Detecting Absorption of Solar Radiation By Clouds With Satellite Imagery Processing

Liuchenghang(Leon) Zhu Changyeon Jo

December 14, 2012

Abstract

Solar irradiance data are crucial to PV system performance analysis. Although lightmeters (or pyranometers) have been applied in utility-level solar farms, they are not economically feasible for commercial/residential-level solar sites. Publicly accessible data, such as meteorological satellite imagery, are potentially an ideal source to meet the emerging demand. This research applied satellite imagery processing and supervised learning algorithms to classify/quantity cloud absorption, and thereby predict solar irradiance.

Prior Work

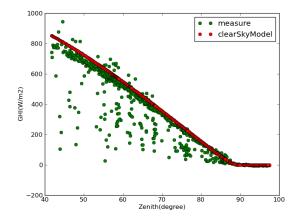
In the past two decades, researches have been conducted on solving the satellite-to-radiance problem both commercially and academically. The handful published researches are either based on data-feed-intensive physical models or simple statistical models, and presented high variance (hourly bias 2.5%-4.5%, hourly variance 20%-40%, Polo et. 2008). Also, the commercial models are kept confidential to the public. This study aimed to apply machine learning classification and regression algorithms to explore and improve analytic approaches to solve this problem.

Introduction

The solar irradiance that reaches the top of Earth atmosphere is fairly deterministic and can be modeled by several mature models (*Matthew et al. 2012*). We name it as clear sky irradiance I_{CS} (here our target irradiance type is Global Horizontal Irradiance (GHI), w/m^2).

When passing through the atmosphere, part of sunlight will be absorbed by clouds. We denote the absorbed fraction as cloud factor CF. Then we have $GHI_{ground} = GHI_{CS} \times (1-CF)$. Thus, our goal is to model CF with satellite data. On the one hand, by re-writing the above equation as $CF = 1 - GHI_{ground}/GHI_{CS}$, we show how to obtain "real" CF as target values (y). On the other hand, we predict CF as CF = h(imagePixels, weatherParameters)

Figure 1: Calibrated Clear Sky Model



Data Processing

 GHI_{ground} were collected by lightmeters at University of Nevada and Nevada Power Clark Station, Nevada. Hauwitz model (1945) was applied to model GHI_{CS} . We found it tended to underestimate clear sky irradiance after checking it with ground measurement (we observed GHI_{model} < measure). Model expression: $GHI_{CS} = 1098 \times cos(Z) \times exp(\frac{-0.057}{cos(Z)})$ where

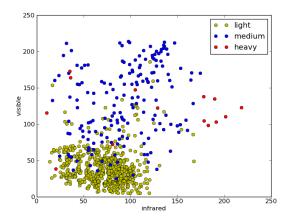
Z is the solar zenith angle. We adjusted the baseline by keeping its model structure while linear-regressing it so that it selectively fitted clear-sky-like points while made nearly all the other points below the modeled results, shown Fig 1. $log(GHI_{CS}) = log(a) + b \times log(cos(Z)) + \frac{c}{cos(Z)}$

Satellite images of multiple bands (visible, infrared, near-infrared, ice and water vapor) were collected from *The National Center Of Atmospheric Research*. They maintain a weekly archive of hourly observations. We kept track of images for 2 months. We first geo-referenced raw images with $ArcGIS^{\textcircled{\tiny{\$}}}$, so that we could query pixel locations (column, row) by geographical locations (latitude, longitude). 8 neighboring pixels, along with the central pinpoint, were extracted and averaged as our raw pixel data for each band. The weather station data were collected from NOAA's $Hourly/Sub-Hourly\ Observational\ Data$.

