Statistical Validation and Power Modelling of Hourly Profiles for a Large-Scale Photovoltaic Plant Portfolio

Statistical Validation and Power Modelling of Hourly Profiles for a Large-Scale Photovoltaic Plant Portfolio

Giuseppe Alba, Gianfranco Chicco, Alessandro Ciocia, Filippo Spertino
ENSIEL Consortium, Torino Branch – Dipartimento Energia "Galileo Ferraris", Politecnico di Torino
Corso Duca degli Abruzzi 24, Torino 10129, Italy
giuseppe.alba@studenti.polito.it, {gianfranco.chicco, alessandro.ciocia, filippo.spertino}@polito.it

***Abstract*—In the last decades, the number of photovoltaic (PV) generators significantly increased over the world. An accurate analysis of the massive data gathered from the meters of the PV plants can improve the management of their intermittent generation. This paper presents a methodology to analyse the production profiles of a portfolio of thousands of PV plants installed in an Italian region. The procedure faces the problem of filtering poor and incomplete data, then uses a stratified sampling technique for the statistical validation of the remaining profiles. Finally, the checked production profiles are used to adjust the energy model to better match the measured data and calculate the whole PV portfolio production.**

Keywords— Photovoltaic generators, photovoltaic plant portfolio, stratified sampling, PV modelling, optimization.

I. Introduction

The increase in photovoltaic (PV) module efficiency, the progressive reduction of their price, and the increasingly higher targets for renewable energy deployment, make the PV technology one of the most installed in the last years, with a new world capacity of more than 100 GW in 2020 [1]. In the last years, also the diffusion of metering systems increased, and the operators could access an ever-increasing amount of measured PV generation power profiles. The measured data can be used for different purposes, among which performance analysis [2], improvement of energy modelling, development of methods to better increase the penetration of renewable energy in the grid [3], forecasts [4], etc. A key issue is the extraction and elaboration of big datasets of power profiles, with their statistical validation. In fact, the PV profiles can be incomplete or affected by bad data, and the search for the main causes of bad data is a topic that is gaining interest [5].

The proposed procedure has the goal to analyse and simulate measured production profiles of a portfolio of thousands of PV plants. It is the case of countries with high penetration of distributed generation, such as Italy, the methodology applies to PV systems with different sizes and characteristics, but the available information is limited to the position of the plants (coordinates), the nominal power of the PV generator, the technology, and the hourly PV generation profiles measured by the meter of the Distribution System Operator (DSO).

Considering that it is not possible to check the status of each plant in a portfolio of several thousands of plants, the procedure isolates the wrong profiles due to failure in the monitoring infrastructure or incorrect operation of the plant. In addition, the procedure faces the issue related to the missing detailed information about the plants, typical in the management of such large portfolios (i.e., tilt and azimuth of the PV modules, data from local irradiance and temperature sensors, datasheet of PV modules and converters). Thus, in the first step of the procedure (Section II), the definition of the minimum requirement in terms of information about the PV generators is discussed. The procedure continues with a multicriteria filtering of the data to remove profiles with errors (due to issues in the measurement or in data transmission and storage), or profiles of PV plants with non-adequate performance. Section III discusses the modelling of the PV generation by a straightforward model, and the parameters that could be adjusted to better match the generation profiles. In Section IV, the second step of the procedure is presented in detail: it consists of the stratified sampling method applied to the portfolio of PV plants. It results in the partition in subgroups of plants for the statistical validation of the generation profiles. In Section V, a case study with the application of the procedure are presented. The last section contains the conclusions.

II. Analysis and Filtering of PV Production Profiles

A. STEP#A – PV plant general information

To analyse the production profiles of a PV plant, geographical location is essential to obtain meteorological data, for example from free GIS databases [6,7]. The tilt and azimuth of the PV modules are necessary for the calculation of the irradiance on their surface; these values can be obtained by inspections, or by viewing the layout of the plants. If the documentation of the PV plants is not available, and the inspections are not possible, a common method to check the general construction information is the use of satellite or street images [8]. In addition, the starting operation date is necessary for assessing degradation losses [9]. Nominal power is essential for the calculation and check of the measured production. It is also the variable used to classify the plants. Regarding the technology, this information permits to define the values of some parameters in the production model (this aspect is clarified in Section IV).

In the present work, the proposed procedure and analysis consider fixed PV plants of any size, both ground-installed and building-integrated. Only plants with tracking systems, concentrators, and storage systems, that are 1% in Italy, and require specific models for their simulation, are excluded.

B. STEP#B – PV plant production data filtering

The energy profile of a PV system usually comes from the electricity meter installed between the DC/AC converter and the point of connection with the grid, as shown in Fig. 1. Regarding the time step, generally profiles from monitoring systems have 1 hour resolution [10]. In fact, in most applications, hourly profiles are easier to obtain and elaborate, involving lower computational cost with respect to data with shorter time steps, e.g., from seconds to minutes [11].

