |

## 5.2 Results

Altogether we collected 60 examples of anomalies, 24 of which are due to snow (TRCA), 18 due to shadow (UW) and 18 due to dirt (UW). Tables 1, 2 and 3 list the PR and CVPR values for all the shadow, snow and dirt anomalies, respectively. Shadow appears to drop the power output to one-half or less of the expected output. The PR values for snow anomalies range from 0.1 to 0.88 depending on the thickness and density of the snow cover. Dirt appears to have less of an effect on the power output than other anomalies: the PR values for our dirt anomalies range from 0.85 to 0.97. However, this may be an artifact of our experimental procedure: the dirt we manually placed on the panels did not stick to the panels for very long and slid off them within several minutes (recall that the UW panels are tilted 15 degrees). In prior work, the effect of dirt and dust has been reported to be higher. Finally, we note that, as expected, the CVPR of shadow anomalies appears significantly higher than that of direct cover anomalies.

### 5.2.1 *Our Classifier*

We now test our simple classification rule: for each time interval in which PR drops below $\tau_{PR}$, if CVPR is below $\tau_{CVPR}$, the power drop is due to direct cover; otherwise, the power drop is due to shadow (then, separating direct cover into snow vs. other cover can be done easily with the help of weather data).

The first task is to determine a value for $\tau_{PR}$. In general, we need to trade off between missed anomalies and false alarms. Our shadow and snow anomalies all had a PR under 0.88, but there were seven dirt anomalies with a PR above 0.9. However, as we mentioned earlier, in practice we expect dirt anomalies to have a lower PR than the PR we obtained in our experiments. Thus, $\tau_{PR} = 0.9$ is a reasonable choice. That is, we identify an anomaly if the actual power output of a panel is 90 percent or less of the expected output.

**Table 2: PR and CVPR values of all 24 snow anomalies**

| PR | CVPR |
|---|---|
| 0.48 | 0.17 |
| 0.5 | 0.18 |
| 0.45 | 0.74 |
| 0.56 | 0.37 |
| 0.58 | 0.6 |
| 0.1 | 1.44 |
| 0.55 | 0.31 |
| 0.77 | 0.11 |
| 0.42 | 0.48 |
| 0.11 | 1.8 |
| 0.88 | 0.02 |
| 0.47 | 0.44 |
| 0.62 | 0.18 |
| 0.62 | 0.34 |
| 0.35 | 0.46 |
| 0.76 | 0.15 |
| 0.74 | 0.08 |
| 0.84 | 0.05 |
| 0.85 | 0.05 |
| 0.78 | 0.19 |
| 0.81 | 0.09 |
| 0.81 | 0.17 |
| 0.23 | 0.67 |
| 0.37 | 0.63 |

Next, we need to choose a value for $\tau_{CVPR}$. Based on our training data, the best thresholds are 0.75 and 1.17. With $\tau_{CVPR} = 0.75$, 50 out of 60 anomalies are classified correctly, with two snow and 8 dirt anomalies misclassified as shadow. With $\tau_{CVPR} = 1.17$, 51 out of 60 anomalies are classified correctly for an accuracy of 0.85, with 3 shadow anomalies misclassified as direct cover, and two snow and 4 dirt anomalies misclassified as shadow. As we mentioned in Section 4.2, there are simple optimizations that may improve accuracy, such as removing PR outliers within the time interval of an anomaly. Furthermore, having access to more labeled data should help choose a better threshold. That said, based on our results so far, we conclude that a $\tau_{CVPR}$ value of around one should work well.

We also point out that only three shadow anomalies had a CVPR value below one, and they happened on cloudy days, on which the solar radiation was not as noisy as that in Figures 2 and 3. As a result the CVPR was lower than it would be had there been periods of sunshine and clouds throughout the day. On the other hand, there are several snow and dirt anomalies with a relatively high CVPR between 1.4 and 1.8. These correspond to thin layers of dirt or snow, which may have allowed more diffuse radiation to reach the panel than a thick and dense cover would.

### 5.2.2 Other Classifiers

For comparison, we also tested several classifiers using the WEKA machine learning toolkit [9]. Each classifier was given two feature variables: PR and CVPR, and the class label, which could be shadow or direct cover. Table 4 shows the accuracy of the tested classifiers using ten-fold cross validation. The algorithms are: the C4.5 decision tree, the Best First (BF) decision tree, the Naive Bayes (NB) decision tree, the Functional Tree (FT), the Simple Cart decision tree algorithm, Support Vector Machines (SVM) with linear and degree-4 polynomial basis, and the k-Nearest-Neighbour algorithm with three different values of $k$.