Classification Although our ultimate goal was to quantify cloud cover or solar irradiance, we found a simple linear regression of raw pixel data wouldn't work. Therefore, we started to classify the clouds to see if it would facilitate quantification, or hopefully if a granular multi-class classification could proximate a quantification. The features were listed as [pixel-intensities=[visible, infrared, water vapor, ice, near-infrared(red), near-infrared(green), near-infrared(blue)], station-data=[temperature, dew point, cloud ceiling, pressure, windspeed]]]

Unlike expected, there seemed to be only 3 distinct classes: light (CF < 0.2), medium (CF 0.2 - 0.8) and heavy (CF > 0.8) clouds, as shown in Fig2. In addition, classifiers had difficulty classifying points with visible pixel intensity < 10(155) out of 1008 samples). Near the lower bound, imagery noises posed significant impacts on distinguishability. Our solution was classifying samples with visible pixel values < 10 and the rest separately. For the non-noisy portion of samples, we classified them as the previously mentioned 3

Figure 2: Infrared - Visible



classes, while we only classifyed the noisy portion as Heavy (CF>0.8) and LightToMedium $(CF\leq0.8)$ two classes.

We applied SVM and Gaussian-Naive Bayes to classify the dataset. For SVM, feature scaling was critical for successful classification. Since it's a distance-based classifier, features with high magnitudes tend to be biased. In addition, according to Perez et al. 2002, all pixel values were corrected with hourly solar zenith angle (normalized by cos(Zen)). Navie Bayes was sensitive to feature distribution. A feature CLG (cloud ceiling, belonging to weather station observations) is a binary-like random variable. Its high value is exclusively 722 while other values randomly spread out between 0 and 100. Based on its physical meaning, CLG would be indicative in cloud cover classification and it did work well for SVM, but poorly for Naive Bayes, due to its non-Gaussian distribution. The results were presented in Table 1. The sample counts of different classes were based on the whole dataset. But the training and testing was based on a 70% vs 30% random split. All features (scaled to 0 to 1) were used in SVM. Selected features (not scaled) for Naive Bayes: [visible, infrared, water vapor, temperature, dew point, pressure

Regression

Perez et al. 2002, proposed a non-linear relation

	vis < 10		$vis \ge 10$		
Class	lightToMid	heavy	light	$_{ m mid}$	heavy
Count	117	38	667	172	14
SVM score	0.851		0.910		
NB score	0.723		0.914		

Table 1: Classification and Cross Validation Scores

between GHI and Cloud Index(CI), which performed generally well for pixel-to-irrdiance quantification.

$$GHI = K \cdot GHI_{CS}(0.0001K \cdot GHI_{CS} + 0.9)$$

with $K = 2.36CI^5 - 6.2CI^4 + 6.22CI^3 - 2.63CI^2 - 0.58CI + 1$

In Perez's model, Cloud Index was not the same as the cloud absorption factor (CF) in our study of classification. CI was simply a normalized value of visible band pixel intensity $(CI = \frac{vis - vis_{max}}{vis_{max} - vis_{min}})$. The quadratic function $GHI = h_G(K \cdot GHI_{CS})$ was assumed with fixed parameters, while the coefficients of the polynomial function $K = h_K(CI)$ came from fitting. These were given without further explanation in the publication. To infer its rationality, we derived CF from Perez's model:

$$CF = 1 - \frac{GHI}{GHI_{CS}} = 1 - K(0.0001K \cdot GHI_{CS} + 0.9) = 1 - 0.9K - 0.0001K^2 \cdot GHI_{CS}$$

As we can see, K was probably defined to describe cloud transmittance. If we plug in a very low GHI_{CS} , then CF will get close to 1 - 0.9K; and if we increase the coefficient of K from 0.9 to 1, we will get $CF \approx 1 - K$. In this case, K will exactly represent the fraction of light that passes through the cloud, namely $GHI = K \cdot GHI_{CS}$ Thus, fitting K verses CI is equivalent to fitting CF verses CI (or generally, features generated from pixel intensity). Interestingly, Perez's model indicates that this only holds at low irradiance level. As we have higher values of irradiance, CF will not be linearly-correlated with K, due to the significant effects from the quadratic term associated with GHI_{CS} . Physically, it means cloud absorption fraction is both cloud

	Full Freedom	Perez	Two Step
Corr	0.953	0.947	0.946
MBE	0.008	8.47	9.31
MBE(%)	0.002	2.30	2.53
RMSE	47.23	59.58	58.33
RMSE(%)	12.84	16.19	15.85

Table 2: Cross Validation of Three Regression Approaches

and irradiance dependent, which explains why we previously weren't able to fit CF simply based cloud informations (pixel intensities).