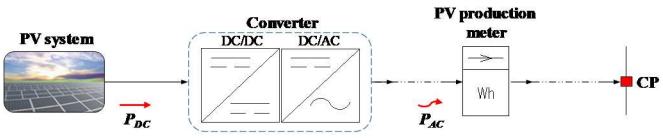

Fig. 1. Grid-connected PV system architecture.

Production profiles obtained by meters of DSOs can be affected by errors in data measurement, transmission, and storage. Also plant shutdown due to maintenance operations and failures must be considered. The resulting profiles, with missing days, weeks, or months of production lead to the underestimation of the yearly performance of the plant, that is one of the criteria for data filtering. Taking into account the above-described issues, four filtering criteria are proposed to remove wrong profiles:

- Step#α - Night production percentage: it may happen to observe abnormal production at night, which is obviously physically impossible, as the possible presence of storage would need separate metering. In this step, the ratio μ between night energy production (occurring during the whole year, between 11 p.m. and 4 a.m.) and the total annual production is calculated, removing only plants that exceed a threshold limit. This limit μ_{max} has to be selected according to the scope of the analysis: the most stringent limit ($\mu_{max} = 0$) leads to the removal of all the plants with at least a single measurement (error) during night hours.
- Step #β - Zero-production days: in some cases, production data could be partial. "Holes" in the profiles (days, weeks or entire months) can be due to failure of the plants, maintenance, or failure of the monitoring infrastructure. In the presented procedure, plants with missing data are removed to avoid wrong contributions in the validation of the energy production. Otherwise, the failure of the monitoring infrastructure could mistakenly lead to underestimate the performance of the plants.
- Step#γ - Typical territorial range production: A plant with optimal installation in a specific location has an annual productivity [kWh/kWp/year] with small interannual variation (generally in the range ±4%). This step eliminates plants that produce beyond the lower and upper bounds limits based on productivity maps, such as those available in the PVGIS database [6]. Obviously, plants are often installed in non-optimal conditions. For example, in building applied plants, modules are installed at the tilt and azimuth of the roof of the building. For these reasons, the limits must be properly selected to exclude only the plants with atypical annual production. For example, in Italy the PV production typically ranges between 1100 and 1600 kWh/kWp/year, typical for Northern and Southern Italy, respectively. In this case, plants with production <900 kWh/kWp/year are assumed to be not well working, or in any case they do not represent a noteworthy case. The maximum limit is restricted by the solar radiation source: plants with production > 1700 kWh/kWp/year are presumably affected by measurement errors.
- Step#δ - Depending on the number of remaining plants, an additional step could be performed. In case of a reduced number of plants, and in the absence of additional information, the manual inspection of satellite images of the plants can be performed. Thus, Step#δ consists of the check of actual installation conditions of the remaining plants. The plants that do not represent a noteworthy case (for example, are affected by shadows from near obstacles or incorrect design), can be removed.

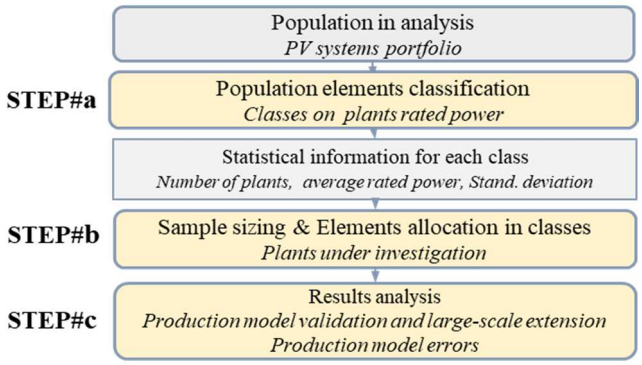

Fig. 2. PV portfolio analysis with stratified sampling technique.

Fig. 3 shows an example of a proper generation profile from a PV plant with rated power of 40 kW (a week in June 2018). There is no generation during night hours or missing data. The shape of production of the sunny days does not show the presence of obstacles (as checked by satellite images); the annual production of the plant (≈1400 kWh/kW) is in the correct range. The same considerations can be applied to the annual production profiles of a 12 kW PV plant (Fig. 4).

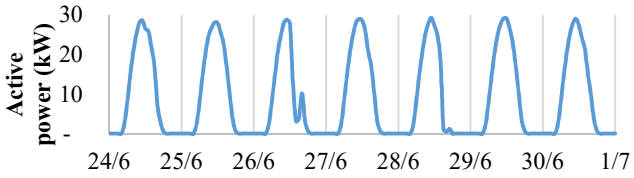

Fig. 3. Example of PV generation profile (40 kW plant, one week).

