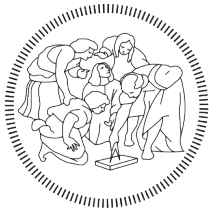
POLITECNICO DI MILANO

Scuola di Ingegneria Industriale e dell'Informazione

Corso di Laurea Magistrale in Energy Engineering

Dipartimento di Energia



POLITECNICO
MILANO 1863

Synthetic Generation of Solar Radiation data for a robust micro-grid design

Relatore: Prof. Marco Merlo

Correlatore: Ing. Matteo Moncecchi

Tesi di Laurea di:

Sara Pistilli, matricola 863568

Anno Accademico 2016-2017

Ringraziamenti

Prima di tutto vorrei ringraziare insieme il Professore Marco Merlo e Matteo per avermi ridato tutta la fiducia in me stessa che avevo perso in questi cinque lunghi anni e per avermi dato la possibiltà di sviluppare questa Tesi.

Ringrazio il Professore Merlo per avermi da subito accolto e mostrato il suo lavoro, lo ringrazio per essere stata una guida e per avermi mostrato il suo approccio determinato nel fare le cose in cui crede, lo ringrazio per la ricerca che porta avanti all'interno del Politecnico, per la passione con cui insegna, per la dedizione che ha per i suoi tesisti e che ha avuto per me e per il suo spirito critico.

Poi vorrei ringraziare Matteo per esserci sempre stato in questi strani mesi di tesi, per essere stato un punto di riferimento fondamentale, un esempio, una sicurezza ed un amico, per avermi incoraggiato e spronato a fare sempre meglio. Lo ringrazio per il suo approccio naturale e solare al suo lavoro e alla ricerca, "se non troviamo una soluzine oggi vedrai che continuando a cercare prima o poi la troveremo", e per avermelo trasmesso.

Inoltre ringrazio quei professori che sono stati un punto di riferimento durante il mio percorso di studi, ringrazio il Politecnico di Milano per aver dato la possibilità a studenti interessati di seguire un percorso magistrale come quello proposto "Energy for Development".

Ringrazio la mia famiglia, tutta e bellissima, i miei amici di Terni che mi mancano tanto, i miei amici di Milano che sono stati la mia casa e la mia dimora, ringrazio la Ginnastica Terni per la grinta che mi ha inculcato, ringrazio gli Scout del TRIX per essere il sottofondo di tutta la mia vita e poi ringrazio Giorgio, perchè è Giorgio.

Ringrazio l'Africa, che ancora non ho mai visto ma che desidero come poche cose al mondo, per essere stata la forza motrice dei miei studi.

Extended Summary

Introduction

1.19 billion people all around the world do not have access to electricity according to the [1] Global Status Report on renewable energy. Rural electrification is one of the issues to be solved in order to provide modern and reliable energy to developing countries. Solar power is one of the fastest growing sources of electricity generation [1] and Distributed Renewable Energy systems represent an affordable solution to energy access.

Due to the highly variable nature of solar resource it is really important to study its behaviour to integrate solar technologies in distributed systems. Several tools already exist to design and optimize distributed renewable energy systems. All these tools require a vast amount of solar data to study the variability and reliability of these systems. Actually, the solar radiation reaching the ground has a random (stochastic) and variable nature due to all the factors such as clouds or gases within the atmosphere that absorb or scatter it. The majority of time solar datasets that exists do not provide data with high resolution and long-temporal coverage. Indeed, to obtain reliable radiation data at ground level requires systematic measurements and high costs.

In particular in developing countries it is really difficult to collect and gather data because of the high cost of installation of measurement systems. Furthermore, to be able to analyze and simulate the behaviour of solar energy facility, data over several years are required (at least 10 or more [2]).

This situation motivates the development of procedures able to calculate and provide solar radiation data for places where measurements are not available, or for places where there are gaps in the measurement records.

A possible approach to the problem is to use stochastics models to synthetically generate solar radiation data. Through the synthetic generation it is possible to reproduce data that have similar statistical characteristics to the measured data and generate multiple years from fewer input years.

In this thesis such an approach has been developed with the goal to be integrated in PoliNRG [3], a tool used for the optimization and dimensioning of off-grid system, located in rural areas.

In literature could be found several approaches to generate data for daily and hourly time step. In these approaches the climatic variable used to simulate the solar radiation is the clearness index [4, 2, 5, 6, 7, 8, 9, 10] and [11].

The proposed procedure for the thesis work is based on both Autoregressive Moving Average (ARMA) and Markov Transition Matrix (MTM) methods. Furthermore the distribution of the daily clearness index has been described by Bendt's correlation.

Looking at the autocorrelation (ACF) and partial correlation function (PACF) of the observed time series data the specific ARIMA model needed for the process has been determined, the same to determine the Markov model order.

Looking at the probability distribution functions of the solar irradiance observed data it has been determined the importance of the local and geographical distribution in high resolution irradiance distributions. The climatic variable clearness index has a fundamental role in differentiating these distributions in function of the geographical area [12].

Finally, the aim of the procedure is to generate multiple years of data with high resolution sampling, starting from scarce and low resolution years of data. The idea to start from low resolution data is due to these data are generalizable and easy accessible also for remote areas. A tool based on Markov Transition Matrices has been implemented using the concept of importance in differentiating local probability distributions and subdividing the world in Climatic Areas based on the Köppen-Geiger definition [13].

To validate the procedure the synthetically generated data obtained are compared with the observed data by means of some statistical characteristics and qualitative analysis. The model validation is based on a comparison of the probability distributions functions, autocorrelation functions and cumulative distributions.

Input Data

The data adopted are based on a website [14] – Solar Radiation Data – that collects data from different web services. In this website it is possible to obtain data with different resolutions and from different locations all around the world.

For the implementation and calibration of the methodology some processes are data-intensive and require a large amount of data.

Models and Correlation

The procedure of generation can be divided in three main sub-processes: Bendt's correlation, ARIMA process and Markov process. The input of the tool are monthly average clearness index values and the output are sub-hourly clearness index values.

Before starting with the description of the methodology it is necessary a little mathematical introduction to the models used.

Bendt's correlation: Bendt's correlation is used to model the probability distribution of the daily clearness index K_t^d for each month [15]. In [16] a parametrical expression of the distributions is presented to model this in function of the monthly average K_t^m, minimum $K_{t,min}$ and maximum $K_{t,max}$ clearness index of each month as presented in equation (1).

$$F(K_t, K_t^m) = \frac{exp(\gamma \cdot K_{t,min}) - exp(\gamma \cdot K_t)}{exp(\gamma \cdot K_{t,min}) - exp(\gamma \cdot K_{t,max})} \tag{1}$$

Where F is the cumulative distribution function or fraction of value in which the K_t^d is lower than a certain given specific value and γ is a particular exponential distribution.

ARIMA Model: The ARIMA (Autoregressive Integrated Moving Average) Models are used to described a series Y in terms of its past values and the current and past values of error terms. This model is called $ARIMA(p, d, q)$ where p, d and q are the model *orders* and are defined as follows:

- p or (AR) term: number of lagged values of Y which represents the autoregressive nature of model.
- q or (MA) term: number of lagged values of the

error term which represents the moving average nature of model

- d: is the number of times Y has to be differences to produce the stationary series y.

Then the ARIMA equation for predicting y has the following form (2):

$$y_t = \mu + \phi_1 \cdot y_{t-1} + ... + \phi_p \cdot y_{t-p} - \theta_1 \cdot a_{t-1} - ... - \theta_q \cdot a_{t-q} \tag{2}$$

Markov Model: Markov models are mathematical representations of a stochastic process under the assumption that its future state depends only on the last n states. Where n is called the Markov order and determine the dependency on the previous $n-$events.

For instance a 1^{st} order model means that, given a process in state i at time $t-1$, the probability that it will be in state j at time t is given by a probability P_{ij} that is fixed. P_{ij} is the transition probability from state i to j and is independent of the states of the process at times $t-2, t-3, \ldots$.

In Equation (3) it is possible to observe how to determine the probability.

$$\begin{aligned} P_{ij} &= \\ &= P(Y_t = j | Y_{t-1} = i, Y_{t-2} = i_{t-2} | \ldots | Y_0 = i_0) = \\ &= P(Y_t = j | Y_{t-1} = i) \end{aligned} \tag{3}$$

Markov models use the Markov Transition Matrices to generate time series. The MTMs are matrices that contain the probability of the events.

Modelling and Synthetic generation of Solar Radiation data

In order to model the solar radiation, it is necessary to use a stochastic variable that can represent the meteorological characteristics along the year. The clearness index has been chosen for such a purpose. Modelling the clearness index K_t with a stochastic approach it is possible to obtain the solar irradiation H through the synthetic generation.

The clearness index is defined as the ratio of the global solar radiation that reaches the Earth's surface to the extraterrestrial solar radiation H_o. It is a non-dimensional variable whose value can vary between 0 and 1.

$$K_t = \frac{H}{H_o} \tag{4}$$

The main task of the tool developed in the thesis it is to obtain high resolution (sub-hourly) solar radiation data from the clearness index. The resolution that can be achieved depends on the resolution of the input data for the different models used.

The choice of using the clearness index as main climatic variable throughout the procedure depends on several reasons. The first reason is that, being non-dimensional, it can be generalized for different areas. The second reason is that it is easier to remove the daily trends than it would be with the solar irradiation.

It is possible to subdivide the tool procedure in three steps, given the goal of each one:

- From monthly clearness index to daily clearness index distribution function.
- From daily clearness index distribution function to yearly sequence of daily clearness index.
- From daily clearness index to sub-hourly resolution clearness index.

The approach needs monthly clearness index input data for at least one year. If data from multiple years are available, the procedure will take into consideration more possible cases and better represent the variable conditions between different years.

Eventually, using the Bend's procedure with the average monthly clearness index it is possible to find the cumulative distribution function of the daily clearness index K_t^d for each month.

The next step is to use the ARIMA model to determine the order of the daily clearness index values obtained from Bendt within each month. An important step is the determination of the ARIMA model that best fits the clearness index time series. To determine the most suitable ARIMA model, it is necessary to make some tests on the data and study the autocorrelation functions (ACF) and the partial autocorrelation function (PACF).

In order to determine the best ARIMA model that can be applied to any general location, multiple years of data from two different locations around the world have been used. First of all a study has been conducted to find some reasonable values of possible orders of the ARIMA model. Once the reasonable ARIMA models are defined, they are tested on the time series and ranked based on the quality of the results, using metrics such as the P-value or statistical likelihood principle values.

Once the ARIMA model has been chosen, the sequence of its values is used to give the sequence to the daily clearness index values obtained through the Bendt's correlation. Each time the ARIMA model is used the output sequence is unique and independent from the other generated profiles.

The increase in the data resolution is obtained by using the Markov model. As for the ARIMA model, more data are needed to determine the order of the model that better describes the clearness index time series. The study to determine the Markov model order than better describe the time series is done throughout the PACF.

Since different days with different average clearness index values tend to have different behaviours, the days have to be clustered based on K_t^d into different classes. Each class represents a particular range of meteorological conditions. For each class, a Markov Transition Matrix is built from the input data. The number of classes is limited because too many classes would lead to empty MTMs. Indeed, it is necessary to create a MTM for each class to synthetically generate high resolution data that can be representative of the observed input values.

The numbers of output (years) obtained from the procedure is determined by the number of years of daily sequence generated with the ARIMA model. Indeed, the output of the Markov process are multiple years with sub-hourly resolution.

The MTMs can be created using sub-hourly clearness index time series from multiple locations close to the location under study, with similar meteorological characteristic along the year.

Many data are necessary for the creation of the MTMs to prevent them from being empty. Depending on the locations under study, the input data with the highest available resolution (10-min, 5-min, 1-min) are used. If multiple years of data are available the quality of the results is increased. To increase the number of data for the creation of the MTMs, the input data can be taken from the studied location and also from other close-by locations.

The choice of these other localities is done in function of the coordinates, considering $\pm 3°$ from the latitude and longitude of interest as in Figure 1. In this way the result is similar to a circular area that can describe the location to have consistent MTMs.

Then, it is possible to build the matrices (one for each class) that will be used in the procedure. Once the MTMs have been created, it is possible to generate the final high-resolution time series for each day

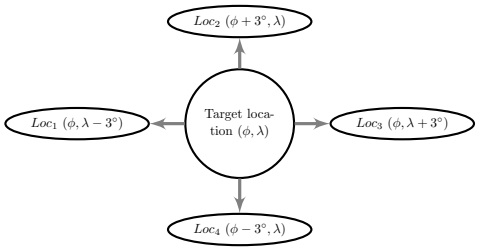


Figure 1: Location of interest and the other four locations considered for the input data.

starting from the daily clearness index values.

The steps for the generation of the values in the day are the following:

1. The K_t are normalized:

$$K'_t = \frac{K_t}{1.031 \cdot exp(\frac{-1.4}{0.9+9.4/AM}) + 0.1} \quad (5)$$

2. Each day is classified based on its K_t^d value and its class is determined.

3. The appropriate MTM is selected in function of the class.

4. The first two values of the day, K'_{1s} and K'_{2s}, are set equal to the daily K_t^d. The MTM row is selected based on the values of K'_{1s}, K'_{2s}.

5. The row represents the probability distribution vector to generate K'_{3s}. K'_{1s}, K'_{2s} correspond to the values at midnight, where H_o is zero. This ensures that the value of K'_{ts} at sunrise is independent at each repetition of the procedure, i.e. for each day.

6. Generating a random number r from a uniform distribution between 0 and 1 k'_{3s} is generated:

$$K'_{ts} = \frac{1}{m} \cdot \left[(i-1) + \frac{r - F_{l-1}(t)}{F_l(t) - F_{l-1}(t)}\right] \quad (6)$$

   Where $F(t)$ is the corresponding cumulative vector of the MTM row, m are the possible states that the Markov process can generate, i is the number of the MTM row selected and l is the number of column that contain a cumulative probability $F_l(t)$ higher than the r value.

7. Subsequent values $K'_{4s}; K'_{5s}; ...; K'_{144s}$ are generated similarly. The process goes on until the last time step of the day.

8. The synthetically generated K'_{ts} sequence is converted to K_{ts} using equation 5.

9. To convert the clearness index into global solar radiation H data the equation (4) is used.

10. The synthetically generated value of the daily clearness index K_{ds} is calculated having H.

11. The obtained values are checked by comparing the absolute distance between K_{ds} and the starting K_t^d.

12. If the distance is lower than a determined tolerance (δ), the values are kept. Otherwise, the procedure is repeated from Step 4.

If before the Markov process multiple years were generated, the same number of years are also obtained as output of the Markov process.

A graphical example of the methodology process and examples of output values are reported in Figure 2.

Cases

To validate the procedure, it has been chosen two different locations, Amsterdam and Ngarenanyuki. Ngarenanyuki, in Tanzania, has been chosen because there is a secondary school located in a rural area where the project Energy4Growing [17] is being developed. A hybrid micro-grid is linked to this school to generate electricity.

The other location, Amsterdam, has been chosen because it has a really different climatic characteristic along the year and it can be used to make a

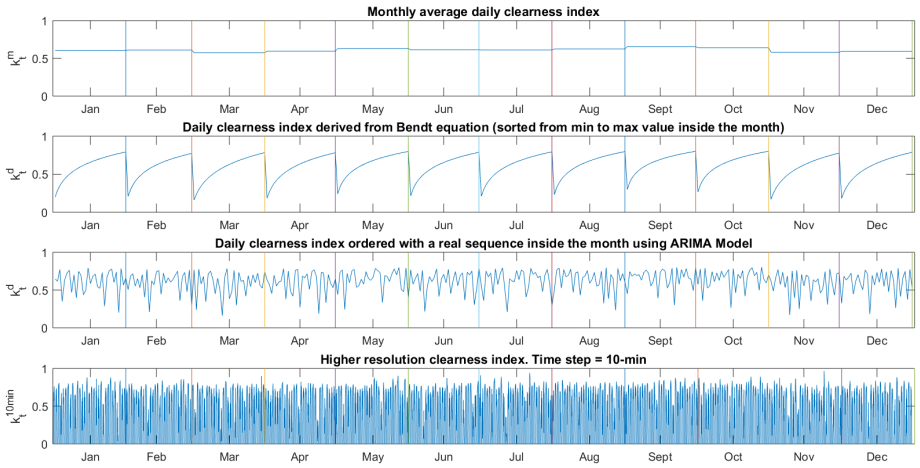


Figure 2: Graphical representation of the proposed methodology.

comparison and to test the procedure in two different conditions.

Indeed for the micro-grid in Tanzania, the meteorological conditions are quite stable along the year and the weather is quite good with an daily average clearness index around 0.62. Instead for Amsterdam the weather conditions are really variable and a strong seasonality is presented along the year with a yearly clearness index of 0.40. In Figure 3 a representation of the solar irradiation for the two locations.

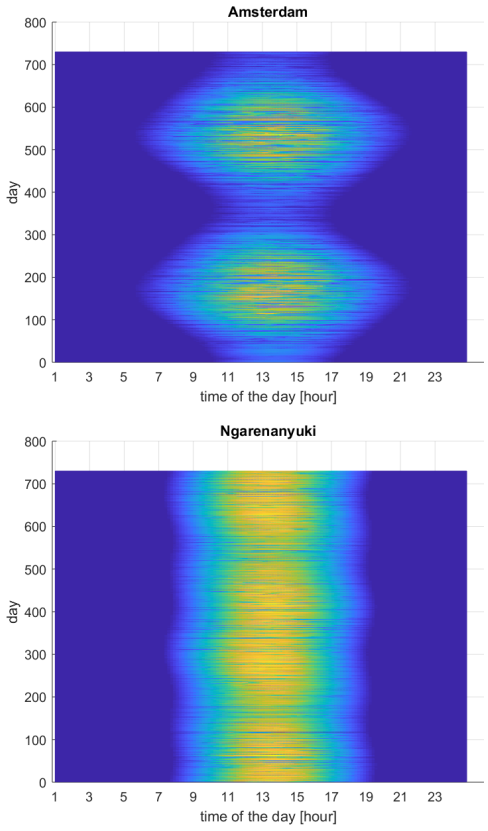


Figure 3: Two years of solar irradiation with view from the top.

Final Tool and Generalization

To reduce the amount of input data required to the user of the tool, some processes have be generalized, without significant losses in the performance of the procedure.

The two procedures that require the highest amount of input data are the determination of the orders of the ARIMA model and the creation of the Markov Transition Matrices.

The ARIMA model generalization can be performed by comparing different locations and choosing a set of model orders p, d, q that can be accu-

Model	p	d	q
1-ARIMA	1	1	0
2-ARIMA	2	1	0
3-ARIMA	0	1	1
4-ARIMA	0	1	2
5-ARIMA	1	1	1

Table 1: Orders of the selected ARIMA Models.

rately represent the behavior of the daily clearness index time series in any general location.

For this purpose, the daily clearness index data from four different locations have been studied, Amsterdam, Ngarenanyuki, Madrid and Valencia. The first step is the guess of some reasonable values for the orders of AR and MA terms.

To use some stochastic models it is necessary to eliminate some trends in the time series and it is important to investigate the degree of differencing to sationarize the time series. For all the years and all the locations studied, the degree of differencing to remove trend was equivalent to 1, so the order d is settled to be 1. To create a list of the possible ARIMA model to test it on the time series data it is necessary to look at ACF and PACF functions plots. An example is reported in Figure 4. The order for both p and q terms from the plots seem to varies between 0 and 2. The models chosen to be tested are presented in Table 1.

Among the five models listed above, the one that best models the clearness index data is the $ARIMA(1,1,1)$, presenting the highest P-value and the lowest error mean. Nevertheless also the other models can give a good representation of the time series.

A similar study has been done on the time series to choose the most suitable the value of the coefficients (ϕ, θ and variance σ). For all the years and for all the locations the coefficient has been computed using the ARIMA(1,1,1). Comparing the results, is it possible to see that the coefficients have very similar values for both the AR term and MA term. As in [6], it has been decided to make an average between the coefficients obtained for the different locations to determine a general value that can properly represent all the time series. The values obtained are 0.3310 for ϕ_1, 0.9493 for θ_1, and standard deviation $\sigma = 0.1436$. These will be the values that are included in the model by default, to synthetically generate the sequence of the daily clearness index all around the world.

Also the Markov model can be generalized. This generalization allows to simplify the procedure and, on top of that, to use the tool for areas where limited input data are available. Before generalizing the Markov process, the classes for the clustering of the daily clearness index have to be selected. These classes are used to differentiate between different days and practically to create different MTMs. Studying the trend of the clearness index in two different locations with very different climatic characteristics, it can be highlighted that daily clearness index values below 0.25 and higher than 0.75 are really unusual. Depending on the location, the curve of the probability density of the data can be shifted slightly to the left (low clearness index) or to the right (higher clearness index). There is a clear trade-off in the choice of the number of classes between how accurately each day is represented and how many values are available to populate the MTMs. It is important to avoid to have empty MTMs and on the contrary to have excessive classes to represent in a detailed way the different daily clearness index. As a compromise, it has been chosen to create five classes that can represents each climatic condition as reported in the Table 2.

This results is a good compromise between having

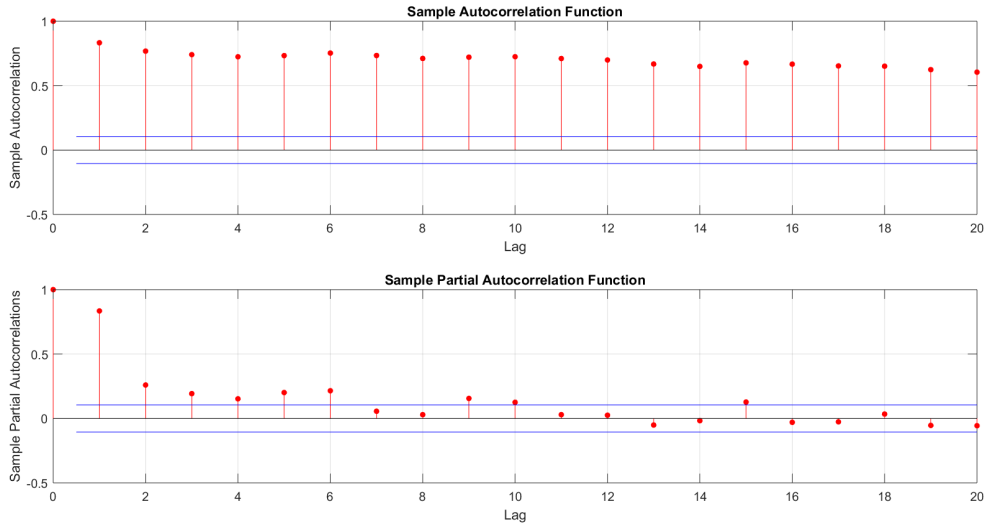


Figure 4: Example of ACF and PACF until lags 20 for a time series.

Class	Classification	Range
1	Extremely cloudy day	$0 \leq k_t^d < 0.25$
2	Cloudy day	$0.25 \leq k_t^d < 0.40$
3	Variable day	$0.40 \leq k_t^d < 0.60$
4	Sunny day	$0.60 \leq k_t^d < 0.75$
5	Extremely sunny day	$0.75 \leq k_t^d \leq 1$

Table 2: Classification of the daily clearness index classification inside the classes.

Index	Area
A	equatorial
B	arid
C	warm temperature
D	snow
E	polar

Table 3: Division of the world in climatic areas proposed by Wladimir Köppen.

enough representative classes and having reasonably populated MTMs.

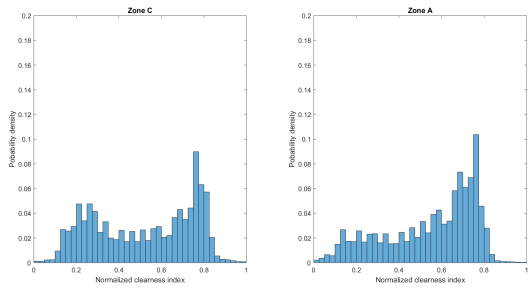
Studying the probability distribution functions of the clearness index and looking at its characteristic for different locations, it has been noticed that the climatic area plays an important role for the determination of the final climatic characteristics [12]. The *pdf* of the different areas could be either unimodal or bimodal, depending on the location. The results from [12] point to the importance of local distribution and type of clouds variability in high-resolution irradiance distributions, and highlight the role of the clearness index in differentiating these distributions.

Looking at the *pdf* of the synthetically generated data and the observed in Figure 5, it is possible to see that the generated data reproduce well the *pdf* of the observed data.

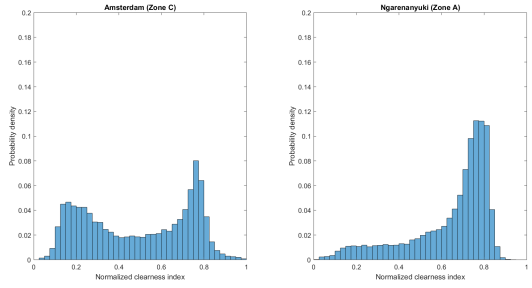
The classification of climatic areas proposed by Wladimir Köppen [13] has been used. The world is divided in five areas as reported in Table 3 and in Figure 6.

Due to the importance of the climatic area, ten locations for each areas have been selected to create the MTMs. These Matrices are included as a library, one for each class for each areas to make the procedure generalizable and suitable also in case of few available data. Each location in the world is associated to a determined are as can be seen in the Figure 6.

The final procedure to use the tool can be represented by the flow chart in Figure 7.



(a) Observed data.



(b) Synthetic generated data.

Figure 5: Probability Density Function for two different Climatic Areas.

Methodology validation

The objectives of a synthetic generation tool is not to generate data that represent some real data point by point, but to generate data that have general characteristics and behaviour similar to the observed data. Therefore, to validate the method it is not useful to measure the accuracy of the method in reproducing the original data but it is more useful to measure other characteristics.

- Total values: yearly and monthly total solar irradiation.
- Distributions of daily and sub-hourly values within the year.
- Behaviour of the fluctuations of daily and sub-hourly values.

The third point has not been directly validated in a quantitative way, because the behaviour of the fluctuations is determined by the ARIMA and Markov model characteristics. Therefore, if these models have been properly calibrated, the fluctuations of daily and sub-hourly values should be properly reproduced. The calibration of the two models has been tested on the locations under study.

Global Validation: The procedure can be validated taking into consideration two different aspects:

- quantitative (statistical analysis and total results);
- qualitative.

The quantitative analysis is focused on some statistical properties of the final results, for instance the Root Mean Squared Error (RMSE) and Mean Bias Error (MBE).

Figure 8b represents the cumulative distribution function of the 10-min irradiance observed and generated data. In Figure 8a the cumulative distribution of daily values is reported. The two cumulative distributions are comparable for both resolutions. The RMSE computed between the observed and generated data with 10-min resolution is 4.5% for Amsterdam and 3.2% for Ngarenanyuki. For the daily data the RMSE is 6% for Amsterdam and 8.5% for Ngarenanyuki.

The generated data have to be statistically representative and must preserve some important quantities, such as the monthly and yearly total energy. These constraints mean that the total solar irradiation, characterized by the monthly means and by the yearly sum, should remain fairly constant when compared to the same quantities obtained from the original time series. As shown in Figure 9 the generated monthly values closely approximate the observed values.