**Table 3: PR and CVPR values of all 18 dirt anomalies**

| PR | CVPR |
|---|---|
| 0.9 | 0.7 |
| 0.91 | 1.2 |
| 0.88 | 1.13 |
| 0.92 | 0.68 |
| 0.82 | 0.57 |
| 0.95 | 0.55 |
| 0.97 | 0.41 |
| 0.94 | 0.43 |
| 0.9 | 0.13 |
| 0.93 | 1.41 |
| 0.9 | 0.98 |
| 0.86 | 1.45 |
| 0.88 | 0.97 |
| 0.85 | 1.56 |
| 0.93 | 1.04 |
| 0.9 | 0.66 |
| 0.89 | 0.53 |
| 0.9 | 0.89 |

**Table 4: Accuracy of other classification algorithms**

| Classifier | Accuracy |
|---|---|
| C4.5 | 0.93 |
| BF Tree | 0.92 |
| NB Tree | 0.93 |
| FT | 0.86 |
| Simple Cart | 0.93 |
| SVM (Linear) | 0.88 |
| SVM (degree 4 polynomial) | 0.88 |
| kNN (k = 1) | 0.95 |
| kNN (k = 3) | 0.93 |
| kNN (k = 5) | 0.88 |

The accuracy of the other classifiers is higher than that of our simple rule, at the cost of over-fitting. For instance, the C4.5 algorithm gave the following tree, which overfits the data by making multiple tests on PR; the numbers in brackets correspond to the number of anomalies covered by each leaf node in the decision tree. Interestingly, PR, not CVPR, is tested at the root of the tree. However, as the tree shows, some direct cover anomalies have low PR whereas others have higher PR (depending on the thickness and density of the dirt or snow).

```
PR <= 0.51
|   CVPR <= 0.74: Direct Cover (8.0)
|   CVPR > 0.74
|   |   PR <= 0.11: Direct Cover (2.0)
|   |   PR > 0.11: Shadow (18.0)
PR > 0.51: Direct Cover (32.0)
```

Similarly, the BF tree also overfit the data by making multiple tests on PR and CVPR. The root node actually tests on CVPR but the threshold is too high and a second test on CVPR is required in the second layer of the tree.

```
CVPR < 1.81
|  PR < 0.525
|  |  CVPR < 0.745: Direct Cover (8.0)
|  |  CVPR >= 0.745
|  |  |  PR < 0.22: Direct Cover (2.0)
|  |  |  PR >= 0.22: Shadow (7.0)
|  PR >= 0.525: Direct Cover (32.0)
CVPR >= 1.81: Shadow (11.0)
```

Simple Cart also overfit the data with similar problems to that of the BF tree:

```
CVPR < 1.81
|  PR < 0.525
|  |  CVPR < 0.745: Direct Cover (8.0)
|  |  CVPR >= 0.745: Shadow (7.0)
|  PR >= 0.525: Direct Cover( 32.0)
CVPR >= 1.81: Shadow (11.0)
```

# 6. CONCLUSIONS AND FUTURE WORK

In this paper, we addressed the problem of identifying and explaining anomalies in the power output of solar panels. We developed and tested a simple classification rule based on the physical properties of solar radiation. The proposed rule can distinguish between power drop due to shadow and power drop due to direct cover such as dust or snow on the panel.

Based on our experimental results, there is room for improvement of our anomaly classifier, both in terms of accuracy and ability to further pinpoint the nature of a direct cover (dust, dirt, leaves, etc.). In general, given the rising popularity of solar panels and the availability of per-panel data, there is much more solar panel data mining that can be done. Examples include clustering the power output time series (and other measurements) to determine similar panels (in terms of performance and/or anomalies), outlier detection, and association rule mining among different panels (e.g., if there is a shadow on panel x then there will be a shadow on panel y within 15 minutes).

# 7. ACKNOWLEDGEMENTS

We would like to thank the Toronto and Region Conservation Authority (TRCA) for giving us a copy of their PV power output and image data, and we thank Bo Hu for setting up the power output monitoring infrastructure on the University of Waterloo PV array.

# 8. REFERENCES

[1] R. W. Andrews, A. Pollard and J. M. Pearce, The effects of snowfall on solar photovoltaic performance, Solar Energy 92 (2013): 84-97.

[2] G. Becker, B. Schiebelsberger, W. Weber, An approach to the impact of snow on the yield of grid-connected PV systems, in Proc. 21st European Photovoltaic Solar Energy Conference (EU PVSEC), 2006.

