e-Prime - Advances in Electrical Engineering, Electronics and Energy 3 (2023) 100120

Contents lists available at ScienceDirect

e-Prime - Advances in Electrical Engineering, Electronics and Energy

journal homepage: www.elsevier.com/locate/prime

On the uncertainty of estimating photovoltaic soiling using nearby soiling data

Leonardo Micheli [a,*], Matthew Muller [b]

[a] Department of Astronautical, Electrical and Energy Engineering (DIAEE), Sapienza University of Rome, Rome, Italy
[b] National Renewable Energy Laboratory, Golden, CO, USA

ARTICLE INFO

Keywords:
Photovoltaic systems
Soiling
Spatial interpolation
Nearest neighbor
Solar energy

ABSTRACT

The accumulation of dust on the surface of photovoltaic modules can reduce their performance and affect the cost competitiveness of this technology. This phenomenon is known as soiling and can be mitigated through appropriate corrective and/or preventive actions. In order to maximize its effectiveness, it is important to plan the soiling mitigation strategy even before the PV system is operational. This is typically done through a nearest neighbor approach, by estimating soiling using data from the nearest operational photovoltaic system. This work focuses on understanding the uncertainty related to this practice. For this purpose, the semi-variance function is used to study the dissimilarity between the soiling losses of two locations in California depending on their distance. The results show that, when the soiling loss at a nearby system is used to estimate soiling of a site, the uncertainty can be approximated to increase linearly at a rate of 0.08–0.10%/km up to 60 or 80 km. After this distance, the use of a nearest neighbor approach is no longer justified, as it produces an uncertainty as big as the average soiling loss of the sites in the dataset used in this study. In some conditions, uncertainties $> 0\%$ are found also for sites located within 25 km, meaning that even close-by systems might soil differently.

1. Introduction

The United Nations has included the deployment of affordable and clean energy among the Sustainable Development Goals for 2030 [1]. Photovoltaic (PV) technologies can directly convert the solar radiation into electricity and have experienced one of the most important growths among renewable energies in the last decade. Pushed by the low cost, the versatility and the easiness of installation, the global PV capacity has achieved a significant 1 TW milestone in early 2022, which is expected to double by 2025 [2].

In addition to installing new PV capacity, the road towards a more sustainable society and a cleaner energy market also requires maximizing the performance of the existing PV power plants. Indeed, it has been estimated that 3 to 7% of the annual global PV energy production is being lost because of *soiling*, a reversible performance issue affecting PV systems worldwide [3]. It consists of the accumulation of contaminants on the surface of PV modules, which decreases the amount of sunlight converted into electricity. This energy loss directly translates into revenue losses, which are even worsened by the increased operations and maintenance (O&M) costs needed to cover the required soiling monitoring and cleaning activities [4].

Soiling is commonly tackled by cleaning the PV modules. This activity, however, comes at a cost that can affect the profitability of the power plants. For this reason, it is important to optimize the cleaning schedule through appropriate soiling monitoring and/or estimation [5]. Soiling mitigation, however, it not just an O&M practice, but can start from the PV site selection and the PV system design [3]. Indeed, when a new PV site is selected, a correct prediction of soiling can help engineers to better design the system and O&M teams to plan an optimal cleaning schedule to mitigate these losses with minimum impact on costs [4]. Soiling of PV systems is generally monitored through the use of soiling stations. These compare the performance of a naturally soiled and a manually or automatically cleaned PV device [6]. Innovative optical sensors [7,8] have been recently launched in the market to reduce the costs of soiling monitoring and, compared to soiling stations, do not require any maintenance or reference clean cell to operate.

Because of the uneven distribution of soiling sources around the plant and of the wind patterns, some areas of a PV system might experience stronger soiling deposition than others [9,10]. In order to identify this nonuniform accumulation, more than one soiling sensor is typically recommended for systems of capacity $\geq$5 MW expected to experience soiling losses $> 2\%$. Despite that, in most cases, soiling is assumed to be

* Corresponding author.
E-mail address: leonardo.micheli@uniroma1.it (L. Micheli).

https://doi.org/10.1016/j.prime.2023.100120
Received 24 May 2022; Received in revised form 20 January 2023; Accepted 30 January 2023
Available online 2 February 2023

uniformly distributed across the strings, and a single soiling loss is considered for the whole PV plant.

Soiling extraction methods can be used as alternatives to sensors to estimate soiling trends directly from PV performance time series [11, 12]. These do not require any specific hardware, as they identify soiling accumulation and cleaning events from the analysis of power output, irradiance, and, when available, additional parameters such as temperature and rain (typically measured by default at utility-scale PV sites).

Nonetheless, when a PV site is not operational yet, monitoring data or PV performance time series are not available, and soiling needs to be estimated in other ways. In some cases, soiling sensors can be deployed at candidate PV locations before the site is selected and the PV plant is designed. However, a reliable campaign would require at least 6 months of data collection, in order to fully characterize the most soiled months, typically in summer. Because of this, and because of the costs associated with acquiring, deploying and operating stations and sensors, the assessment of soiling is conducted in most cases using the closest available data or using soiling models. For example, several authors have suggested to estimate soiling from particulate matter, precipitation and other environmental parameters [13–16] which are typically available in long-term datasets. However, these models have been generally developed using data from a limited number of sites and are specific to particular regions. Therefore, soiling at a potential PV site is still commonly estimated from either measured or extracted data available at nearby systems.