The average of the relative distance for the yearly total values is 2.4% for Ngarenanyuki and 3% for

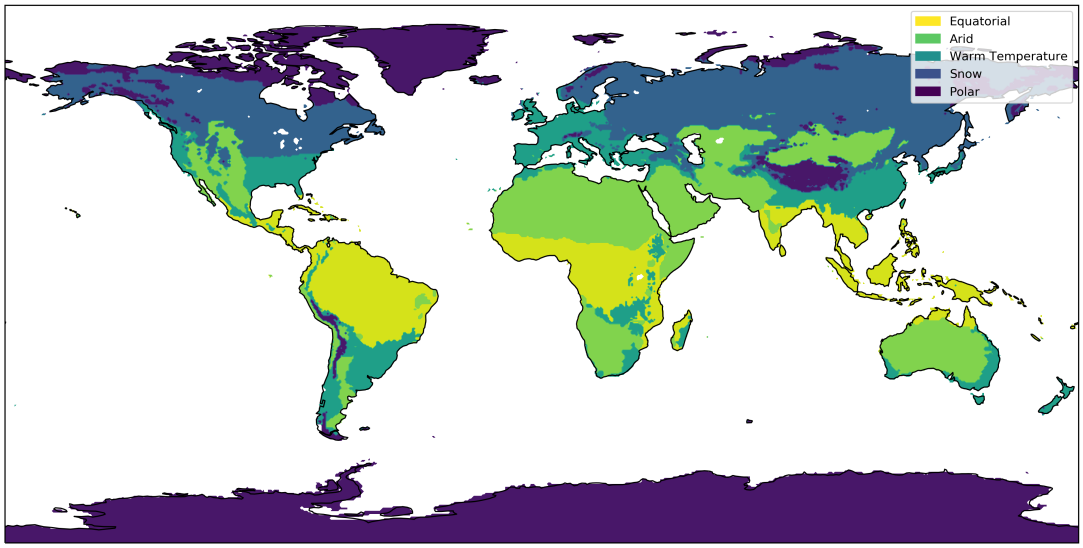


Figure 6: World map of climatic areas as proposed by Köppen-Geiger.

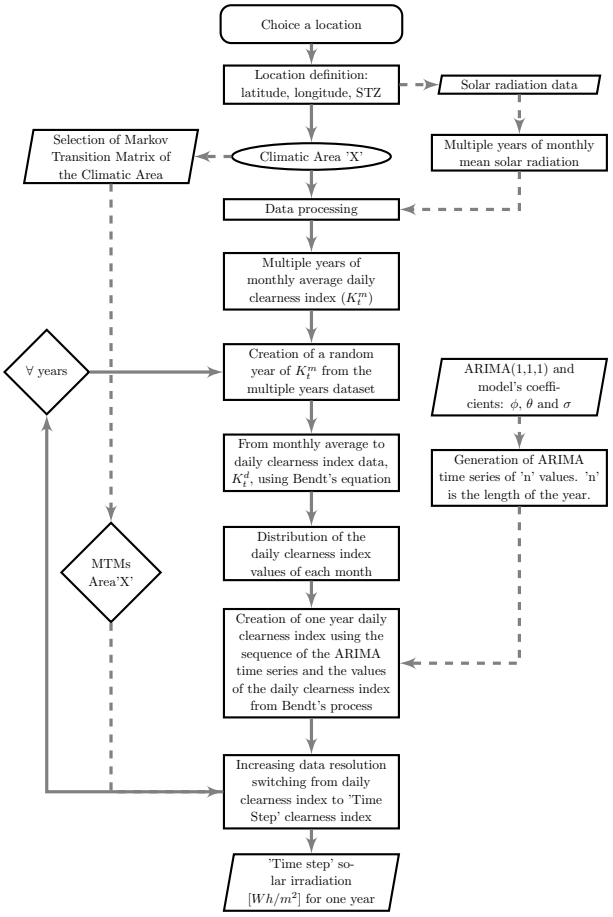
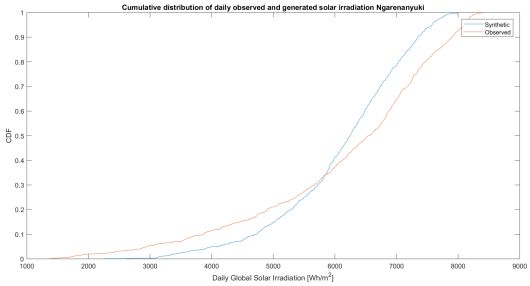
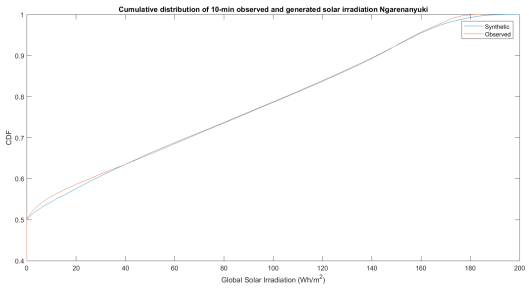


Figure 7: Flow chart of the final procedure that includes all the generalizations applied.



(a) Daily resolution.



(b) 10-min resolution.

Figure 8: Comparison of cumulative distribution function of observed and generated daily values for Ngarenanyuki.

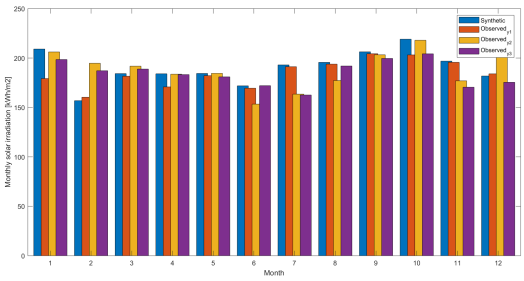


Figure 9: Monthly solar irradiation observed (three different years) and generated values for Ngarenanyuki.

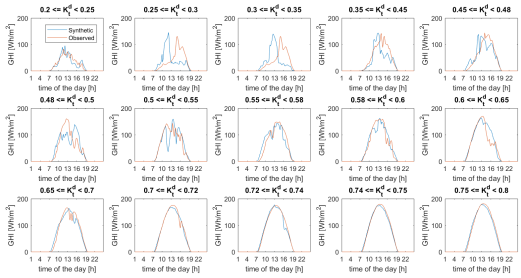


Figure 10: Comparison of different daily profiles with a resolution of 10-min.

Amsterdam.

For the qualitative analysis, it is possible to look at the dynamics of the fluctuations for 10-min time step data. The dynamic of the observed and synthetically generated time series should be comparable. Figure 10 shows a comparison of generated and observed global solar radiation daily profiles for selected days with similar K_t^d values for Ngarenanyuki.

The plots shows that the method is able to preserve the expected behavior under complete and partial clear sky conditions. Indeed, higher fluctuations are showed for partial overcast or overcast sky conditions, where the clearness index is lower, for both observed and generated days. The overall dynamic behavior is coherent with what is observed in the real data.

Figure 11 is a comparison of daily values between observed and generated data. From these result it is possible to see that the general trend along the year

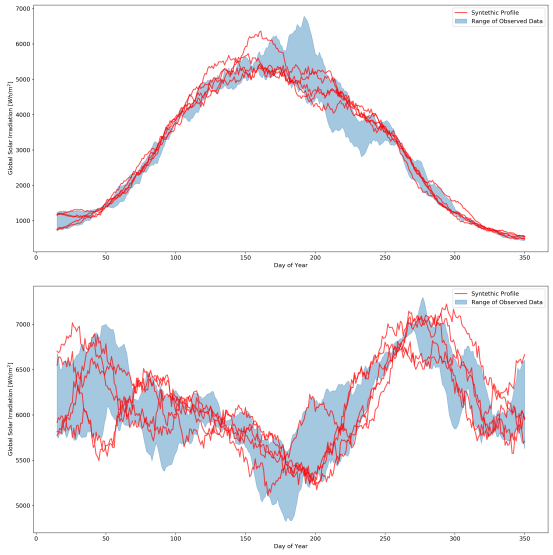


Figure 11: Comparison of observed and synthetic daily Irradiation for Amsterdam and Ngarenanyuki.

is maintained by the synthetic data.

Figure 12 shows the distribution of 10-minutes values of each class for generated and observed data. The distributions are very well approximated for all classes. The smoother distribution in the generated data is due to the higher number of years.

Conclusion

The tool seems to correctly synthetically generate clearness index time series. Furthermore, it can be used in any location of the world with a good approximation to generate multiple years of solar radiation data.

These data are very important to properly optimize and design distributed energy systems when the variability affects significantly the optimal design of the energy system.

This procedure has been developed to be integrated in the PoliNRG software [3], a software for the optimization of micro-grid system, but can be used also to provide inputs for the others energy modelling software.

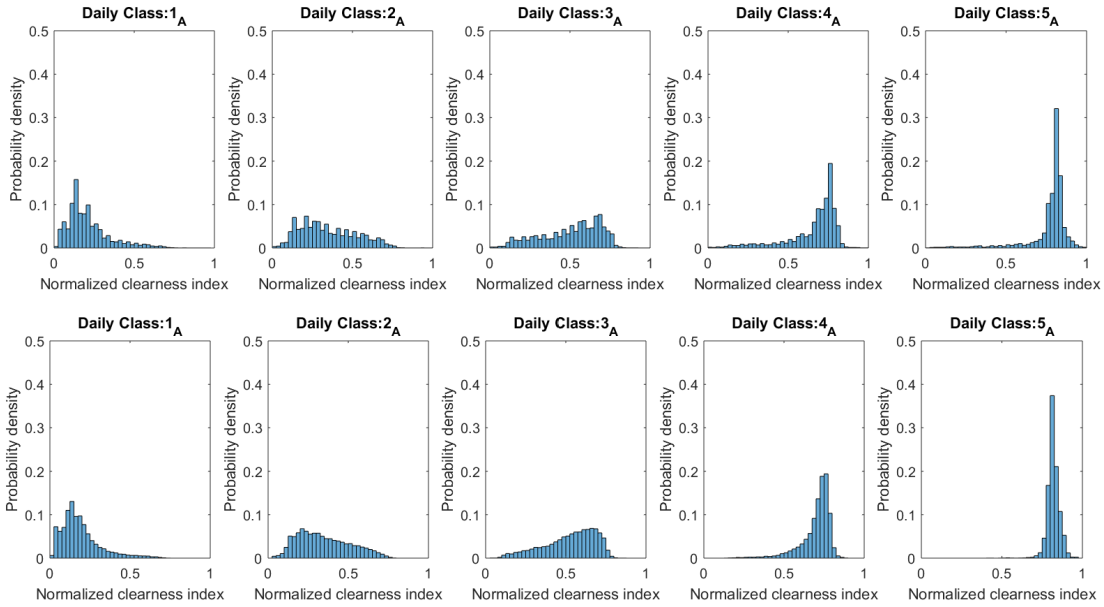


Figure 12: Probability Density Function for Observed(top) and Synthetically Generated(bottom) Zone A.

Finally it is important to highlight that, although the methodology can be applied to areas with very limited availability of measured data, the quality of the results improves if a larger amount of data is available. The same applies to the resolution of input data.

The main advantage of the methodology is its flexibility. Indeed, it can be applied to any location, enabling the users to choose the best trade-off between the quality of the results and the amount of input data.

Bibliography

[1] REN21. *Renewables 2017 global status report 2017*. 2017. ISBN: 9783981810769.

[2] Viorel Badescu. *Modeling Solar Radiation at the Earth ' s Surface*. 2008, p. 536. ISBN: 978-3-540-77454-9. DOI: 10.1007/978-3-540-77455-6. URL: http://link.springer.com/10.1007/978-3-540-77455-6.

[3] Claudio Brivio et al. "A novel software package for the robust design of off-grid power systems". In: *Journal of Cleaner Production* 166.August (2017), pp. 668–679. ISSN: 0959-6526. DOI: 10.1016/j.jclepro.2017.08.069. URL: http://dx.doi.org/10.1016/j.jclepro.2017.08.069.

[4] Jamal Hassan. "ARIMA and regression models for prediction of daily and monthly clearness index". In: *Renewable Energy* 68 (2014), pp. 421–427. ISSN: 0960-1481. DOI: 10.1016/j.renene.2014.02.016. URL: http://dx.doi.org/10.1016/j.renene.2014.02.016.

[5] J. Boland. "Time-series analysis of climatic variables". In: *Solar Energy* 55.5 (1995), pp. 377–388. ISSN: 0038092X. DOI: Doi10.1016/0038-092x(95)00059-Z. URL: http://www.sciencedirect.com/science/article/pii/0038092X95000059Z.

[6] J. M. Santos, J. M. Pinazo, and J. Canada. "Methodology for generating daily clearness index index values Kt starting from the monthly average daily value Kt. Determining the daily sequence using stochastic models". In: *Renewable Energy* 28.10 (2003), pp. 1523–1544. ISSN: 09601481. DOI: 10.1016/S0960-1481(02)00217-3.

[7] V. A. Graham, K. G. T. Hollands, and T. E. Unny. "A time series model for Kt with application to global synthetic weather generation". In: *Solar Energy* 40.2 (1988), pp. 83–92. ISSN: 0038092X. DOI: 10.1016/0038-092X(88)90075-8.

[8] V. A. Graham and K. G. T. Hollands. "A method to generate synthetic hourly solar radiation globally". In: *Solar Energy* 44.6 (1990), pp. 333–341. ISSN: 0038092X. DOI: 10.1016/0038-092X(90)90137-2.

[9] J. Polo et al. "A simple approach to the synthetic generation of solar irradiance time series with high temporal resolution". In: *Solar Energy* 85.5 (2011), pp. 1164–1170. ISSN: 0038092X. DOI: 10.1016/j.solener.2011.03.011. URL: http://dx.doi.org/10.1016/j.solener.2011.03.011.

[10] J. M. Bright et al. "Stochastic generation of synthetic minutely irradiance time series derived from mean hourly weather observation data". In: *Solar Energy* 115 (2015), pp. 229–242. ISSN: 0038092X. DOI: 10.1016/j.solener.2015.02.032. URL: http://dx.doi.org/10.1016/j.solener.2015.02.032.

[11] B. O. Ngoko, H. Sugihara, and T. Funaki. "Synthetic generation of high temporal resolution solar radiation data using Markov models". In: *Solar Energy* 103 (2014), pp. 160–170. ISSN: 0038092X. DOI: 10.1016/j.solener.2014.02.026. URL: http://dx.doi.org/10.1016/j.solener.2014.02.026.

[12] Carlos M. Fernández-Peruchena and Ana Bernardos. "A comparison of one-minute probability density distributions of global horizontal solar irradiance conditioned to the optical air mass and hourly averages in different climate zones". In: *Solar Energy* 112.February 2015 (2015), pp. 425–436. ISSN: 0038092X. DOI: 10.1016/j.solener.2014.11.030.

[13] Köppen-Geiger. *World map of the Köppen-Geiger climate classification update*. http://koeppen-geiger.vu-wien.ac.at/present.htm. Accessed: 2017-03-11.

[14] Solar Radiation data SoDa. *Solar Radiation data*. http://www.soda-pro.com/. Accessed: 2017-03-11.

[15] B.Y. Liu and R.C. Jordan. "The interrelationship and characteristic distribution of direct, diffuse and total solar radiation." In: (1960).

[16] P. Bendt, M. Collares-Pereira, and A. Rabl. "The frequency distribution of daily insolation values." In: (1981).

[17] E4G. *Energy4Growing*. https://www.facebook.com/energy4growing2014/. Accessed: 2017-03-11.

Abstract

Energy access is a challenge for developing countries, and is widely considered a fundamental enabler for economic, social and cultural development. Distributed Renewable Energy systems represent one of the key solutions to provide energy access in developing countries. Indeed, the majority of developing countries have abundant availability of renewable energy resources, in particular solar and wind. Micro-grids are particularly well suited for rural areas, which are cut off from national electricity grids and other major infrastructure. Due to its modularity, ease of maintenance and affordability, PV power has emerged as the most common renewable source for micro-grids.

Due to the high variability of renewable energy resources, a proper design of a micro-grid requires extensive simulations with sub-hourly resolution. Delivering, installing and maintaining equipment for measuring solar irradiation is particularly difficult in rural areas, which are also very sensitive to investment costs. In addition to this, collecting multiple years of data is intrinsically time-demanding, representing a major obstacle in applications where time-to-market is critical. The lack of large amounts of data needed for such simulations is often a major roadblock in rural areas, leading to increased costs, delays and reduced performance of these systems. Whereas hourly and sub-hourly data are rarely available, multiple publicly available databases provided monthly average data for most locations around the world.

The goal of this thesis work is to develop a stochastic, site-independent, methodology to synthetically generate sub-hourly solar irradiation data, based on limited inputs. By site-independent it is meant that the same methodology can be applied to different locations, generating location-specific outputs without major adjustments to the procedure. This methodology should provide inputs to software for the design and optimization of Distributed Energy Systems, particularly for solar technologies.

A tool based on this methodology has been implemented in MATLAB with the final goal to be integrated in PoliNRG (POLItecnico di Milano – Network Robust desiGn), a software package for the robust design of Distributed Energy Systems.

The methodology is composed by a combination of stochastic models, including Autoregressive Integrated Moving Average and Markov models, and empirical correlations.

Rather than directly modeling solar irradiation, the proposed procedure uses the clearness index as auxiliary climatic variable. Using a non-dimensional variable has several advantages and makes generalizing the procedure easier.

The methodology is designed to turn monthly data into sub-hourly values. This transformation is achieved through several steps. For each step, the most suitable model has been chosen, based on the resolution of input and output values. It is possible to decompose the methodology in three main steps: the Bendt's block, the ARIMA block and Markov block.

The first block is used to turn monthly average clearness index data into daily values. The Bendt's correlation is used to model the cumulative distribution of daily clearness index values within each month. This step returns the unordered daily values.

In the second step, the sequence of daily values within each month is determined through an ARIMA(1,1,1) model. This model is used to order the values obtained in the first step through the Bendt's correlation.

In the third step, a second order Markov model has been used to increase the resolution from daily to sub-hourly. In order to obtain reliable results, daily clearness index values are clustered in Markov classes. Each class is characterized by a specific Markov Transition Matrix.

To further reduce the dependency on location-specific data, the calibration procedure of both the ARIMA and Markov models has been generalized. This generalization reduces the input required from the user and enables the application of the model in areas where hourly and sub-hourly data are not available.

It has been shown that the ARIMA model can be safely calibrated at a global level. Both the model orders and its parameters can be determined based on measurements of other locations.

In addition to this, it has been shown that the calibration of the Markov model can be performed for Climatic Areas. Climatic Areas have been defined as proposed by Köppen-Geiger [1]. For each climatic area, a corresponding group of Markov Transition Matrices has been generated. Pre-computed MTMs also decrease the overall computational time of the

procedure.

The methodology and its implementation have been validated by applying them to two specific cases, quite different from each other, Amsterdam and Ngarenanyuki (Tanzania). The results seem to well reproduce the probability distribution of the observed data for both locations, and the statistical parameters are well approximated.

Although there is a slight loss of accuracy due to the generalization of the models, the main advantage of the proposed methodology is its flexibility. Indeed, it can be used to obtain maximum detail when local data are available, or to achieve a good approximation when local data would otherwise not sufficient to model Distributed Energy Systems.

Keywords: Synthetic generation, Solar energy, Renewable energy, Rural electrification, Stochastic models, Distributed Energy Systems

Sommario

L'accesso all'energia è una delle principali sfide per i paesi in via di svilppo, ed è considerato un passaggio fondamentale per lo sviluppo economico, sociale e culturale del pease. I sistemi distribuiti di energia rinnovabile rappresentano una delle principali soluzioni per garantire l'accesso all'energia in queste aree. Infatti la maggior parte dei paesi in via di sviluppo ha una grande disponibilità di risorse energetiche rinnovabili, in particolare sole e vento. Le micro-grids reppresentano una soluzione particolarmente adatta per aree rurali, che generalmente non sono raggiunte dalla rete elettrica nazionale e da altre infrastrutture. I sistemi fotovoltaici, proprio per la loro modularità sono facili da manutenere e economici; essi stanno emergendo come la risorsa rinnovabile più comune da inserire in micro-grids in queste aree rurali.

Per la loro alta variabilità, le risorse rinnovabili, rendono più difficile la progettazione di micro-grids e inoltre richiedono simulazioni estensive con frequenza intra-oraria. Le apparecchiature per le misurazioni di irraggiamento solare richiedono sforzi particolari per una corretta istallazione e mantenimento in aree rurali, in cui i costi di investimento sono particolarmente critici. Inoltre, la raccolta di molti anni di dati richiede un grande dispendio di tempo, questo rappresenta il maggior ostacolo per applicazioni in cui il tempo di attuazione è cruciale. La mancanza di un grande quantità di dati necessari per simulazioni energetiche è spesso un ostacolo per lo sviluppo di progetti in aree rurali che porta all'aumento dei costi e ritardi alla produzione di questi sistemi.

Mentre dati orari ed intraorari sono difficilmente disponibili, molti databases publici hanno a disposizione dati medi mensili per la maggior parte delle località in tutto il mondo.

L'obiettivo di questo lavoro di tesi è quello di sviluppare una metodologia stocastica e non basata sulla località specifica per generare in modo sintetico dati di irraggiamento intraorario, basata su un numero limitato di inputs. Il fatto di non essere basata su una località

specifica permette alla metodologia di poter essere applicata a siti differenti, generando outputs specifici senza ulteriori aggiustamenti alla procedura. Questa metodologia ha l'obiettivo di fornire inputs per software di design e ottimizzazione di Sistemi Energetici Distribuiti, in particolare per la tecnologia del solare.

La procedura proposta è stata implementato in MATLAB con lo scopo di essere poi integrato in PoliNRG (POLItecnico di Milano – Network Robust desiGn), un software per il design robusto di Sistemi Energetici Distribuiti. La metodologia è composta da una combinazione di modelli stocastici, inclusi modelli Autoregressive Integrated Moving Average e Markov, e correlazioni empiriche.

Invece di modellare direttamente l'irraggiamento solare, è stato scelto il clearness index come variabile climatica ausiliaria. Usando una variabile adimensionale si possono riscontrare diversi vantaggi e soprattutto si può rendere la generalizzazione della procedura più semplice.

La metodologia è strutturata in modo da generare dati intra-orari partendo da valori mensili e questa trasformazione si è ottenuta tramite diversi passaggi. Per ogni passaggio si è scelto il modello che meglio potesse riprodurre la trasformazione in funzione della risoluzione dei valori in input ed in output. I passaggi principali sono tre: la sezione di Bendt, la sezione dell'ARIMA e la sezione di Markov.

La prima sezione converte dati medi mensili di clearness index in dati giornalieri tramite la corrrelazione di Bendt, questa modella una distribuzione cumulata di clearness index giornalieri all'interno di ogni mese senza dargli una sequenza reale. Il secondo passaggio è utile a dare una sequenza ai valori giornalieri, ottenuti nel primo passaggio, all'interno dell'anno ed è determinata usando un modello ARIMA(1,1,1). Nel terzo passaggio si usa un modello di Markov per incrementare la risoluzione dei valori da giornalieri ad intra-orari. Con lo scopo di ottenere valori affidabili i valori giornalieri del clearness index sono raggruppati in diverse classi di Markov corrispondenti a diverse condizioni atmosferiche ed ogni classe è caratterizzata da una una specifica Markov Transition Matrix.

Al fine di rendere la procedura meno possibile basata su località specifiche, la calibrazione dei modelli ARIMA e di Markov sono state generalizzate. Questa generalizzazione permette di ridurre gli input richiesti all'utente. Inoltre la generalizzazione da la possibilità di applicare la metodologia in aree dove dati orari e intra-orari non sono disponibili.

La calibrazione del modello di Markov è effettuata per Aree Climatiche, con la divisione proposta da Köppen-Geiger [1]. Per ogni area climatica è stato generato un gruppo specifico di Markov Tansition Matrices e la possbilità di pre-determinare le MTMs diminuisce

ulteriormente il tempo richiesto alla procedura.

La metodologia e la sua implementazione sono state validate applicandole a due casi specifici, Amsterdam e Ngarenanyuki (Tanzania). I risultati sembrano riprodurre bene le probabilità di distribuzione dei dati osservati per entrambe le località, inoltre i parametri statistici sono ben approssimati.

Sebbene ci sia una leggera perdita di accuratezza dovuta alla generalizzazione dei modelli, il più grande vantaggio che ne deriva sta nella flessibilità del modello finale. Infatti, la metodologia può essere usata per ottenere il massimo del dettaglio quando i dati lo permettono, o alternativamente per raggiungere un buon livello di approssimazione nei risultati quando non sono disponibili dati sufficienti.

Parole chiave: Generazione Sintetica, Energia solare, Energia ronnovabile, Elettrificazione rurale, Modelli stocastici, Sistemi energetici distribuiti

Contents

List of Figures

List of Tables

Abbreviations

ACF	**A**uto **C**orrelation **F**unction
AIC	**A**kaike's **I** **C**riteria
ARMA	**A**uto **R**egressive **M**oving **A**verage
ARIMA	**A**uto **R**egressive **I**ntegrated **M**oving **A**verage
CSP	**C**oncentrated **S**olar **P**ower
DRE	**D**istributed **R**enewable **E**nergy
GHI	**G**lobal **H**orizzonatal **I**rradiance
GNI	**G**ross **N**ational **I**ncome
HDI	**H**uman **D**evelopment **I**ndex
INEP	**I**ntegrated **N**ational **E**lectrification **P**rogramme
MTM	**M**arkov **T**ransition **M**atrix
NOCT	**N**ominal **O**perating **C**ell **T**emperature
NREL	**N**ational **R**enewable **E**nergy **L**aboratory
PACF	**P**artial **A**uto **C**orrelation **F**unction
PoliNRG	**P**olitecncio di Milano **N**etwork **R**obust **D**esign
PV	**P**hoto **V**oltaic
SDGs	**S**ustainable **D**evelopment **G**oals
SG	**S**ynthetic **G**eneration
SHS	**S**olar **H**ome **S**ystem
STC	**S**tandard **T**est **C**onditions
STZ	**S**tandard **T**ime **Z**one
TMY	**T**ypical **M**eteorological **Y**ears
UNDP	**U**nited **N**ations **D**evelopment **P**rogramme
UTC	**U**niversal **T**ime **C**oordinated

List of Symbols

Variables	Unit	Description
F_{ps}	$[-]$	View factor from the photovoltaic panel to the sky
F_{pg}	$[-]$	View factor from the photovoltaic panel to the ground
G	$[W/m^2]$	Global irradiance on the horizontal surface
G_b	$[W/m^2]$	Beam irradiance on the horizontal surface
$G_{b,n}$	$[W/m^2]$	Beam irradiance on the normal surface
G_d	$[W/m^2]$	Diffuse irradiance on the horizontal surface
G_o	$[W/m^2]$	Extraterrestrial irradiance on the horizontal surface
G_{sc}	$[W/m^2]$	Solar constant
G_T	$[W/m^2]$	Global irradiance on the tilted surface
H	$[Wh/m^2]$	Global irradiation on the horizontal surface
H_b	$[Wh/m^2]$	Hourly beam irradiation on the horizontal surface
H_d	$[Wh/m^2]$	Diffuse irradiation on the horizontal surface
H_o	$[Wh/m^2]$	Extraterrestrial irradiation on the horizontal surface
H_T	$[Wh/m^2]$	Global irradiation on the tilted surface
AM	$[-]$	Air mass
K_t	$[-]$	Clearness index
K_t^d	$[-]$	Daily clearness index
K_t^m	$[-]$	Monthly average daily clearness index
$K_t^{\Delta t}$	$[-]$	Time-step clearness index
$K_{t,min}$	$[-]$	Minimum daily clearness index
$K_{t,max}$	$[-]$	Maximum daily clearness index
K_t'	$[-]$	Normalized clearness index
K_{ds}	$[-]$	Synthetically generated daily clearness index
K_{ts}'	$[-]$	Normalized synthetically generated clearness index
R_b	$[-]$	Geometric factor

Variables	Unit	Description
β_{PV}	$[^\circ]$	Surface tilt angle
day	$[day]$	Day of the year
δ	$[^\circ]$	Declination angle
DS	$[h]$	Daylight saving hours
E_n	$[minutes]$	Equation of time
ϵ	$[^\circ]$	Terrestrial axial tilt angle
λ	$[^\circ]$	Local observer longitude
λ_{tz}	$[^\circ]$	Longitude of the standard time zone
N_{day}	$[h]$	Number of daylight hours
ω	$[^\circ]$	Solar hour angle
ω_{ss}	$[^\circ]$	Sunset hour angle
ω_{sr}	$[^\circ]$	Sunrise hour angle
ϕ	$[^\circ]$	Local observer latitude
ψ	$[^\circ]$	Solar azimuth angle
ψ_{PV}	$[^\circ]$	Azimuth angle of the surface
ρ	$[\%]$	Reflectivity of the ground
STZ	$[-]$	Standard time zone
θ	$[^\circ]$	Angle of incidence
θ_z	$[^\circ]$	Zenith angle
t_c	$[h]$	Clock time
t_s	$[h]$	Solar time
t_{sc}	$[h]$	Time Correction Factor

Chapter 1

Introduction

1.1 General Overview

Solar energy is one of the fastest growing sources of electricity generation. It can be harnessed directly from the sun, even in cloudy weather. Solar energy is used worldwide and is increasingly popular for generating electricity and heating.