[3] J. R. Caron and B. Littmann, Direct Monitoring of Energy Lost Due to Soiling on First Solar Modules in California, IEEE Journal of Photovoltaics 3.1 (2013): 336-340.

[4] C. Deline, Partially shaded operation of a grid-tied PV system, in Proc. 34th IEEE Photovoltaic Specialists Conference (PVSC), 2009.

[5] A. Drews, A. C. De Keizer, H. G. Beyer, E. Lorenz, J. Betcke, W. Van Sark, W. Heydenreich, E. Wiemken, S. Stettler and P. Toggweiler, Monitoring and remote failure detection of grid-connected PV systems based on satellite observations, Solar Energy, 81.4 (2007): 548-564.

[6] M. S. El-Shobokshy and F. M. Hussein, Effect of dust with different physical properties on the performance of photovoltaic cells, Solar Energy 51 (1993): 505-511.

[7] D. Frankel, K. Ostrowski and D. Pinner, The disruptive potential of solar power, MicKinsey Quarterly, April 2014.

[8] D. Goossens and E. Van Keschaever, 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.

[9] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, The WEKA Data Mining Software: An Update, SIGKDD Explorations (2009) 11(1):10-18.

[10] B. Hu, Solar Panel Anomaly Detection and Classification, University of Waterloo M.Math Thesis, 2012.

[11] J. K. Kaldellis, P. Fragos and M. Kapsali. Systematic experimental study of the pollution deposition impact on the energy yield of photovoltaic installations, Renewable Energy 36.10 (2011): 2717-2724.

[12] J. M. Kuitche, R. Pan and G. TamizhMani, Investigation of Dominant Failure Mode(s) for Field-Aged Crystalline Silicon PV Modules Under Desert Climatic Conditions, IEEE Journal of Photovoltaics 4.3 (2014): 814-826.

[13] D. H. W. Li, G. H. W. Cheung and J. C. Lam, Analysis of the operational performance and efficiency characteristic for photovoltaic system in Hong Kong, Energy Conversion and Management 46 (2005): 1107-1118.

[14] A. Luque and S. Hegedus, Eds., Handbook of photovoltaic science and engineering. John Wiley & Sons, 2011.

[15] M. Mani and R. Pillai, Impact of dust on solar photovoltaic (PV) performance: research status, challenges and recommendations, Renewable and Sustainable Energy Reviews 14 (2010): 3124-3131.

[16] S. Mau and U. Jahn, Performance analysis of grid-connected PV systems, in Proc. 21st European Photovoltaic Solar Energy Conference (EU PVSEC), 2006.

[17] T. Oozeki, T. Izawa, K. Otani and K. Kurokawa, An evaluation method of PV systems, Solar Energy Materials and Solar Cells, 75.3 (2003):687-695.

[18] L. Powers, J. Newmiller and T. Townsend, Measuring and modelling the effect of snow on photovoltaic system performance, in Proc. 35th IEEE Photovoltaic Specialists Conference (PVSC), 2010.

[19] S. J. Ransome, J. H. Wohlgemuth, S. Poropat and E. Aguilar, Advanced analysis of PV system performance using normalised measurement data, in Proc. of 31st IEEE Photovoltaic Specialists Conference (PVSC),

2005.

[20] E. Skoplaki 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.

[21] S. Stettler, P. Toggweiler, E. Wiemken, W. Heydenreich, A. C. de Keizer, W. van Sark, S. Feige, M. Schneider, G. Heilscher and E. Lorenz, Failure detection routine for grid-connected PV systems as part of the PVSAT-2 project, in Proc. 20th European Photovoltaic Solar Energy Conference (EU PVSEC), 2005.

[22] I. H. Witten, E. Frank and M. A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, 3rd ed., Morgan Kaufmann Publishers, 2011.

[23] A. Woyte, J. Nils and R. Belmans, Partial shadowing of photovoltaic arrays with different system configurations: literature review and field test results. Solar Energy 74 (2003): 217-233.

[24] A. Woyte, M. Richter, D. Moser, S. Mau, N. Reich, U. Jahn, Monitoring of photovoltaic systems: good practices and systematic analysis, in Proc. 28th European Photovoltaic Solar Energy Conference (EU PVSEC), 2013.

[25] J. Zorrilla-Casanova, M. Piliougine, J. Carretero, P. Bernaola, P. Carpena, L. Mora-Lopez and M. Sidrach-de-Cardona, Analysis of dust losses in photovoltaic modules, World Renewable Energy Congress, 2011