Perez is credited for the found of this empirical quadratic relation, though we feel sceptic about the model parameterization. As long as $GHI = h_G(K, GHI_{CS})$ is not strictly derived from physics, its parameters can actually be fitted too. Effort had been paid mainly to engineer cloud features while leaving this quadratic function untrained. Our idea is that we can simultaneously train the two equations in a more balanced manner. Firstly, a simple approach is to plug $K = h_K(CI)$ into h_G , expand the equation and fit all the coefficients. Specifically, assuming the intercepts are unfixed, we can expand the overall model as $GHI = a_{10}GHI_{CS}^2CI^{10} + a_9GHI_{CS}^2CI^9 + ... + a_1GHI_{CS}^2CI + a_0GHI_{CS}^2 + b_5GHI_{CS}CI^5 +$ $b_4GHI_{CS}CI^4 + ... + b_1GHI_{CS}CI + b_0GHI_{CS} + c$ Now we have 17 features and 18 coefficients (one intercept) for linear regression.

Two-Step Regression

If the empirical equations bear some unrecognized physical meanings, we should expand the hypothesis set more cautiously. For example, when constructing the function K(CI), Perez had probably taken into account that K(CI=0) should equal 1, and hence made the intercept fixed. If we arbitrarily unlock dimensions of freedom for a simple linear regression, we won't have clear idea what physical bounds might be missing. Here, the problem really is we want of fit the quadratic function, but at the meantime, we don't want to unlock other 9 extra dimensions of

freedom, in concern of over-fitting.

We experimented with a two step linearregression approach. Loop until converge { G step:

with
$$K^{(i)} = g_5CI^{(i)^5} + g_4CI^{(i)^4} + g_3CI^{(i)^3} + g_2CI^{(i)^2} + g_1CI^{(i)}$$
, fit $GHI^{(i)} = a(K^{(i)}GHI^{(i)}_{CS})^2 + b(K^{(i)}GHI^{(i)}_{CS})$,

K step: with
$$K^{(i)} = \frac{-b + \sqrt{b^2 - 4a(c - GHI^{(i)})}}{2aGHI_{CS}^{(i)}}$$
, fit $K^{(i)} = g_5CI^{(i)^5} + g_4CI^{(i)^4} + g_3CI^{(i)^3} + g_2CI^{(i)^2} + g_1CI^{(i)}$ }

The adjustable parameters were a, b in $h_G(K \cdot$ GHI_{CS}) and the other 5 linear coefficients in $h_K(CI)$. Also, we fixed the intercepts of the two linear regressors as 0 and 1, respectively. By doing so, we only unlocked two dimensions of freedom (a and b) that we were interested in, while strictly keeping other model structure unchanged. Mathematically, this algorithm is not guaranteed to converge. In addition, two possible roots of the quadratic equation makes K step ambiguous. However, let's think about the physical bounds of our satellite-to-irradiance problem. K was constructed to describe cloud transmittance factor, with a range of 0 - 1. Clear-Sky-Irradiance GHI_{CS} is a positive value (or 0), so is the real irradiance GHI (we have to pay attention, since measurement noise and baseline shifting may result in negative values). Thus, as one of the quadratic solution, our $K \cdot GHI_{CS}$ should theoretically be non-negative. At the mean time, according to the equation, the sign of the two roots depends on -GHI which is negative or zero, then we'll always have a positive root (if not zero). If the data is of good quality, and given the system is self-consistent, this approach is likely to work. All the coefficients of h_q was initialized as 0.