Solar power is mainly generated through photovoltaic (PV) devices or concentrated solar power (CSP) technologies.

Photovoltaics (PV), are electronic cells that convert sunlight directly into electricity. Today PV is ready to play a major role in the global electricity generation mix. Solar PV installations can be combined to provide electricity on a large scale, or arranged in smaller configurations for mini-grids or personal use. In order to bring electricity access to people who do not live near power transmission lines, PV-powered mini-grids represent an appealing option. This is particularly relevant in developing countries with significant energy resources.

The cost of PV panels has decreased sharply in the last decade, making them not only affordable but often the cheapest source of electricity. The Figure 1.2 shows the benchmarked values for all the sectors and drivers of cost decrease and increase as showed by the National Renewable Energy Lab [2].

Concentrated solar power (CSP) uses mirrors to concentrate solar radiation on a collector. The concentrated solar rays heat up a fluid, which creates steam to drive a turbine and generate electricity. CSP is most commonly used to generate electricity in large-scale power plants.

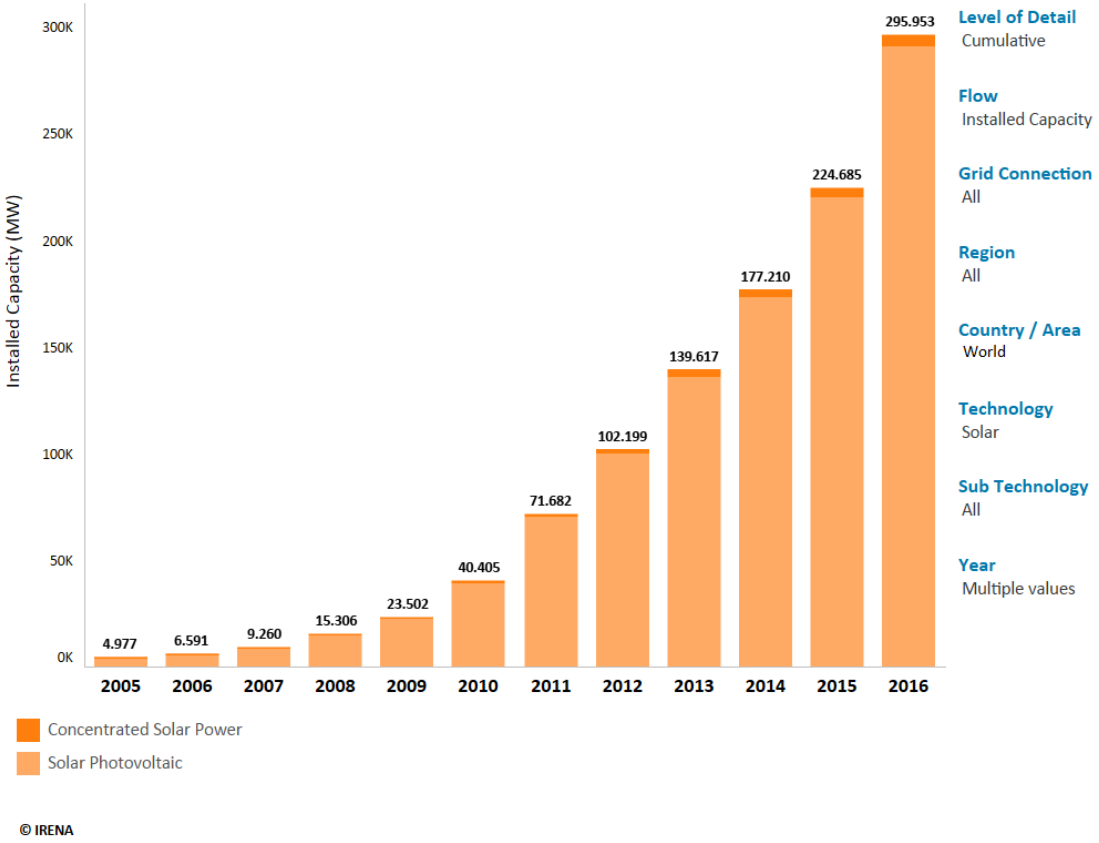


Figure 1.1: Solar renewable energy cumulative installed capacity

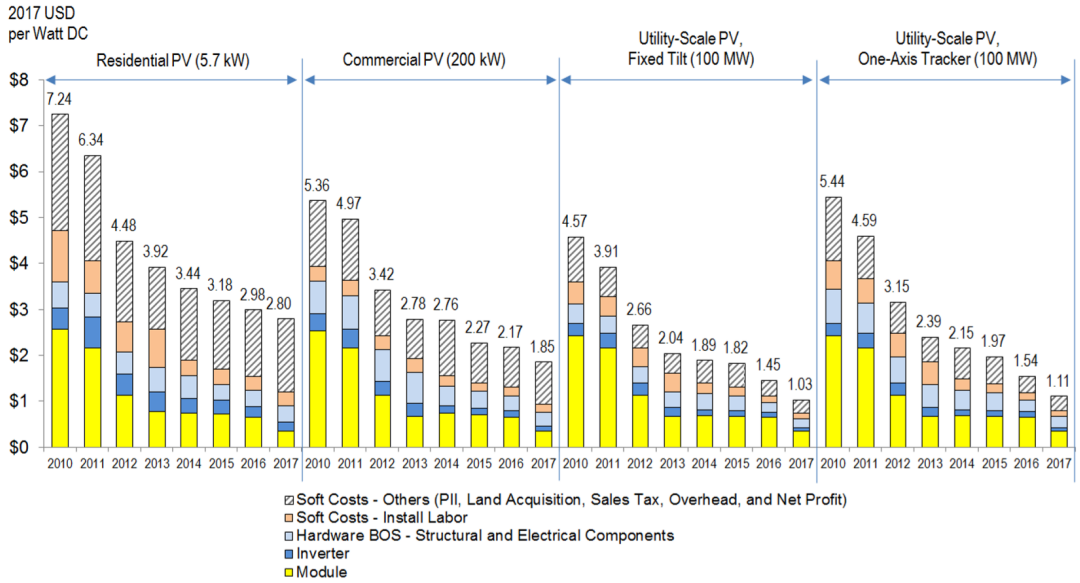


Figure 1.2: NREL PV system cost benchmark summary, 2010–2017

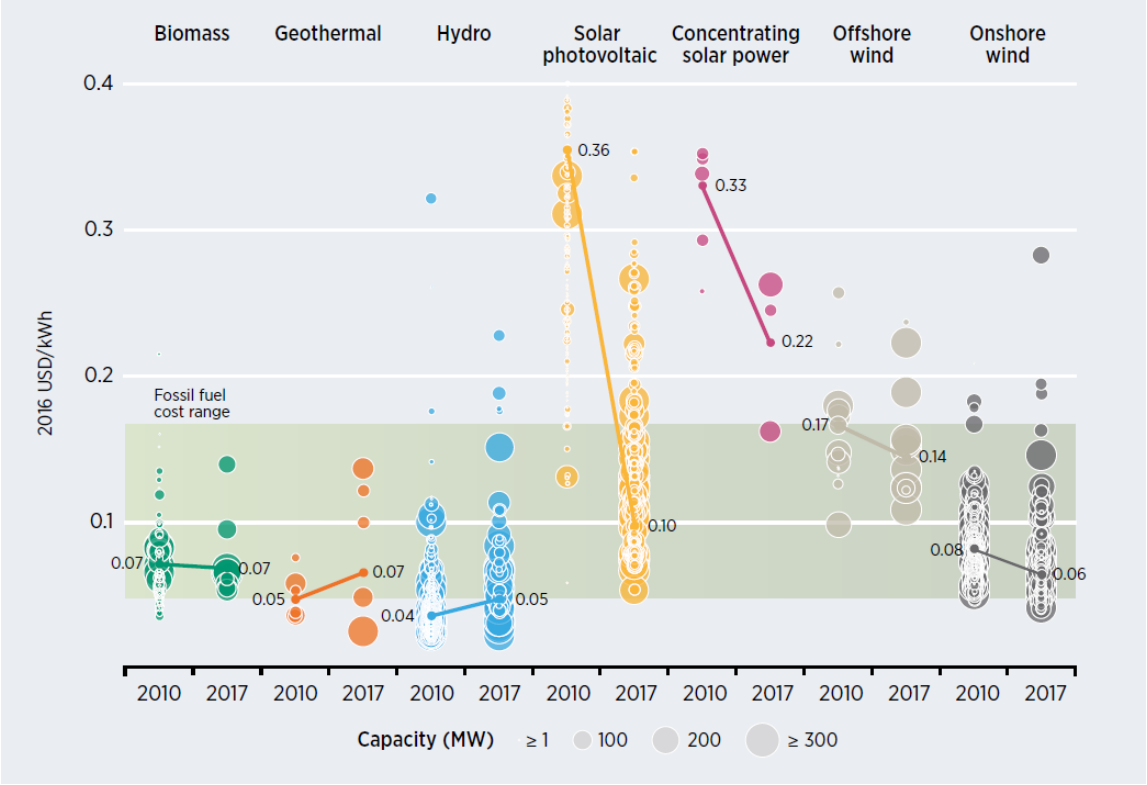


Figure 1.3: Global levelised cost of electricity from utility-scale renewable power generation technologies, 2010-2017. Diameter of the circle: size of the project, with its centre the value for the cost of each project. Thick lines: global weighted average LCOE value for plants commissioned in each year. The band represents the fossil fuel-fired power generation cost range.

One of the main advantages of CSP power plants over PV power plants is that storing heat is usually less expensive, more efficient and easier to scale than storing electricity. This allows the effective decoupling of electricity production from sun availability On the other hand, PV power is currently cheaper than CSP, as shown in Figure 1.3, from the Renewable Cost Database from [3].

In addition to this, PV systems are modular and can be tailored to specific applications, whereas CSP systems are usually very large and require significant investments.

Due to the variability of solar resources and the impossibility to dispatch PV power production, dimensioning and sizing solar plants requires extensive simulations with hourly and sub-hourly resolution; this requires significant amounts of high-resolution, location-specific data. Although monthly or daily simulations can be used for high-level analyses, they are not suitable for detailed evaluations and optimizations of PV powered systems.

The need for multiple years of high-resolution data is particularly critical in developing countries and remote areas, where installing the necessary equipment is often difficult and

expensive. For the reasons mentioned above, in many of the areas with the highest potential for these solar technologies, data (if available) are scarce, expensive and present significant gaps. Furthermore, the investment needed to obtain these measurements can represent a significant share of the total cost for small projects, such as micro-grids and distributed energy systems.

Finally, the required data need to be collected throughout multiple years. This can significantly affect the time needed to complete the project.

The purpose of this thesis is to develop a simple, affordable and widely applicable methodology that can compensate this lack of data and enable faster and more accurate evaluation and design of distributed energy systems. Furthermore, a site-independent procedure that is applicable worldwide is sought.

The procedure, from a limited input, allows to generate multiple years of meteorological data also with the high resolution for a better optimization of solar energy systems. Indeed, the methodology is specific for the case of the generation of solar profiles and often, data available for this kind of task, are not only limited but do also have a low temporal resolution (daily or monthly).

This work is developed using well known stochastic models, such as autoregressive, moving average and Markov models, that can be useful to create a complete model. Complete model means that the solar radiation data can be represented as a combination of two components, deterministic and stochastic.

The deterministic component, that takes into consideration the daily or yearly cycles, and the stochastic one that take into account the variable nature of solar power.

The model introduced above allows to achieve two things:

- generating multiple years of data;
- increasing the resolution from monthly to sub-hourly.

To validate the results, the generated data have been compared to satellite-derived solar radiation data obtained from the SoDa database, [4]. This dataset contains a fundamental limitation: it doesn't provide a vast amount of free data with a high resolution, but only three years. Then, an improvement of the model could be to find a more extensive dataset.

Having a more extensive dataset could mean also increase the resolution of the data, for instance from 10-min until 1-min or higher time step.

1.2 Description of the work

PV systems rely on time-varying solar resources their power production cannot be dispatched. Consequently, extensive simulations are needed in order to assess the potential, limitations and optimal ways to integrate PV technologies in energy systems. Models used to perform aforementioned simulations rely on high-resolution, location-specific data for multiple years, which are often not available.

In order to compensate this lack of data, this thesis aims at developing a site-independent methodology to generate weather data from limited inputs. This would enable simulations of PV systems in locations that lack detailed weather records.

A number of models for the generation of solar radiation profiles have been proposed in scientific literature [5, 6, 7]. A majority of these use the clearness index as climatic variable, a ratio of the global solar radiation on the earth's surface to the extraterrestrial solar radiation. Most of these models use either Autoregressive Moving Average (ARMA) or Markov Transition Matrix (MTM) methods. Often they have been constructed to generate daily or hourly radiation profiles, because of the availability of input data.

Due to the fact that only one year of data are not sufficient to simulate solar powered energy systems, some methodologies have been excluded and the research for the approach has been focused on methodologies that can generate multiple years of data, commonly defined synthetic generation techniques.

The methodology for synthetic generation of solar radiation profiles has been based on [6] and [5]. [5] proposes a methodology to calculate synthetic daily solar radiation values using as input the monthly average clearness index. Subsequently, an ARIMA stochastic model is used to determine the sequence of the daily clearness index values within each month. One of the advantages of the ARIMA model is that its calibration can be generalized to accurately reproduce clearness index values for any location. This allows to use daily clearness index data from several locations around the world to build a model that is applicable to locations where daily measurements are not available.

As discussed in [7], clearness index profiles have significantly different statistical properties at daily and hourly resolution if compared to 1-min, 5-min or 10-min resolution. Indeed, the distribution observed using 5 min data in [7] is bimodal, and this property disappears as the data are aggregated over longer time intervals. Therefore, different statistical models are needed when dealing with data at higher resolutions.

In order to convert daily clearness index values obtained through [5] into sub-hourly data, the methodologies proposed in [6] and [7] have been used. In particular, [6] presents a methodology to pass from daily clearness index data to 1-min resolution data using a Markov model. Although the Markov model requires input data with the same resolution as its outputs, it has the advantage that it can be calibrated with data from other locations in the same climatic area [6, 8].

The ability to base the calibration of the ARIMA and Markov models on data obtained in other locations enables the generation of multiple years of sub-hourly data based on monthly average data spanning a lower number of years.

Although the proposed methodology can define data up to a resolution of 1 minutes, the MATLAB implementation has been designed to achieve a 10-minute resolution because of the available dataset for the calibration of the Markov model. Each model or process used to build the methodology has been internally validated and subsequently inserted as part of the overall procedure.

The final procedure is general and site-independent, so it can be applied in any part of the world to obtain high resolution solar radiation profiles. Using the clearness index as climatic variable is key to make the process site-independent. A validation of the procedure has been performed on different localities across the world. The relative error computed on the total energy produced in one entire year of data lies between 1% and 2% and the cumulative distributions correspond.

After obtaining the global solar irradiation throughout the clearness index, the measure needed for the simulation of energy systems is the power output of the PV panels. In this part of the work a correlation demonstrated in [9] is used to split the global solar irradiation in its beam and diffuse components. The global radiation is split into its components based on the clearness index value. Using these two components in an isotropic sky model and some design data of the PV panel, it is possible to compute the power production.

The proposed model could be improved in the future thanks to a more accurate correlation to split the global solar irradiation in the two components. Some more detailed correlations exist and have been tested, but they requires site-specific parameters. This would make the overall procedure harder to generalize [10]. Therefore, it has been decided to use a more general relationship.

The entire methodology has been implemented in MATLAB, a commercial numerical computing programming language, but could be carried out in any other language.

1.3 Thesis' Framework

The thesis is structured as follows.

Chapter 2 presents a review of scientific literature pertaining to this topic, and a general overview of the related research area is presented. In addition to this, the motivation and relevance of this thesis work are thoroughly discussed. Furthermore, an overview of available options to generate radiation profiles is presented, arriving to the description of the methodology chosen for the thesis. At the end of this section, an overview of the mathematical models applied in this work is provided. Furthermore this chapter shows a general overview of several fundamental correlations and climatic variables that will be used in the proposed methodology. The overview begins with the description of the solar angles used to derive some geometrical solar quantity, such as the solar extraterrestrial irradiance. Subsequently, some correlations used to derive power output from PV panels are introduced.

Chapter 3 describes the proposed methodology step by step from a mathematical point of view to pass from monthly input data to sub-hourly resolution output data. Starting with the input data, continuing with Bendt's correlation and its derivation, and concluding with the description of how the two stochastic models (ARIMA and Markov) have been used to give the sequence to the data along the year and to increase the resolution of the data.

In *Chapter 4* the calibration of ARIMA and Markov models and its MATLAB implementation are presented. In this part the calibration procedure is outlined in a general but detailed way, providing a guide to its application. A general calibration that is suitable for any location has been implemented, but the same procedure can be followed to create a more specific calibration if enough data are available for the location under study.

Chapter 5 describes how the calibration procedure has been generalized to allow the methodology to be used in any location of the world with very limited input data. Indeed, this generalization provides a way to synthetically generate a large amount of output data with a high resolution only based on monthly average data.

In *Chapter 6* the results are reported for two specific locations. Some important modeling choices are discussed for these specific cases, depending on the available data sets. Subsequently, the methodology is validated by analyzing the results for the two specific locations. The validation procedure is divided in two parts. The first one describes some internal validation for each model used. In the second part, the general validation is described.

In *Chapter 7* the conclusions are illustrated, together with an analysis of some limitations

of the procedure and the future developments.

Chapter 2

Goals and Mathematical Models

2.1 Goals

First of all, it is important to have in mind the status of energy access around the world, to have a real awareness and to fully understand the work done. To do this some data from [11] will be reported.

Approximately 1.19 billion people, 16% of the global population, did not have access to electricity in 2014, and about 2.7 billion people, 38% of the global population, live without clean cooking facilities, [11]. The vast majority of people without access to electricity and clean cooking facilities live in sub-Saharan Africa and Oceania, and most of them live in rural areas. In Figure 2.1 the world's map reports some percentage of the energy access around the world.

In Asia, countries such as China and Malaysia have made great strides towards electrification. Elsewhere in the region, however, high percentages of national populations remain without access to modern energy.

In addition, the number of people using on firewood, charcoal or crop residue to meet their household needs is more than 63% in India, 33% in China, 89% in Bangladesh and 38% in Indonesia. Although the Middle East and Northern Africa regions have electrification rates of almost 92% and 99%, respectively, in some individual countries high percentages (from 30% to 60%) of the population still lack access to clean and reliable energy. Similarly, in Latin America and the Caribbean, although 95% of inhabitants have access to grid electricity, millions of people without access are concentrated in six countries: Bolivia, Colombia,

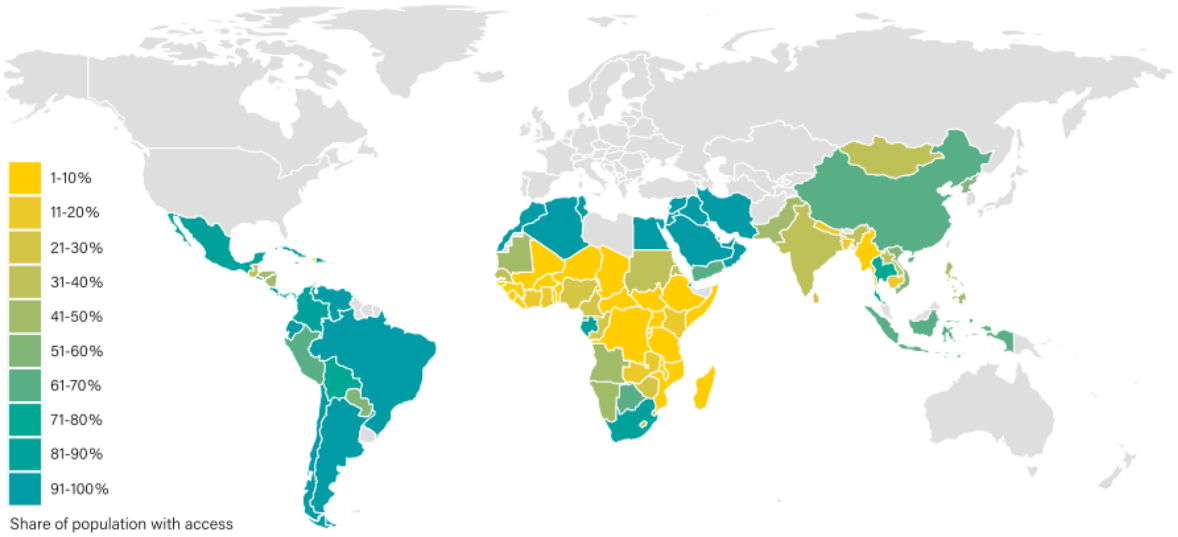


Figure 2.1: Electricity Access in Developing Countries, 2014

Guatemala, Haiti, Nicaragua and Peru.

People in rural and remote regions generally have three possibilities to improved access to energy.

- Household-level: use of isolated devices and systems to generate power and heat for space and water heating, cooking and productive uses.
- Community-level: renewable energy-based mini-or micro-grid systems.
- Grid-based electrification, where the grid is extended beyond urban and peri-urban areas.

Focusing on the first two distributed approaches of improving energy access the possibilities include small-scale solar PV and stand-alone systems. Distributed energy use varies by price, resource availability and type of household.

In recent years, off-grid solar energy has been one of the fastest growing industries in providing energy access in rural areas. Between 2010 and 2016, about 23.5 million off-grid solar systems were sold worldwide. In 2016 alone, nearly 8.2 million off-grid solar systems were sold, representing a global increase of 41% compared to 2015. By 2016, more than 100 companies worldwide actively focused on stand-alone solar lanterns and solar home system (SHS) kits. Sales were highest in sub-Saharan Africa and roughly 10% of the 600 million people living off-grid on the African continent are supplied with energy through DRE (Distributed Renewable Energy Technologies) systems. More than 23 MW of mini/micro-grid

Figure 2.2: Goal 7: Ensure access to affordable, reliable, sustainable and modern energy for all [14].

projects based on solar PV and wind power were announced in 2016 in Africa.

Looking at these data and the high penetration in power generation of solar technologies, it is important to study and understand the effects they would have on the energy system, for better planning system distribution operations.

2.2 Energy Access and Development

The Sustainable Develolpment Goals [12] specifically mention access to modern energy and clean cooking facilities among its goals. The world energy outlook shows that about 1.19 billion people have no access to electricity and about 2.7 billion people still rely on biomass for cooking. In Sub-Saharan Africa 500 millions people live in rural areas without access to electricity [13].

Referring to the SDGs, the objective of most rural electrification programs in the developing world is to bring socioeconomic development to households by 2030. The Goal 7 [14] is "Ensure access to affordable, reliable, sustainable and modern energy for all" in Figure 2.2.

Access to energy for all is essential for every aspect of the life and it's the base for increasing incomes, providing security, food production and employment particularly in rural areas. Indeed energy access is one of the major challenges and opportunities the world faces today. This is particularly true taking into consideration the central importance of the issue of climate change.

After the ratification of the Kyoto Protocol by most of the countries in the world, the global energy policy framework has undergone a paradigm shift. Indeed the human soci-

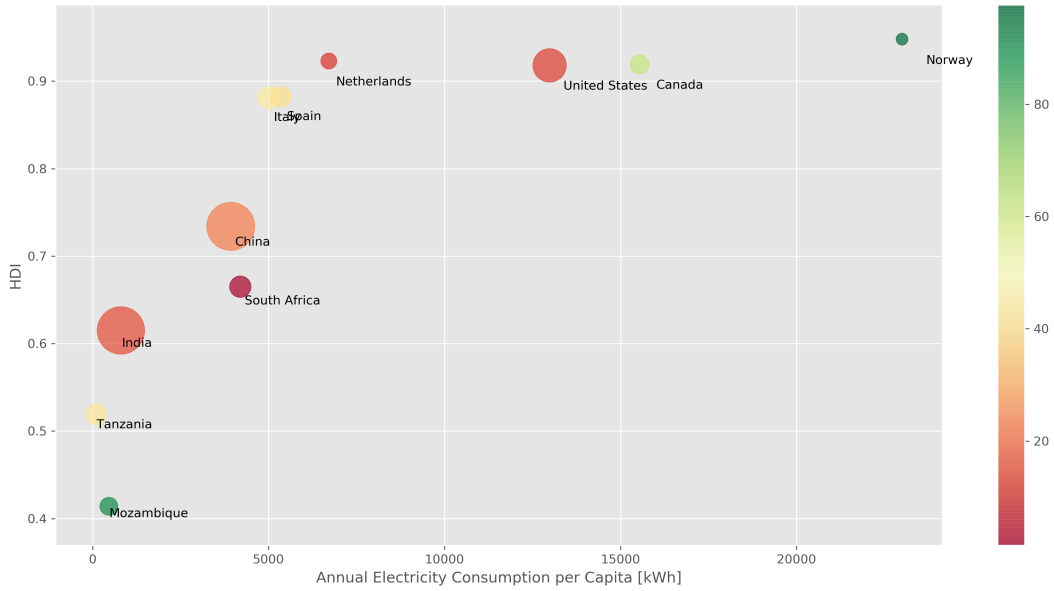


Figure 2.3: Electricity consumption and human development index. Bubble size: population. Color: percentage of renewable generation in electricity mix. Data source: World Bank, World Development Index (2013).

eties have been leaving aside the sole economic growth to embrace the broader concept of sustainable development [13]. In this new vision the electricity ensures enhancement of the qualitative criteria of the human development.

Indeed today the level of development of a country can be defined by an indicator, the human development index (HDI), ranging from 0 to 1, which is yearly calculated by the United Nations Development Programme (UNDP) [15]. The HDI is computed taking into consideration three indices and the level of quality of health, education and income. The three indicators are: the life expectancy index (based on the life expectancy at birth), the education index (based on the mean and expected time of schooling) and the gross national income (GNI) index (based on the GNI per capita).

In the Figure 2.3 it is possible to see that higher electricity consumption is correlated with higher development and human welfare indicators.

In a framework in which the that total primary energy supply has grown more than twice in 40 years [16], mainly for supplying the economic growth, the access to electricity holds a central role for the human development.

The Sustainable Energy for All [17] initiative has been lead from the UN Secretary-General

Ban Ki-moon to ensure universal access to modern energy services, improve efficiency and increase use of renewable sources.

In the following paragraphs, a review of relevant scientific works is provided.

For integrating renewable resources in power systems in developing countries two possible strategies can be considered: the grid-connected architecture for meeting the sustainable development requirements, and the rural off-grid electrification for settling the lack of human development [13].

Despite the preference for grid extension, alternative approaches to electrification have been pursued as well in order to meet the universal electrification objectives by 2030. The disadvantages in term of cost of grid extension are particularly elevated in achieving remote area [18]. Indeed, Photovoltaic (PV) technology is one of the first among several renewable energy technologies that was adopted globally for meeting basic electricity needs of rural areas that are not connected to the grid [19].

South-Africa is a particularly interesting case for assessing the advantages and disadvantages of micro–grid systems compared to the expansion of national grid. Indeed [20] investigate the possibility of a hybrid mini-grid as a solution for rural development in South Africa. The study was developed for two rural villages in South Africa Lucingweni and Thlatlaganya, where the mini-grid and solar home stand-alone systems have been introduced, respectively. [21] tries to argue against the use of the mini–grid system in South Africa because of its high electricity production cost [21].

As an alternative, the solar home system (SHS) was chosen as the preferred technology to electrify rural households from Integrated National Electrification Program (INEP) [22]. This program was aimed at increasing energy access to deprived households. The SHS is not capable of supporting the level of demand of the entire village despite its cost-effectiveness.

[20] reaches the conclusion that, with adequate planning and optimization of available resources, the cost of electricity production can be reduced and mini-grid projects are feasible in rural areas. It does so by investigating the optimal energy mix needed to produce electricity at the lowest cost in two rural villages.