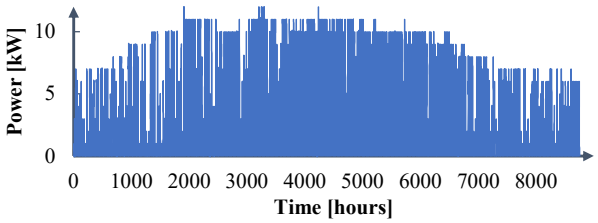

Fig. 4. Example of PV generation profile (12 kW plant, one year).

III. PV Modelling: Calculation and Optimisation

A. Modelling of PV energy production

The production of a generic grid connected PV plant (as shown in Fig. 1), can be calculated by considering a model proportional to the rated power of the generator P_{PV} defined at Standard Test Conditions (STC, G_{STC}= 1000 W/m^2, T_{STC} = 25°C) and irradiance G. As a reasonable compromise between simplicity and accuracy for a large-scale population of PV plants, this straightforward (STR) model considers the effect of temperature and several loss parameters [12]:

$$P_{AC} = P_{PV} \cdot \frac{G-G_0}{G_{STC}} \cdot C_T \cdot \eta_G \cdot \eta_{DC/AC} \qquad (1)$$

- $(G-G_0)/G_{STC}$: is the modelling used to take into account the nonlinearity effects of the semiconductor as irradiance changes. The low irradiance limit G_0 is the value below which the PV modules do not produce; it is in the range 10÷50 W/m^2 [13].
- $C_T = 1+\gamma_T(T-T_{STC})$: is the thermal factor that describes the linear dependence of production on cell temperature against STC conditions; γ_T is the thermal coefficient of power and is in the range -(0.3÷0.5) %/°C for the crystalline silicon (c-Si) technology [14].

- $\eta_G = \eta_{life}\,\eta_{dirt}\,\eta_{refl}\,\eta_{mis}\,\eta_{cable}$: the overall performance η_G of the DC side of the PV plant is calculated as the product of efficiencies taking into account different sources of losses. Production decreases due to dirt deposited on the glass of the modules η_{dirt}, the mismatch phenomenon of the current-voltage characteristics η_{mis}, reflection η_{refl} and joule effect losses in cables η_{cable}. The loss by ageing $\eta_{life} = 1-\gamma_{life}\cdot n$ depends on the age of plant n (years) and the annual loss coefficient γ_{life} (typically, -0.8%/year for c-Si) [15].
- $\eta_{DC/AC}$: the DC/AC efficiency is a nonlinear term that takes into account the overall performance of DC/AC converter including the tracking of the maximum power point [15]. In commercial devices, the maximum efficiency exceeds 98%, with much lower performance when the device works at low power levels [16].

The module temperature T is calculated with (2) as a function of air temperature T_a, irradiance G and the nominal operating cell temperature (*NOCT*). This value, from manufactures of the modules, is 42–50 °C for c-Si [17]:

$$T = T_a + \frac{NOCT - 20°C}{800\ {}^{W}/_{m^2}} \cdot G \tag{2}$$

B. Adjustment of the model to match actual PV production

The goal of the present work is to calculate the energy production of a whole large portfolio of installed plants. Thus, the STR model has to be adjusted, because it is created for generic plants. Nevertheless, actual plants are affected by sources of losses not considered in the STR: e.g., poor quality modules leading to high mismatch losses, shadows from near constructions that reduce the production especially in winter, or high thermal losses (e.g., in case of building integrated PV plants). Moreover, in case of missing detailed information about the generators, such as the exact tilt and azimuth, the STR model could not be able to calculate the energy production with an acceptable accuracy. For this reason, the STR model is upgraded by optimising some of the parameters and introducing an adaptation coefficient C_A.

The optimisation is performed by defining the set of parameters $\mathbf{x} = (\gamma_{T\%},\ G_0,\ C_A)$ and searching for their best set to match the measured profiles. The optimisation is written in the classical form, where $f(\mathbf{x})$ is the objective function, while **lb** and **ub** are the lower and upper limits of the parameters, respectively:

$$\min f(\mathbf{x}) \qquad \text{subject to}: \ \mathbf{lb} \le \mathbf{x} \le \mathbf{ub} \tag{3}$$

By varying G_0, the efficiency of the PV modules changes in case of low irradiance, especially early in the morning and at late afternoon. On the contrary, γ_{TH} affects the production particularly at midday, when the temperature is the highest. The adaptation coefficient C_A is multiplied by the rated power of the PV generator and takes in consideration all the uncertainties that are not characterised by the previous terms. In the absence of optimisation, the default is $C_A = 1$.

