# Machine Learning Prediction of Photovoltaic Energy from Satellite Sources

Alejandro Catalina, Alberto Torres-Barrán, José R. Dorronsoro

Dpto. Ing. Informática, Universidad Autónoma de Madrid

**Abstract.** Satellite–measured irradiances can be an interesting source of information for the nowcasting of solar energy productions. Here we will consider the Machine Learning based prediction at hour $H$ of the aggregated photovoltaic (PV) energy of Peninsular Spain using the irradiances measured by Meteosat's visible and infrared channels at hours $H, H-1, H-2$ and $H-3$. We will work with Lasso and Support Vector Regression models and show that both give best results when using $H-1$ irradiances to predict $H$ PV energy, with SVR being slightly ahead. We will also suggest possible ways to improve our current results.

**Keywords:** Photovoltaic energy, EUMETSAT, SEVIRI Channels, Lasso, Support Vector Regression, Nowcasting.

## 1 Introduction

Global concerns on climate change, the extremely fast development of nations such as China and India and the obvious interest of clean and affordable energy are pushing forward the worldwide use of renewable energies, particularly solar. Thus, it is crucial to constantly improve PV energy production forecasting, the subject of a large research effort (see [6,1] for comprehensive reviews on recent work) and also be the focus of this work. PV energy prediction has two distinct horizons. For day ahead or longer prediction horizons, Numerical Weather Predictions (NWP), such as those of the European Center for Medium Weather Forecasts (ECMWF) are the key input. However, NWP are refreshed at best each 6 hours with ECMWF runs starting at hours 00, 06, 12 and 18 UTC and which are widely available about six hours later. This means that, in Spain, the 00 run will be useful to predict PV energy for approximately the period from 05 UTC to 12 UTC and the 06 run for the period 12 UTC to 19 UTC; the forecasting improvement of the other runs would affect basically to night hours. Concrete hours in other countries will change but it is nevertheless clear that other inputs are needed for better short term or intraday forecasts of PV energy.

Section 7.2 in [1] gives a complete overview of several approaches for the intraday forecasting problem. A first approach is to work in an endogenoeus, i.e. a pure time series (TS) scenario using for nowcasting purposes the latest PV energy readings; a relevant example here is [12] for a single PV plant in California, showing that the best results are obtained by a mixed Genetic Algorithm/Artifical

Neural Network (GA/ANN) model. The pure time series information can be enlarged with other past information such as irradiances or temperatures [13] or combining it with NWP forecasts for the hours ahead, where a large number of contributions have been made. This approach was followed in [2] for the PV production of Peninsular Spain, where past readings are combined with the corresponding NWP–based day ahead predictions emitted the day before to try to correct these day ahead predictions for the next hours. It was shown that this can improve on NWP–based day ahead forecasts for about 2–3 hours, depending on the hour of the day. Images from sky cameras [8], Ch. 9, have also been considered; for instance in [10] they are used for short term (up to 15') prediction of Direct Normal Irradiance (DNI). While very interesting locally, it is not clear how to exploit them over large geographic areas. This leaves us with satellite–based measurements [8], Ch. 3, that can offer at least in some of their channels relevant information with adequate update frequencies.

Satellite information has been used for nowcasting radiation values and PV energy having at its basis the HELIOSAT method to estimate solar irradiance from satellite images [4]. More precisely, images are first used to compute a dimensionless cloud index value which describes for a given hour $H$ the influence of cloudiness on atmospheric transmittance; then this cloud index is used to estimate the ratio between actual global irradiance and the output of a clear-sky model, that can be computed for any given hour. Next, a Motion Vector Field approach is used to forecast cloud images for the following hours $H+1, H+2, \ldots$, which, in turn, allows to forecast cloud indices, irradiance ratios and, finally, irradiance values for these hours. While targeting initially irradiance forecasts, this approach has been also used to derive short term PV energy forecasts for PV plants and regional aggregations (see [7,9,15]).

In this work we are also going to use satellite images to derive short time PV forecasts but we will follow a different approach, purely Machine Learning (ML) based. We will have as the target the aggregated PV production of Peninsular Spain. Spain's PV landscape is quite fragmented, with 4,786.6 MW of installed power at the end of 2014, fairly concentrated in its southern half but with very few plants with peak power above 20MW. This makes an individual plant–based approach to aggregated PV prediction hard to come by and therefore we will try to predict aggregated PV energy directly. Initially we will consider satellite images for METEOSAT's 11 spectral bands, i.e., from visible to long-wavelength infrared, with wavelength ranges 0.6 $\mu$m to 13.4 $\mu$m. METEOSAT also has a High Resolution Visible (HRV) channel; however we will not consider it here because of the homogeneity with other channels. In addition, its higher spatial resolution may not be so important in large area PV forecasts and its relevant information should also be captured by other visible channels considered.