Energy supply to rural areas in developing countries is a complex issue that transcends the simple selection of a best technology. Frequently, the methods for evaluating alternatives for the electrification focus only on few aspects. Indeed, in most cases the factors that influence rural electrification are neglected, given that the majority of the proposals do not consider the needs of the population [23]. It is really important to identify an energy-planning process

that is coherent and understandable based on the resident's requirements and conditions of the local environment.

Related to this, [24] tries to obtain the best possible formulation of the multi-criteria decision-making problem related to the alternatives for a robust planning that is coherent to the local environment and needs of resident. This multi-criteria decisional tool makes electrification planning a multidimensional process with the aim to better respond to the requirements of decentralized energy planning in rural areas.

The 7th SDG is also a significant policy challenge [25]. Often, rural area have weak institutional frameworks and lack regulatory policies. There are a set of policy recommendations for helping the extension of local mini-grid systems in non-electrified areas of the developing world. An important general criterion is "one size does not fit all", indeed each solution needs to be tailored to the local context.

To conclude, in developing countries there is a growing demand for distributed systems and in particular off-grid PV systems. These systems, that can substitute diesel generators or can act as back-up of weak national grids, are recognized to be a reliable and cost-effective solution which can supply electricity in remote and un-electrified areas.

The study of the sizing optimization and designing is an important technical issue about off-grid PV systems [26]. Indeed, the unpredictable energy resources have to be matched with unknown or uncertain load demands with the aim to size the components of these systems (i.e. PV array, battery bank, inverter,etc.) and to provide the most favorable conditions in terms of reliability and costs.

Due to all these above mentioned aspects is really important optimize system in a clear way and evaluate all the various possible scenarios. In the optimization and design the amount and type of data available plays a fundamental role. In particular it is important have multiple years of data with an high resolution to model better than possible the uncertainties.

2.3 Modelling Distributed Renewable Energy technologies

The most common techniques for the analysis and design of distributed energy systems are based on their simulation. Indeed, it is not possible to approach these problems analytically.

Several tools have been developed for the simulation and optimization of distributed energy systems. Such tools can be classified based on a number of criteria; as listed in the following.

- Scale of application: regional, country-level, local.
- Level of detail in the simulation: aggregated systems, single technologies.
- Time-resolution of the simulation: yearly, seasonal, daily, hourly, sub-hourly.
- Extent of the simulation: lifetime of the system, typical year, typical days.

In particular, the last two points determine the required input data. In general, the longer and the more detailed the simulation, the larger the required input dataset. Some tools include functionalities aimed at adjusting the resolution and extent of the input time series, whereas others require input data with specific properties. In addition to this, some commercial tools have access to databases or internal resources while others entirely depend on user provided input. Regardless of the specific case, obtaining or generating such input data represents a critical issue for the reliability and scalability of these tools.

This thesis work is focused on the simulation of solar-related technologies. More specifically, it aims at providing a technique to generate solar radiation profiles required as input by the tools discussed above.

Below, a non-comprehensive overview of such commercial tools is provided, with a brief description of how solar input data are obtained or generated.

HOMER (Hybrid Optimization of Multiple Electric Renewables) is a micro-grid software for optimizing micro-grid design in all sectors. HOMER has been developed in the U.S. Department of Energy's National Renewable Energy Laboratory (NREL). The scale of application varies from village power and island utilities to grid-connected campuses. For whet concern the solar power HOMER requires as inputs the monthly average clearness index. Through a times series model for the clearness index it synthetically generates hourly solar global radiation data.

TRNSYS TRaNsient SYstems Simulation Program and has been developed by Solar Energy Lab [27]. This software is used to study passive solar heating systems and to model solar energy applications with very precise unit size.

PoliNRG (Politecnico di Milano - Network Robust Design) is a novel software package for the robust design of off-grid electric power systems [28]. PoliNRG has been developed in the Department of Energy in Politecncio di Milano and CESI s.p.a. by some researchers ad professors. In the input processing block the input data are elaborated to obtain

load and sources profiles over the entire lifetime of the plant to modelling and simulate different off-grid system configurations and evaluates the related techno-economic performances.

Openmod The Open Energy Modelling is an initiative that promotes open energy modelling in Europe [29]. This software has been developed from energy modelers from various institutions. "Open" refers to model source code that can be studied, changed and improved as well as freely available energy system data. The majority of inputs data can be found in the link where are collected links to data relevant for the modelling of energy and electricity systems and markets.

The proposed methodology has been implemented as a component of PoliNRG, but it could be used to provide solar inputs to any other tool.

Furthermore, sizing distributed energy systems based renewable energy resources means matching unpredictable energy renewable sources with uncertain demands while providing the best reliability and costs. Nevertheless, no particular attention has been devoted so far in the scientific literature concerning specific approaches for daily load profiles.

In PoliNRG the effect of rural users' energy consumptions uncertainty on the sizing of these systems has been appropriately investigated [26]. The effect of load profile uncertainty on the off-grid and distributed PV systems are analyzed to obtain profiles as input data for the design of off-grid systems taking into consideration the specific features of rural electrification areas. Indeed the optimum configurations are significantly affected by load profiles. Due to this issue it has been introduced an innovative stochastic method to formulate different possible realistic daily load profiles for un-electrified rural areas. Consequently an approach to identify the robust solution with regards the assumed uncertainty is proposed.

The procedure is based on a set of data that can be surveyed and/or assumed in rural areas, and it relies on a stochastic bottom-up approach with correlations between the different load profile parameters (i.e. load factor, coincidence factor and number of consumers) in order to build up the coincidence behavior of the electrical appliances. The procedure has been implemented in a software tool which can eventually support the off-grid systems design process for rural electrification [30].

2.4 Problem Introduction

The amount of solar radiation reaching the Earth upper atmosphere is a quantity rather constant over time. On the contrary, the radiation reaching the Earth surface is random in nature, due to the gases, clouds and dust within the atmosphere, which absorb or scatter radiation at different wavelengths. Due to the highly variable nature of solar power, understanding the outcomes of its penetration levels requires in-depth analyses. Indeed at least ten years (standard estimated cycle of atmospheric conditions [5]) of record data are needed in order to be able to analyse the simulated behaviour of energy system.

Obtaining reliable and detailed records of radiation data at ground level requires long-term and high-frequency measurements. Furthermore, it is difficult to collect recorder data for a long time window (years) without any losses in the information. However, there is a dearth of measured long-term solar radiation with spatial density resolution data in a lot of countries, even in developed countries. This situation prompted the development of mathematical and statistical calculation procedures to provide radiation estimates for places where measurements are not exhaustive and there are gaps in measurement records.

2.4.1 Actual Solar Radiation Profile

A large number of solar radiation computation models were developed, spanning from computer codes to empirical relations, widely tested.

For practical purposes, computer codes sometimes are complicated, unpractical and unusable.

In order to predict radiation based on historical weather it is possible to use site-specific radiation models. A site-specific model is based on empirical relationships of solar radiation in terms of recorded weather station variables. Although using a site-specific equation, to determine appropriate coefficients, requires an extensive data set with actual solar radiation data, this kind of approach is simpler and may be more accurate than other complicated models.

These simple approaches can sometimes be extended to nearby locations, to those interested in sites near to where these models are developed.

When choosing among the available models it is necessary to take into account two fundamental factors:

- The availability of meteorological and other data used as input by the model

- The model accuracy.

A method to generate long-term meteorological data, is the generation of data thet can be study as time series.

There are three main types of approach to generate long-term data: measured values, typical meteorological years (TMY) and mathematically generated time series:

Measured values Ground and satellite data. Although they give the most precise informations, there are only few ground stations existing worldwide with the necessary set of information. It's needed at least 10 years of data and it's expensive and time consuming to get them. From satellite data only radiation parameters can be derived.

TMY Typical Meteorological Years of data include variations of several years in one year, consequently a data set of one year (or greater) is necessary for simulation. The idea is to reduce some years of data in a single one process. In general, TMY reproduce well the 10-years statistical distributions of the original data; in particular, extreme values are included. They are site dependent and generally, TMY are only representative for small areas and could reproduce carefully every single conditions.

Mathematically generated time series . These methods can be classified into two subgroups: stochastic methods (Markov chains, autoregressive processes) and Fourier analyses. The advantage of mathematically generated time series techniques is that they are in principle site independent and can be used for any place. The aim of this technique is to designed time series that fit measured values as much as possible in features as the statistical distribution or autocorrelations. One year includes information of several ones as in TMYs. Mathematically generated time series could provide data for any site in the world and the time series distribution correspond to typical values.

2.4.2 Models Based on Synthetic Generation

The requirements in terms of temporal resolution and accuracy depend on the availability of data used to built the model, the considered technology and on the final purpose of the model.

Time series of solar irradiance for energy systems and electrical plant analyses usually come from ground measurements or are derived from satellite images. However, there could be situations that motivate to the use of mathematically or synthetically generated time

series. The latter employ stochastic methods, Markov chains or autoregressive models to build a synthetic time series of irradiance data. All these methods try to fill the gaps in the available data resources obtaining the parameters requested by solar applications.

Deterministic studies, that do not take into account the variable nature of solar power, are only suitable to draw only general conclusions about the effects of increased integration of solar generators in power systems, especially in distribution systems. Probabilistic studies, using solar irradiation as an input to the simulation program, can provide much more detailed insight.

Therefore, it is important to combine these two aspects to generate solar data that can be useful for a large scale of studies, for instance in determining the effects of increased integration of photovoltaic generators in power systems.

2.4.3 Synthetic Generation

Synthetic Generation, SG, is a methodology with the purpose to obtain, from a limited set of data, an amount of data, as meteorological variables, for a period higher than the input one. Synthetic data could be developed in minutes, daily or hourly scale and the general aim of SG is to express daily/hourly distributions and variations without (or with as little as possible) depending on the location. An algorithm for generating synthetic data, can produce any number of yearly data sets. Having a multitude of possible input sets will enhance the testing performances. A number of models for global solar radiation have been proposed in scientific literature. Most of these models have been constructed for daily or hourly radiation, but some are also designed to work with higher resolutions.

Collectioning some articles, it's possible getting to the idea to create a methodology that can generate the sequence of solar radiation through Synthetic Generation.

For this purpose, some stochastic models such as ARIMA/ARMA and Markov Model have been studied and developed to model some stochastic variables related to solar energy.

The research done in this field can be divided in two different categories: methodologies based on hourly and daily resolution, and methodologies based on higher temporal resolution as 1-min, 5-min and 10-min. This distinction is due to the fact that the two resolutions have deeply different statistical properties. Indeed, it has be proven in [8] that high frequency solar irradiance distributions are site-dependent, even if hourly/daily/monthly distributions show universal properties.

In [8] one-minute global horizontal solar irradiation distribution characteristics have been

studied through the meteorological variable clearness index that accounts the atmospheric transmittance. The results of this study highlight the importance of the local distribution for high resolution such as 1-min time step solar irradiance distributions. Furthermore the clearness index is fundamental for differentiating the different distributions

The universal properties of lower resolution distributions are debated in [31] where the statistical properties of hourly global radiation are analysed for various climates and locations. In [31] analytical expressions for the probability distribution, Gaussian functions, of the hourly clearness index are proposed dependent on the daily clearness index and the solar altitude angle.

In order to derive a mathematical model of a physical system, one should know as much as possible about the inputs to the system to be able to decide the different forms of possible responses. This is true for estimating the performance of systems like photovoltaic cells or solar hot water heaters. A high number of studies on probabilistic modeling and synthetic generation of solar radiation data have been focused on daily and hourly data.

In the article [32] has developed a method to identify the cyclical components of some climatic variables, including solar irradiance and ambient temperature. After decoupling the steady state periodic part, analysing the residuals time series shows that the daily solar irradiance residuals are a first-order autoregressive process. This result allows to build a methodology for generating synthetic series for these variables which are statistically indistinguishable from the original time series. Input data modellings different devices performances can be generated.

In [33], the work of J. Boland has been further developed, proposing a classical time series modelling structure to first describe the behaviour of global solar radiation on both daily and hourly time scales. Subsequently, procedures for generating synthetic sequences are presented when only daily values are available. The deterministic component comprises cycles at different frequencies of cycle per year and per day. The residual time series, formed by subtracting the contributions at these significant frequencies from the original measured time series, can be represented by a first-order autoregressive progress (AR(1)).

This study done by J. Hassan, [34], is a first step toward the objective of developing different classical one-parameter-based regression models and time series ARIMA model for the estimation of daily and monthly global solar radiation. The article is divided in two parts, the first one analyses several empirical equations to determine monthly mean daily global and diffuse solar radiation through a regression model. In the second part a time series analysis

has been performed, using an ARIMA(2,1,1) model, to predict the daily clearness index. Although the ARIMA(2,1,1) model is not the best model between the two options, it needs only one weather parameter, the clearness index, to estimate the irradiance, and the result are satisfactory.

J.M. Santos et al., [5], have presented a methodology to develop and generate sequences of synthetic daily global solar radiation values using as input the monthly mean average radiation. Furthermore a stochastic model, ARIMA(1,1,1) is presented and used to define the sequence of the daily clearness index time series.

Increasing the resolution of the data, HOMER [35], a software for microgrid optimization generate synthetic hourly solar data from monthly average data in case of lack of measured solar radiation data. HOMER synthesizes hourly solar radiation data using an algorithm based on the work of V.A. Graham,[36] and [37]. The algorithm, through a stochastic procedure for generating synthetic sets, produces realistic hourly solar irradiation data. The procedure requires only the latitude and the twelve monthly means values of daily events. All the work is based on analyses of meteorological records.

These works reveal that the probability and stochastic characteristics of sets of hourly events within an individual day can be closely predicted using the clearness index of the day. Disaggregating the clearness index allows to create a parameter which is independent of the geographical location. The values of k_t can be modeled by a Beta distribution in the month. A simple autoregressive process AR(1), of the time series of transformed variables it is used to increase the resolution to have a hourly stochastic model.

The results suggest universal behavior, indeed the autoregressive parameter obtained for the locations studied were all very close so they can be used to generalize the procedure at a global scale. In order to better model the performance of electrical generation by solar energy systems, it could be useful to have input data with higher temporal resolution. This would help taking into consideration the impact of weather transient effects.

Some studies are proposed to have a general idea of how it is possible to increase the resolution.

Firstly, a simple approach by J. Polo et al. [38], to generate synthetically irradiance is proposed.

Starting from the hourly mean values, adding a random fluctuation, a resolution of 10-min intervals can be achieved. The characteristic amplitude of the random fluctuation depends on the sky conditions, Comparing the results, despite the noticeable uncertainty in the 10-min

synthetic irradiance values, the dynamic behavior of the fluctuations is comparable to the original data. The boundary conditions imposed to the above method are that the overall potential of the solar resource should remain reasonably constant and the dynamics of the fluctuations should be coherent with the state of the sky.

Hourly data fail to capture the intermittent nature of solar irradiance, therefore some models use hourly weather datasets to generate minutely irradiance time series. In order to increase the resolution to 1-min, J.M. Bright et al. [39], proposes a methodology to generate a synthetic minutely irradiance time series from a large database of hourly weather observation. The weather observation data are used to produce a set of Markov chains taking into account seasonal, diurnal, and cloud dynamics. The use of multiple Markov chains are shown to be fundamental for the generation of realistic minutely irradiance time series over a typical meteorological year.

A model for the synthetic generation of 1-min global solar radiation data starting based on the daily clearness index is proposed by B.O. Ngoko et al. [6]. The database is composed by three-year global solar radiation data taken at 1-min intervals from two locations in Japan used to construct the Markov transition matrices. The Markov transition matrices, built with a second order Markov model, are used to synthetically generate 1-min global solar radiation data through the clearness index.

2.5 Mathematical approaches

It is necessary a short introduction of the models and correlations mentioned above and used in this thesis for the Synthetic Generation of solar radiation data profiles.

- Bendt's Correlation
- ARIMA (Autoregressive Integrated Moving Average) and ARMA (Autoregressive Moving Average) Model
- Markov Model

The introduction is focused on the mathematical aspect of each model and correlation that is at the base of the work done.

2.5.1 Bendt's Correlation

Definition

Bendt's correlation has been proposed in [40] to model the probability distribution of daily clearness index values depending on monthly clearness index.

In [40], the authors study the statistical characteristics of solar radiation using atmospheric transmittance, also called clearness index, as a random variable. The researchers show how the daily clearness index is related to the monthly average clearness index. In [41] and [42], a parametrical expression of the distributions obtained in [40] is presented.

Below it is reported a description of the process.

- In [40], the authors develop an expression for the frequency distribution of the daily clearness index, K_t^d, based on monthly average clearness index values, K_t^m.
- Studying the trend of the clearness index for 90 locations in U.S.A. over a period of 20 years, they reach the conclusion that K_t^d has an exponential distribution throughout the month, ranging between minimum and maximum values, $K_{t,min}$ and $K_{t,max}$ of each moth.
- The cumulative distribution function of the occurrence frequency curve F, or the fraction of value in which the K_t^d is lower than a certain given specific value, can be computed as follows:

$$F(K_t, k_t^m) = \frac{exp(\gamma \cdot K_{t,min}) - exp(\gamma \cdot K_t)}{exp(\gamma \cdot K_{t,min}) - exp(\gamma \cdot K_{t,max})} \tag{2.1}$$

  Where γ is a particular exponential distribution. The value of γ, which defines the particular exponential distribution, can be determined if the minimum value $K_{t,min}$, the maximum value $K_{t,max}$ and the average monthly value K_t^m are known.

  For the continuous exponential distributions of clearness index values, the minimum is greater than zero (minimum theoretical value), and the maximum value will be less than one, (maximum theoretical value). Using these limits and the average value of the continuous exponential distribution with the expression (2.2) it is possible derive the value of γ:

$$K_t^m = \frac{(K_{t,min} \cdot \frac{1}{\gamma}) \cdot exp(\gamma \cdot K_{t,min}) - (K_{t,max} \cdot \frac{1}{\gamma}) \cdot exp(\gamma \cdot K_{t,max})}{exp(\gamma \cdot K_{t,min}) - exp(\gamma \cdot K_{t,max})} \tag{2.2}$$

The derivation of the formula about the occurrence frequency, F, come from Bouger-Lambert's exponential law of radiation absorption as described in [43].

Bouger-Lambert's law: When radiation passes through a homogeneous medium, this produces an exponential distribution of the clearness index if the extinction coefficient of the medium is considered to be a variable with a linear distribution. The variable can vary between a maximum value, different from $+\infty$, and a minimum one, greater than 0. The extinction coefficient is the sum of factors that absorb solar radiation like clouds, aerosols, gases and drops of water.

However, Bouger-Lambert's law can only be applied to direct radiation through an uniform medium with radiant properties independent from the wavelength.

Extreme values

A lot of study has been done to generalize as much as possible the value of the minimum and maximum clearness index for each month and to find a formulation independent on the specific locality. For what concern $K_{t,min}$, Bendt et al. [40] determined a value of 0.05, which can be assumed to be constant and independent of the location. Instead, for $K_{t,max}$ multiple relationships have been proposed, starting from a fixed value as 0.864 passing through an expression dependent on the area studied in [42], arriving to a procedure generalizable for each geographical area. With this last correlation, proposed in [44], it is possible to obtain an accurate maximum clearness index for each month having the monthly mean daily clearness index, the solar declination, δ, the latitude of the location, ϕ and the altitude of the area, z.

$$K_{t,max} = 0.51585 + 0.34847 \cdot k_t^m + 2.302810^{-4} \cdot \delta + 3.410810^{-4} \cdot \phi - 9.570910^{-6} \cdot z \quad (2.3)$$

It is now possible to obtain the exponential distribution of the daily clearness index F from the value of its monthly average K_t^m. The daily clearness index doesn't have a real sequence inside the month but only an estimation of the values.

2.5.2 ARMA/ARIMA Model

Definition

When a series Y is only expressed in terms of its past values and the current and past values of error terms this model is called Autoregressive Integrated Moving Average or $ARIMA(p, d, q)$. Where p, d and q are the model *orders* and are defined as follows.

- p or (AR) term: is the number of lagged values of Y which represents the autoregressive nature of model.

- q or (MA) term: is the number of lagged values of the error term which represents the moving average nature of model.

- d: is the number of times Y has to be differences to produce the stationary series y.

Sometimes in order to obtain a forecast of Y, we have to integrate the values. If no differencing is involved, this model is called an Autoregressive Moving Average or $ARMA(p, q)$, with the order p and q as above.

Forecast for y at time t is equal to a *constant* μ plus a weighted sum of the last p values of y plus a weighted sum of the last q forecast errors. In most cases, $p + q$ is less than or equal to 3.

Then the ARMA/ARIMA equation for predicting y has the following form:

$$y_t = \mu + \phi_1 * y_{t-1} + ... + \phi_p * y_{t-p} - \theta_1 * a_{t-1} - ... - \theta_q * a_{t-q} \tag{2.4}$$

where:

- ϕ: Auto Regressive parameters.

- θ: Moving Average parameters.

- μ: Constant.

- a: Forecast error.

Before using the model, it must be calibrated. This means defining the ARIMA model orders (p, d, q) and parameters (ϕ, θ). Several qualitative and quantitative techniques can be used to calibrate the model. After this step, it is always necessary to validate the obtained ARIMA model, by observing the residuals.

Choice of Model orders

Firstly, determine the value of p and q that should be used in the equation for predicting the stationaries series y. Then, try some standard combinations of p and q and look at plots of the autocorrelations and partial autocorrelation of y:

- The AutoCorrelation, (ACF Autocorrelation Function), of y at lag k is the correlation between y and itself lagged by k periods.

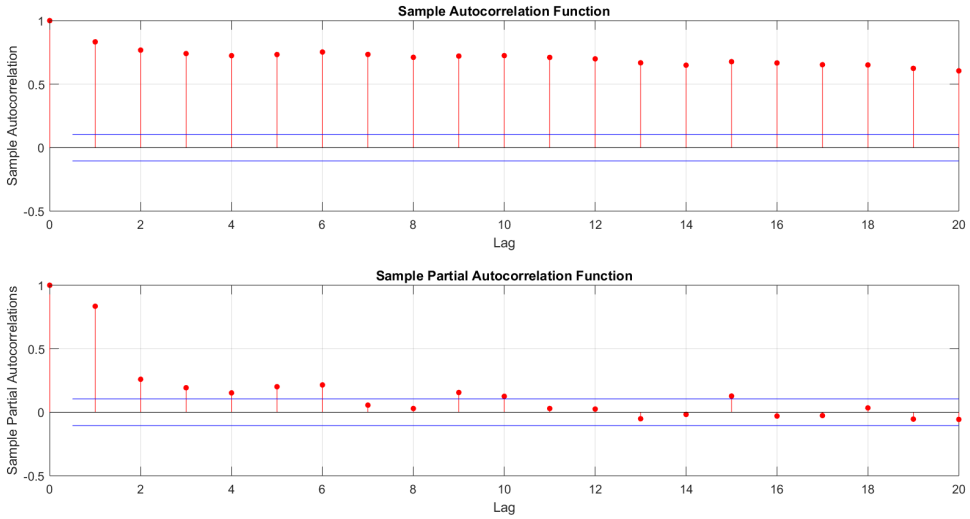


Figure 2.4: Example of ACF and PACF until lags 20 for a time series.

- The Partial Autocorrelation (PACF, Partial AutoCorrelation Function) measure the link between y_t and y_{t-k} net of the influence exercised by all the intermediate variables. PACF is also called conditional correlation and is really important for the identification of stochastic process.

Some examples of AutoCorrelation and Partial Autocorrelation Function are plotted in Figure 2.4.

Some rules for determining p and q from the are reported below.

- ACF plot "cuts off sharply" at lag k, while in the PACF plot decays gradually, then set $q = k$ and $p = 0$.
- The PACF plot "cuts off sharply" at lag k while there is a more gradual decay in the ACF plot, then set $p = k$ and $q = 0$.
- If there is a single point at lag 1 in both the ACF and PACF plots, higher two times than the limiting line, then set $p = 1$ and $q = 0$ if it is positive, and set $p = 0$ and $q = 1$ if it is negative.

In addition to the qualitative criteria mentioned above, several quantitative techniques are available:

- Methods based on the likelihood based inference, that determines the maximum likelihood estimates under both null and alternative hypothesis. Likelihood is used in statistic after data are available to describe plausibility of a parameter value [45].

- Methods based on the determination of statistical values such as P-Value and T-student levels.
- Methods based on AIC, Akaike's Information criteria. This test estimates the quality of each model, relative to each of the other models. AIC does not provide a test of a model in the sense of testing a null hypothesis.

None of the methods above is conclusive, and each of them has advantages and disadvantages. Depending on the specific application, multiple tests can be used to reach a conclusion about the best model orders.

Non-stationary time series

In the following steps a general approach (proposed by [45]) to built an ARIMA model is reported:

1. The first step is the identification of the need to do any non-linear transformation such as differencing, logging or raising-to-power on the original time series, Y. These transformations have been done in order to convert the series to a form where its local random variations are consistent over time and symmetric in appearance.
2. If Y is still non-stationary at this point of the procedure, i.e., exhibits random-walk behavior, then apply the first-difference transformation.
3. If it still looks "non-stationary" after a first-difference transformation then apply another first-difference transformation. The total number of differences, d, could be in the range between 0 and 2.

y is the "stationarized" time series. If $d = 0$, then y is the same as Y. A "stationarized" time series has no trend and a constant variance over time.

The methodology used to determine the stationarity of the time series is the Unit Root Test, which is defined below.

Unit Root A unit root process, or a difference stationary process, is a stochastic trend in a time series. The time series can be described as a "random walk with drift". If a time series has a unit root, it shows a systematic pattern that is unpredictable. Therefore, if a time series has a "unit root" it is non-stationary [46] and [47].

Unit Root Test Unit root tests are tests applied to determine if a time series is stationary or not. If a shift in time does not produce a change in the shape of the distribution, the time series can be considered stationary.

Many tests exist to determine if a series is a Unit Root process or not, but these tests have a low statistical power. This is why so many tests exist: none of them stand out as having the most power, and several tests are usually applied to reach a conclusion. Tests include:

Dickey Fuller Test. This test is based on linear regression. Serial correlation can be an issue, in which case the Augmented Dickey-Fuller (ADF) test can be used.

Elliott–Rothenberg–Stock Test. This test has two subtypes: The P-test takes the error term's serial correlation into account, while the DF-GLS test can be applied to de-trend data without intercept.

Schmidt–Phillips Test. This test includes the coefficients of the deterministic variables in the null and alternate hypotheses. Subtypes are the rho-test and the tau-test.

Phillips–Perron (PP). This test is a modification of the Dickey Fuller test, and corrects for autocorrelation and heteroscedasticity in the errors.

Zivot-Andrews test. This test allows a break at an unknown point in the intercept or linear trend.

ARIMA model checking

The residuals are an important criterion for the selection and evaluation of the model used. Ideally, the residuals should be independent and identically distributed, i.e. acting as white noise. They should also have mean equal to zero, otherwise the model is biased.

2.5.3 Markov Model

Markov models – or Markov chains – are mathematical representations of a stochastic process under the assumption that its future state is only depends on the last n states. In other words, if the last n states are known, the probability distribution of the next state is known. Where n is called the Markov order and can be determined studying the properties of the time series being modeled.

Order	Dimension
1	$[m \times m]$
2	$[m^2 \times m]$
$\vdots$	$\vdots$
n	$[m^n \times m]$

Table 2.1: Matrix Dimensions.

Definition

As described above, Markov models can have different orders. The order is a really important characteristic of the model because it determines the dependency of an event on the previous n-events. The order of Markov models can be explained as follows:

- 1^{st} order. Given a process in state i at time $t-1$, the probability that it will be in state j at time t is given by a probability P_{ij} that is fixed. P_{ij} is the transition probability from state i to j and is independent of the states of the process at times $t-2, t-3, \ldots$. A Markov model is described through the Memoryless property: the conditional distribution of Y_t given Y_0; Y_1; ...Y_{t-1} depends only on Y_{t-1}. In Equation (2.5) it is possible to observe this property.

$$P_{ij} = P(Y_t = j | Y_{t-1} = i, Y_{t-2} = i_{t-2} | \ldots | Y_0 = i_0) = P(Y_t = j | Y_{t-1} = i) \tag{2.5}$$

- 2^{nd} order: the probability that the process will be in a particular state k at time t depends not only on its state at time $t-1$ but also on the states at time $t-2$.