In order to provide soiling information that could be used in pre-feasibility studies, a soiling map has been presented in recent years by the National Renewable Energy Laboratory (NREL), collecting data extracted through referenced techniques from 42 PV systems and 41 soiling stations installed in the USA [17,18]. This map can be a useful tool to predict soiling losses at sites where soiling or PV data are not available. The study where it was presented showed that, by using spatial interpolation techniques, the data on the map can be used to estimate soiling at a location with coefficients of determination as high as 78% [18]. That study also showed that a simple method as the Nearest Neighbor, where the soiling loss of a site is assumed equal to that at the closest location with available data, can return good estimations if the two sites are within 50 km. Even if mentioned, the study did not detail how the soiling estimation quality varies in relation to the distance between two sites. In a different work, the correlations between monthly soiling accumulation rates, in %/month, at various sites and their distances were extracted [19]. The results are of interest, showing a decreasing and non-linear relation between the correlations of the soiling profiles and the distances of two sites. That work, which focused on monthly soiling rates while the present aims to investigate annualized soiling losses, did not provide any quantitative information on the relation that could be of use for the community. The aim of the present study is to further investigate the uncertainty in soiling estimation based on the Nearest Neighbor approach, by investigating the spatial correlation between soiling of nearby sites.

The results of this study can be of interest even when the aforementioned models are used to estimate annual, seasonal, and daily soiling losses from environmental variables. Indeed, even in these cases, experts might make use of data that are not locally sourced. For example, data of the MERRA-2 dataset are available on a $0.5^\circ \times 0.65^\circ$ grid, which corresponds to approximately 55 km × 72 km. Alternatively, even ground-monitors, such as the particulate matter sensors of the U.S. Environmental Protection Agency (EPA) network [20], can be distributed unevenly over the land and might be located several km away of the site of interest. Previous works have already shown that the distance between the PV system and the measurement location can impact the soiling estimation [9,19]. This work aims to identify the qualitative and quantitative correlation between uncertainty and distance between the training and the test locations.

For all the above-mentioned reasons, it is essential to understand the soiling estimation uncertainty introduced by non-locally-sourced data. In this light, this work analyzes, using a spatial-statistical method named semi-variogram, the profile of soiling dissimilarity between two sites as a function of their distance. The result of this work is an easy and reliable numerical tool that the community can use to calculate the uncertainty on the estimation of soiling.

2. Methodology

2.1. Soiling data

Soiling is commonly quantified using the soiling ratio ($r_{s,w}$), a metric expressing the ratio between the electrical output of a soiled PV device and that same output in clean conditions [21]. The soiling ratio, which corresponds to the difference between one and the fractional power loss due to soiling, has a value of 100% in clean conditions and tends to decrease with the accumulation of soiling. The methodologies used in the present work to extract the soiling ratio, here expressed in terms of daily ratios measured over the data collection period, are detailed in Ref. [18].

Of all the sites available on the soiling map [17], the present study considers only the 32 sites (14 soiling stations and 18 PV systems) installed in California, the most densely populated region on the map. The sites have soiling ratios ranging between 93% and 100%, with an average of 96%. The average nearest-neighbor distance is 39 km, while the furthest site is 94 km away from its neighbor.

The histogram in Fig. 1 shows the distribution of the distances between each pair of data points. It should be highlighted that each site is used in combination with all the other data points to generate the distance-dependent relation. So, the 32 sites produced 992 (32×31) data points. The average and median distances between sites in the dataset are 339 km and 300 km, with the closest sites being 5 km apart, and the maximum distance being 926 km instead.

2.2. Spatial analysis

This study is based on the analysis of a semi-variogram, a spatial-statistical function that measures the spatial autocorrelation between a pair of points located at a certain distance. It compares the similarity between pairs of points at a given distance and direction apart (the lag) and expresses mathematically the average rate of change of a property with separating distance. In simple words, it is assumed that two locations nearby are more likely to have similar features than farther apart locations. The difference between the values of a specific feature (the soiling loss, in this case) of two given locations is therefore depicted as a function that increases with the distance. This function reaches a maximum value at a certain distance, above which there is no longer spatial autocorrelation.