Since current silicon PV cells cannot transform infrared rays, it is clear that infrared channels, particularly those farther away, won't be as relevant as the visible ones. Moreover, EUMETSAT measures reflected radiance which can only act as a proxy to the incoming irradiance that actually produces PV energy. Nevertheless, we will use channel information to predict PV energy at hour $H$

from channel readings at hours $H$, $H-1$, $H-2$ and $H-3$. We will consider hourly PV and satellite data for the years 2013, 2014, and 2015, with 2013 used for model training, 2014 for validation when selecting model hyper-parameters and 2015 as the test set. We will use two different approaches, working first with linear Lasso sparse models and then with Gaussian kernel Support Vector Regression (SVR). SVR is of course more powerful and will indeed give better results but Lasso not only gives a base benchmark model but may allow, in principle, an easier model interpretation as its non-zero coefficients are associated to concrete geographic points. We can summarize our paper contributions as follows:

1. We analyze the capabilities and channel relevance of satellite measured irradiances to predict aggregated PV energy production over Peninsular Spain.
2. We discuss the application of two standard, widely used ML methods, Lasso and SVR, to predict PV energy up to 3 hours ahead from the most relevant irradiance channel values.
3. We discuss model results and point out to ways to improve on them.
4. We give a first analysis of model interpretability, spatially for Lasso and temporally for SVR.

We emphasize that while we believe them interesting and useful, our results are to be seen as a first step towards a better use of satellite–measured irradiances in PV energy prediction. The paper is organized as follows. In Section 2 we will briefly review the EUMETSAT system, the satellite measures and other information it provides and the concrete channel information which can be obtained essentially in real time from EUMETSAT. These channels information is analyzed in Section 3 in the context of predicting the PV energy of Peninsular Spain. In principle it is to be expected that the shorter wavelength channels are the most related to PV energy and our analysis will confirm this. In turn, of these, the `IR016`, `VIS008` and `IR039` channels have the strongest correlation with PV energy and in Section 4 we will first study the quality of Lasso and SVR models to predict PV energy at a given hour $H$ from the radiance measures at hours $H, H-1, H-2$ and $H-3$. As we shall see, best results are achieved when using irradiances at hour $H-1$, with SVR slightly outperforming Lasso. Finally, we will discuss further ways to improve on these results in Section 5.

## 2 Meteosat Satellite Data

The European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT) operates a series of geostationary Meteosat satellites that cover, among others, Europe, Africa and the Atlantic Ocean. Their goal is to provide information on the radiance emitted and reflected by the Earth's surface and atmosphere to be used in meteorological forecasting, as well as in climate monitoring and research, often after further processing is done by the centers on EUMETSAT's Satellite Application Facility (SAF) network.

The Meteosat Second Generation (MSG) satellites are equipped with the Spinning Enhanced Visible and Infrared Imager (SEVIRI) technology and provide near real–time radiance values every 15 minutes over eleven spectral bands,

going from the visible (with a wavelength 0.6 $\mu$m at Channel 1) to the long infrared (with a wavelength 13.4 $\mu$m at Channel 11); the spatial resolution is about $3 \times 3$ km for most of the covered regions. Some of these channels measure concrete properties, such as absorption of water-vapor (Channel 5), ozone (Channel 8) or CO2 (Channel 11). There is also a high-resolution channel with visible radiance on the 0.6 $\mu$m–0.9 $\mu$m wavelength range with a $1 \times 1$ km resolution over Europe and parts of Africa. Another product of interest is the Cloud Mask, a floating point number within the range [0, 5], where 0 means clear and 5 totally covered skies.

The basic pixel counts that SEVIRI collects over each channel are further processed by several calibration procedures. First, readings are converted into radiances assuming a linear relation, that is, by adding an offset to the pixel count and multiplying the resulting number by an appropriate calibration factor (as described in EUMETSAT's inter-calibration documentation). For the visible channels a reflectance percentage is computed as the fraction of their radiance to the maximum solar irradiance. The infrared radiances are converted into equivalent brightness temperatures through an empirical formula [7]. This means that the 11 channel radiances give rise to another 11 variable set, with 3 reflectances and 8 brightness temperatures. Thus, excluding the high resolution channel and adding the cloud mask we would have in principle 23 values at each grid point. Obviously, their use and interest will depend on the concrete objectives pursued.

This information has been usually applied to modeling solar irradiance at concrete locations, often after a detailed analysis of the physical processes involved. However, our goal, wide area PV energy forecasting, is quite different and while surface irradiance is certainly a key information, the very large number of installations involved, their different rated powers and their technological variations suggest to follow other paths. Here we will pursue an "agnostic", Machine Learning (ML) based approach, considering in principle all channel information as potentially relevant and focusing on some channels and discarding others not by physical reasons but by their information content with respect to the PV energy target. As we shall see, this results in discarding some a priori obvious channels but also in retaining others that at first glance would seem less relevant. We discuss next these issues together with the data we will use.