- n^{th} order: the probability that the process will be in a particular state at time t depends not only on its state at time $t-1$ but also on the states at times $t-2, \ldots, t-n$.

The transition matrix P_n holds the transition probabilities for the $n-th$ order Markov process. The transition matrices are built based on the input data.

If m states are allowed, the transition matrix of 1^{st} order, P, is a matrix of dimensions $[m \times m]$. For a Markov process of the n^{th} order, the Markov matrix has a dimensions $[m^n \times m]$. Look at Table 2.1.

Markov Model Order

The order of a Markov model determines how many previous observations are used in generating the next state in a Markov process. The order of the model identifies the serial

correlation characteristic of the modeled data. The following steps are use to determine the order:

1. Assume that observation the at time t is statistically independent of the observations at times $t-n-1; t-n-2; ...$

2. If the observation at time t has a significant conditional dependence on the observations at times $t-n-1; t-n-2$, an n^{th} order model would not be able to correctly capture this characteristic.

Markov Transition Matrix

For a first order process, the transition probability from state i to j can be easily estimated by counting the number of times the sequence i, j is observed and dividing by the total number of times i is observed, as shown in equation (2.6).

$$P_{i,j} = \frac{f_{i,j}}{f_i} \tag{2.6}$$

A second order process adds complication to the model. The probability of moving to state k immediately after the observation of states i and j can be estimated by counting the number of times the sequence i, j, k is observed and diving by the total number of times the sequence i, j is observed, as shown in equation (2.7).

$$P_{i,j,k} = \frac{f_{i,j,k}}{f_{i,j}} \tag{2.7}$$

Higher orders follow the same logic.

With sufficient solar radiation data, the data can be discretized into radiation states from which Markov transition matrices can be constructed. Structure of first and second order Markov Transition Matrices:

$$P^1 = \begin{bmatrix} f_{11} & f_{12} & \dots & f_{1m} \\ f_{21} & f_{22} & \dots & f_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ f_{m1} & f_{m2} & \dots & f_{mm} \end{bmatrix}$$

$$P^2 = \begin{bmatrix} f_{111} & f_{112} & \cdots & f_{11m} \\ f_{121} & f_{122} & \cdots & f_{12m} \\ \vdots & \vdots & \ddots & \vdots \\ f_{1m1} & f_{1m2} & \cdots & f_{1mm} \\ f_{211} & f_{212} & \cdots & f_{21m} \\ \vdots & \vdots & \ddots & \vdots \\ f_{mm1} & f_{mm2} & \cdots & f_{mmm} \end{bmatrix}$$

Using Markov models

Once the MTMs are generated, they are used to generate a sequence of values.

In the i^{th} row of the MTM, corresponding to the n previous observations, it is possible to find the probability distribution of the observation at time t. The row corresponds to a probability distribution vector, $P(t)$, of the observation at time t. $F(t)$, is the corresponding cumulative distribution vector if:

$$F_j(t) = \sum_{h=1}^{j} P_h(t) \tag{2.8}$$

where $P_h(t)$ is the h-th entry in $P(t)$, and $F_j(t)$ is the j-th entry in $F(t)$. The state Y_t can therefore be generated by picking a random number r from a uniform distribution between 0 and 1 and mapping it onto $F(t)$. For instance, if $F_{l-1}(t) < r < F_l(t)$, then $Y_t = l$. By assuming that the cumulative distribution curve is linear between states $l-1$ and l, the synthetically generated value is obtained using the transformation 2.9.

$$Y_s = \frac{1}{m} \cdot \left[(i-1) + \frac{r - F_{l-1}(t)}{F_l(t) - F_{l-1}(t)} \right] \tag{2.9}$$

Where $m = 100$ in 2.9 is the number of possible states.

2.6 Solar Irradiance and Climatic Variables

In the following sections, some important parameters and variables used for modelling PV systems are introduced. The discussion begins with a mathematical formulation of solar irradiance and arrives to the power production of PV panels.

2.6.1 Solar Irradiance Components

Before starting it is important to have in mind some basic relationship and definitions about solar energy that could be useful to fully understanding of the work.

The data available in general come from satellite measurements, with a variable spatial resolution that provide some components of the solar radiation. The most common and the most used solar variable is the Global Horizzontal component. The Solar Global Horizzontal Irradiance, GHI, is the total amount of solar radiation hitting the horizontal surface on the earth which differs from the power output of the PV array. The latter depends on the amount of radiation striking the surface of the PV array, which in general is not horizontal. So it is necessary, for each time step, to compute the global solar irradiance on the surface of the PV array. The process to do this is described in the following section.

2.6.2 Solar geometry

Some of the solar angles are fixed and describe the reciprocal position plane-earth, others describe the relationship sun-earth that vary with time, and a last set of angles describe the reciprocal position of sun-plane. The solar position model is presented in the figure:

Relevant factors for the derivation of the solar components are the latitude, longitude, day of the year and time of the day. The time of year affects the solar declination, which is the latitude at which the sun's rays are perpendicular to the earth's surface at solar noon. The following equation is used to calculate the declination angle, angular position of the sun at solar noon with respect to the plane of equator, with North positive and South negative:

$$\delta = \epsilon \cdot \sin\left(\frac{360 \cdot (day + 284)}{365}\right) \tag{2.10}$$

Where ϵ is the terrestrial axial tilt angle and corresponds to 23.45° and δ can vary between -23.45 ° and +23.45°. The time of day affects the location of the sun in the sky, which can be described by the hour angle, ω. The hour angle is the angular displacement of the sun with respect to the local meridian and can vary between -180 ° and +180°. By convention, it is zero at solar noon, time of day at which the sun is at its highest point in the sky, negative in the morning and positive in the afternoon. The equation (2.11) describe how to compute the hour angle ω:

$$\omega = (t_s - 12) \cdot 15^{\circ}; \tag{2.11}$$

All the relationship used to calculate the sun-earth and sun-plane angles use the solar time, (t_s). Solar time, that does not coincide with the clock time (t_c), is the time based on the apparent angular motion of the sun across the sky. The solar noon is the time in which the sun crosses the meridian of the observer.

The solar time, t_s, corresponds to hours 12 at solar noon. The equation (2.12) describes the longitude angle correspondent to the standard time zone:

$$\lambda_{tz} = STZ \cdot 15^{\circ}; \tag{2.12}$$

Where STZ is the Stardard Time Zone, that can be used to describe the local time of a region or a country. The local time within a time zone is defined by its offset from Coordinated Universal Time (UTC), the world's time standard. The equation above is based on the fact that each meridian is 15 degrees, so the sun moves across the sky at 15 degrees per hour.

All time-dependent data, such as solar radiation data and electric load data, are specified in clock time, (t_c), and not in solar time. Then it is necessary to calculate the data in solar time from data in clock time, also called standard time. When switching from t_c to t_s it is important to take into consideration some factors:

- Solar time is measured according to the observer position, whereas the clock time is constant within a given time zone.
- Solar day does not have a constant duration throughout the year, whereas the clock day does.
- Clock time shifts one hour forward during the warm season, spring and summer, in some locations: daylight savings.

Two corrections are applied for the conversion. The first one is a constant correction to delete the difference in longitude between the meridian on which the local standard time is based and the meridian that runs through the observer. This relations 2.13 shows that sun takes constantly 4 minutes to cover 1 ° longitude that corresponds to 15 ° longitude in one hour:

$$4\Big[\frac{minute}{longitude^{\circ}}\Big] \cdot 60^{-1}\Big[\frac{hour}{minute}\Big] = 15^{-1}\Big[\frac{hour}{longitude^{\circ}}\Big] \tag{2.13}$$

So the first correction is 2.14:

$$\frac{(\lambda - \lambda_{tz})}{15} \tag{2.14}$$

where λ_{tz} is the standard time zone longitude and λ represents the observer's longitude.

Solar day does not have a constant duration throughout the year, because the Earth's rotational speed varies along the trajectory around the Sun because the effects of obliquity and the eccentricity of the earth's orbit. So with the Equation of time (2.15) that represent

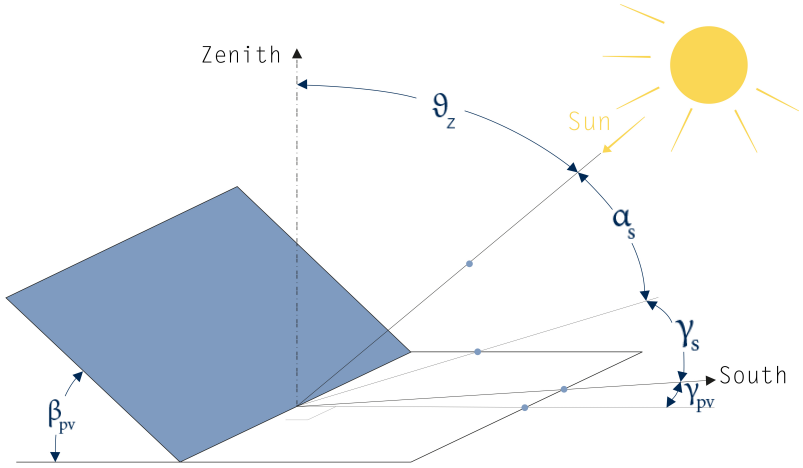


Figure 2.5: PV panel position and solar angles.

the second correction, it is possible to take into account this difference in minutes for each day:

$$\begin{aligned} En = 229.18 \cdot \Big[& 0.000075 + 0.001868 \cdot \cos(D) - 0.03277 \cdot \sin(D) \\ & - 0.014615 \cdot \cos(2D) - 0.040849 \cdot \sin(2D) \Big] \\ D = \quad & 360 \cdot \frac{day - 1}{365} \end{aligned} \tag{2.15}$$

In (2.15) *day* represent the day of the year and can be a value $1 \leq day \leq 365$.

The equation (2.16) represent the Time Correction Factor t_{sc}. It is used for the transformation from clock time to solar time and takes into account the factors introduced above:

$$t_{sc} = \frac{\lambda - \lambda_{tz}}{15} + \frac{En(day)}{60} - DS(h); \tag{2.16}$$

Equation of time is converted in hour and DS is the daylight saving hours, its value is 1 when daylight saving time is in operation (usually spring and summer), otherwise $DS = 0$.

The final equation is (2.17):

$$t_s = t_c + t_{sc} \tag{2.17}$$

Furthermore, any surface could have different orientations, so it's important to define an angle of incidence, θ. The latter is the angle between the sun's beam radiation and the normal to the surface and it can be calculated through the law of cosines as in (2.18a) and

(2.18b) :

$$\cos\theta = \cos\theta_z \cdot \cos\beta_{PV} + \quad \sin\theta_z \cdot \sin\beta_{PV} \cdot \cos(\psi - \psi_{PV}); \tag{2.18a}$$

$$\theta = \cos^{-1}(\cos\theta_z \cdot \cos\beta_{PV} + \quad \sin\theta_z \cdot \sin\beta_{PV} \cdot \cos(\psi - \psi_{PV})); \tag{2.18b}$$

When θ is higher than 90°, means that the sun is behind the surface.

In the equation (2.18b), θ_z is the zenith angle, ϕ is the latitude, ψ is the azimuth angle of the location and ψ_{PV} is the azimuth angle of the PV panel. These angles are visible in Figure 2.5. For a horizontal surface, tilted angle of zero, the angle of incidence corresponds to the zenith angle. The zenith angle, in equation (2.19a) and (2.19b), is the angle between a vertical line and the line to the sun. The zenith angle is 0° when the sun is directly overhead, and 90° when the sun is at the horizon.

$$\cos\theta_z = \sin\delta \cdot \sin\phi + \quad \cos\delta \cdot \cos\phi \cdot \cos\omega; \tag{2.19a}$$

$$\theta = \cos^{-1}(\sin\delta \cdot \sin\phi + \quad \cos\delta \cdot \cos\phi \cdot \cos\omega). \tag{2.19b}$$

The PV array position can be described using two parameters, the tilt angle and the azimuth angle. The tilt angle is the angle that describe the slope of the panel respect to the horizontal plane, if tilt angle is zero means horizontal panel. The azimuth is the direction towards which the surface faces, so it is the angle between projection of the sun on horizontal plane and geographical South. It can be derived in equation (2.20):

$$\psi = sgn(\omega) \cdot \left|\cos^{-1}\frac{\cos\theta_z \cdot \sin\phi - \sin\delta}{\sin\theta \cdot \cos\phi}\right| \tag{2.20}$$

The solar azimuth angle ψ can assumes value in the range of $-180°$ to 180°. By convention, zero azimuth corresponds to south, and positive values refer to west-facing orientations.

To describe the total amount of radiation during the day the last three variables used are the sunset hour angle ω_{ss}, sunrise hour angle ω_{sr} and number of daylight hours N_{day}. Solving the equation (2.20) for $\theta_z = 90°$, gives the sunset hour angle, in (2.21a). The negative of this angle is the sunrise hour angle (2.21b).

$$\omega_{ss} = \arccos(-\tan\phi \cdot \tan\delta) \tag{2.21a}$$

$$\omega_{sr} = -\omega_{ss} \tag{2.21b}$$

$$N_{day} = \frac{2}{15} \cdot \omega_{ss} \tag{2.21c}$$

2.6.3 Irradiance and Irradiation

The amount of solar radiation arriving at the top of the atmosphere over a particular point on the earth's surface, varies over the year due to the variation in the radiation emitted by the sun, negligible, and due to the variation of the earth-sun distance due to the eccentricity of earth's orbit. These two factor affect the solar constant $Gsc = 1367[W/m^2]$ in the range of $\pm 3\%$. To calculate the global extraterrestrial radiation, the amount of solar radiation incident on a horizontal plane at the top of the earth's atmosphere, it has been used the following equation:

$$\begin{aligned} G_o = G_{sc} \cdot [1 + 0.033 \cdot \cos(\frac{360 \cdot day}{365})] \\ \cdot (\cos\phi \cdot \cos\delta \cos\omega + \sin\phi \cdot \sin\delta) \end{aligned} \tag{2.22}$$

The global extraterrestrial radiation outside of the atmosphere is measured in W/m^2.

The integral value of the global extraterrestrial radiation incident on a horizontal plane outside of the atmosphere over the period from sunrise to sunset, is defined as the daily extraterrestrial radiation on a horizontal surface, H_o and is measured in $J/m^2/day$:

$$\begin{aligned} H_o = G_{sc} \cdot \frac{24 \cdot 3600}{\pi} \cdot [1 + 0.033 \cdot \cos(\frac{360 \cdot day}{365})] \\ \cdot (\cos\phi \cdot \cos\delta \cos\omega_{ss} + \frac{\pi \cdot \omega_{ss}}{180} \cdot \sin\phi \cdot \sin\delta) \end{aligned} \tag{2.23}$$

2.6.4 Clearness Index

The clearness index, K_t, is the ratio of the global surface radiation to the extraterrestrial radiation. Numerous studies about Synthetic Generation process has done based on Clearness index. The following equation defines the clearness index:

$$k_t = \frac{HorizontalGlobalSolarRadiation}{HorizontalGlobalExtraterrestrialSolarRadiation} = \frac{H}{H_o} \tag{2.24}$$

The clearness index is the fraction of the solar radiation that is transmitted through the atmosphere that achieve the Earth's surface, so it is a measure of the clearness of the atmosphere. It is a dimensionless number between 0 and 1 and has a high value under sunny conditions, and a low value under cloudy conditions.

The clearness index can be defined on an instantaneous, hourly, daily or monthly character. Modelling a stochastic variable as the clearness index, k_t, the Global Solar Radiation,

H, can be obtained. Typical values of the clearness index vary from 0.25, a very cloudy day, to 0.75, a very sunny day.

2.7 From Global to Beam and Diffuse Radiation

The global solar radiation on the earth's surface can be decomposed into beam radiation and diffuse radiation.

Beam radiation is the solar radiation that coming from the sun hitting earth's surface without any scattering by the atmosphere. The rest of the radiation is the diffuse radiation, that it is the direct radiation scattered by the earth's atmosphere. For the reason why that it is scattered it comes from all parts of the sky. The sum of these two component is the global solar radiation, G:

$$G = G_b + G_d \tag{2.25}$$

The subdivision between beam and diffuse radiation is fundamental when the amount of radiation incident on an inclined surface must be calculated. The orientation of the surface has a stronger effect on the beam radiation, which comes from only one direction, than it does on the diffuse radiation, which comes from all directions. For this purpose, it has been used a correlation tested by Erbs et al. [9], and subsequently used by HOMER [48], a software for the optimization of energy systems. The correlation calculates the diffuse fraction as a function of the clearness index as follows:

$$\frac{G_d}{G} = \begin{cases} 1 - 0.09 \cdot K_t, & K_t \leq 0.22 \\ 0.9511 - 0.1604 \cdot K_t + 4.388 \cdot K_t^2 - 16.638 \cdot K_t^3 + 12.336 \cdot K_t^4, & 0.22 < K_t \leq 0.8 \\ 0.165, & K_t > 0.8 \end{cases} \tag{2.26}$$

This formulation has been validated for a hourly time step. Although other formulations have been proposed for higher resolution time steps (up to 1-minute or higher), they require an amount of additional data that are difficult to collect and not free to everyone. Ultimately, the time step is determined by the time step of the input data. Since data with higher frequencies than hourly are rarely available, it is common practice to use equation (2.26) for higher resolution time steps. Using the global horizontal radiation to calculate the clearness index,

the diffuse radiation can be computed with a good approximation. Finally, the beam radiation is calculated by subtracting the diffuse radiation from the global horizontal radiation.

2.7.1 Incident Radiation on the PV panel

To calculate the global radiation striking the tilted surface of a PV panel, it has been used a isotropic sky model. It relies on some assumptions that simplify the model:

- The PV system does not influence the radiation fields.
- The radiation is uniform around the object.
- The incident radiation is separated in beam, isotropic diffuse and solar radiation diffusely reflected from the ground.

In contrast to the anisotropic sky model, the circumsolar diffuse and horizon brightening components are assumed to be zero. Before applying this model, three more factors must be defined:

- R_b, geometric factor
- F, view factors
- ρ_g, ground average reflectance

The geometric factor, R_b is defined as the ratio of beam radiation on the tilted surface to beam radiation on the horizontal surface:

$$R_b = \frac{\cos\theta}{\cos\theta_z} \tag{2.27}$$

The view factor to the ground, F_{pg}, and to the sky, F_{ps},of a plane with a tilt angle β can be calculated by integrating the solid angle formed with the two areas, respectively:

$$F_{ps} = \frac{1}{2} \cdot [1 + \cos\beta] \tag{2.28}$$

$$F_{pg} = \frac{1}{2} \cdot [1 - \cos\beta] \tag{2.29}$$

Therefore, the final formulation used to compute the radiation incident on a PV panel is:

$$G_T = (G - G_d) \cdot \frac{\cos\theta}{\cos\theta_z} + G_d \cdot F_{ps} + G \cdot \rho_g \cdot F_{pg} \tag{2.30}$$

As shown in equation (2.30), in an isotropic model the global irradiance on a tilted surface can be calculated based on global horizontal and diffuse horizontal irradiance.

2.8 PV Panel Power

Once computed the radiation incident H_T on the PV panel, using the set of formula available in section 2.6.1, it is necessary determine the power provided by the PV panel.

To determine the power output $[W]$ from the PV panel the formula (2.31) has been used. Instead of to obtain the electrical energy $[Wh]$ from the PV panel the formualtion (2.32) has been used.

$$P_{pv} = P_{pv,nom} \cdot d_{pv} \cdot \left(\frac{G_T}{G_{ref}}\right) \cdot [1 + \gamma \cdot (T_{cell} - T_{ref})] \tag{2.31}$$

$$E_{el,pv} = P_{pv,nom} \cdot d_{pv} \cdot \left(\frac{H_T}{G_{ref}}\right) \cdot [1 + \gamma \cdot (T_{cell} - T_{ref})] \tag{2.32}$$

In the Equations (2.31) and (2.32) the $P_{pv,nom}$ correspond to the nominal power of the PV array, $[kW]$, so its power output under standard test conditions, STC. The d_{pv}, [%], correspond to the photovoltaic derating factors. The derating factor is a scaling factor that takes into account the reduction of the PV panel output in real-world operating conditions instead of the rated conditions. In his value are factors as soiling of the panels, wiring losses, shading, snow cover, aging are taken into account.

The G_T $[kW/m^2]$ is the total solar radiation incident on the PV array in the current time step. The temperature coefficient of power $\gamma[\%/^\circ]$ takes into account the losses temperature-related effects.

The G_{STC}, $1[kW/m^2]$, and the T_{ref} or $T_{cell,STC}$, $25[^\circ C]$, are the incident radiation and the PV cell temperature, respectively, under standard test conditions.

The $T_{cell,NOCT}$ nominal operating cell temperature, NOCT, is the surface temperature that the PV array reaches if it is exposed to a solar radiation, G_{NOCT}, of $0.8[kW/m^2]$, an ambient temperature, $T_{amb,NOCT}$, of $20[^\circ C]$, and a wind speed of $1[m/s]$. This quantity provides a measure of how the surface temperature of the PV panel varies with the ambient temperature and solar radiation. Sometimes this quantities is reported in the product data but for the most of the cases her values varies between $45 - 48[^\circ C]$.

To resume these parameters, its values are reported in Table 2.2.

The $T_{cell}[^\circ]$ is the PV cell temperature in time step at which the solar radiation incident is computed. The cell Temperature, depends on the ambient temperature T_{amb} and on others parameters. In the Equation (2.33) the T_{cell} is computed and the others dependence are noted.

$$T_{cell} = \frac{T_{amb} + (T_{cell,NOCT} - T_{amb,NOCT}) \cdot \frac{G_T}{G_{T,NOCT}} \cdot \left[1 - \frac{\eta_{mp,STC} \cdot (1 - \gamma T_{cell,STC})}{\tau\alpha}\right]}{1 + (T_{cell,NOCT} - T_{amb,NOCT}) \cdot \frac{G_T}{G_{T,NOCT}} \cdot \frac{\gamma \eta_{mp,STC}}{\tau\alpha}} \tag{2.33}$$

Variable	**Value**	
G_{STC}	1000	$[W/m^2]$
$T_{cell,STC}$	25	$[°C]$
G_{NOCT}	800	$[W/m^2]$
$T_{amb,NOCT}$	20	$[°C]$
$T_{cell,NOCT}$	45-48	$[°C]$

Table 2.2: Values of STC and NOCT parameters.

In the Equation (2.33) the product $\tau\alpha$ represents the product of the solar transmittance of any cover over the PV array for the solar absorptance of the PV array and is approximated to a constant. Both values, solar transmittance and absorptance are measured in [%]. Assuming this product as equal to 0.9 doesn't introduce any significant errors, [49].

The efficiency $\eta_{mp,STC}$ represent the photovoltaic efficiency at standard test conditions at the maximum power point of the PV panel. This efficiency can be found in the characteristics of the products, alternatively it can been computed as follow, (6.1).

$$\eta_{mp,STC} = \frac{P_{pv,nom}}{A_{pv} \cdot G_{STC}} \tag{2.34}$$

Where A_{pv} is the surface area measured in $[m^2]$ of the PV module. In case of lack of other data to compute $\eta_{mp,STC}$, some general database are available. For instance, in [50] is provided a table with some average values of the efficiency for some type of PV modules.

To obtain the electrical energy output from the PV panel some further input are necessary, than the below mentioned.

- Reflectivity of the ground, ρ.
- Time step of the system.
- Tilt angle of the PV panel, β_{PV}.
- Azimuth angle of the PV panel, ψ_{PV}.
- Data design that depends on the manufactures of the PV panel
- Ambient temperature, T_{amb} with the resolution chosen for the time-step.

The data design of the PV panel includes the nominal power, $P_{pv,nom}[W]$, the maximum power point efficiency $\eta_{mp,STC}$, the coefficient of temperature-power, γ, the derating factor, d_{pv} and the balance of system efficiency, η_{BOS}.

Chapter 3

Proposed Methodology

As already introduced the objective of thesis work is to design a tool that, based on limited input data, can generate multiple years of high resolution solar irradiation data. This, in turn, allows extensive simulations of solar powered energy systems in areas where limited solar measurements are available.

The scarcity of high resolution data is particularly critical in rural and isolated areas, as discussed in Chapter 2. The tool should provide a direct and complete procedure for the Synthetic Generation, from the measured input data to the generated solar radiation profiles.

The most data-intensive steps of the methodology should be able to leverage data from other locations to compensate the lack of measurements in the location under study. This enables the application of the methodology in areas where only limited amounts of data are available.

The methodology is designed to be integrated in PoliNRG [28], a tool for the optimization of micro-grid systems.

The models and correlations used in the procedure have been chosen depending on their requirements in terms of input data and their ability to generate high-resolution data.

3.1 Methodology Overview

As discussed in the previous section 3, the goal is the generation of multiple years of data with an high-resolution time step.

The proposed model uses as stochastic models an Autoregressive Moving Average (ARMA) method and a Markov Transition Matrix (MTM) method to model the chosen climatic vari-

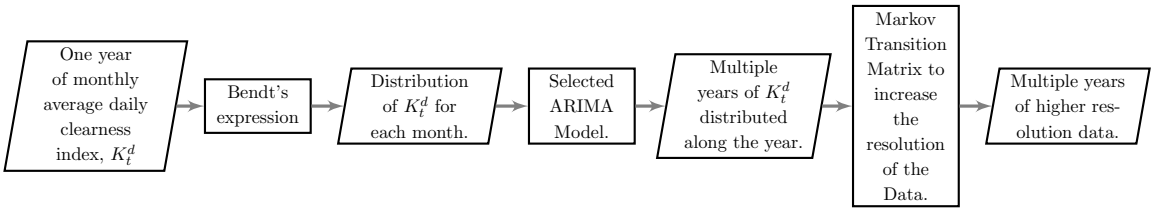


Figure 3.1: Flow chart of the overall procedure.

able. In the thesis work, the climatic variable selected for the generation of solar profiles is the clearness index K_t.

The procedure starts from monthly average clearness index data to obtain sub-hourly clearness index values. The number of generated years can be larger than the number of input years.

In addition to this the calibration of the various models used in the procedure can be based on data from other locations, as described in chapter 5, if not enough measurements are available for the location under study.

The steps of de procedure are described below.

1. The first step of the procedure is the determination of the cumulative distribution function of the daily clearness index inside each month through the Bendt's procedure. The input requested from this step is the monthly average clearness index for at least one year.

2. The second step allows to determine the sequence of the daily clearness index inside each month using an ARIMA model.

3. The last step is performed to increase the resolution of the daily clearness index until sub-hourly time step using a Markov model.

All the steps are presented in the sections below. The overall procedure starting from the input until the output is shown in the flow chart in Figure 3.1.

3.2 Input data

This methodology has been developed with the aim to require as little input data as possible. The input required are monthly average clearness index values, for one or multiple years. The required input can be easily obtained from multiple sources for most locations around the world. Moreover, missing data can be replaced with measurements obtained in neighboring

locations at the same latitude. For instance, [51] provides a data set for monthly clearness index values at different latitudes in each continent. The values are based on [49], or on TMY values for some regions.

Ideally, it is desirable to have access to several years of data for the location under study. Having several years of data instead of one could avoid the disproportionate influence of a particular meteorological year. A method to avoid the influence of a particular year is picking up a random monthly average daily clearness index among the years available. Using this procedure, the generated year will be less influenced by the specific condition of a single year and generating multiple years will capture a wider range of meteorological conditions for the location under study.

One might wonder why monthly average clearness index values have been used as input instead of daily values, that are often available. The developed procedure, that goes from monthly to sub-hourly values, is meant to be as general as possible, therefore the least demanding input is preferred. In addition to this, starting from monthly aggregated values it is possible to add at each step a stochastic component to the input data, generating a different output each time. For instance, when turning monthly data into daily values the sequence is different each time, depending on the random component of the ARIMA model. Then, passing from daily to sub-hourly data, the output series is different for each day depending on the path followed by the Markov chain.

On the other hand, if the procedure was not started from monthly values input, but instead from daily values, the sequence of the output series would be pre-determined. Nevertheless, the procedure also allows the use of daily input data, and an option is to start from daily values and "disturb" them.

The following sections discuss the mathematical models used to generate sub-hourly values based on monthly input data.