The semi-variogram is expressed as the half of the average of the

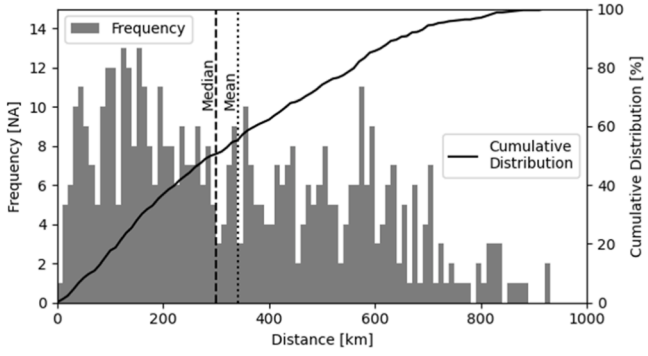

Fig. 1. Distribution of the distances between each pair of data points in the dataset. A 10 km bin size has been considered.

square differences between each pair of points separated by a distance h, and can be experimentally calculated by using the following expression:

$$\gamma(lag) = \frac{1}{2 \cdot N(lag)} \cdot \sum_{i=1}^{N(lag)} \left[r_{s,w,i1} - r_{s,w,i2} \right]^2 \quad (1)$$

where $r_{s,w,i1}$ and $r_{s,w,i2}$ are the soiling ratios of each pair of data points located at a distance within $lag - \frac{h}{2}$ and $lag + \frac{h}{2}$ of each other. N(lag) is the number of pairs of data points in this distance range. The lags are incremented at intervals of fixed dimension (*lag size*, expressed in km), whose value will be discussed in the next section. The semi-variance is calculated for each lag considering all the pairs of data points located at distances falling within the range $lag - \frac{h}{2}$ to $lag + \frac{h}{2}$. The $\pm\frac{h}{2}$ tolerance allows to use in the calculation all the available pairs of data points, and to employ each pair once and in only one lag. The first data point is calculated in the range from 0 km to *lag size*.

A semi-variogram of the Californian soiling data collected in the NREL soiling map is shown in Fig. 2. In this case, the data points, plotted as black markers, are calculated for lag sizes of 25 km, up to a maximum distance of 450 km, which corresponds to about half of the maximum distance between a pair of data points in the dataset. The 25 km lag size makes it possible to include at least 10 data points per lag.

The experimental semi-variograms can be modelled by using different equations [22]. In the present work, three common fitting functions have been considered, expressed by the equations reported in Table 1: Exponential, Gaussian, and Spherical, all plotted in Fig. 2. Each semi-variogram can be described through certain characteristics, labeled in the right plot of Fig. 2:

- Sill (c + c0): maximum value approached by the semi-variogram.
- Practical Range (a_{95}): the distance at which the function achieves 95% of the sill.
- Actual Range (a): the distance at which the function reaches the sill. This is not available for asymptotical functions (i.e., the exponential and the gaussian models).
- Nugget (c0): the value of the function when it intercepts the y-axis.

In this work, the curve fitting has been performed through the curve_fit function in the SciPy library for Python 3 [23]. The initial guesses at each minimization have been set as shown in Table 2, and only values ≥ 0 have been accepted for each variable.

The quality of the fit of each model has been assessed by calculating the coefficient of determination (R^2) and the root-mean-square-error (RMSE) of the experimental and modelled semi-variance. The coefficient of determination has a value of 100% if a linear relation, with no error, exists between the modelled values and the experimental values. The root-mean-square-error represents the square root of the quadratic difference between each pair of measured and modelled values, and it is expressed in the same units as the soiling ratio (%).

3. Results and discussion

3.1. Experimental semi-variogram

The equations describing the three functions used to model the semi-variograms in Fig. 2, as well as the values of each parameter, are reported in Table 1, along with the R^2 and RMSE describing the quality of the fit. The models return R^2 between 48 and 55% (and p-value < 0.05), with the best results achieved with the gaussian and spherical models.

All the models tend to a maximum value (*sill*, labeled as c + c0) of 4.3%. The spherical model reaches it at a distance of 72.5 km (*actual range*, labelled as a). The exponential and the gaussian models, which are based on asymptotical functions, reach 95% of the sill at a distance named *practical range*, here labelled as a_{95}. In order to consistently compare the models, the practical range of the spherical model, corresponding to the distance at which the model achieve 95% of the sill, has also been studied. All the models achieve the practical a_{95} at distances between 58 km and 77 km. This means that, if the closest data point is located further than this practical range, the estimation of soiling would be subject to the highest uncertainty and would return non-trustworthy results, as the uncertainty would be as big as the average soiling loss. For this reason, the application of other spatial interpolation methods should be preferred in these cases [18].

In addition, it is of interest to evaluate the value of the nugget for the various functions. It should be reinstated that the fitting procedure is allowed to return c0 ≥ 0 for all the functions. As it can be seen, however, while the exponential is fitted to start from zero, the gaussian and spherical experimental semi-variograms are fitted to start from a *nugget* (*c0*) of 1.0% and 0.2% respectively. This suggests that even two systems located close-by may soil differently, because of local tilling, harvesting or building activities, or because of the emissions of factories or other pollutant sources that might affect only PV modules located in proximity [9,19]. Moreover, different geometries or orientations of the modules within the same site can lead to different accumulations of soiling. Also,