## 3 Satellite Data Analysis

We will use Meteosat channel readings at UTC hours 0 to 23 for the years 2013, 2014 and 2015 that we have downloaded from the EUMETSAT Data Centre. Given the large area we are concerned with, we have downsampled Meteosat's initial resolution to that of a 0.125° grid (the same resolution that has been standard until recently by the NWP forecasts of the ECMWF). We will work only with the grid points of Peninsular Spain; this results in a total of 3,391 points. We will consider first the information for all channels with hourly resolution, from January 1, 2013 to December 31, 2015. Although this should result in 26,304 hours, the actual number is 25,161 hours because of missing data. Of course,

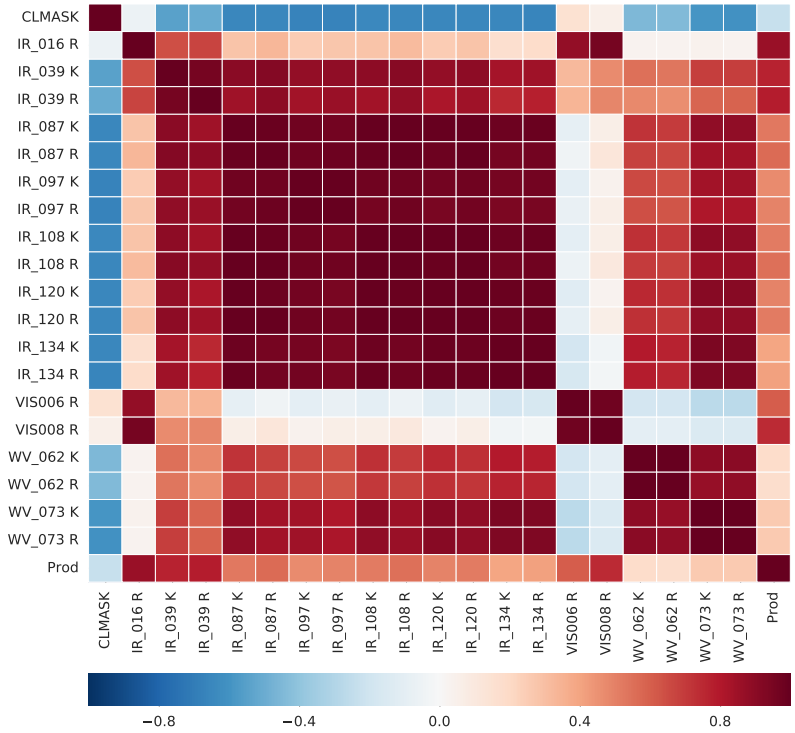



**Fig. 1.** Correlations between cloud index, satellital radiances and PV energy.

only daylight hours are relevant and we have simply selected at the 15-th day of each month the UTC sunrise and sunset hours of Girona (at the easternmost of Spain) and Santiago de Compostela (at the westernmost) respectively and rounded them to the closest hour (daily sunrise and sunset times usually vary by about 15' from those at the month's 15–th day). After dropping the non selected hours that we take as dark, we are finally left with 5,475 hours per year.

Recall that we have in principle 23 satellite measures for each grid point that, if taken on their entirety, would result in input patterns with an extremely large dimension of $3,391 \times 23 = 77,993$. However, since our aim is PV energy nowcasting by Machine Learning methods, it is likely that the first visible or near visible channels would be of the greatest interest. Moreover, since their reflectance percentages are given essentially as dilations, they may not add useful information to the models. In any case, a basic measure of variable relevance is its correlation with the target variable, here PV energy.

To get a manageable first measure, we have computed the average of each variable at every hour over the grid points and then the correlation of these averages with the corresponding PV energy reading excluding, as mentioned, the reflectance percentages of the three visible channels. The resulting correlation matrix is graphically depicted in Figure 1. As it can be seen, infrared radiances