Monthly average daily clearness index values

If the clearness index data are available, they can be fed into the model directly. Otherwise, global horizontal irradiation measurements can be used to determine it.

Indeed, having the coordinates of the locations and the STZ (Standard Time Zone), it is possible to calculate the extraterrestrial solar irradiation, H_o, and also the extraterrestrial solar irradiance, G_o, as shown in equation (2.23) and equation (2.22). Both are computed in monthly average, daily, hourly and sub-hourly time step resolution. The global solar

irradiation is calculated for all the resolutions of H_o mentioned above.

An important step is the transformation of the global irradiation data from UTC (Coordinated Universal Time) to Solar Time. This procedure is done by shifting the data by the number of time steps, based on the Standard Time Zone of the location and its distance from the reference meridian. Subsequently, the global solar irradiation and the extraterrestrial irradiation are time-aligned.

The next step is the derivation of the clearness index, K_t for the entire procedure. Remembering that the clearness index is defined by equation (3.1).

$$K_t = \frac{H}{H_o} \tag{3.1}$$

3.3 From monthly values to daily distribution function

Input Multiple years of monthly average daily clearness index. Twelve values for each year.

Output One year of daily clearness index sorted from minimum to maximum value inside each month.

The first step of the procedure turns monthly clearness index data into the distribution of its daily values in each month. For this step it has been chosen to use Bendt's correlation, a deterministic correlation proposed in [40] and introduced in section 2.5.1. This choice respects the approach of the tool to achieve a good compromise between the reliability of the correlation and the amount of input data required.

Using this procedure, the probability distribution function of the daily clearness index K_t^d inside each month is computed based on three variables for each month:

- the monthly average daily clearness index K_t^d
- the minimum daily clearness index, $K_{t,min}$
- the maximum daily clearness index, $K_{t,max}$

The minimum clearness index and maximum clearness index for each month are determined as described in section 2.5.1. $K_{t,min}$ is imposed as a constant and equivalent to 0.05. As discussed in section 2.5.1 lower values are extremely rare in any location of the world. For what concerns the $K_{t,max}$, it is computed for each month with an equation dependent on K_t^m, the latitude, the solar declination and altitude of the location. In case of absence of some of these values, there is an alternative option that is only function of K_t^m, proposed in [40].

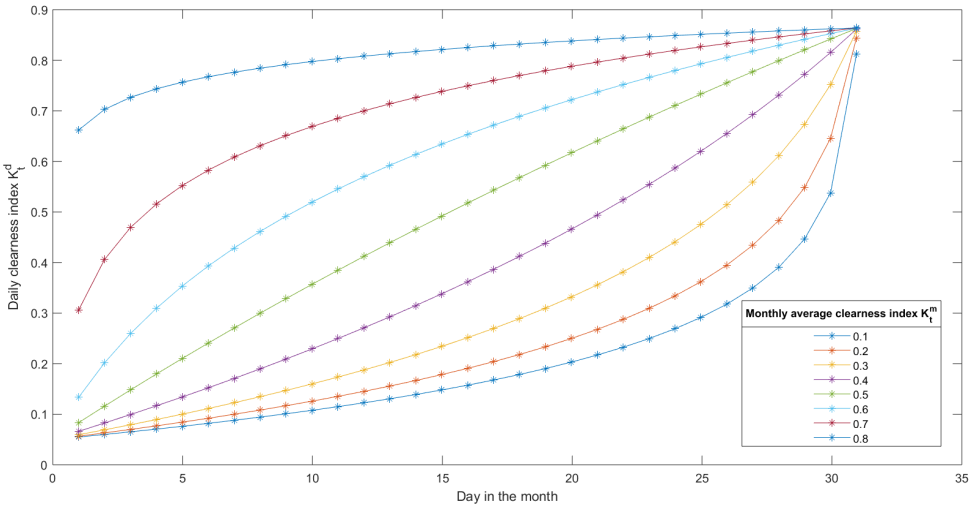


Figure 3.2: Cumulative distribution of the clearness index as a function of the K_t^m for the correlation of Bendt, et al.

The fundamental assumption behind Bendt's correlation is that the daily clearness index has an exponential distribution throughout the month, ranging between the minimum and maximum values recorded. The distribution of K_t^d along the month is obtained by modeling the cumulative frequency distribution F through the correlation proposed in section 2.5.1 by [5]. Each month differs from the others for the monthly average daily clearness index and for a coefficient, γ, that defines a particular exponential distribution. The γ coefficient is obtained through equation (2.2). This way, the cumulative distribution of K_t^d is obtained for each month.

Figure 3.2 represents the cumulative distribution of daily clearness index values for several values of K_t^m.

Only one year of monthly data is necessary as input, but multiple years can be used as input to increase the ability of the model to represent different meteorological conditions. If the user of the model would like to generate more than one year of synthetic data, but there are only twelve values at disposal, the K_t^m used for a specific month will be always the same. Nevertheless, the following steps of the procedure have stochastic components that make each generated year independent.

If the user has access to more than one year of input data, each time a year is synthetically generated a different value of K_t^m is randomly extracted from the available values for each month. Each monthly value is independently chosen from a random year. In Figure 3.3 it is possible to see this step of the procedure. This allows the creation of a new year at

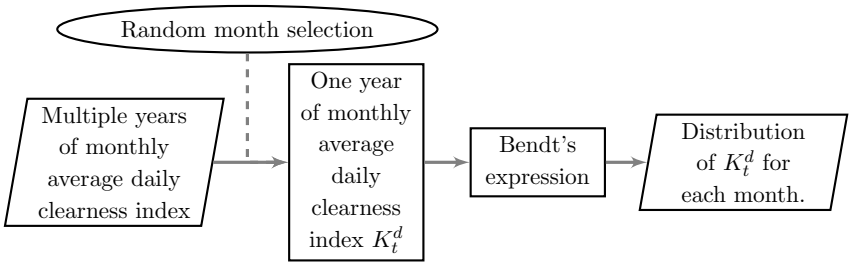


Figure 3.3: Flow chart of the process from the input of the model to the output of the Bendt's process.

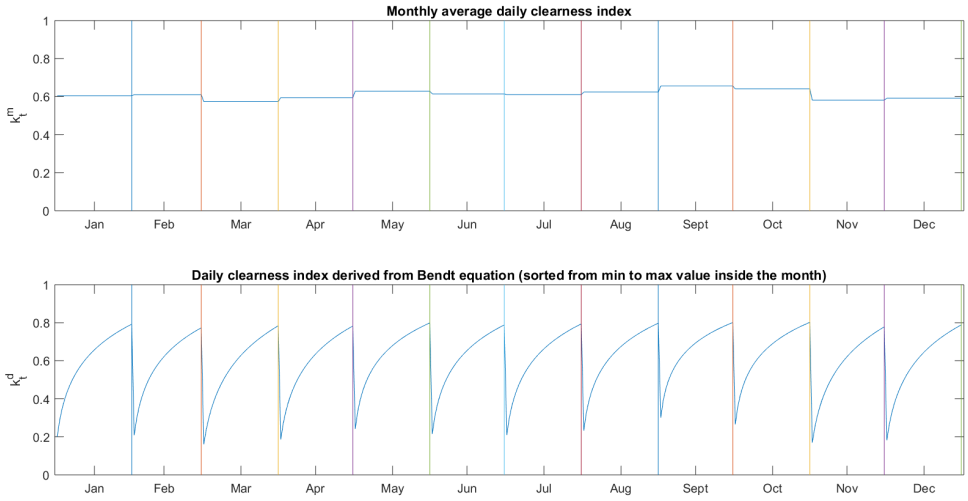


Figure 3.4: Plot of the process, from monthly average daily clearness index to daily clearness index for each month

each iteration of the procedure. The year created in this way is independent of the specific meteorological conditions verified in a specific year.

Figure 3.4 is a graphical and intuitive example of what happens in this step of the procedure passing from monthly to daily values.

3.4 From distribution to sequence of daily values

Input ARIMA orders, values of AR and MA coefficients and one year of daily clearness index sorted from minimum to maximum value inside each month.

Output N-years of synthetically generated daily clearness index values.

It has been chosen to use an ARMA/ARIMA model to pass from the distribution of daily values (obtained in the previous section) to their actual sequence. Indeed, among other applications, ARMA/ARIMA processes can be used to determine the sequence of a time series when the distribution of its values is known.

The choice to use ARIMA models to determine the sequence of the values depends on several factors. Multiple authors use ARIMA to model solar radiation data, specifically the author of [5] uses it to order daily clearness index values. In addition to this, ARIMA models are also used to forecast solar radiation profiles, as in [52, 53]. Having an autoregressive and random component, ARIMA models are particularly well suited to capture the variability of time series.

In this specific application, only the sequence of the values of the time series is extracted from the ARIMA model, reducing the sensitivity on the estimation errors. Furthermore, since the goal of this thesis is the generation of solar radiation profiles – rather than forecasting them – the main concern is to reproduce a realistic behavior. ARIMA models are extremely reliable for this purpose, whereas they have weaknesses in forecasting due to their random component.

This chapter is focused on how the ARIMA model is used to generate the sequence of daily values. For a discussion of how the ARIMA model characteristics are determined see section 4.2.

3.4.1 ARIMA Model Application

For this step of the procedure it is important to use an ARIMA model suitable to give the sequence to the daily clearness index values. From the literature it is possible to find some suggestions about the best model to use in case of clearness index time series [52, 53]. In section 4.2 is available a description of how this process has been performed in this thesis work and can be performed to do a location-specific calibration. The ARIMA model orders are chosen according to the ranking performed in 4.2.

Once the ARIMA model characteristics have been determined, the model chosen is used to give the sequence of daily clearness index values inside each month. The sequence of daily clearness index values is determined according to the ARIMA model as described in equation (5.1). As discussed in 4.2.1, the ARIMA model is defined by its orders (p, d and q), its parameters (ϕ and θ), the constant value and the standard deviation (σ). The model errors a_t are a normal random series with null mean and standard deviation that corresponds to σ, previously determined. All of these values have been determined in the previous step of the procedure.

Using equation (5.1), the model produces a sequence of daily values for one year. For each month, the order of these values is used to determine the sequence of the values obtained

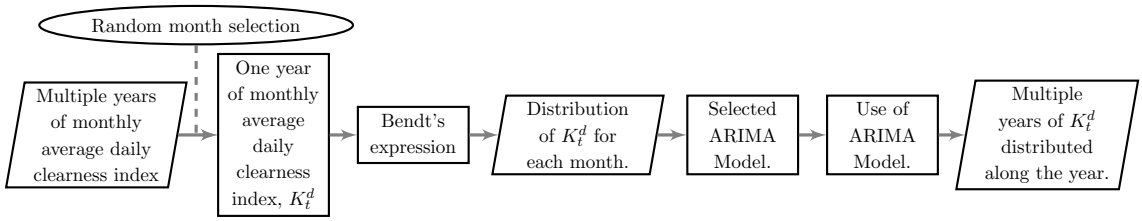


Figure 3.5: Flow chart from input data until the output obtained using ARIMA model.

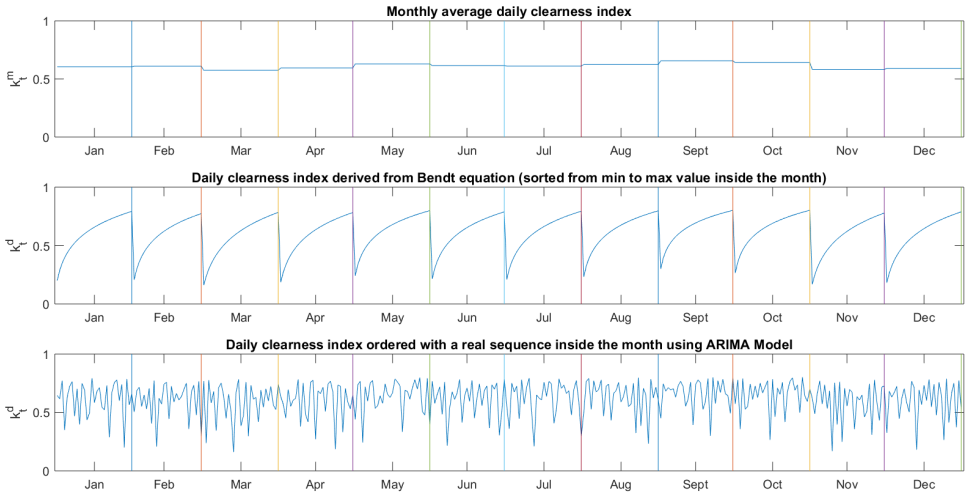


Figure 3.6: Plot of the process, from monthly average daily clearness index to daily clearness index in the year

through the Bendt's correlation.

Due to the random component of the ARIMA model, it is possible to simulate multiple years, with each year being unique.

It is important to underline that only the order is determined through the ARIMA model, whereas the values are generated from Bendt's correlation.

For each month, a series of daily values is generated through the ARIMA model. Each value in decreasing order is replaced with the corresponding value from the distribution obtained from Bendt's correlation. For instance, the largest value of a given month in the ARIMA-generated series is replaced by the largest value of the same month from the Bendt's values.

Figure 3.5 describes the process from the input to the output obtained using Bendt's correlation and ARIMA model, while Figure 3.6 represents the evolution of the intuitive example showed in Figure 3.5.

3.5 From daily to sub-hourly values

Input Multiple years of synthetically generate daily clearness index values and MTMs.

Output Multiple years of synthetically generate sub-hourly clearness index values.

For the stochastic generation of a climatic variable, such as the clearness index, it is possible to use the transition matrix approach of the Markov chain process. As explained below, it is possible to increase the resolution of the data by using the MTM correspondent to the daily clearness index.

In the Synthetic Generation of data, for a n^{th} order Markov model, the observation at time t is dependent on the n previous observations. This dependency is modelled through a Markov Transition Matrix. As discussed in section 2.5.3, each row of the matrix corresponds to a given combination of the last n states, in other words the starting state. Each column corresponds to the next state of the system, in other words the arrival state. The value of P_{ij} represents the probability of state j happening after state i. The way MTMs are created is discussed in section 2.5.3.

Markov models are a mathematical representation of the conditional dependence between two events. Indeed, they are a simple way to model the dependence of an observation, in a given time series, on the n previous and adjacent observations. They have been used in the modelling of various stochastic variables including the modelling of wind time series [54] and in the modelling of global solar radiation data [6].

The choice to use a Markov process to model the sub-hourly clearness index time series depends on several factors. First, Markov models are particularly well-suited for the synthetic generation of multiple scenarios due to their probabilistic component.

In addition to this, when modelling distributed energy systems, it is important to properly represent the fluctuations of power production between adjacent states. Markov chains have been developed to well represent the behaviour of time series, and in particular to model the dependency between consecutive states of the system.

Finally, a Markov process can be used to model time series without trends. The choice of using the clearness index as climatic variable allows to easily remove any trends from the data, as explained in subsection 3.5.1. On the other hand solar irradiation has a very strong trend across the day which would have made a Markov process unsuitable for his modelling.

The sub-hourly solar irradiation values can be obtained through a Markov process whose properties depend on the daily value K_t^d.

To model the clearness index K_t it has been chosen to use $m = 100$ to represent the possible states. Each value ok K_t can be approximated to one of the 100 allowed states. These models can then be used to synthetically generate sets of solar radiation data, H, using equation (3.2):

$$H = K_t \cdot H_o \tag{3.2}$$

3.5.1 Normalized Clearness Index

In modelling a time series, the first step is the removal of any trends in the data. In [55], the dependence of K_t on the zenith angle is outlined, and this dependence is removed by transforming K_t into a new variable, the normalized clearness index K'_t, through 3.3:

$$K'_t = \frac{K_t}{1.031 \cdot exp(\frac{-1.4}{0.9+9.4/AM}) + 0.1} \tag{3.3}$$

In the formulation 3.3, AM is the Air Mass. The Air Mass is the path length which sun rays take through the atmosphere, normalized to the shortest possible path length. The shortest path length is when the sun is directly overhead and the Air Mass is equal to 1. The Air Mass is a way to quantify the reduction of solar radiation as it passes through the atmosphere. The Air Mass can be modeled as:

$$AM = \frac{1}{\cos\theta_z} \tag{3.4}$$

The above calculation for Air Mass assumes that the atmosphere is a flat horizontal layer, but because of the curvature of the atmosphere, the air mass is not quite equal to the atmospheric path length when the sun is close to the horizon. At sunrise, the angle of the sun from the vertical position is 90° and the air mass is infinite, whereas the path length is not. An equation which take into account the curvature of the earth is:

$$AM = \frac{1}{\cos\theta + 0.50572 \cdot (96.07995 - \theta)^{-1.6364}} \tag{3.5}$$

Normalizing the clearness index through equation (3.3) allows to have a K'_t independent of the zenith angle, θ_z. Modelling in a stochastic way the normalized variable K'_t, it is possible to obtain K_t from equation (3.3), and subsequently the solar radiation.

3.5.2 Markov Classes

Due to the variability of weather conditions, each day can have really different statistical properties. In the creation of a stochastic model specific for global solar radiation it is of fundamental importance to take into account this weather variability. For instance, in a clear

day with a high daily clearness index, the probability of having a high clearness index along the day is higher than in a cloudy day. In addition to this, the volatility of the weather conditions might be strongly dependent on the average daily conditions. Indeed, sunny or overcast days typically have more stable conditions than partially cloudy days. This means that using a unique Markov Transition Matrix for all days in a year would not lead to accurate results.

To overcome this issue, several MTMs can be created to model different meteorological conditions. Each MTM corresponds to a *class* of meteorological conditions. Therefore, the first step for the construction of the model is to cluster the days into classes based on the daily clearness index values.

Although a higher number of classes would lead to a better model of the meteorological conditions, the number of classes is inherently limited by the size of the input dataset. Indeed, a higher number of classes requires a higher number of data in order to populate the corresponding MTMs. For a detailed discussion of this trade-off see section 4.3.2.

3.5.3 Markov Model Application

In the standard procedure the MTMs are already provided by the tool. These MTMs have been created as described in chapter 4, and are meant to be valid for any location, as discussed in chapter 5.

Once the MTMs are available, they can be used to generate the sub-hourly values based on the daily clearness index. The first step is to determine the Markov class of each day, based on K_t^d. Subsequently, the appropriate transition matrix is chosen in function of the class each day belongs to.

Then, using equations (2.8) and (2.9), sequences of normalized clearness index values K'_{ts} are synthetically generated for each day. The subscript s is used to indicate that this is a synthetically generated variable.

The next step is to derive the non-normalized values K_{ts} using equation (3.3). Using solar geometry equations, the extraterrestrial solar radiation H_o is calculated. H_o is greater than zero between sunrise and sunset and equal to zero otherwise. Finally, the sub-hourly synthetically generated solar radiation data H_s can be obtained by multiplying the extraterrestrial solar radiation H_o for the clearness index, as shown in equation (3.2).

Example of Markov model application

It has been chosen to describe an example of the application of a second order Markov model because the notation is relatively simple and it is easy to generalize this example to $n-$order models. The procedure is described step-by-steps to provide a better understanding of the synthetic generation algorithm. The algorithm is used to generate one day at a time, with its n_d normalized elements, where

$$n_d = \frac{24\,[hours/day] \cdot 60\,[minutes/hour]}{\Delta t\,[minutes]} \tag{3.6}$$

The steps of the procedure are the following:

1. The clearness index values K_t are normalized as described in 3.3. The normalized values are referred to as K'_t.

2. Each day is classified based on its daily clearness index K_t^d value.

3. The appropriate MTM is selected depending on the class determined at the previous step.

4. The algorithm is initialized by setting the first two values of the day, K'_{1s} and K'_{2s}, equal to K_t^d of the day. In the appropriate MTM, the row is selected based on the values of K'_{1s}, K'_{2s}.

5. The row is used as the probability distribution vector to generate K'_{3s}. K'_{1s}, K'_{2s} correspond to the values at midnight, where H_o is zero. Hence, a number of simulations steps will happen before sunrise. This ensures that the value of K'_{ts} at sunrise is independent at each repetition of the procedure, i.e. for each day.

6. Generating a random number from a uniform distribution between 0 and 1 K'_{3s} is generated as described in 2.9.

7. The row in the MTM corresponding to the values of K'_{2s} and K_{3s} is used as the probability distribution vector for generating K'_{4s}. Subsequent values $K'_{5s}; K'_{6s}; ...; K'_{144s}$ are generated similarly. The process goes on until the last time step of the day, at midnight.

8. The synthetically generated K'_{ts} sequence is converted to K_{ts} using equation (3.3).

9. To convert the clearness index into global solar radiation H data the equation (3.2) is used.

10. The synthetically generated value of the daily clearness index K_{ds} is calculated having H.

11. The obtained values are checked by comparing K_{ds} to K_t^d.

12. If the test is satisfactory, the values are kept. Otherwise, the procedure is repeated.

Control of the generation procedure

It is important find a way to control the procedure, to avoid obtaining data completely different from the starting data. To control the generation procedure, the distance between the daily clearness index generated and the starting clearness index is evaluated. If $|K_d - K_{ds}| > \delta$, the procedure is restarted from Step 1. δ is a selected tolerance that stops the procedure.

The procedure has to be repeated for each day until the end of the year. If, the years generated from the Markov procedure are more than one, it is possible to repeat the procedure for each year until the end using the MTMs built for the selected location.

The output for two example locations is reported in Figure 3.7.

The overall procedure from the input until the output of the Markov model is shown in the flow chart of Figure 3.1 at page 42.

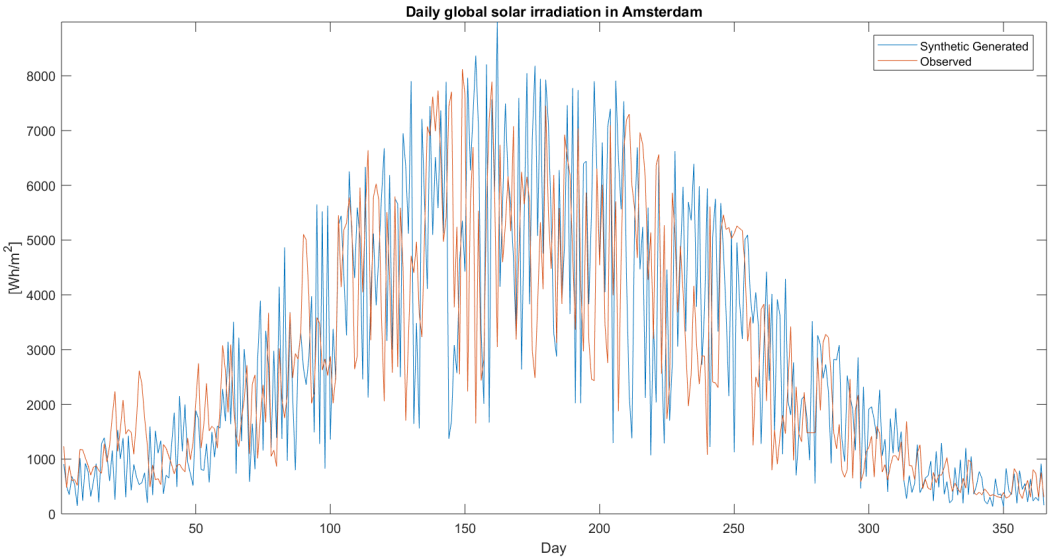
A graphical example of real output values is reported in Figure 3.8.

3.6 Validation

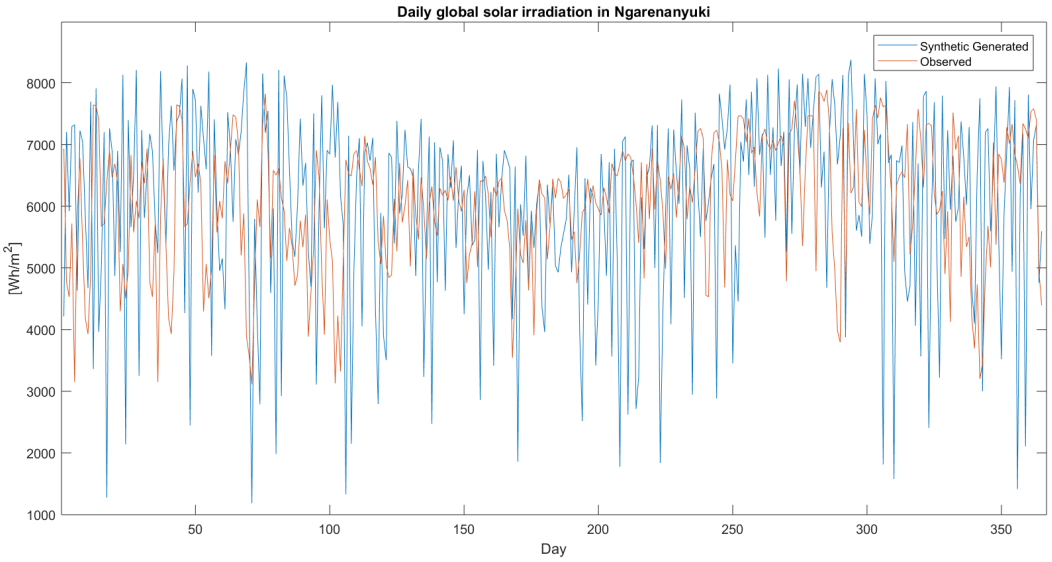
The ARIMA and Markov models presented above are stochastic methodologies used for the synthetic generation of data. They are not intended to forecast the modeled data exactly. The general idea of the procedure is to generate values statistically similar to the observed data. Therefore, the validation of the methodology consists in testing its ability to generate synthetic values with statistical properties similar to the observed data.

The detailed discussion of the validation procedure is presented in chapter 6. The validation is divided in two parts, internal and external. The internal validation is performed for each model used in the procedure (Bendt, ARIMA and Markov).

For Bendt's correlation, the MAPE and RMSE error on the monthly distributions have been calculated. For the ARIMA and Markov models, particular attention has been devoted to the analysis of ACF and PACF functions. Indeed, the main requirement for these models is that they reproduce accurately the behaviour of the fluctuations.



(a) Amsterdam.



(b) Ngarenanyuki.

Figure 3.7: One year of daily clearness index for observed and generated data.

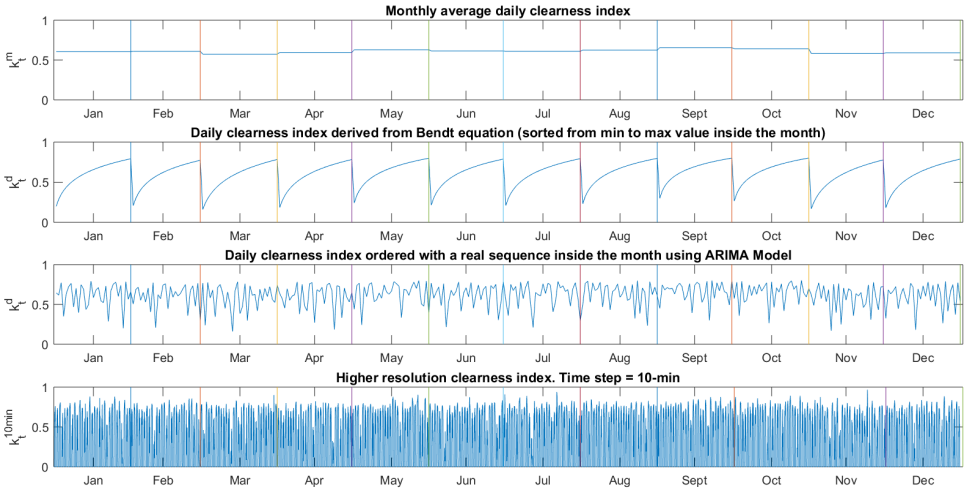


Figure 3.8: Plot of the process, from monthly average daily clearness index to higher resolution clearness index.

The external validation relies on comparisons of total values and distribution functions of sub-hourly, daily, monthly and yearly values of solar irradiation between generated and observed data. In addition to this, several statistical parameters have been analyzed. These include standard deviation, skewness and kurtosis.

Chapter 4

Models Calibration

The methodology presented in Chapter 3 has been entirely developed in MATLAB. This chapter deals with the ARIMA and Markov models calibration, analyzing limitations and the possible future developments.