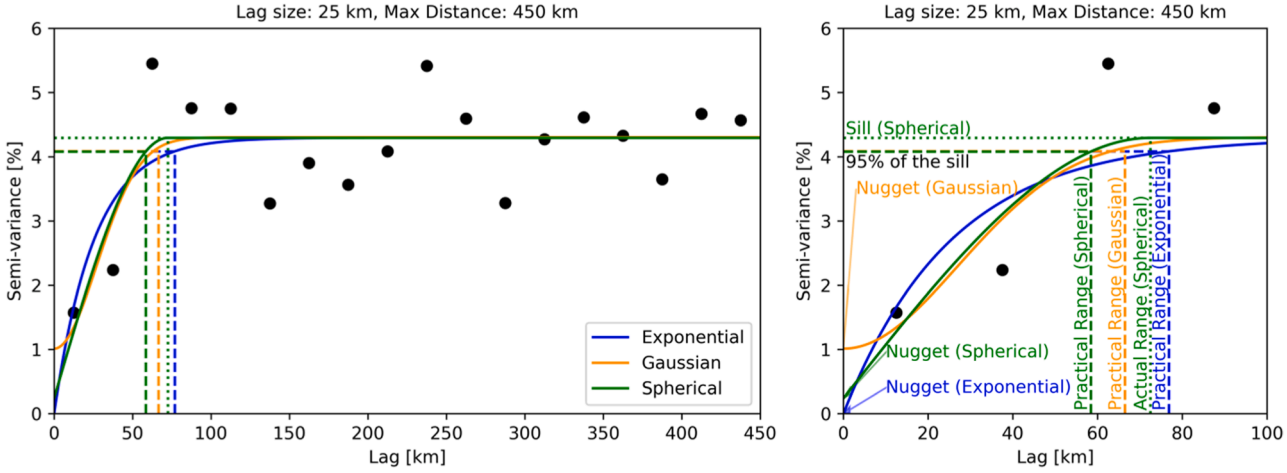

Fig. 2. Semi-variograms of the data points located in California collected on the NREL soiling map, obtained for lag intervals of 25 km and a maximum distance of 450 km (left plot). On the right, a zoom of the same semi-variograms is reported. The continuous lines represent the best fit returned by the three models. The dashed lines show the 95% of the sill (i.e., 95% of the maximum value approached by the function) and the practical range (i.e., distance at which the function reaches the 95% of the sill) for each model. The green dotted line shows the sill (i.e., maximum value achieved by the spherical model) and the actual range (i.e., distance at which the function reaches the sill) of the spherical model.

Table 1
Equations used to model the semi-variogram in Fig. 2, for lag size of 25 km, maximum distance of 450 km. The results of each model are also shown and are expressed through the actual range (a), the practical range (a_{95}), the nugget (c0), the sill ($c + c0$), the coefficient of determination (R2) and the root-mean-square-error (RMSE).

Model	Equation	a [km]	a_{95} [km]	c0 [%]	$c + c0$ [%]	R^2 [%]	RMSE [%]
Exponential	$c\cdot(1 - e^{-3\cdot h/a_{95}}) + c0$	N.A.	76.8	0.0	4.3	48.5	0.705
Gaussian	$c\cdot(1 - e^{-3\cdot h^2/a_{95}^2}) + c0$	N.A.	66.5	1.0	4.3	54.0	0.664
Spherical	$c\cdot\left(\frac{3\cdot h}{2\cdot a} - \frac{1}{2}\left(\frac{h}{a}\right)^3\right) + c0$ if $h \le a$ $c + c0$ if $h > a$	72.5	58.4	0.2	4.3	55.0	0.656

Table 2
Initial guesses for minimization of a, a_{95}, c and c0.

a [km] (Spherical) a_{95} [km] (Exponential, Gaussian)	c [%]	c0 [%]
Lag size	4.0	0.5

the nugget probably accounts for the uncertainty in extracting soiling and some potential non-uniform soiling distribution across the PV plants [9,19]. Therefore, system designers and O&M teams should always consider a minimum uncertainty when estimating soiling for a site even if using data from remarkably close and similar locations. The limited number of sites available for this study does not allow further investigate the spatial correlations of soiling at the shortest distances.

3.2. Limits and confidence interval

The reliability and the shape of the semi-variogram can be affected by a number of factors [22], such as the modeling method, the lag size, the maximum distance, and the number of points. All these factors are the focus of the present section where we aim to discuss how their variation might affect the semi-variogram.

First, it is worth discussing the modeling techniques used to fit the experimental variogram. The results reported in Table 1, obtained by modeling a semi-variogram produced for lags of 25 km and a maximum distance of 450 km, showed that the gaussian and the spherical models return the best curves, with the highest and similar R^2. The exponential model is instead the worst performing approach, because it seems to overestimate the semi-variance at shorter distances and to underestimate it at higher distances, returning the longest practical range among the three models. The exponential and the gaussian models return the same errors. The main difference between the two models is the profile at extremely low distances, with the gaussian model slowly rising at short distances.

In order to understand the reliability of the three models, the same analysis conducted previously has been repeated by using shorter maximum distance (100 km) and lag interval (20 km) to highlight the behaviors of the three curves at short distances. The results of the analysis are shown in Fig. 3 and reported in Table 3. The three models return good fits, with $R^2 > 97\%$. The spherical model is still found to be slightly better performing than the gaussian one. Therefore, all the results reported in the rest of this paper are those obtained using the spherical fitting.