The quality of the matching is obtained by minimising the Standard Deviation (SD) between measured and simulated profiles on hourly level. The SD is weighted on the nominal power $P_{PV,j}$ of each plant in the sample under analysis. This is calculated for all the portfolio of plants, or for subsets defined in Section IV. In (4), j is the plant, k is the time step, and $P^j_{AC,k}$ and $P^j_{m,k}$ are the calculated and measured hourly average powers, respectively:

$$f(\mathbf{x}) = \sum_{j=1}^{J} \frac{\sqrt{\frac{1}{\Delta T}\sum_{k=1}^{K}\left(P^j_{AC,k} - P^j_{m,k}\right)^2}}{P_{PV,j}} \tag{4}$$

IV. Stratified Sampling of the PV Portfolio

In case of large portfolios of PV systems, the surveying of data and its processing can be computationally expensive. Inferential statistical methods allow to generalise the results of a whole population with an acceptable degree of confidence, by analysing a selected sample of elements. In this work, in which the population under analysis is a large portfolio of PV systems, the Neyman's Stratified Sampling (SS) technique [18] is applied. The variable used for the grouping is the nominal size of the PV generators.

A. STEP#a – Population elements classification

The SS is applied to the whole population of PV plants. The basic requirement of the SS is the creation of homogeneous classes (or subgroups), with any variable that allows elements with similar characteristics to be grouped in each class. The hypothesis is that each class can be represented by the Gaussian Probability Distribution (GPD). This makes it possible to use some statistical properties of the GPD to simply select a subgroup of elements (PV plants) to analyse and validate the whole portfolio.

In the case of a PV portfolio, the rated (or nominal) power may be selected as classification variable to create the classes, thanks to the direct relation with the energy production. The number of classes H and their size limits in the entire portfolio are defined as inputs. Then, statistical information is calculated for each class h = 1, …, H of the population composed of N elements*:* the number of plants N_h, the average rated power μ_h, and its standard deviation σ_h.

It is noteworthy that, to improve the production estimation, the criterion for selecting n_h plants from the whole population N_h (for each class h) should be carefully selected. In case of PV plants, considering the almost direct relationship between nominal power and energy production, the proposed criterion is the selection of plants with nominal powers close to the average power of the class.

B. STEP#b – Sample sizing and elements allocation

After the division of the whole portfolio in classes, the next step is the analysis of a sample of plants with reliable data. In the best case, the sample is the whole portfolio; nevertheless, generally, the sample is smaller after the filtering (Section II). Then, the sample, containing n elements, is divided into H classes; the results is the calculation of n_h elements to statistically represent all the N_h plants of each h class.

To calculate n_h, the optimal allocation method used consists of minimising the standard deviation $\hat{\sigma}$ of the nominal powers of the plants in the sample; it is solved with the Lagrangian method, leading to:

$$n_h = n \cdot \frac{N_h \cdot \sigma_h \cdot \sqrt{\frac{N_h}{N_h - 1}}}{\sum_{h=1}^{H} N_h \cdot \sigma_h \cdot \sqrt{\frac{N_h}{N_h - 1}}} \tag{5}$$

$$min\ \hat{\sigma}^2 = min\ \frac{1}{N^2}\sum_{h=1}^{H}\left[\frac{N_h}{n_h}\cdot(N_h - n_h)\cdot \hat{\sigma}_h{}^2\right] \tag{6}$$

where $\hat{\sigma}_h$ are the standard deviations of the nominal powers of each class.

The link between the sample size and the accuracy of the estimation is given by using (5) and (6):

$$\hat{\sigma}^2 = \frac{1}{N^2 \cdot n}\left[\left(\sum_{h=1}^{H} N_h \cdot \hat{\sigma}_h \sqrt{\frac{N_h}{N_h - 1}}\right)^2 - n \cdot \sum_{h=1}^{H} N_h {\sigma_h}^2 \frac{1}{N_h - 1}\right] \quad (7)$$

It is assumed the hypothesis that the rated power values inside each class follow a Gaussian distribution. In this case, is possible to size the sample with n elements starting from a desired interval of confidence d:

$$d_{\%} = \frac{d}{\hat{\mu}} \cdot 100 = \frac{k \cdot \hat{\sigma}}{\hat{\mu}} \cdot 100 \quad (8)$$

where $\hat{\mu}$ is the average value of the nominal power of the PV plants in the sample, and k is the coverage factor associated with the distribution based on various levels of confidence. The interval of confidence d is the desired error on the estimation (for example, ±5%).

In this work, dimensioning the sample with (7) can be useful for carrying out surveys after validation of the model, such as the calculation of the production of the whole portfolio of PV plants. It is noteworthy that the sample sizing and the achievable confidence interval are limited by the data quality. The Gaussian distribution itself is a theoretical assumption that enables simplifications in the representation, while the real data could be distributed differently, for example with a prevailing set of rated values. After the filtering procedure, the remaining plants could not be sufficient to fully validate all the classes. From the filtering step, defining the available systems for each class as n_{vh}, the condition that allows to validate a class is $n_{vh} \geq n_h$.