4.1 Input data for the calibration procedure

The variables needed for the calibration of the models are:

- K_t^d: daily clearness index (4.2a), used to calibrate the ARIMA model
- $K_t^{\Delta t}$: clearness index with sub-hourly resolution (Δt time step) (4.2b), used to create the Markov Transition Matrices

Where:

$$n_h \coloneqq 60/\Delta t \qquad \text{number of time steps in one hour} \tag{4.1a}$$

$$n_d \coloneqq 60 \cdot 24/\Delta t \qquad \text{number of time steps in one day} \tag{4.1b}$$

and:

$$K_t^d = \frac{\sum_1^{n_d} H_{\Delta t}}{\sum_1^{n_d} H_{o,\Delta t}} = \frac{H_d}{H_{o,d}}; \tag{4.2a}$$

$$K_t^{\Delta t} = \frac{H_{\Delta t}}{H_{o,\Delta t}}; \tag{4.2b}$$

All these data are collected for several years.

4.1.1 Input of the ARIMA model

The input to built the ARIMA process is a vector that contains the daily clearness index for the entire year. To have a better understanding of the best ARIMA model that fit the daily clearness index time series it is possible to choose as input multiple years data and multiple locations.

4.1.2 Input of the Markov Process

For the implementation of the Markov process, a vast amount of observed data are needed. If the amount of data for the location under study is not sufficient to populate the MTMs, data from other "similar" locations can be used as described in section 5.

For each year of data a matrix with the day of the year as column index and the time step as row index must be provided. For instance, if a time step of 10 minutes is used, 144 steps per day 4.3 are considered.

$$\frac{24[\frac{hour}{day}] \cdot 60[\frac{min}{hour}]}{10[\frac{min}{step}]} = 144[\frac{step}{day}] \tag{4.3}$$

The resolution of the input data used for the MTMs determines the resolution of the output data of the entire procedure. Indeed, for the construction of Markov Transition Matrices the input data needed are the data with the highest resolution of the overall process. For instance if the data input of the Markov process have a sub-hourly resolution also the synthetically generated output of the model have the same sub-hourly resolution.

4.2 ARIMA model calibration

This section describes how the most suitable ARIMA model for this application has been determined. If sufficient input data are available, this procedure called "Model Building" can be followed to calibrate the ARIMA model for a specific location. In order to provide a self-sufficient tool, a generalized calibration of the ARIMA model that can suit any location has been performed, as described in Chapter 5. If not enough input data are available it is possible to use the ARIMA model provided by the tool in Chapter 5.

Input Multiple years of observed daily clearness index values for the location of interest (and possibly other relevant locations)

Output Ranking of the models and coefficients of the ARIMA/ARMA model

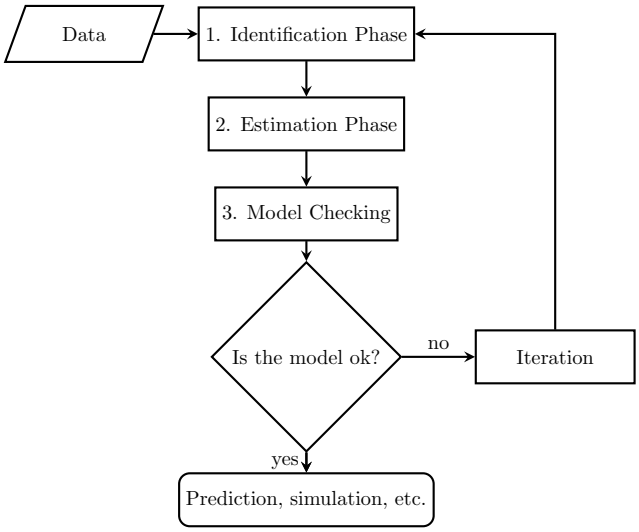


Figure 4.1: Flow chart of the Model Building.

4.2.1 ARIMA Model Building

Multiple approaches to build an ARIMA model have been proposed in scientific literature. In this thesis work the ARIMA model is built through a procedure called Model Building proposed in "Identification, estimation, and model checking" [56, Chapter 6].

The Model Building is a procedure designed to re-elaborate and process the time series. Furthermore this process is used to built the ARIMA model that suit the available data and can give the real sequence along the year to the data. In the thesis case the input data are one year of daily clearness index values. The Model Building is an iterative process, divided in three main stages and it is reported in Figure 4.1:

1. Identification Phase
2. Estimation Phase
3. Model Checking

In the subsections below, each phase is described in detail.

Identification Phase

The purpose of the identification phase is to identify a range of reasonable values for the ARIMA model orders p, d, q. This phase is based on observed data.

The steps of this phase are described below.

- Stationary or not stationary: the first step is the determination of the stationarity or not of a series. Stationarity ensures that inference of a stochastic process can be based on a single time series, for example one year of data is enough, which is the only realization available of the stochastic process.

- Degree of differencing d: if a time series is non-stationary or has very slow variations compared with sampling frequency, the estimated autocorrelation will decrease very slowly to zero. This non-stationarity should be removed using a suitable order of differentiation. It is necessary to choose d as the lowest order of differencing for which the estimated autocorrelation decreases sufficiently quickly to zero. Generally, d is equal to $0, 1, 2$.

- Identification of p and q: evaluate some hypothetical combinations of p and q that can yield theoretical autocorrelations and could approximate the estimated autocorrelation function. The combinations can be more that two. The more they are the better it is to determine the model that fits best the data. There are several characteristics of ARIMA process that can be verified in the ACF and PACF plots, such as damped exponential and/or sine functions after lag $q - p$ and damped exponential and/or sine functions after lag $p - q$, respectively.

- Seasonality: once the differencing process has been done, if lags different from $(p - q)$ are still present, the seasonal differencing is needed. This is used to remove the slowly decrease from lag different from value $(p-q)$. In this case the ARIMA is in the seasonal case [57], a multiplicative seasonal model. Indeed, in this model it is possible to add the a seasonal periodicity to it. The seasonality is the first lag, different than $(p - q)$, whose multiples correspond to significant spikes in the autocorrelation function. If, for example, the significant autocorrelation will be at lag 12 and in all the lags multiple of 12, the seasonality is 12. Furthermore, the order of p is equal to 13 (corresponding to the sum of the non-seasonal and seasonal differencing degrees, $(1 + 12)$) and the order of q is also equal to 13 (corresponding to the sum of the degrees of the non-seasonal and seasonal MA polynomials, $(1 + 12)$).

In Figure 4.2 some examples of ACF and PACF plot are reported.

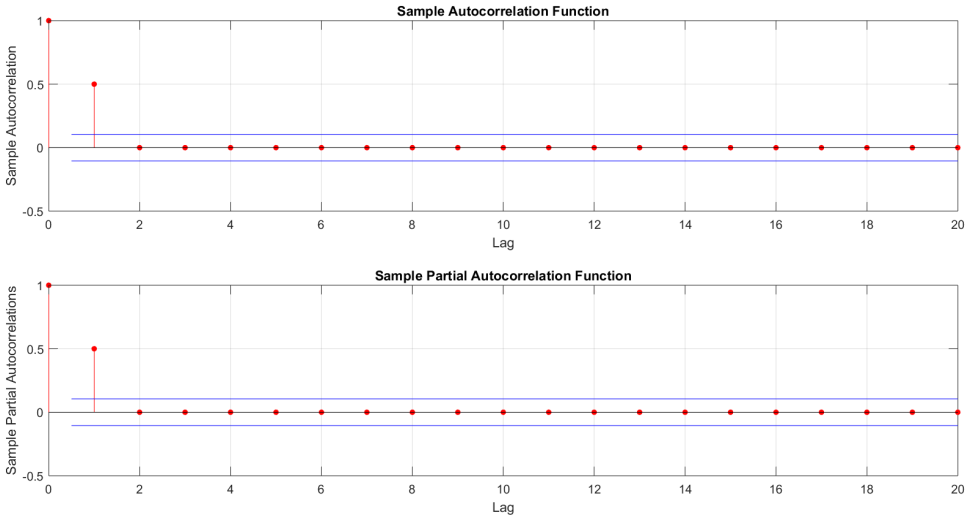


Figure 4.2: Example of ACF and PACF for an ARIMA Series.

Estimation Phase

In this phase the selection of the most suitable model orders is performed. The idea is to test in parallel different models with different combination of values of p and q. For each model, or better for each combination of the AR and MA terms, the parameters and coefficients needed for the simulation are determined and the subsequent model is generated.

If daily clearness index values along the year are being processed, the possible models that can be taken into consideration are reported below:

- $ARIMA(0, d, 1)$

$$y_t = \mu - \theta_1 * a_{t-1} \tag{4.4}$$

- $ARIMA(0, d, 2)$

$$y_t = \mu - \theta_1 \cdot a_{t-1} - \theta_2 \cdot a_{t-2} \tag{4.5}$$

- $ARIMA(1, d, 0)$

$$y_t = \mu + \phi_1 \cdot y_{t-1} \tag{4.6}$$

- $ARIMA(2, d, 0)$

$$y_t = \mu + \phi_1 \cdot y_{t-1} + \phi_2 \cdot y_{t-2} \tag{4.7}$$

- $ARIMA(1, d, 1)$.

$$y_t = \mu + \phi_1 \cdot y_{t-1} - \theta_1 \cdot a_{t-1} \tag{4.8}$$

The choice of the order of the AR and MA terms reported above depends on two factors. The first one is that looking at the ACF plot and PACF plot, these orders look like the most suitable for the given time series. The second reason is that several papers, [5, 33, 38], that deal with this procedure in relation with climatic variable, such as the clearness index, use the parameters reported above.

The Estimation Phase can be performed using input data from multiple years and multiple locations around the location under study, to improve the robustness of the results. The output of this phase is a ranking of the most adequate models for the chosen location. Below the description of the procedure is reported.

- Determine the number of parameters which best describe the observed variation of the series (i.e. the model order).
- Search the most suitable model among the chosen classes of ARMA processes.
- Determine p, q.

Among the several techniques to determine the most adequate model orders discussed in 2.5.2 the following have been chosen for the proposed methodology.

- Zero Criterion: Computing the likelihood based inference.
- First Criterion: Looking for P-Values higher than 5% or looking for the maximum T-Student level among the outputs.
- Second Criterion: AIC. Choosing (p, q) such that $AIC(p, q)$ is minimized.

None of these three criterions is better than the others. Indeed, each of them takes into account different parameters and different boundary conditions. Thus, the idea is to generate a ranking among the models based on the number of criterions they satisfy. Each model will be classified in the ranking from the position one to five. The first model is the one that can better represent the time series. From this point onwards, only the first two models will be used. When dealing with clearness index time series as the main climatic variable, generally the more adequate models are $ARIMA(1, 1, 1)$ and $ARIMA(0, 1, 2)$. These two models have been chosen because they seem to be coherent with the data, are the first two models of the ranking, and with the collected articles related to this subject [5, 37].

Model Checking

After this complete analysis of the model order it is also important to observe the values of the Residuals and the Residuals Autocorrelation function.

This *Model Checking* analysis provides a good guess to definitively determine if the estimated models are adequate or not.

In the list below are reported some steps to describe how the residuals properties are evaluated:

- Plot a_t;
- Test for changes in sign;
- Test in autocorrelation function, analyzing lags. If the value at several lags is larger that two standard deviations the model is not adequate.

Finally, adopting the remained models or model, depending on the residuals analysis, it is possible to compute the coefficients for the AR part, ϕ, and for MA part, θ.

If multiple years of input data are available, the selected ARIMA model is calibrated on each year and the coefficients are averaged across the years to obtain a single model [5]. The next step is the simulation and generation part of the time series.

Implementation of the Model Building

After guessing some ARIMA/ARMA models that could be ideal to describe the daily clearness index time series, these will be tested. This is the first part of the Model Building procedure and is called "Identification Phase". To determine the models that will be tried two directions are followed. The first direction is experimental and relies on the analysis of the ACF and PACF functions. The second direction considers what has been done before in other works, in particular [33, 5] and [37], to derive some information and examples.

Before determining the final model, it is also important to look at the stationarity of the time series. Studying daily clearness index time series for different locations it is possible to reach the conclusion that generally one degree of differencing, $d = 1$, is enough to achieve the stationarity of the time series in most locations. In any case, the `diff_degree` can be tailored to specific cases, varying between 0, 1 and 2.

The models selected are reported in the Table 4.1, with the order of Autoregressive term, p, and the Moving average term, q.

Model	p	d	q
ARIMA	0	0,1,2	1
ARIMA	0	0,1,2	2
ARIMA	1	0,1,2	0
ARIMA	2	0,1,2	0
ARIMA	1	0,1,2	1

Table 4.1: Selected ARIMA Models.

After a first analysis on different locations around the world to have a general idea about the order of the terms, the implementation part of the ARIMA model is carried out only for the location of interest.

The second part of the Model Building procedure is the "Estimation Phase" and is used to determine the best model to represent the actual time series. To determine the best model the analysis has carried forward for the location under study. The dataset, in which to implement this phase, is for three different years of the location.

The objective of this phase is the estimation of the model orders and the determination of the best ARIMA, among different models, that can represents the time series. The determination is done in function of "Zero Criterion", "First Criterion" and "Second Criterion", as discussed in chapter 2.5.

In the following section, some relevant functions used in MATLAB are described, in order to illustrate some important steps of the implementation.

4.2.2 Finding the differencing degree

This subsection describes how `Stat_NonStat_Function` has been implemented. This function gets as input one year of daily clearness index values K_t^d. It returns as output d, the degree of differencing needed to make the time series stationary. Inside the function, a `Test_Correlograms` sub-function is used to graphically display the autocorrelation and partial autocorrelation of the time series. The degree of differencing is determined through a `while` cycle, which goes on until the stationarity of the time series is confirmed by two tests. At each iteration of the loop, the degree of differencing is increased, i.e. the time series is differentiated.

The first test used to determine the stationarity is the Unit Root Test (`Test_UnitRoot`). Depending on the output of the Unit Root Test, three cases can happen:

1. the time series is stationary and the `while` loop is interrupted;

2. the time series is not stationary and the `while` loop is repeated;

3. the stationarity of the time series is not determined and the Variance Ratio test (`Test_Vratio(Y)`) is performed.

The Unit Root Test (`Test_UnitRoot`) gets as input the time series and returns two Boolean variables, `h1` and `h2`. A review of implementations of the Unit Root Test is provided in subsection 4.2.1. It has been chosen to perform the test using both a graphical and statistical approach:

- PACF and ACF are used to graphically assess stationarity. After inspecting a plot of the time series, the plots of the sample autocorrelation function (ACF) and sample partial autocorrelation function (PACF) are evaluated. For what concerns the PACF plot, it is necessary to check if the slope of the plot indicate a unit root process as described in section 2.5.2. For what concerns the ACF plot, it is necessary look at if the length of the line segments on the ACF plot gradually decay and if this pattern continue for increasing lags. These behaviors indicates a non-stationary series.

- The MATLAB Econometrics Toolbox is used to statistically assess stationarity. This is A MATLAB tool with four formal tests to choose from to check if a time series is non-stationary: `adftest, kpsstest, pptest, and vratiotest`. For an overview of these tests, see subsection 4.2.1.

Inside `Test_UnitRoot`, the two following tests are carried on:

- Dickey-Fuller test: `adftest` is used to perform the Dickey-Fuller test on the data. The Dickey-Fuller test assesses the null hypothesis of a unit root for a univariate time series. The output `h1` of the model has the following meaning: `h1 = 1` indicates rejection of the unit-root null in favor of the alternative model. `h1 = 0` indicates a failure to reject the unit-root null.

- Kwiatkowski, Phillips, Schmidt and Shin (KPSS) Test: `kpsstest` is used to perform the test of Kwiatkowski, Phillips, Schmidt and Shin (KPSS). This test assesses the null hypothesis that a univariate time series is trend stationary against the alternative hypothesis that it is a non-stationary unit-root process. The output `h2` of the model has the following meaning: `h2 = 1` indicates rejection of the trend-stationary null hypothesis in favor of the unit-root alternative. `h2 = 0` indicates a failure to reject the null hypothesis.

The outputs of the Unit Root Test are evaluated through `Test_Stat_or_NonStat`. This function gets as input `h1` and `h2` and determines the next step of the procedure. Four combinations of `h1` and `h2` can arise.

Case 1 Dickey-Fuller test: it is not possible to reject the null hypothesis (`h1=0`). KPSS test: the null hypothesis is rejected (`h2=1`). Both imply that the time series has a unit root. The time series is not stationary and the `while` loop is repeated increasing the degree of differencing.

Case 2 Dickey-Fuller test: rejection of the null-hypothesis (`h1=1`). KPSS test: the null-hypothesis cannot be rejected (`h2=0`). Both tests indicate that the time series is stationary. The degree of differencing has been determined and the `while` cycles is interrupted.

Case 3 it not possible to reject any of the tests (`h1=0, h2=0`). This means there are not enough observations in the input data. Increasing the degree of differencing would not bring any measurable benefits. Therefore the `while` cycle is interrupted and a warning message is displayed.

Case 4 In both tests the null hypothesis is rejected (`h1= 1, h2=1`). This indicates heteroskedasticity in the time series. In this case both are component hypotheses. Therefore, it is necessary to perform a further test. In this case, the Variance Ratio test (`Test_Vratio`) is used to determine the stationarity of the time series.

The variance Ratio Test (`vratiotest`) gets as input the time series and returns a of Boolean decision variable `h`. `h = 1` indicate rejection of the random-walk null hypothesis in favor of the alternative. `h = 0` indicate a failure to reject the random-walk null hypothesis. Based on the value of `h`, the `while` cycle is repeated or interrupted.

If `diff_degree` is higher than two, the model drops into the "Seasonality Case" and a seasonal ARIMA model is used. The order of seasonality is determined based on the ACF plot, as described in subsection 4.2.1.

4.2.3 Finding the ARIMA model coefficients

This subsection describes how `TestModels_Function` has been implemented. This function gets as input the differencing degree d and the stationary time series, and returns the coefficients for each of the ARIMA/ARMA models selected (Table 4.1). Furthermore, it return

the model that best fits each year of data, based on the three criteria described in 4.2.1. The two main functions used in this stage are available in MATLAB.

- `Arima`: this function creates an $ARIMA(p, d, q)$ model by specifying the degrees p, d, and q. If the model has no integration and no seasonal components, the only properties included in the standard Box and Jenkins notation for a $ARIMA(p, 0, q) = ARMA(p, q)$ model are p and q.
- `Estimate`: this function is used to determine the coefficients of the AR term and MA terms. In addition to this, it returns the variance, the model constant, the p-value, the residuals values and their covariance.

Based on the output values of the `estimate` function, it is possible to perform the test to determine the best model for each time series, using the three criteria described in 4.2.1: p-value, maximum likelihood and Akaike-Bayesian.

For practical reasons, in the thesis implementation only the best two models are selected and used in the next steps of the procedure.

The models are ranked based on each criteria, and subsequently ranked based on how many times each model is the best according to each criterion.

4.2.4 Testing the selected models

This subsection describes how `Check_Model` has been implemented. This function is used to perform the third phase of the model building, "Model Checking". In this phase an analysis is performed on the residuals of each model determined in the previous step.

Firstly, it is important to visualize the autocorrelation plots to determine if the model is adequate or not. If the ACF plot presents values larger than two standard deviations, the model is not adequate.

The second factor to verify is assessing whether the residuals are autocorrelated by conducting a Ljung-Box Q-test. The `lbqtest` test confirms if the residuals are uncorrelated using a Boolean indicator.

Based on these two tests, the models ranked in 4.2.3 are selected or discarded.

4.3 Markov Model Calibration

Input Multiple years of sub-hourly clearness index values from source dataset.

Output Order of the Markov model, number of classes and corresponding range.

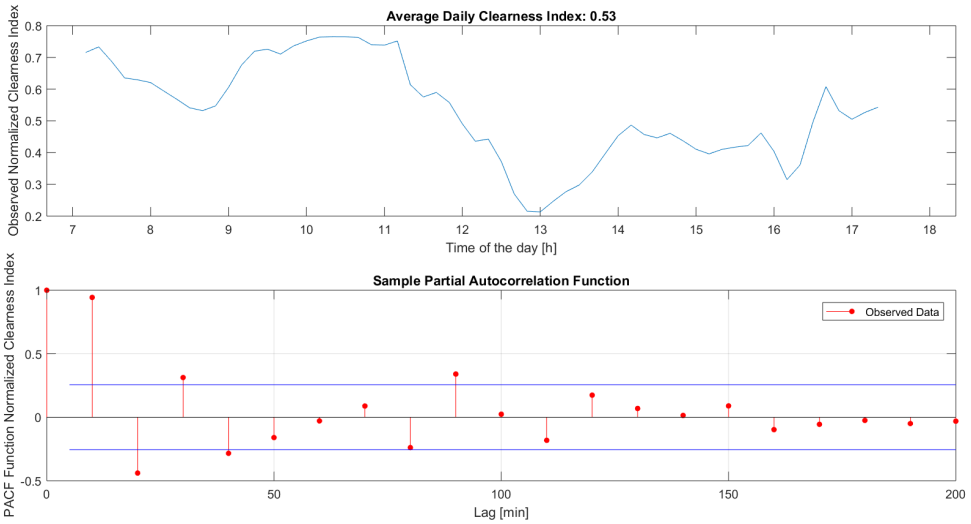


Figure 4.3: Example of PACF for a day of Normalized Clearness Index with time step between data of 10-min

The Markov model is used to achieve higher resolution in the data, passing from a daily resolution to a sub-hourly resolution. To create the MTMs, a really large dataset is needed. A precise number of the amount of data required to fill the matrices have not been determined. To satisfy this requirement, it is possible to use data from nearby locations and from multiple years [6] as discussed in Chapter 5.

In this section are presented some suggestions that can be used to build a Markov model specific for the location to study. In case of lack of data in 5 it is possible to find MTMs that can be directly used for each locations.

4.3.1 Markov Model Order Selection

The first step is choosing the model order for the construction of the MTMs. The Markov model order is defined in section 3.5.

This step is partially based on a qualitative analysis and is left to the user. The order of the Markov model can be selected checking the conditional dependency in the modeled dataset between observations at various lags using partial autocorrelation function (PACF). A significant spike in the partial correlation between k_t and k_{t-h} suggests that the observation at time t is significantly dependent on the observation at time $t-h$. PACF is also used to select the order of an Autoregressive (AR) model, as discussed previously in section 2.5.

The blue lines in Figure 4.3 identify the 5% significance level, and any values of PACF falling within these bounds can be determined not to be statistically significant at the 5%

level. Identifying the statistically significant lags of the PACF allows to determine the order of the model. Generally, the partial autocorrelation values for lags higher than two are very close to zero hence not statistically significant.

4.3.2 Markov Classes Formation

As discussed in 3.5.2, using a unique MTM for all days in a year would not lead to satisfactory results. Instead, several MTMs are used to model different types of meteorological conditions, with each MTM representing a *class* of meteorological conditions.

The first step to build the Markov model is the identification of the classes in which the daily clearness index values have to be clustered. This procedure is highly dependent on the characteristics of the data under study. This classification is fundamental for the next steps.

The limitations in the choice of the number of classes are described in 3.5.2. The choice of the number of classes and their range is extremely difficult to automate, and is left to the modeler. The main criteria are:

- Markov Transition Matrices should be sparse but should have enough values around the main diagonal. An excessive number of classes will lead to some MTMs being almost empty.
- Each single Markov Transition Matrix should not represent extremely diverse conditions. A very low number of classes forces very different meteorological conditions to be clustered in the same class.
- Although it is normal that some MTMs are less sparse than others, an excessive difference means the range of each class should be adjusted.

A more practical description of the procedure is presented in section 5.2.3.

In order to have a larger data set for the construction of the matrices, it is possible to take the data either from close-by locations or locations with similar weather characteristics, further details are discussed in Chapter 5. If more than one locations are used, the final model will combine the statistical characteristics of all the locations.

The choice of the Markov classes can be tailored to the distribution of meteorological conditions in the location under study. For instance, if the area under study tends to be particularly cloudy and the sunny days are rare, the lower K_t^d will be more relevant than the higher. In this case there will be more groups representative of low K_t^d values than of higher values.

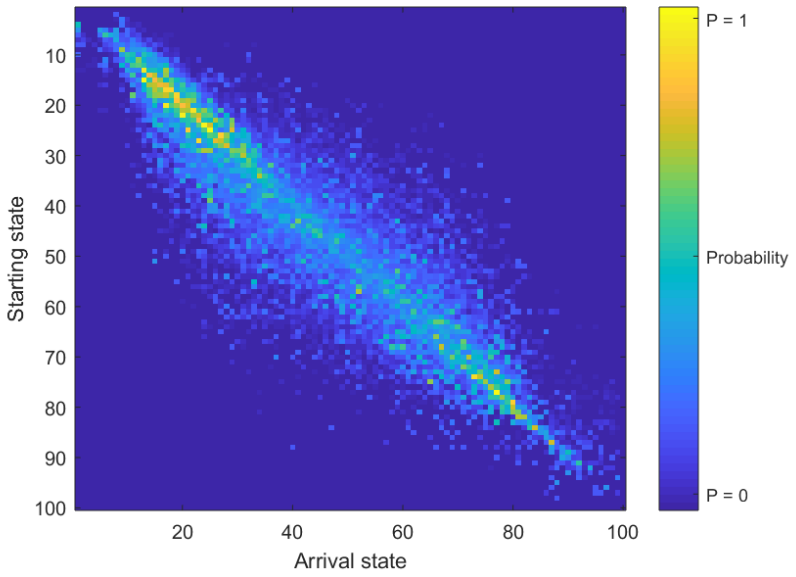


Figure 4.4: Example of Markov Transition Matrix.

After the creation of the classes, the Markov Transition Matrices can be constructed for each class. To construct the MTMs all the normalized clearness index values are transformed into the m determined states, approximating each K_t' to the closest state. In the thesis case the number states m used is 100, this is a reasonable approximation for the clearness index. From here onwards the sequence of K_t' is transformed into a sequence of states.

The MTMs are built in function of the order selected for the Markov models. Furthermore, the MTMs are empirically generated from the solar radiation and are used in the synthetic generation of solar radiation data. Constructing a first order matrix is rather simple. The procedure to build the MTMs is described in section 2.5.3.

Higher order MTMs are constructed in a similar way to the first order MTMs. The difference is that each row of the $n-th$ order MTMs describes the probability of observations depending on the n previous observations. The MTMs are generally very sparse matrices with most non-zero values lying around the diagonal. This shows a strong correlation between consecutive observations. Indeed, clearness index values are strongly autoregressive. In Figure 4.4, an example MTM is reported.

4.3.3 Building Markov Transition Matrices

The MTMs are built in MATLAB using the function `mtm_Break` that takes as input all the normalized sub-hourly time step clearness index values and the classes of daily clearness index. This function return as output the Markov Transition Matrices and the cumulative

MTMs.

The sub-hourly time step clearness index are initialized as a matrix. The indexes of the row represent the day of the year, the indexes of the columns represent the moment of the day. If the input come from multiple years, the days are ordered in sequence. The second input is a vector that contain the numbers of class at which each day belongs to.

For each class, chosen at the step before a MTM is created. All the MTMs have the same dimension, that depends on the number of possible states, and on the order chosen for Markov. The example is done for m equal to 100 and in case of second order Markov model. In this case, all the MTMs are a tridimensional matrix of dimension $[100 \times 100 \times 100]$.

The normalized time-step clearness index have to be transformed into states. Each values have to corresponds to one of the 100 allowed states and this can be done rounding the clearness index values and multiplying them for 100.

The first thing to do is the identification of the days that belongs to each class. In function of the vector of the daily class is possible to cluster the days in different set. Once the days that belongs to a determined class are identified, the normalized time-step clearness index transformed into the allowed states are taken as input of the function `Markov_Matrix_Class`. The function `Markov_Matrix_Class` is devoted to create the MTMS. This function creates the vectors of the probability transition $Pijk$ and in case of second order Markov model the function is composed by two `for` cycles, one inside the other. These two cycles go through the values of the time-step clearness index to find the sequence ijk and count how many times this sequence is verified. The probability transition is the ratio between the number of time the sequence ijk is verified on the number of time the sequence ij is verified.