In addition to that, it is interesting to note that, within the practical range, the shape of the semi-variograms in Figs. 2 and 3 can be approximated with a straight line (Fig. 4). The slope of the line quantifies the raise in uncertainty in the estimation of soiling depending on the distance between the investigated site and the known data point. If available, the community could use that slope to easily determine the uncertainty in the estimation of soiling when the nearest neighbor approach is used. For the experimental variogram in Fig. 2, the slope of the best fitting line for the points within the actual range of the spherical model (72.5 km) is 0.08%/km, with R^2 of 87%. This means that, within the practical range, the uncertainty on the estimation of soiling increases

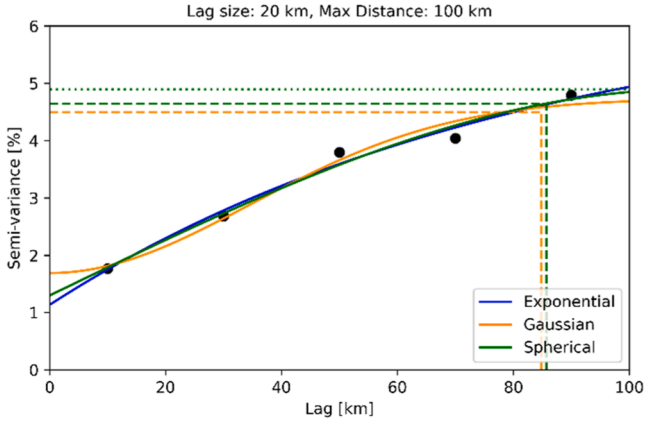

Fig. 3. Semi-variograms of the data points on the map, obtained for lag intervals of 20 km and a maximum distance of 100 km.

Table 3
Results from modeling the semi-variogram, for lag of 20 km, maximum distance of 100 km.

Model	a [km]	a_{95} [km]	c0 [%]	$c + c0$ [%]	R2 [%]	RMSE [%]
Exponential	N.A.	249.4	1.1	6.6	98.4	0.135
Gaussian	N.A.	84.8	1.7	4.7	97.6	0.166
Spherical	110.1	85.7	1.3	4.9	98.1	0.145

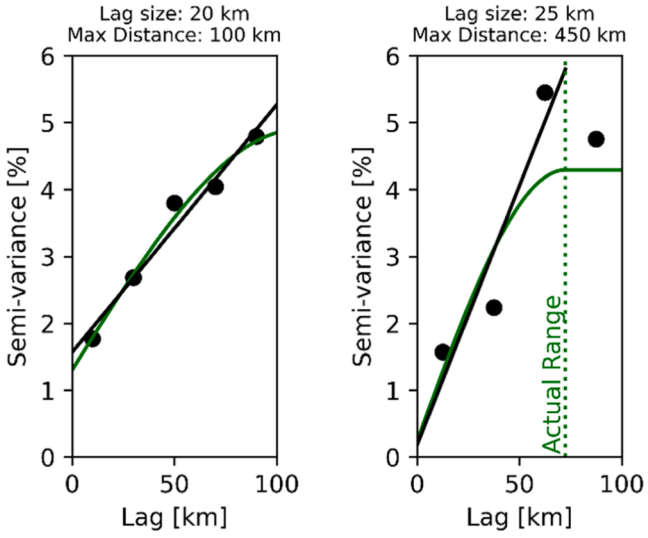

Fig. 4. Line fitting for the semi-variograms within the practical range. On the left, the semi-variogram shown in Fig. 2 obtained by considering lags of 25 km and a maximum distance of 450 km (slope: 0.08%/km, R2: 87.4%). On the right, the semi-variogram shown in Fig. 3 obtained by considering lags of 20 km and a maximum distance of 100 km (slope: 0.04%/km, R2: 96.5%).

by 0.8% for each 10 km of distance from the closest available data point. For the experimental variogram in Fig. 3, instead, the slope is 0.04%/km, with R^2 of 97%.

As shown, the results of the fit can change depending on the parameters in input. So, in order to understand the robustness of the parameters found in Table 1 to the user-inputs, these have been recalculated by varying the lag size from 15 to 30 km, at steps of 5 km. The results are listed in Table 4. It can be seen that the sill is found to be steady at 4.2% to 4.3%. The curve always reaches the actual range between 72 and 83 km and the practical range between 54 and 65 km, keeping a slope between 0.04 and 0.08%/km before that.

The fit for a lag size of 30 km each has a R^2 of 100% (Table 4) because there are only two data points available for the line fit (the practical range is less than twice the distance of the lags). For this reason, lags of 30 km and above have not been further considered. The 25 km lag has been preferred, as it is the maximum lag size with at least three data points within the practical range (12.5 km, 37.0 km, and 62.5 km). Ideally, however, if more points were available, it would be important to further decrease the lag size in order to model even better the shorter distances.