C. STEP#c – Analysis of the results

For the validation of the energy production calculated by the adjusted STR model, the errors can be calculated in different periods Δt of the year. For each class h of the sample, the average energy deviation is:

$$\Delta E_{\%_h}^{\Delta t} = \frac{1}{n_h} \cdot \sum_{i=1}^{n_h} \left|\Delta E_{\%_{i,h}}^{\Delta t}\right| \quad (9)$$

where $\Delta E_{\%_{i,h}}^{\Delta t}$ is the relative energy deviation between the energy calculated by the model $E_{i,h}$ and the measurement for a single i plant of each sample class.

After the validation of the model, the upscaling of the production of the sample can be performed, for estimating the production of the whole population of PV systems inside each class. The average energy produced by the PV plants belonging to the class h during the period Δt is calculated as:

$$\bar{\hat{E}}_h^{\Delta t} = \frac{1}{n_h} \cdot \sum_{i=1}^{n_h} E_{i,h}\Big|_{\Delta t} \quad (10)$$

The up-scaling of the results for each class is carried out by calculating the entire PV portfolio production $\hat{E}_{pop}^{\Delta t}$ by:

$$\hat{E}_{pop}^{\Delta t} = \sum_{h=1}^{H} N_h \bar{\hat{E}}_h^{\Delta t} \quad (11)$$

To evaluate the energy deviation on the total production of the photovoltaic portfolio, obtained after the up-scaling procedure, the total measured production $E^{\Delta t}{}_{pop}$ in the period Δt is used as the reference:

$$\Delta E_{\%_{pop}}^{\Delta t} = 100 \cdot \frac{\hat{E}_{pop}^{\Delta t} - E_{pop}^{\Delta t}}{E_{pop}^{\Delta t}} \quad (12)$$

Finally, the tolerance interval associated with the estimation is calculated using (8),

$$\hat{\sigma}_{pop}^2 = \sum_{h=1}^{H}\left[\frac{N_h}{n_h}(N_h - n_h) \cdot \frac{1}{n_h - 1}\sum_{i=1}^{n_h}\left(E_{i,h}^{\Delta t} - \bar{\hat{E}}_h^{\Delta t}\right)^2\right] \quad (13)$$

where

$$\hat{\mu} = \hat{E}_{pop}^{\Delta t} \quad (14)$$

$$\hat{\sigma} = \sqrt{\hat{\sigma}_{pop}^2} \quad (15)$$

The analysis will be statistically acceptable if the measured value $E_{pop}^{\Delta t}$ lies within the confidence band defined with (16) associated with the estimated value:

$$\hat{E}_{pop}^{\Delta t} \pm d \quad (16)$$

V. Case Study

The PV portfolio under analysis includes plants installed in Lazio, a region in central Italy. Currently, there are 54,323 plants in this territory: the total PV nominal power is ≈1.3 GW, that is, 6.5% of the national power. Only 1% of these plants has power higher than 200 kW_p, but they reach about 70% of the total installed power. Regarding the technology, 85% of the plants have polycrystalline silicon (p-Si) modules, while 14% have monocrystalline silicon (m-Si) modules. Only 0.5% of the plants has concentrators or tracking. The Italian Transmission System Operator (TSO) [19] provided the general information (site coordinates, nominal power, and technology) for all the plants (year 2018). The information related to the whole portfolio is used in the stratified sampling procedure. The hourly production profiles are provided for a part of this portfolio (9,096 plants); these profiles are used in the filtering procedure.

A. Plants production patterns in the filtering step

The 9,096 production profiles (year 2018, one-hour time step) have been analysed following the criteria described in Section IV. The filtering is essential for the model validation and removes most of the data, due to incomplete or unreliable profiles, as shown in Table I. The thresholds have been set in the most drastic case, namely, the PV plant is excluded when even a single data point during the year that does not satisfy the filtering conditions.

TABLE I. Data Filtering Procedure

Filter	Filter description	Removed plants	Remaining plants
A	Night production percentage: 0% between 23:00 and 04:00	2,188	6,908
β	Missing-production days: 0 days/year	6,808	100
γ	Typical territorial production outside 900 < kWh/kWp/y <1700	15	85
Δ	Check of installation conditions	12	73

With the first filter, 24% of the plants are eliminated: the procedure does not admit production during night hours. An example of a 200 kW_p plant with obvious errors in the annual profile is given in Fig. 5. This case is selected because it includes three errors: first, production is not zero during night hours; second, the maximum production (80 kW) is much lower than the rated power; finally, the daily profile is identical for each day of a single month. In most of the cases the anomalies are not so extended.

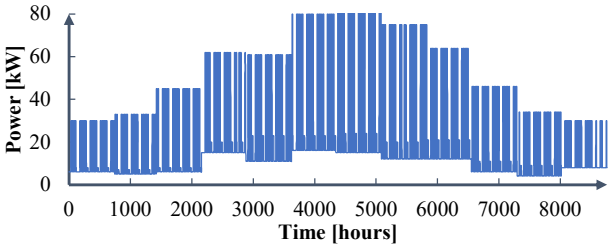

Fig. 5. Example of a PV profile with abnormal night production.