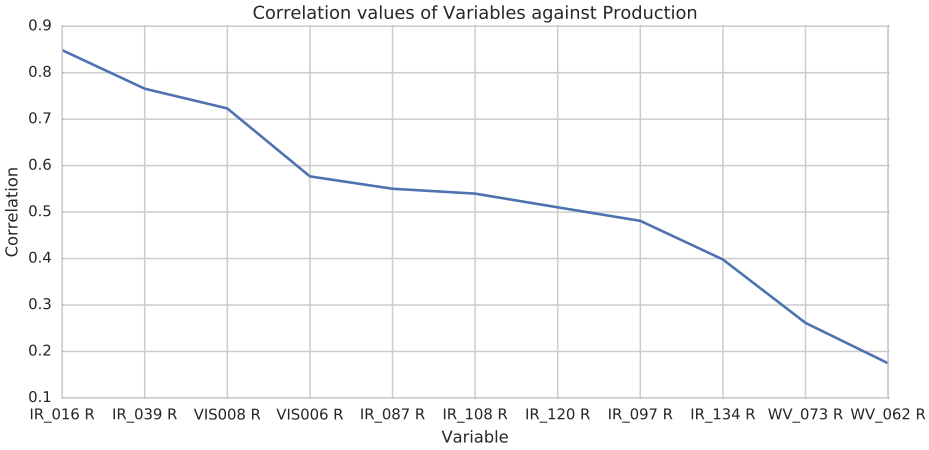



**Fig. 2.** Correlations between cloud index, satellital radiances and PV energy.

are highly correlated with their corresponding brightness temperatures. Besides, all variables but Cloud Mask have positive correlation with PV energy. Somewhat surprisingly, the highest correlations correspond to the infrared channels `IR016` and `IR039` followed by the visible channel `VIS008`. The remaining correlations are quite lower, as shown in Figure 2, where we only give correlation values for radiances and sort channels by decreasing correlation.

Because of this, in what follows we will work only with the radiances of the `IR016, IR039` and `VIS008` channels plus the brightness temperature of channel `IR039`. Note that since channel data consists of radiances while PV data are energies, a possibly better analysis could be done considering for each hour the average radiances of the previous four 15' satellite readings. In any case, we recall that this is a first analysis to be extended in further work.

## 4 Modelling PV Energy from Channel Information

In this section we will use data from the selected channels to model PV energy at hour $H$ with the radiance values over Peninsular Spain at $H, H-1$, $H-2$ and $H-3$. To do so we will consider two well established ML models, Lasso and Support Vector Regression (SVR) which have received some recent attention in the literature [11,3,13]. We briefly describe them next.

### 4.1 The Lasso and Support Vector Regression Models

Given an $N$ pattern sample $\{(x^1, y^1), \ldots, (x^N, y^N)\}$ with $d$–dimensional inputs $x^p$ and 1–dimensional targets $y^p$, the Lasso solution [5] $w^*, b^*$ minimizes the $L_1$ regularized loss

$$\ell_L(w, b) = \frac{1}{2N} \sum_p (w \cdot x^p + b - y^p)^2 + \lambda \|w\|_1.$$

**Table 1.** Hyper–parameters of the Lasso and SVR `m1N` models.

| Model | Parameter | `m0M` | `m0N` | `m0E` | `m1M` | `m1N` | `m1E` | `m2M` | `m2N` | `m2E` | `m3M` | `m3N` | `m3E` |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Lasso | $\lambda$ | 0.033 | 0.017 | 0.027 | 0.022 | 0.023 | 0.055 | 0.106 | 0.065 | 0.040 | 0.157 | 0.073 | 0.036 |
| SVR | $C$ ($\times 10^3$) | 0.170 | 0.155 | 0.178 | 0.321 | 0.474 | 0.096 | 0.184 | 0.109 | 0.054 | 0.102 | 0.090 | 0.049 |
| | $\epsilon$ | 0.034 | 0.112 | 0.019 | 0.123 | 0.026 | 0.042 | 0.132 | 0.075 | 0.032 | 0.053 | 0.168 | 4.117 |
| | $\gamma$ ($\times 10^{-5}$) | 3.064 | 3.085 | 3.061 | 3.060 | 3.053 | 3.350 | 3.054 | 3.057 | 4.683 | 3.770 | 3.069 | 6.108 |

The $L_1$ regularization introduces sparsity in the model with two consequences. First, it avoids possible singularities in the sample covariance matrix; notice that sample size, 5,475, is smaller than input dimension, $13,564 = 3,391 \times 4$. Second, $L_1$ regularization drives many model coefficients to zero, and the positions of the non-zero coefficients may yield first interpretation of the model.

An obvious drawback of Lasso is its possibly poor modeling results because of its linear nature. This motivates our choice of the non linear and possibly more powerful Gaussian kernel SVR [14], one of the workhorses in non linear regression, as our second model. Assuming first for simplicity a linear SVR model, the SVR cost function is

$$\ell_S(w,b) = \sum_p [y^p - w \cdot x^p - b]_\epsilon + \frac{1}{C}\|w\|_2^2; \tag{1}$$

here we use now the $\epsilon$-insensitive loss $\ell(y,\hat{y}) = [y-\hat{y}]_\epsilon = \max\{|y-\hat{y}| - \epsilon, 0\}$ and $L_2$ regularization. We thus allow an $\epsilon$-wide, penalty-free "error tube" around the model. To find the optimal $w^*, b^*$, (1) is rewritten as a constrained minimization problem which is then transformed using Lagrangian theory into the dual problem, the one actually solved; see [14] for more details. To improve on a linear model, the kernel trick [14] is used to take advantage of the fact that only dot products are involved when solving the dual problem. This allows to replace the initial products $x \cdot x$' with a positive definite kernel $k(x,x')$ that can be written as $k(x,x') = \phi(x) \cdot \phi(x')$, where the $x$ are mapped through $\phi(x)$ into a larger, possibly infinite dimensional Hilbert space $H$. We obtain thus a non linear model $f(x) = W \cdot \phi(x) + b$ and, in turn, the optimal $W^* \in H$ can be written as $W^* = \sum \alpha_p^* \, \phi(x^p)$, where the $x^p$ for which $|\alpha_p^*| > 0$ are the Support Vectors (SVs). We thus have

$$f(x) = b^* + W^* \cdot \phi(x) = b^* + \sum \alpha_p^* \, \phi(x^p) \cdot \phi(x) = b^* + \sum \alpha_p^* \, k(x^p, x).$$