If the variation in the lag interval is found to have a limited effect on the shape of the curve, changing the maximum distance does vary the value of the sill, making it range between 4.2 and 5.0% (Fig. 5). The most consistent R^2 are obtained for maximum lag distances between 150 and 450 km, where the sill is limited between 4.2 and 4.5% and achieved at an actual range between 71 and 75 km (with a_{95} of 57 to 60 km). On the other hand, the slope of a linear approximation is found to be consistent across the maximum distance range, with a value of 0.08%/km, while the nugget ranges between 0.1% and 0.3%.

So far, the analysis has shown that both the lag size and the maximum distance can have an effect on the shape of the semi-variogram. The maximum distance seems to have mainly an effect on the sill value, which we also expect to change with the addition of new data points with higher soiling losses. On the other hand, the lag interval seems instead to have a limited, but still visible, impact on the nugget and on the slope of the curve, which we believe are the parameters of most interest for the community.

One of the main limits of the current investigation is represented by the sample size, which consists of 32 data points. In order to investigate the impact of the number of data points on the results, we repeated the analysis by considering subsets containing only 28 randomly selected sites (87.5% of the total). The analysis was repeated 35,960 times, to allow for all the combinations of sites to be accounted at list once. For each lag, the median and the extremes of the confidence interval (set at 5% and 95%) have been calculated. The best fits of the median and the extreme data series have then been calculated. As it can be seen in Fig. 6, the sill strongly changes depending on the data in input, while the actual and the practical ranges are fairly insensitive to them (60 km to 63 km, and 74 km to 78 km respectively). The nugget is 0.0% and the slopes ranges between 0.08% and 0.12%.

The results shown in Fig. 6 suggest that the addition of more sites to the map might vary the values of the sill and might also have an effect on the nugget and on the slope. The results of this work should therefore be considered valid for the current set of data points, located all in California and available on the NREL soiling map, and should be repeated in future with a larger number of sites and for different regions. For the current dataset, an uncertainty of 0.10% per km of distance from the closest data point should be considered when soiling at a site is estimated using a nearest neighbor approach. The maximum uncertainty is achieved at 74 km – 78 km and 95% of the sill at 60 km – 64 km. This confirms that, if no soiling data are available within these radii, other spatial interpolation techniques should be taken into consideration [18].

Overall, the results of this work show how the uncertainty increases with the distance from the soiling measurement. They also warn PV owners, designers, and operators that nearby sites might soil differently and establish preliminary boundaries for a reliable application of a nearest neighbor approach. However, even if the present dataset is one of the largest available on PV soiling, it is limited from a statistical point of view both in terms of population and covered area. For this reason, more studies are strongly recommended in order to (i) increase the number of data points and (ii) diversify the investigated climates and conditions. Moreover, a larger number of data will allow filtering sites by specific characteristics, such as tracking mechanism or land cover for example. Previous works have shown indeed that the accuracy of soiling

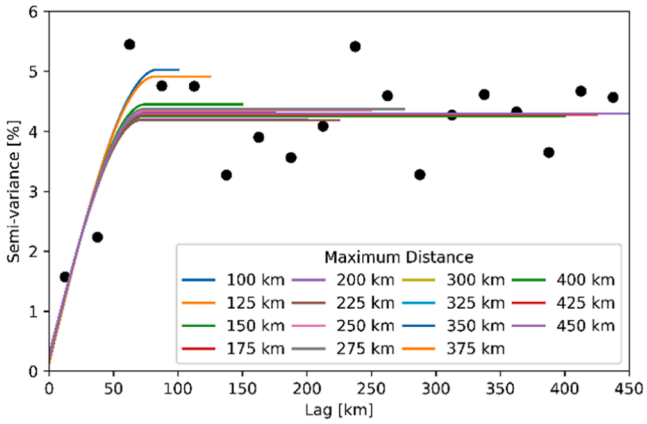

Fig. 5. Spherical semi-variograms for different maximum distances at lags of 25 km and a maximum distance of 450 km.

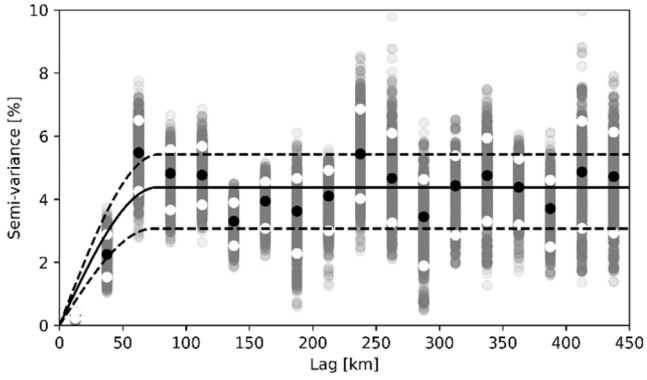

Fig. 6. Spherical models of the median and the confidence interval semi-variograms at lags of 25 km of 87.5% of the sites after 35,960 iterations.

Table 4
Results obtained for different lags at a maximum distance of 450 km. The coefficients of determination marked with an asterisk ("*") are those obtained by fitting only two data points and therefore should be discarded.