In the second step, 75% of the plants are removed from the database due to missing data. As anticipated, the most conservative criterion is applied: at least one day of missing production is sufficient to remove the plant. In this work, this criterion is not relaxed to avoid elaboration of profiles with missing data. An example of a plant with rated power of 440 kW with missing data is visible in Fig. 6. In this selected example the days of missing data are several. Moreover, there are maximum production peaks well beyond the nominal capacity of the plant. As in the previous example, this wrong profile is due to the failure of the monitoring infrastructure.

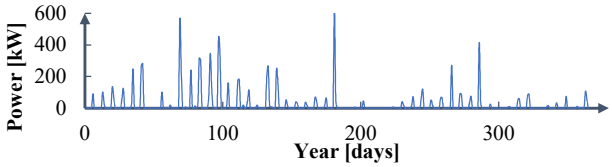

Fig. 6. Example of a PV profile with missing production days.

The third filter removes 15 plants with annual production much lower than the minimum limit of 900 kWh/(kW·year). Finally, the manual check of the satellite images leads to 12 plants affected by wrong installation condition, big obstacles close to the PV plant (e.g., trees and other buildings), etc.

B. Data classification and statistical analysis

The whole portfolio of 54,323 plants has been considered. For all these plants, the nominal power is known, and it is used as the classification variable. The limits of each class have been selected to allocate the largest number of plants in the middle of each range. Following this criterion, 11 classes have been defined, where, for each class h = 1, ..., 11 the rated power of the plant P_h is expressed in kW: $P_1 \leq 3.5$, $3.5 < P_2 \leq 6.5$, $6.5 < P_3 \leq 12.5$, $12.5 < P_4 \leq 25$, $25 < P_5 \leq 70$, $70 < P_6 \leq 120$, $120 < P_7 \leq 500$, $500 < P_8 \leq 1200$, $1200 < P_9 \leq 3600$, $3600 < P_{10} \leq 20000$, and $P_{11} > 20000$. The probability density function has been calculated for each class, under the Gaussian hypothesis for using it in the stratified sampling. This hypothesis just approximates what happens in practice, e.g., a peak in the distribution appears just below 100 kW_p, because PV plants with nominal power ≥ 100 kW_p obtained lower feed in tariffs, resulting less cost-effective [20].

The data about the PV plants in each class are shown in Table II. The power share $PS_{h\%}$ is the percentage weight of the class compared to the total installed power; the classes with the largest power shares include large plants, although they are less numerous. The validation of the sample of plants after the filtering (n=73) is carried out by a comparison made class by class. If the number of available plants n_{vh} is higher than the minimum n_h, the sample has statistical validity.

As shown in Table III, in the case study there are no sufficient plants for a complete validation; thus, validation is possible for six out of eleven classes. Regarding the tenth class, there is not a sufficient number of plants for the statistical validation, but it includes plants with the highest power share. Therefore, the next subparagraph presents also the results of the simulation performed in this class.

TABLE II. STATISTICAL INFORMATIONS FOR EACH POWER CLASS

Class	N_h (-)	PS_h (%)	μ_h (kW)	σ_h (%)
1	22,205	4.3	2.6	27.7
2	22,377	8.6	5.0	16.3
3	4,32	3.1	9.5	16.8
4	2,923	4.0	18	13.6
5	1,086	3.7	44	25.4
6	589	4.2	94	11.7
7	439	8.3	248	44.3
8	248	16.2	860	18.9
9	81	15.1	2,448	24.9
10	52	27.0	6,832	42.9
11	3	5.4	23,680	2.9

TABLE III. STATISTICAL VALIDATION FOR EACH POWER CLASS

Availability of plants with respect to stratified sampling allocation											
Class	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
n_{vh}	-	-	-	1	6	15	11	18	15	7	-
n_h	3	4	1	1	2	1	10	8	10	31	1
$n_{vh} \geq n_h$	no			Yes						no	

C. Model adjustment to match the measured profiles

The results of the optimization, i.e., the best sets of parameters for each class, are presented in Table IV. The starting values are $\mathbf{x}_0 = (\gamma_{T\%},\ G_0,\ C_A) = (-0.5\%/°C,\ 17\ W/m^2,\ 1)$ with boundaries set to be always far from the optimal value.

The increase of G_0 in every class demonstrates that the STR model needs an adjustment in case of low irradiance level, when the efficiency of PV modules falls. The coefficient C_A is higher than unity, so the degradation of the PV generators is lower than the values declared by the manufacturers. The coefficient γ_T lies in the range [-0.43; -0.51], in agreement with the literature.