Using a Gaussian kernel $e^{-\gamma\|x-x'\|^2}$ gives a model $f(x) = b^* + \sum_p \alpha_p^* \, e^{-\gamma\|x-x^p\|^2}$. Note that the SVR model also lends itself to an interpretation from a temporal point of view, as the SVs correspond to the day–hour pairs whose radiances define the centers of the different Gaussians the model is made of.

Both models require careful hyper–parameter selection, $\lambda$ for Lasso and $C$, $\epsilon$ and $\gamma$ for SVR, which we will do using the year 2013 as a training set and 2014 as the validation set. Notice that the data have a strong temporal structure that

**Table 2.** Lasso and SVR monthly average test errors for year 2015.

| | H | | H-1 | | H-2 | | H-3 | |
|---|---|---|---|---|---|---|---|---|
| Month | Lasso | SVR | Lasso | SVR | Lasso | SVR | Lasso | SVR |
| January | 3.39 | 2.83 | 2.42 | 2.31 | 4.37 | 3.53 | 6.42 | 5.17 |
| February | 4.17 | 3.65 | 3.37 | 3.15 | 4.48 | 4.14 | 6.83 | 6.77 |
| March | 3.98 | 3.46 | 3.38 | 3.01 | 4.91 | 4.79 | 6.87 | 6.30 |
| April | 4.14 | 3.48 | 3.80 | 3.00 | 4.80 | 3.29 | 5.37 | 4.38 |
| May | 3.16 | 2.33 | 2.80 | 2.36 | 3.84 | 2.87 | 4.11 | 3.43 |
| June | 2.99 | 2.71 | 2.57 | 2.47 | 3.26 | 2.45 | 3.64 | 2.99 |
| July | 2.19 | 2.48 | 2.41 | 1.97 | 3.33 | 2.20 | 4.03 | 2.62 |
| August | 3.09 | 2.80 | 2.90 | 2.67 | 3.09 | 3.00 | 3.66 | 3.28 |
| September | 3.03 | 2.82 | 2.87 | 2.58 | 3.94 | 3.14 | 4.49 | 3.86 |
| October | 3.18 | 3.63 | 3.12 | 3.67 | 4.10 | 4.28 | 5.36 | 5.71 |
| November | 2.98 | 3.32 | 2.20 | 2.94 | 3.86 | 3.62 | 6.12 | 5.90 |
| December | 3.39 | 4.07 | 2.75 | 3.47 | 4.29 | 4.39 | 6.00 | 5.81 |

might be partially lost if standard $k$–fold CV were used. On the other hand, error estimates over an entire year should be robust and significant enough.

### 4.2 Lasso and SVR Results

Recall that we will predict PV energy for an hour $H$ using four variables at each grid point from hours $H, H-1, H-2$ and $H-3$, which we will indicate with the indices 0, 1, 2 and 3. In other words, we consider four problems according to the distance between the hour for which PV energy is sought and the hour from which readings are used. While in principle we could use a single model for all day hours, notice that this could work fine for the $H$ problem, but it will deteriorate for the other problems, as similar satellite readings might have to predict higher hour–ahead PV energy values in the morning but lower values in the afternoon. Because of this, we will build for all these problems three submodels adjusted to different day hours which roughly approximate morning (M), noon (N) and evening (E). For the models that relate radiances at hour $H$ with PV energy at the same hour, we will use the following hour subsets and notation:

- Model `m0M` for hours `[5, 6, 7, 8, 9, 10]`;
- Model `m0N` for hours `[11, 12, 13, 14]`;
- Model `m0E` for hours `[15, 16, 17, 18, 19, 20]`.

For the other models relating $H-1$, $H-2$ and $H-3$ radiances with hour $H$ PV energy, models `m1N, m2N, m3N` and `m1E, m2E, m3E`, will use the same hour subsets of models `m0N` and `m0E` respectively, while model `m1M` will use hours `[6, 7, 8, 9, 10]`, `m2M` hours `[7, 8, 9, 10]` and `m3M` hours `[8, 9, 10]`.