Lag [km]	R^2 [%]	RMSE [%]	a [km]	a_{95} [km]	c0 [%]	Sill [%]	Slope of line fit [%/km]	R^2 of line fit [%]
15	21.5	1.29	82.6	65.2	0.8	4.3	0.05	76.5
20	25.5	0.98	79.3	61.7	1.2	4.2	0.04	94.9
25	55.0	0.66	72.5	58.4	0.2	4.3	0.08	87.4
30	51.5	0.67	66.9	54.0	0.2	4.3	0.07	100.0*

estimation through spatial interpolation increases if only sites with similar features are considered [18]. For this reason, one can expect the uncertainty rate to decrease if only similar PV systems are employed for the estimation.

4. Conclusions

A statistical analysis of the spatial correlation of the Californian soiling data points collected on the NREL soiling map is presented here. A semi-variogram function is computed to understand the accuracy of estimating the soiling at a site through the nearest neighbor approach (i.e., assuming the soiling loss equal to the closest soiling measurement) depending on the distances between the sites.

It is found that the experimental semi-variogram of the 32 PV soiling data investigated can be modelled using a spherical function that achieves a maximum value of 4.4% (*sill*) after 74 km (*actual range*). The curve rapidly grows until a distance of 60 km at which it achieves a value equal to 95% of the sill (*practical range*). Within that distance, the semi-variogram can be approximated to a line of slope 0.08%/km. After that distance, the estimation is exposed to the maximum uncertainty. This means that, in lack of soiling data available within 60 km of the investigated site, the estimation of soiling should not be conducted by using the nearest neighbor approach, but potentially through an alternative spatial interpolation method. Although, even if available, it should be noted that, for a group of sites with an average soiling loss of 4%, estimating soiling from a single site located at 25 km or 50 km would be subject to absolute uncertainties of 2% or 4% respectively.

In some cases, the semi-variogram intercepts the y-axis at a value greater than 0%, suggesting, in agreement with previous literature, that PV systems, even if located within the same sites, can soil differently due to the exponential spatial decay of pollutants emitted from local sources or to different geometries and orientations of the modules.

The experimental equation used to fit the semi-variogram is reported in the text, in order to help PV investors and O&M teams to better understand the uncertainty in their soiling predictions. It is important to highlight that the study also shows how the choice of the input data might slightly affect the results of the analysis. Despite that, the actual and the practical ranges are found to be consistent if only randomly selected subsets of the data points are considered, with values between 60 and 65 km, and between 74 and 80 km, respectively. The slope of the line fit is found to vary between 0.08 and 0.12% per km depending on the data in input. This suggests that the addition of new sites might lead to a variation of the results and therefore, the conclusions of this work should therefore be considered valid only for the investigated dataset, restricted to only Californian sites. The analysis should be repeated when more data points from climatically different locations are available.

CRediT authorship contribution statement

Leonardo Micheli: Conceptualization, Methodology, Software, Investigation, Visualization, Formal analysis, Writing – original draft. **Matthew Muller:** Conceptualization, Methodology, Validation, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data are available on the NREL soiling map (https://www.nrel.gov/pv/soiling.html)

Acknowledgments

The work of Leonardo Micheli was supported by Sole4PV, a project funded by the Italian Ministry of University and Research under the 2019 «Rita Levi Montalcini» Program for Young Researchers.

This work was in part authored by Alliance for Sustainable Energy, LLC, the manager and operator of the National Renewable Energy Laboratory for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy (EERE) under Solar Energy Technologies Office (SETO) Agreement Number 38258. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.