TABLE IV. OPTIMIZED PARAMETERS

	Class						
Class	**4**	**5**	**6**	**7**	**8**	**9**	**10**
$\gamma_{T\%}$ [%/°C]	-0.46	-0.50	-0.50	-0.51	-0.54	-0.44	-0.42
G_0 [W/m²]	30	20	20	24	14	29	31
C_A	1.026	1.039	0.942	1.046	1.017	1.009	1.014

D. Simulation results

Fig. 7 and Fig. 8 show daily production profiles in a clear sky day and in a cloudy day, respectively. It is a 5 MW plant from the dataset of accepted profiles. The PV generation profiles calculated after the adjustment of the parameters is labelled "STR-adj". With respect to the standard STR model, STR-adj better matches the measured data, especially during sunny days. During cloudy days, both STR and STR-adj cannot match well the measured data, because the formula (1) was created for clear sky days, and the optimisation is performed on a yearly basis. In fact, with respect to the daily profile, the energy calculation at monthly and annual level is more accurate. Fig. 9 shows annual energy deviations between the models and the measurements for the 73 selected PV plants. STR-adj model improves production estimation.

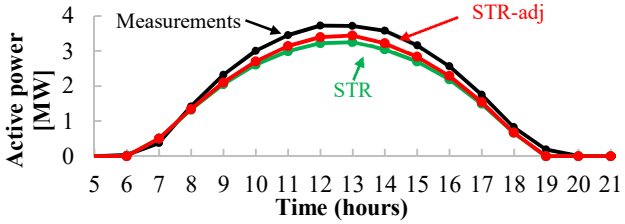

Fig. 7. Daily production profile in a sunny day of June.

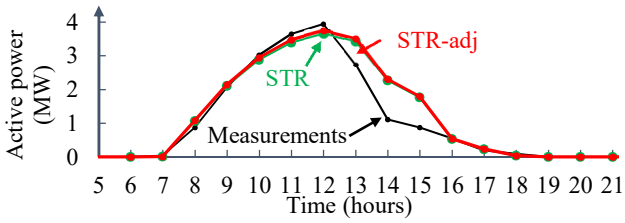

Fig. 8. Daily production profile in a cloudy day of March.

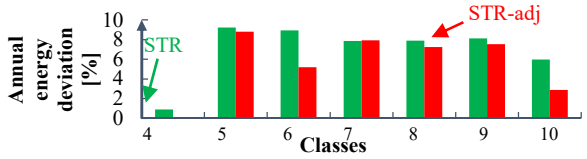

Fig. 9. Yearly energy deviation between calculated and measured energies.

Table V shows the annual production of the entire photovoltaic portfolio, obtained by extending the results with (10) on the basis of the average annual production for each class. The actual confidence intervals are calculated with (8) and $k = 2$. The deviations $\Delta E_{\%,s,pop}$ are less than 5.5%, that is, the expected value obtained from the sizing of the sample, thus demonstrating the effectiveness of the procedure. The model optimization reduces the deviation from -0.77% to -0.6%.

TABLE V. YEARLY PV PRODUCTION: ESTIMATION VS. ACTUAL DATA

	Measurements	**STR**	**STR-adj**
$E_{s,pop}$ [GWh]	1619	1606.7	1609.5
$d_{\%}$	-	3.64	3.63
$\Delta E_{\%,s,pop}$	-	-0.77	-0.60

VI. CONCLUSIONS

In this work, a procedure has been presented to analyse the generation profiles of thousands of PV plants gathered from the meters of the DSOs. The procedure is applied to a case study with a portfolio of 54,323 PV systems in Italy. The information related to the plants are minimal: the site, the nominal power, and the technology. For a subgroup of 9,096 plants the hourly power profiles are also present. First, the whole portfolio has been analysed by using the stratified sampling technique, based on the rated powers. The results are 11 subgroups of plants. Secondly, a multicriteria filtering has been applied based on the shape of the profiles and on the expected productivity. Starting from the 9,096 hourly profiles, more than 75% of them are not usable due to failures of the monitoring infrastructure, demonstrating the potential benefits in monitoring upgrades. At the end of the filtering, only 73 energy productions have been found that satisfy the proposed criteria and can be used for further analysis. Then, for each class, the statistical validation is performed by comparing the minimum number of required profiles and the results of the filtering procedure; only 6 classes out of 11 exhibit statistical validity. For these classes, the measured profiles have been compared with the results of an energy model, with and without the optimisation of its internal parameters. The energy results have been upscaled to the whole portfolio of PV plants, showing an energy deviation for the usual energy model of -0.77% and an even smaller deviation for the optimised model (-0.60%).

Future work will investigate the relaxation of the filtering criteria and the problem of elaboration of the profiles with missing data. It will also include the scalability of the filtering procedure, e.g., to avoid the manual check of satellite images. Future work will also consider other models for the calculation of the PV production, including different formulas for the PV module temperature. The optimisation will be performed on multi-steps for the different seasons, to better simulate the power production also in cloudy days.

REFERENCES

[1] IEA, International Energy Agency, Renewables 2020, Available online: https://www.iea.org/reports/renewables-2020.