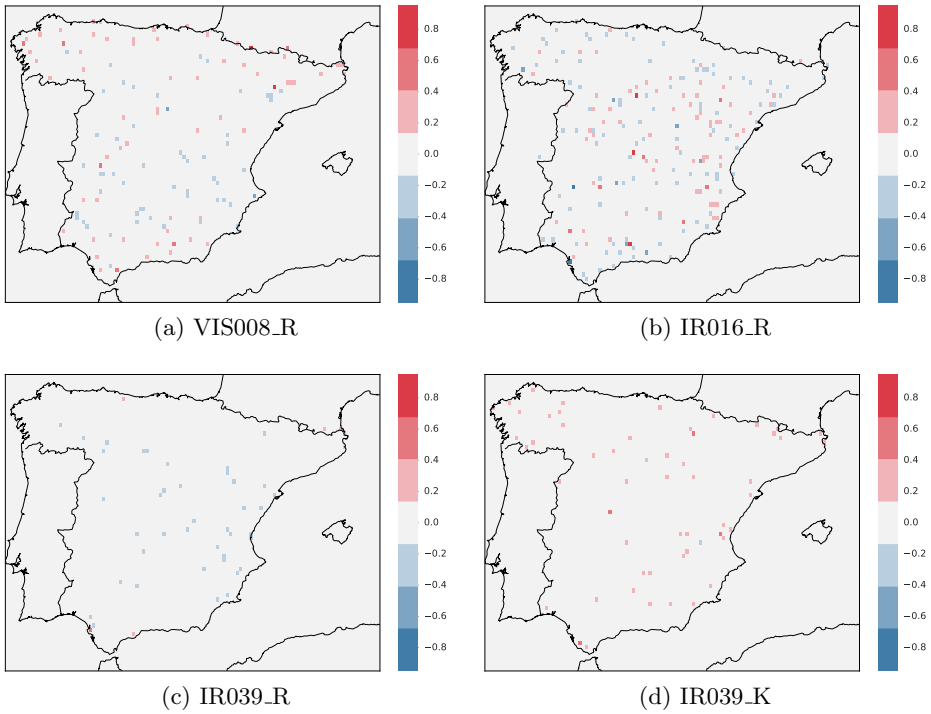
We normalize each feature to 0 mean and 1 standard deviation. While this is not actually needed for Lasso, as it is a linear model, it is crucial (and customarily done) for Gaussian SVR, to control for large value effects over the Gaussian kernel. The hyper–parameters for the twelve models we have used are in Table 1; notice the fairly big $C$ and small $\epsilon$ values, that suggest a tight, small training error model. Also, the $\gamma$ values are close to SVR's default $1/d = 7.37 \times 10^{-5}$.

**Table 3.** Average SVR test errors of `m1` models per hour and month.

| | Hour | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Month | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| January | 0.00 | 0.43 | 0.39 | 1.61 | 2.68 | 3.49 | 3.60 | 3.64 | 3.71 | 2.82 | 2.34 | 0.96 | 0.19 | 0.00 | 0.00 |
| February | 0.01 | 0.34 | 1.01 | 2.80 | 3.29 | 4.78 | 3.96 | 4.25 | 5.80 | 4.65 | 4.35 | 2.16 | 0.76 | 0.01 | 0.00 |
| March | 0.24 | 0.57 | 2.56 | 3.25 | 3.43 | 4.04 | 4.02 | 4.19 | 4.08 | 4.03 | 2.96 | 2.13 | 1.59 | 0.23 | 0.00 |
| April | 0.47 | 1.57 | 2.06 | 3.09 | 3.98 | 3.81 | 3.90 | 4.05 | 3.35 | 5.03 | 3.87 | 2.71 | 2.13 | 1.11 | 0.01 |
| May | 0.85 | 2.53 | 2.52 | 2.52 | 3.14 | 2.62 | 2.50 | 2.30 | 2.41 | 2.30 | 2.54 | 2.65 | 2.46 | 2.20 | 0.37 |
| June | 1.33 | 3.59 | 3.46 | 2.70 | 1.80 | 1.79 | 1.48 | 1.36 | 1.66 | 2.29 | 3.00 | 4.31 | 4.47 | 3.07 | 0.67 |
| July | 0.96 | 2.07 | 1.88 | 2.19 | 1.72 | 2.05 | 1.21 | 1.23 | 1.12 | 1.68 | 2.10 | 3.03 | 3.45 | 3.05 | 0.82 |
| August | 0.61 | 2.40 | 3.01 | 2.69 | 2.04 | 2.17 | 2.09 | 2.18 | 2.55 | 2.57 | 4.05 | 4.71 | 4.23 | 1.99 | 0.48 |
| September | 0.24 | 1.49 | 2.04 | 2.78 | 2.78 | 2.71 | 2.70 | 2.51 | 2.85 | 3.31 | 3.51 | 3.12 | 2.17 | 0.45 | 0.00 |
| October | 0.44 | 0.56 | 2.51 | 3.69 | 4.45 | 4.31 | 4.40 | 5.15 | 5.78 | 4.20 | 3.30 | 2.01 | 0.64 | 0.00 | 0.00 |
| November | 0.01 | 0.40 | 0.86 | 2.56 | 3.55 | 4.66 | 4.76 | 4.87 | 4.86 | 3.04 | 2.54 | 0.80 | 0.04 | 0.00 | 0.00 |
| December | 0.00 | 0.41 | 0.47 | 2.25 | 3.00 | 7.26 | 5.20 | 5.29 | 6.05 | 2.99 | 1.79 | 0.73 | 0.00 | 0.00 | 0.00 |
| Average | 0.43 | 1.36 | 1.90 | 2.68 | 2.99 | 3.64 | 3.32 | 3.42 | 3.68 | 3.24 | 3.03 | 2.44 | 1.84 | 1.01 | 0.20 |