References

[1] Department of Economic and Social Affairs, The sustainable development goals report, New York, NY, 2016. 10.1177/000331979004100307.
[2] SolarPower Europe, Global market outlook for solar power 2022-2026, 2022.
[3] K. Ilse, L. Micheli, B.W. Figgis, K. Lange, D. Daßler, H. Hanifi, F. Wolfertstetter, V. Naumann, C. Hagendorf, R. Gottschalg, J. Bagdahn, Techno-economic assessment of soiling losses and mitigation strategies for solar power generation, Joule 3 (2019) 2303–2321, https://doi.org/10.1016/j.joule.2019.08.019.
[4] R.K. Jones, A. Baras, A. Al Saeeri, A. Al Qahtani, A.O. Al Amoudi, Y. Al Shaya, M. Alodan, S.A. Al-Hsaien, Optimized cleaning cost and schedule based on observed soiling conditions for photovoltaic plants in Central Saudi Arabia, IEEE J. Photovolt. 6 (2016) 730–738, https://doi.org/10.1109/JPHOTOV.2016.2535308.
[5] E. Urrejola, J. Antonanzas, P. Ayala, M. Salgado, G. Ramírez-Sagner, C. Cortés, A. Pino, R. Escobar, Effect of soiling and sunlight exposure on the performance ratio of photovoltaic technologies in Santiago, Chile, Energy Convers. Manag. 114 (2016) 338–347, https://doi.org/10.1016/j.enconman.2016.02.016.
[6] M. Gostein, T. Duster, C. Thuman, Accurately measuring PV soiling losses with soiling station employing module power measurements, in: Proceedings of the 42nd IEEE Conference on Photovoltaic Specialists, 2015.
[7] M. Korevaar, J. Mes, P. Nepal, G. Snijders, M.X. van, Novel soiling detection system for solar panels, in: Proceedings of the 33rd European Photovoltaic Solar Energy Conference and Exhibition, 2017, https://doi.org/10.4229/EUPVSEC20172017-6BV.2.11.
[8] M. Korevaar, T. Bergmans, J. Mes, X. van Mechelen, A.A. Merrouni, F. Wolfertstetter, S. Wilbert, Field tests of soiling detection system for pv modules, in: Proceedings of the 36th EU PVSEC, Marseille, France, 2019.
[9] M. Gostein, K. Passow, M.G. Deceglie, L. Micheli, B. Stueve, Local Variability in PV Soiling Rate, in: Proceedings of the 35th European Photovoltaic Solar Energy Conference and Exhibition, Bruxelles, Belgium, 2018, pp. 1979–1983.
[10] V. Etyemezian, G. Nikolich, J.A. Gillies, Mean flow through utility scale solar facilities and preliminary insights on dust impacts, J. Wind Eng. Ind. Aerodyn. 162 (2017) 45–56, https://doi.org/10.1016/j.jweia.2017.01.001.
[11] A. Kimber, L. Mitchell, S. Nogradi, H. Wenger, The effect of soiling on large grid-connected photovoltaic systems in California and the Southwest Region of the United States, in: Proceedings of the IEEE 4th World Conference on Photovoltaic Energy Conference, 2006, pp. 2391–2395.
[12] A. Skomedal, M.G. Deceglie, Combined estimation of degradation and soiling losses in photovoltaic systems, IEEE J. Photovolt. 10 (2020) 1788–1796, https://doi.org/10.1109/jphotov.2020.3018219.
[13] M. Coello, L. Boyle, Simple model for predicting time series soiling of photovoltaic panels, IEEE J. Photovolt. 9 (2019) 1382–1387, https://doi.org/10.1109/JPHOTOV.2019.2919628.
[14] S. You, Y.J. Lim, Y. Dai, C.H. Wang, On the temporal modelling of solar photovoltaic soiling: energy and economic impacts in seven cities, Appl. Energy 228 (2018) 1136–1146, https://doi.org/10.1016/j.apenergy.2018.07.020.
[15] M.H. Bergin, C. Ghoroi, D. Dixit, J.J. Schauer, D.T. Shindell, Large reductions in solar energy production due to dust and particulate air pollution, Environ. Sci. Technol. Lett. 4 (2017) 339–344, https://doi.org/10.1021/acs.estlett.7b00197.
[16] W. Javed, B. Guo, B. Figgis, Modeling of photovoltaic soiling loss as a function of environmental variables, Sol. Energy 157 (2017) 397–407, https://doi.org/10.1016/j.solener.2017.08.046.
[17] National Renewable Energy Laboratory, Photovoltaic modules soiling map, (2018). https://www.nrel.gov/pv/soiling.html (accessed May 18, 2018).
[18] L. Micheli, M.G. Deceglie, M. Muller, Mapping photovoltaic soiling using spatial interpolation techniques, IEEE J. Photovolt. 9 (2019) 272–277, https://doi.org/10.1109/JPHOTOV.2018.2872548.
[19] M. Gostein, K. Passow, M.G. Deceglie, L. Micheli, B. Stueve, Local variability in PV soiling rate, IEEE (Ed.), in: Proceedings of the 7th World Conference on Pholtovoltaic Energy Conversion, Waikoloa, HI, 2018, pp. 3421–3425.

[20] US Environmental Protection Agency, Air quality system data mart [internet database], (n.d.). https://www.epa.gov/airdata (accessed June 1, 2021).
[21] International Electrotechnical Commission, Photovoltaic system performance – Part 1: monitoring (IEC 61724-1, Edition 1.0, 2017-03), (2017).
[22] M.A. Oliver, R. Webster, A tutorial guide to geostatistics: computing and modelling variograms and kriging, Catena 113 (2014) 56–69, https://doi.org/10.1016/j.catena.2013.09.006.
[23] E. Jones, E. Oliphant, P. Peterson, et al., SciPy: open source scientific tools for python, (2001). http://www.scipy.org/.

Dr. Leonardo Micheli is a "Rita Levi Montalcini" Assistant Professor (RTDB) at Sapienza University of Rome, Italy. He graduated in 2015 with a PhD in Renewable Energy from the University of Exeter, UK. His current research interests include the monitoring, the analysis and the prediction of photovoltaic (PV) performance, and the optimization of PV loss mitigation strategies.

Dr. Matthew Muller is an Engineer within the Photovoltaic Performance and Reliability group of the National Renewable Energy Laboratory (NREL), CO, USA. In 2021, he graduated with a PhD in Renewable Energy from the University of Jaén, Spain. He has 15 years of experience working on photovoltaics and concentrator photovoltaics.