The results were presented in Table 2 along with other two regression approaches, where Corr, RMSE and MBE stand for correlation coeffi-

\overline{a}	0.0001	0.001	0.01	0.0001	0.0001
b	0.9	0.9	0.9	0.09	0.009
Corr	0.947	0.932	0.826	0.888	0.786
MBE	8.47	17.22	54.64	34.18	71.55
MBE(%)	2.30	4.68	14.85	9.29	19.44
RMSE	59.58	69.66	137.62	107.52	183.93
RMSE(%)	16.19	18.93	37.41	29.23	49.99

Table 3: Changing Parameters of Perez' Model

cient, root mean square error (w/m^2) , mean bias error (w/m^2) , respectively. RMSE(%) and MBE(%) were also normalized by the mean irradiance value to percentages. Three models all had good performance. Two-step regression performed no better than Perez's model. This indicated that Perez made an excellent estimate of the quadratic coefficients. The quadratic parameters optimized by the two-step model were a=0.00024 b=0.73, which were of the same magnitudes of Perez's ((a = 0.0001, b = 0.9)). We've also experimented changing a and b of Perez's model and found the magnitudes really mattered, shown in Table 3. Therefore, the two-step regression can be seen as a validation of Perez's quadratic parameterization.

Classification-Linear Regression

As we have discussed, we were able to classify clouds as three classes in terms of their absorption factors. We were interested in exploring how well each class of clouds was regressed by the model. As shown in Fig 3 and 4, we believed it necessary to perform regression separately for 3 classes. The training results were presented in Table 4.

We constructed an integrated model with a combination of classification and regression, shown in Fig 5. A comparative cross-validation was performed on our model and Perez's. One dataset was randomly split as 70% and 30% to train and test both models. The results were shown in table 5. As we can see, with similar correlation coefficients, the integrated model presented significantly smaller bias and variance than Perez's.

Figure 3: A Universal Regression on 3 Classes

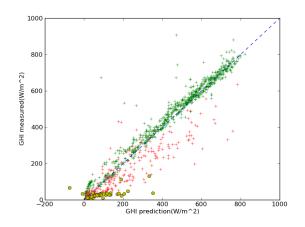


Figure 4: Separate Regressions on 3 Classes

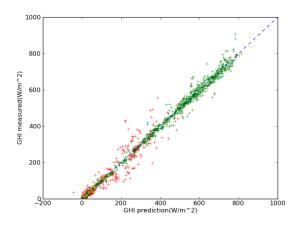


Figure 5: Model Work Flow

2-class Class_light regressor

Class_medium regressor

GHI

Class_heavy regressor

Conclusion

In this study, we found that a 3-class classification facilitated cloud absorption quantification. The parameterization of Perez's model was validated by a two-step regression. An integrated classification-regression model was developed to solve the satellite-to-irradiance problem, and it presented higher performance than the Perez's

	One For All			Sperate		
Class	L	M	Н	L	M	Н
Corr	0.979	0.887	0.541	0.994	0.919	0.983
MBE	22.5	-53.69	-59.57	1.19	-1.92	-0.13
MBE(%)	4.87	-31.99	-198.17	0.25	-1.14	-0.43
RMSE	32.97	88.48	76.44	16.60	43.53	1.87
RMSE(%)	7.13	52.72	254.29	3.59	25.94	6.22

Table 4: Regression Training Results

Model	Integrated	Perez
Corr	0.937	0.943
MBE	3.04	11.27
MBE(%)	0.83	3.05
RMSE	51.47	65.02
RMSE(%)	14.0	17.7

Table 5: Integrated Model vs Perez Model

model. We expect future works can be done on validating the model with more data, especially for more medium and heavy cloud conditions, for they are not common in our current study site, Las Vegas. In addition, this model can probably be extended to perform hour-ahead forecast of solar irradiance, which will benefit grid dispatching market in the near future. We would like to acknowledge scikit-learn for their powerful machine learning modules on Python.

References

Matthew J. Reno, Clifford W. Hansen, Joshua S. Stein, 2012. Global Horizontal Irradiance Clear Sky Models: Implementation and Analysis

B. Haurwitz, 1945. Insolation in Relation to Cloudiness and Cloud Density

Jesus Polo, Luis F. Zarzalejo and Lourdes Ramrez, 2008. Solar Radiation Derived from Satellite Images

Richard Perez, Pierre Ineichen, Kathy Moore, Marek Kmiecik, Cyril Chain, Ray George and Frank Vignola, 2002. A New Operational Satellite-To-Irradiance Model-Description and Validation