To achieve a more homogeneous comparison across problems, we will report average errors for UTC hours between 8 and 20 (i.e., those considered for the $H-3$ problem). These are given for Lasso and SVR in Table 2; for each month we compute errors only on its daylight hours to discard the trivial prediction of no energy at night hours. As it can be seen, the best results for both approaches are achieved for the more stable mid year months and the $H-1$ problem. While at first sight one should expect similar results for the $H$ problem (that we took as a control problem, as it has no practical interest), this is not the case, as its errors are closer to those of the $H-2$ problem. This is most likely due to the fact that before noon radiance at hour $H$ overestimates energy production ending at hour $H$, while underestimates it after noon; similarly, radiance at hour $H$ overestimates radiance at hour ending at $H+2$. (PV energy readings for Peninsular Spain are available within the 10 minutes following each hour.) Worst errors are obtained, as expected, for the $H-3$ problem, but they are too heavily influenced by large errors of the submodel `m3M`, that has to predict PV readings late in the morning from radiances near sunrise, something that is obviously very difficult for winter days, with very small irradiances at dawn hours and where the model prediction essentially reduces to the bias term. In general, the more powerful SVR models improve on the Lasso predictions. Table 3 gives the average test errors per hour and month of the `m1` SVR models (the zero errors in early and late December and January correspond to night hours).

As mentioned, both ML approaches lend themselves to further interpretation. In Figure 3 we show the Lasso coefficients for each channel used by the `m1N` submodel (i.e., noon hours in the $H-1$ problem) as a heat map, where dark blue means large negative coefficients and dark red large positive ones. Since we normalize each feature to 0 mean and 1 standard deviation, the linear coefficient values give a rough idea of the variable influence. For better visualization all figures have the same colormap scale. First, as expected, Lasso imposes sparsity on the models, as clearly non-zero coefficients only appear at few grid points.




(a) VIS008_R (b) IR016_R

(c) IR039_R (d) IR039_K

**Fig. 3.** Lasso coeffcients for the `m1N` submodel.

Notice also that there are more non-zero coefficients for `IR016` radiances which also attain the largest scale values; the situation is more or less the opposite for the `IR039` ones. In any case, this should be studied further. For instance there are quite a few coefficients for grid points in northern Spain, which has a much smaller number of PV plants; moreover compensating effects among variables seem to appear as there are quite a few negative coefficients, some of them with large values and often quite close to the locations of the positive coefficients.

Recall that we can interpret SVR coefficients from a time perspective looking at how many patterns at the possible day-hour pairs are taken as Support Vectors (SV) by the different `m1, m2` and `m3` models. However no sample sparsity is achieved in this way; for instance, out of a maximum of 5,475 possible SVs for the `m1` models, a total of 4,977 have been selected and similar values appear for the `m2` and `m3` models. This is opposite to the Lasso models where, for instance, the `m1N` model has 874 non-zero coefficients while the number of Lasso variables is 13,564, i.e., 4 variables $\times$ 3,391 grid points. Note, however, that the data matrix has a maximum rank of 1,460, i.e., 4 noon hours $\times$ 365 days, which is much closer to the number of non-zero Lasso coefficients.

## 5 Discussion and Conclusions

In this work we have studied the use of Meteosat irradiances to nowcast the PV energy production of Penisular Spain from a Machine Learning point of view. More precisely, we have applied two well known ML methods, Lasso and Support Vector Regression, to predict energy production at hour $H$ from satellite readings at hours $H, H-1, H-2$ and $H-3$. The best forecasts for both models are those from readings at hour $H-1$, followed more or less evenly by those at hours $H$ and $H-2$ and with those at hour $H-3$ being the worst; moreover, SVR forecasts were better than Lasso ones. These results were to be expected except, perhaps, the poorer results of the $H$ readings although, as mentioned, being the readings irradiances at the hour, they overshoot energies in the first half of the day and undershoot them in the second half. More important may be the concrete error values, although they are difficult to compare, as PV energy is highly dependent on geography, and errors for Spain simply will not be comparable with those given in the literature for, say, Germany; the concrete temporal periods studied may also have a noticeable influence.