[2] F. Bizzarri, M. Bongiorno, A. Brambilla, G. Gruosso, and G.S. Gajani, "Model of Photovoltaic Power Plants for Performance Analysis and Production Forecast," *IEEE Trans. on Sust. Energy*, vol. 4, no. 2, pp. 278-285, 2013.

[3] A. Ciocia, G. Chicco, and F. Spertino, "Benefits of On-Load Tap Changers Coordinated Operation for Voltage Control in Low Voltage Grids with High Photovoltaic Penetration," 2020 Intern. Conf. on Smart Energy Systems and Technologies (SEST), 2020.

[4] J. Liu et al., "Research of Photovoltaic Power Forecasting Based on Big Data and mRMR Feature Reduction," 2018 IEEE Power & Energy Society General Meeting (PESGM), 2018.

[5] M. Matam and J. Walters, "Data-integrity Checks and Balances in Monitoring of a Solar PV System," Proc. IEEE 46th Photovoltaic Specialists Conference (PVSC), Chicago, IL, 2019, pp. 1276–1281.

[6] Photovoltaic Geographical Information System (PVGIS). Available online: https://ec.europa.eu/jrc/en/pvgis (accessed on 15 May 2021).

[7] SODA Solar Radiation Data. Available online: http://www.soda-pro.com (accessed on 15 May 2021).

[8] Google Street View, Available online (Accessed on May 2021), https://en.wikipedia.org/wiki/Google_Street_View.

[9] Y.R. Golive et al., "Analysis of Field Degradation Rates Observed in All-India Survey of Photovoltaic Module Reliability 2018," *IEEE Journal of Photovoltaics*, vol. 10, no. 2, pp. 560-567, 2020.

[10] P. Ollas, J. Persson, C. Markusson, and U. Alfadel, "Impact of Battery Sizing on Self-Consumption, Self-Sufficiency and Peak Power Demand for a Low Energy Single-Family House With PV Production in Sweden," Proc. IEEE 7th World Conference on Photovoltaic Energy Conversion (WCPEC), Waikoloa Village, HI, pp. 0618–0623, 2018.

[11] A. Ciocia, A. Amato, P. Di Leo, S. Fichera, G. Malgaroli, F. Spertino, S. Tzanova, "Self-Consumption and Self-Sufficiency in Photovoltaic Systems: Effect of Grid Limitation and Storage Installation", *Energies*, vol. 14, no. 6, art. 1591, 2021.

[12] P. Di Leo, F. Spertino, S. Fichera, G. Malgaroli, and A. Ratclif, "Improvement of Self-Sufficiency for an Innovative Nearly Zero Energy Building by Photovoltaic Generators", Proc. IEEE PowerTech, Milano, Italy, 2019.

[13] A. Ciocia, J. Ahmad, G. Chicco, P. Di Leo, F. Spertino, "Optimal size of photovoltaic systems with storage for office and residential loads in the Italian net-billing scheme", Proc. of the 51st Intern. Universities Power Engineering Conference (UPEC), Coimbra, Portugal, 2016.

[14] S.B. Schujman, J.R. Mann, G. Dufresne, L.M. LaQue, C. Rice, J. Wax, D.J. Metacarpa, and P. Haldar, "Evaluation of protocols for temperature coefficient determination," Proc. IEEE 42nd Photovoltaic Specialist Conference (PVSC), New Orleans, LA, USA, June 2015.

[15] F. Spertino, E. Chiodo, A. Ciocia, G. Malgaroli, and A. Ratclif, "Maintenance Activity, Reliability, Availability, and Related Energy Losses in Ten Operating Photovoltaic Systems up to 1.8 MW," *IEEE Transactions on Industry Applications*, vol. 57, no. 1, pp. 83-93, 2021.

[16] F. Spertino, A. Amato, G. Casali, A. Ciocia, G. Malgaroli, "Reliability Analysis and Repair Activity for the Components of 350 kW Inverters in a Large Scale Grid-Connected Photovoltaic System," *Electronics*. Vl. 10, no. 5, art. 564, 2021.

[17] J.H. Bae, D.Y. Kim, J.W. Shin, S.E. Lee, and K.C. Kim, "Analysis on the Features of NOCT and NMOT Tests With Photovoltaic Module", *IEEE Access*, vol. 8, pp. 546-554, 2020.

[18] J. Neyman, "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection," *Journal of the Royal Statistical Society*, vol. 97, no. 4, pp. 558–625, 1934.

[19] Terna, Italian Transmission System Operator, Available online: www.terna.it/en

[20] V. Di Dio, S. Favuzza, D. La Cascia, F. Massaro and G. Zizzo, "The evolution of the FIT mechanism in Italy for PV systems: A critical analysis", Proc. 2013 International Conference on Renewable Energy Research and Applications (ICRERA), 2013, pp. 890–895.