In any case, there are several ways to improve on the results presented. To begin with, a better choice than the single readings at hour $H$ is given by the averages of the irradiance readings at the four 15' periods that end at that hour (these readings are also provided by EUMETSAT). We can also adjust more closely the Meteosat grid points to the Spanish areas with substantial PV capacity and another clear improvement is to better adapt the training data to the forecast period to take account of seasonal effects. For instance, in [2] the forecast for month $M$ of year $Y$ was obtained using as training data months $M-1, M-2$ of the same year and months $M, M+1$ and $M+2$ of the previous year. This results in month-adjusted models and should give better results than our full year approach, which should have some difficulties balancing the very different solar regimes of, say, summer and winter. Finally, methods such as Heliosat give modelling alternatives and other sources of information can be put to play. For instance, one can contemplate energy nowcasting as a correction of previous energy forecasts, for which day ahead NWP irradiance forecasts and past PV energy readings can be put to good use, as shown in [2]. We are currently pursuing these and other related ideas.

#### Acknowledgments

With partial support from Spain's grants TIN2013-42351-P (MINECO), the UAM-ADIC Chair for Data Science and Machine Learning and S2013/ICE-2845 CASI-CAM-CM (Comunidad de Madrid). The first author is kindly supported by the UAM-ADIC Chair for Data Science and Machine Learning and the second author by the FPU-MEC grant AP-2012-5163. We gratefully acknowledge the use of the facilities of Centro de Computación Científica (CCC) at UAM and thank Red Eléctrica de España for kindly supplying PV energy data.

## References


1. Antonanzas, J., Osorio, N., Escobar, R., Urraca, R., de Pison, F.M., Antonanzas-Torres, F.: Review of photovoltaic power forecasting. Solar Energy 136, 78 – 111 (2016)
2. Fernández-Pascual, Á., Gala, Y., Dorronsoro, J.R.: Machine Learning Prediction of Large Area Photovoltaic Energy Production. In: Data Analytics for Renewable Energy Integration DARE 2014. ECML PKDD Workshop. pp. 38–53 (2014)
3. Fonseca, J.G.S., Oozeki, T., Takashima, T., Koshimizu, G., Uchida, Y., Ogimoto, K.: Photovoltaic power production forecasts with support vector regression: A study on the forecast horizon. In: Photovoltaic Specialists Conference (PVSC), 2011 37th IEEE. pp. 002579–002583 (2011)
4. Hammer, A., Heinemann, D., Hoyer, C., Kuhlemann, R., Lorenz, E., Müller, R., Beyer, H.G.: Solar energy assessment using remote sensing technologies. Remote Sensing of Environment 86(3), 423–432 (Aug 2003)
5. Hastie, T., Tibshirani, R., Friedman, J.: The elements of statistical learning. Springer (2009)
6. Inman, R., Pedro, H., Coimbra, C.: Solar forecasting methods for renewable energy integration. Progress in energy and combustion science 39(6), 533 – 576 (2013)
7. Kühnert, J., Lorenz, E., Heinemann, D.: Satellite-based irradiance and power forecasting for the german energy market. In: Kleissl, J. (ed.) Solar Energy Forecasting and Resource Assessment. pp. 267–297. Academic Press (2013)
8. Kleissl, J.: Solar Energy Forecasting and Resource Assessment. Academic Press (2013)
9. Lorenz, E., Kühnert, J., Wolff, B., Hammer, A., Kramer, O., Heinemann, D.: PV power predictions on different spatial and temporal scales integrating PV measurements, satellite data and numerical weather predictions. In: Proceedings of the 29–th European Photovoltaic Solar Energy Conference and Exhibition (EU PVSEC). pp. 22–26 (2014)
10. Marquez, R., Coimbra, C.F.: Intra-hour DNI forecasting based on cloud tracking image analysis. Solar Energy 91, 327 – 336 (2013)
11. Mohammed, A.A., Yaqub, W., Aung, Z.: Probabilistic forecasting of solar power: An ensemble learning approach. In: Neves-Silva, R., Jain, C.L., Howlett, J.R. (eds.) Intelligent Decision Technologies: Proceedings of the 7th KES International Conference on Intelligent Decision Technologies (KES-IDT 2015), pp. 449–458. Springer (2015)
12. Pedro, H.T., Coimbra, C.F.: Assessment of forecasting techniques for solar power production with no exogenous inputs. Solar Energy 86(7), 2017–2028 (2012)
13. Rana, M., Koprinska, I., Agelidis, V.G.: 2d-interval forecasts for solar power production. Solar Energy 122, 191 – 203 (2015)
14. Schölkopf, B., Smola, A.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press (2001)
15. Wolff, B., Kühnert, J., Lorenz, E., Kramer, O., Heinemann, D.: Comparing support vector regression for PV power forecasting to a physical modeling approach using measurement, numerical weather prediction, and cloud motion data. Solar Energy 135, 197 – 208 (2016)