The outside `for` cycle goes through the index i from state 1 to 100 and for each state, the intern `for` cycle goes through the index j from state 1 to state 100. Entering in the intern cycle, after the i value is determined, all the j that follow i are counted. If there aren't j values that follow the determined i value, the `for` cycle goes through the following i value. Alternatively, if there are a number of j values that follow the determined i all the k values are counted. Using the MATLAB function `tabulate` is possible to have as output a table that returns all the values that follow the sequence ij with their frequencies to occur.

Once all the frequency of each sequence ijk has been computed, it is possible to fill in the vectors of the MTMs. Each of the probability transition vectors correspond to the ij-row of the matrix.

The final step is the creation of the cumulative MTMs. These matrices are the cumulative

sum of the Markov Transition Matrices and are creating using the function `cumsum` in the third dimension of the matrix. These matrices are used to compute the cumulative distribution function such as in the equation (2.8).

4.4 Configuration

The methodology configuration that includes the steps discussed in chapter 3 and the steps discussed in this chapter, is reported in the flow chart in Figure 4.5. The central flow chart represents exactly the step described in chapter 3 from the monthly average daily input clearness index until the high-resolution clearness index output.

The lateral blocks represent the calibration steps described in this chapter that can be further generalized in chapter 5.

The dashed line represents the possibility to synthetically generate multiple years from the same inputs data.

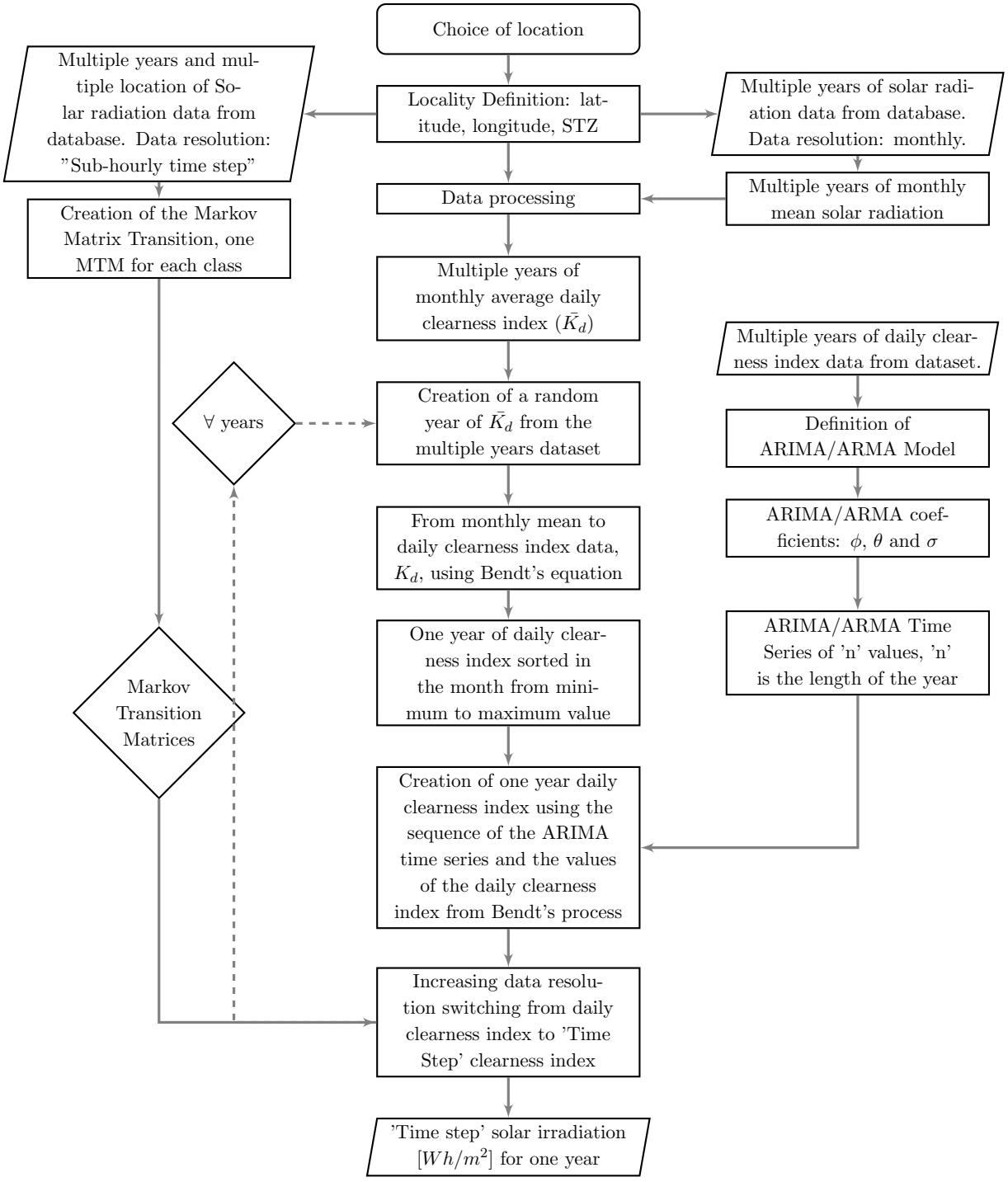


Figure 4.5: Flow chart that describe the entire procedure.

Chapter 5

Generalizing the Model Calibration

The objective of the thesis is to create a tool that can be used by everyone and in every location of the world. To achieve this goal it is necessary that all the steps of the process are made as general as possible and it is also important to make all the coefficients and decision applied in the model less site-dependent as possible. Both stochastic models used, ARIMA model and Markov model, can be generalized to have a more general configuration of the procedure.

5.1 ARIMA Generalization

Generalizing means to define an ARIMA model that could be a good compromise for every considered location. A general ARIMA model has specific values for its orders p and q, and pre-defined values for all the coefficients ϕ, θ and σ. This generalization is done to avoid to the user to repeat every time the model building procedure detailed in section 4.2. Furthermore, applying a generalization to the model allow to the user to use the tool also in case of absence of a large amount of data.

5.1.1 Data used for the generalization

Since the goal of the generalization of the ARIMA model is to find a unique model that can be applied in every place, it is important to look at the optimal ARIMA model for daily clearness index time series for different locations.

Amsterdam, Ngarenanyuki (Tanzania), Valencia and Madrid have been chosen for this purpose. The first two locations represent very different meteorological conditions and have been used for the validation of the general methodology, as described in chapter 6. The other

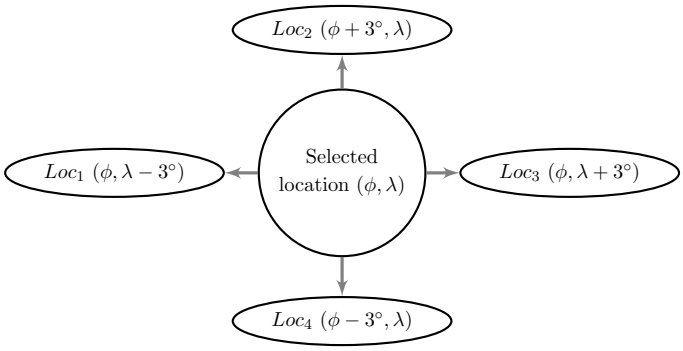


Figure 5.1: Location of interest and the other four locations considered for the input data.

two locations, Valencia and Madrid, have been chosen because they are used in [5] to apply the ARIMA model, and it is useful to make a comparison between the results. These two locations have been used in the identification phase to guess reasonable values of the ARIMA model orders. In addition to this, they have been used to generalize the model coefficients as described in section 5.1.3.

For each location it is better to use more than one year of input data to perform the analysis and to obtain significant results.

For the validation of the generalization, it has been chosen to conduct the analysis using three years of daily clearness index data, from 2004 to 2006, for the locations of interest. In addition to this, data from other four locations close to each of the two main locations have been used. The choice of four locations around each location of interest is aimed at increasing the number of data and the reliability of the results. Figure 5.1 shows an example of how the locations are determined.

Increasing or decreasing $\pm 3°$ of latitude and longitude with respect to the coordinates of the locality of interest, it is possible to get a circular area around the place under study. Each section is then composed by five different locations, with the similar deterministic climatic characteristics along the year to the location of interest. For each of these locations, three years of data are used. In total, thirty yearly times series are analyzed.

5.1.2 Model Order Generalization

To select the model order the estimation phase (already detailed in section 4.2) is performed on five locations close to the target one. The two locations studied and on the four auxiliaries locations for each main location. In Table 5.1 the results obtained for the location of Ngarenanyuki are reported, while in Table 5.2 the reported results are related to Amsterdam.

With the results obtained, through procedure described in section 4.2, it is possible to

Year	Stochastic model	Mean	St. Deviation	P-Value	Likelihood
2004	ARIMA(0,1,1)	3.0912e-04	0.10884	0.54897	290.81
	ARIMA(0,1,2)	3.1007e-04	0.10483	0.56608	304.49
	ARIMA(1,1,0)	-3.853e-05	0.12841	0.00922	230.63
	ARIMA(2,1,0)	-2.067e-04	0.11991	0.07136	255.57
	ARIMA(1,1,1)*	2.2839e-04	0.10289	0.62782	311.27
2005	ARIMA(0,1,1)	-7.559e-05	0.09993	0.7775	321.91
	ARIMA(0,1,2)	-5.375e-05	0.09054	0.5444	357.83
	ARIMA(1,1,0)	-8.245e-05	0.12778	0.0188	232.41
	ARIMA(2,1,0)	5.851e-05	0.11448	0.0218	272.41
	ARIMA(1,1,1)*	-4.877e-05	0.09255	0.7854	349.82
2006	ARIMA(0,1,1)	6.5812e-05	0.10392	0.7667	307.64
	ARIMA(0,1,2)*	2.0015e-05	0.09686	0.4539	333.23
	ARIMA(1,1,0)	4.8667e-06	0.14011	0.0009	198.88
	ARIMA(2,1,0)	-4.049e-05	0.12654	0.0118	235.97
	ARIMA(1,1,1)	5.2771e-05	0.10132	0.9262	316.87

Table 5.1: Statistical parameters of the residuals of the different studied stochastic models for Ngarenanyuki. (*) indicates the best model.

Year	Stochastic model	Mean	St. Deviation	P-Value	Likelihood
2004	ARIMA(0,1,1)	7.0713e-04	0.2093	2.38e-08	52.875
	ARIMA(0,1,2)	8.3752e-04	0.1943	0.2676	20.53
	ARIMA(1,1,0)	1.289e-05	0.2560	0.00	20.53
	ARIMA(2,1,0)	-1.416e-05	0.2183	1.41e-9	37.44
	ARIMA(1,1,1)*	0.0003037	0.1873	0.00023	93.09
2005	ARIMA(0,1,1)	1.4648e-04	0.2159	0.00	41.51
	ARIMA(0,1,2)*	7.9026e-05	0.1844	0.3459	98.87
	ARIMA(1,1,0)	-0.000669	0.2803	0.00	53.49
	ARIMA(2,1,0)	-0.000289	0.2486	1.99e-8	34.82
	ARIMA(1,1,1)	1.6759e-05	0.1945	1.14e-6	89.45
2006	ARIMA(0,1,1)	-3.439e-04	0.1795	8.41e-8	108.59
	ARIMA(0,1,2)	-3.533e-04	0.1669	0.00537	135.15
	ARIMA(1,1,0)	-3.938e-04	0.00	0.00089	31.82
	ARIMA(2,1,0)	-1.632e-04	0.00	0.01182	65.37
	ARIMA(1,1,1)*	-2.255e-04	0.0001	0.92616	136.20

Table 5.2: Statistical parameters of the residuals of the different studied stochastic models for Amsterdam. (*) indicates the best model.

generate a ranking of the models. Only the first two ranking techniques are used to verify the adequacy of the different models. The characteristics that are compared for this step are the P-Value, the error mean and the standard deviation.

For both locations, Amsterdam and Ngarenanyuki, the first two models that better perform the time series are the models ARIMA(1,1,1) and ARIMA(0,1,2). ARIMA(1,1,1) and ARIMA(0,1,2) are the most suitable models, since they are the model with the largest Box-Pierce P-value [57] and one of the lowest error mean.

Through an analysis of the residuals behavior, it is possible to achieve a conclusion about the most suitable model. The two models are adequate to represent the residuals of the time series because of they show little or no residuals coefficient correlation.

Relying also on the bibliography [33, 5], both ARIMA(1,1,1) and ARIMA(0,1,2) are confirmed to be suitable, but the most complete and suitable model for the case of daily clearness index values seem to be ARIMA(1,1,1). In the light of the results, we can conclude that the ARIMA(1,1,1) model can be accepted.

Once this analysis is performed and concluded, the generalized ARIMA model the ARIMA model is uniquely relate to model ARIMA(1,1,1).

5.1.3 Coefficients Generalization

Once the order of the model has been chosen, it is possible to proceed with the generalization of the coefficients. In order to obtain a more general model, in addition to the two main locations – and the four auxiliary locations for each of them – also the results for Valencia and Madrid have been considered. The ARIMA(1,1,1) model establishes that the daily clearness index value, K_t^d, depend on the value of the previous day K_{t-1}, with a certain ratio ϕ_1 of the difference between values K_{t-1} and K_{t-2}, plus a normal random variable, a_t, and its reminder θ_1 of the sample before. This random variable a_t gives a specific mean m, and standard deviation σ.

$$K_t = K_{t-1} + \phi_1 \cdot (K_{t-1} - K_{t-2}) + a_t - \theta_1 \cdot a_{t-1} \tag{5.1}$$

The coefficients ϕ_1, θ_1, and σ, that best describe each series of data are reported in Table 5.3 (i.e. for each year from 2004 to 2006 for both the location considered).

Furthermore, it is possible to conclude summarizing the coefficients for all the years erasing all the extreme values for each city. All the coefficients in Table 5.3 can be summarized for all the years, transforming into mean coefficient for each city. In Table 5.4 are reported the

Year	Amsterdam - ARIMA(1,1,1)	Ngarenanyuki - ARIMA(1,1,1)
2004	$\phi = -0.4821$ $\theta = -0.9483$ $\sigma = 0.172$	$\phi = -0.3435$ $\theta = -0.9169$ $\sigma = 0.102$
2005	$\phi = -0.3826$ $\theta = -0.99$ $\sigma = 0.163$	$\phi = -0.38133$ $\theta = -0.9871$ $\sigma = 0.098$
2006	$\phi = -0.3864$ $\theta = -0.9621$ $\sigma = 0.168$	$\phi = -0.2243$ $\theta = -0.9968$ $\sigma = 0.0988$

Table 5.3: Coefficients of the model ARIMA(1,1,1) in three years for locations of Amsterdam and Ngarenanyuki.

Model	Amsterdam	Ngarenanyuki	Madrid & Valencia
ARIMA(1,1,1)	$\phi = -0.4171$ $\theta = -0.9701$ $\sigma = 0.1735$	$\phi = -0.31639$ $\theta = -0.96697$ $\sigma = 0.09890$	$\phi = -0.2955$ $\theta = -0.9305$ $\sigma = 0.151$

Table 5.4: Coefficients of the model ARIMA(1,1,1) in an average year for locations of Amsterdam, Ngarenanyuki and the average values of Madrid and Valencia from [5].

averaged coefficients for the two locations studied and the values obtained in [5] for Valencia and Madrid. It is possible to see that the order of magnitude of the coefficients is similar and there are not strong dependences on the locality. From that observation and due to the ARIMA model is used only to define the order of the values obtained from Bednt's correlation (not to defined the absolute values), it has been chosen to define some average coefficients that are suitable for each localities. Finally, mean values between the two localities studied, Amsterdam and Ngarenanyuki, and from the locations from the paper [5], Valencia and Madrid are computed.

These coefficients are reported in Table 5.5 and they provide a site independent model ARIMA(1,1,1).

So having the values of the coefficients, the mean, the standard deviation and the seasonality or not, inferred, it is possible to simulate the time series and generate the daily sequence values.

Generalized parameters	ϕ	θ	σ
	-0.3310	-0.9493	0.1436

Table 5.5: Coefficients of the site-indipendent model ARIMA(1,1,1).

5.2 Markov Model Generalization

5.2.1 Data Generalization

Concerning the model generalization, the input data have to be available with the same resolution that is wanted for the output data (if the output that it wants to be achieved has a high sub-hourly resolution, i.e. 10-min, 5-min or 1-min, also the input need to have a sub-hourly resolution).

It should be pointed out that having an adequate series of data for several years with systematic measurements of horizontal global solar radiation with high resolution, proves to be really problematic.

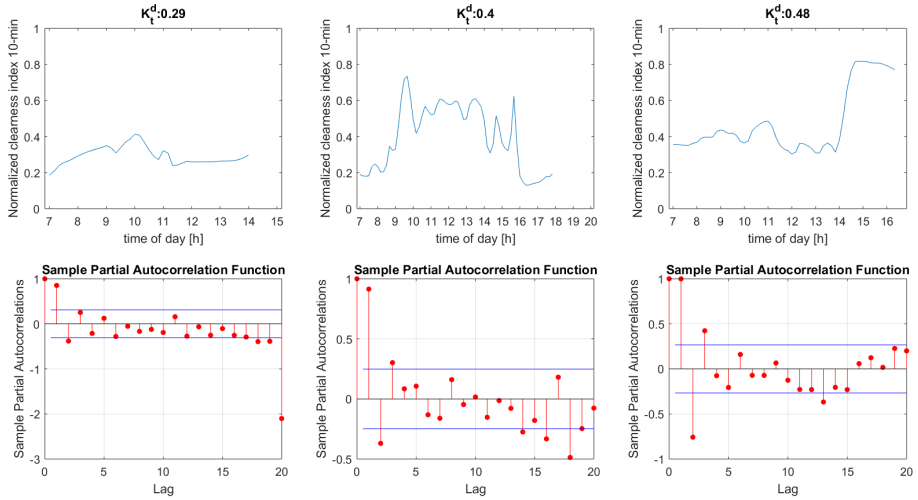
In the thesis' case three years of observed data with 10-min time step for each location studied have been used to built the Markov Transition Matrices, one for each class. The classes are better defined in section 4.3.2.

To increase the number of data without impacting the reliability of the procedure, data from other locations around the location of interest (or with similar meteorological characteristics) have been selected. The choice of the number of locations depends also on the number of classes in which the daily clearness index is divided to built the MTMs, and on the order selected for the Markov model. The order of the Markov model influences the dimensions of the matrices.

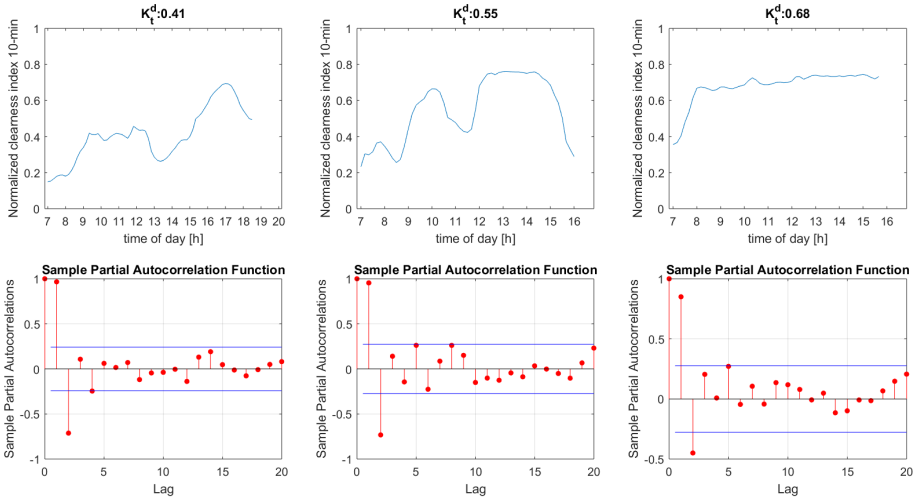
The three choices are strongly interdependent, and the choice of the number of locations and the number of classes has to been done in parallel, once the order of the Markov model has been selected. A good reference to have some general guidelines about the choice of the number of classes is 4.3.2.

5.2.2 Markov Model Order

The Markov model generalization has been carried out for the two locations already considered in the previous ARIMA model generalization phase. In order to define the order of the Markov model, the PACF of normalized sub-hourly clearness index has been evaluated. Looking at the results, for Ngarenanyuki and for Amsterdam, similar trends can be noticed. Figures 5.2a and 5.2b show that the first two lags are statistically significant, especially for days with low values of K_t^d. The statistical significance of lags 1 and 2 lead to the selection of a second order Markov model.



(a) Amsterdam.



(b) Ngarenanyuki.

Figure 5.2: Plots of $k'_{t,10min}$ and corresponding partial autocorrelation plots.

Class	Classification	Range
1	Extremely cloudy day	$0 \leq k_t^d < 0.25$
2	Cloudy day	$0.25 \leq k_t^d < 0.40$
3	Variable day	$0.40 \leq k_t^d < 0.60$
4	Sunny day	$0.60 \leq k_t^d < 0.75$
5	Extremely sunny day	$0.75 \leq k_t^d \leq 1$

Table 5.6: Classification of the daily clearness index classification inside the classes.

5.2.3 Markov Classes

Having data from two different locations, it is useful to visualize more particular cases that can occur in the daily clearness index yearly values. This is a really fundamental thing in the determination of the classes.

For instance, Ngarenanyuki has a more stable daily clearness index around the values of 0.4 and 0.55 with some extreme values before and after. Amsterdam has daily clearness index in general lower than Ngarenanyuki, but with more variability and volatility.

The final choice, to have classes that can represent each climatic condition, is to have five classes divided as reported in Table 5.6.

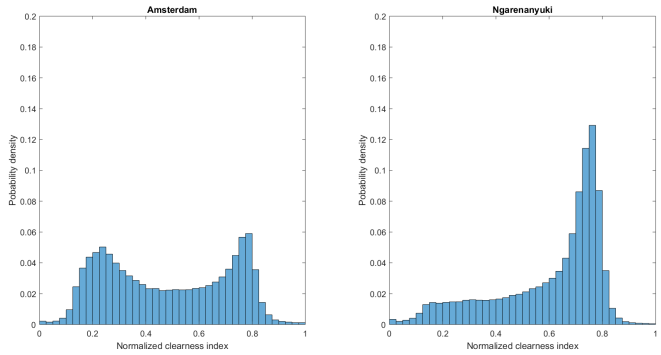
The first and the last classes represent the extreme values, indeed these ranges are very uncommon to have in every locations. The second class represent the case of a general cloudy day. The third class a variable day and the fourth class a sunny day.

5.2.4 Predefined Markov Transition Matrix

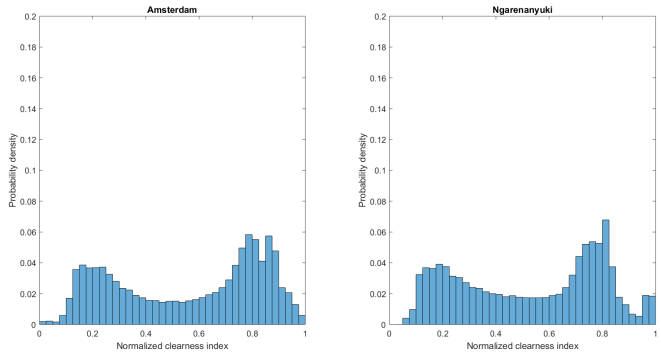
Assuming to have five classes, a resolution of 10-min time step and only three years of data for each studied location, the second Markov Transition Matrices will contain on average, 31 536 data equally distributed. Remembering that the second order Markov matrices have a dimension of 100^2x100. Making a comparison with [6] the Matrices contain in average 210 240 data for each. In sight of this is necessary increase the number of data. Data from others four locations around each location of interest have been selected, as it has been done for the ARIMA model. Increasing or decreasing $\pm 3^\circ$ of latitude and longitude with respect to the coordinates of the locality of interest you get a circular area around your place.

The final dataset is composed by five different locations with similar climatic characteristics to each place of interest, and each of them for three years. That means to have fifteen years of data for each location, on average 157680 for each Matrix.

For instance, the MTMs of Ngarenanyuki will be created with the dataset of Ngarenanyuki



(a) Observed data.



(b) Generated data.

Figure 5.3: Comparison between pdf using the inputs data of the locality of interest and other four localities around it.

(latitude ϕ, longitude λ), plus the dataset of the $Loc_1(\phi, \lambda{-}3°)$, $Loc_2(\phi{-}3°, \lambda)$, $Loc_3(\phi, \lambda{+}3°)$ and $Loc_4(\phi{+}3°, \lambda)$. The same thing will be done for Amsterdam and the four localities around it.

Second order MTMs were empirically generated from the solar radiation data sets and these MTMs are used in the synthetic generation of sub-hourly solar radiation data.

An important verification to see if the generated values reproduce correctly the observed data is looking a the probability density function. In Figure 5.3, it is possible to see a comparison between the pdf of the observed data of five locations and the synthetically generate data. To conclude that the amount of data used to generate the MTMs are not sufficient to reproduce the observed values. The pdf of the entire set of data are a little bit distorted, and more data are needed for the creation of the MTMs.

Trying to simplify as much possible the entire procedure, it has been observed that the process that requires the most time is the Markov's process. The creation of the MTMs and the elaboration of the big amount of data required takes 50% of the total time. The principal

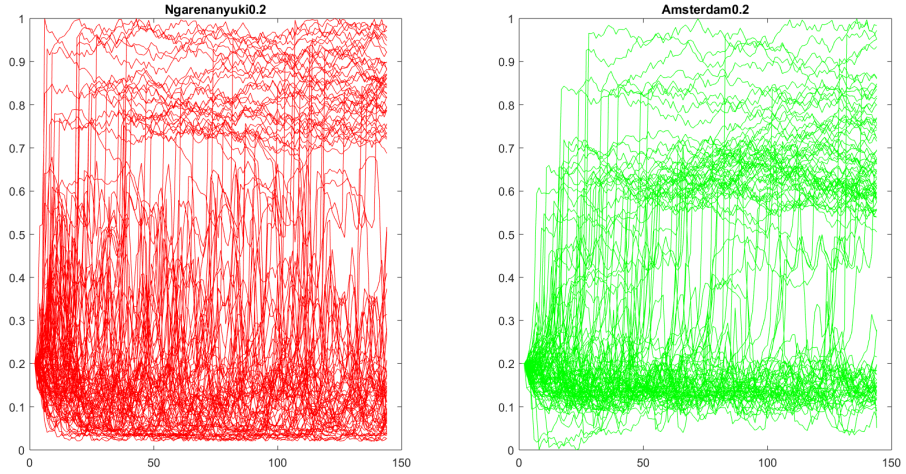


Figure 5.4: Overview of results obtained through different MTMs corresponding to two different locations, starting from the same distribution of daily data.

limitation of the procedure, in the current state, is exactly the fact that vast amounts of data are required and not always, making its use challenging is some areas.

A possible option to simplify the procedure is the generalization of the MTMs, checking if it is possible to have only a group of matrix for all the world. This means to have a single group of MTMs for all the locations where the synthetic generation is performed. Each location exploits the matrix in a different way coming from different daily clearness index meteorological characteristics and variability.

To validate this possibility, it is fundamental to look at the effects that the different groups of MTMs for the two locations have on the same input data.

A way to do this, a well accepted approach, is to generate several (e.g. five) different time series, one for each class, as uniform random values. For Class 1, generating a series of 365 values uniformly distributed between 0 and 0.25, for Class 2 a series of 365 values uniform distributed between 0.25 and 0.40 and so on for the successive classes. Then these five time series are used as five different years of daily clearness index. These sequences can be adopted to synthetically generate five years of 10-min daily clearness index, one for each input time series class. To highlight the differences this process has to be done using the MTMs of both locations, Amsterdam and Ngarenanyuki. Figure 5.4 shows a comparison of the values obtained through the MTMs of Amsterdam and Ngarenanyuki for class two $[0.25 - 0.40]$.

Looking at the output generated data in Figure 5.4, it is possible to see that using different MTMs leads to very different distributions even for the same input series. For instance, in the plot for Amsterdam it can be observed that the majority of the data are concentrated

between 0.1 and 0.2, with another group of values around 0.6. Concerning Ngarenanyuki, the clearness index values are more uniformly distributed between 0.05 and 0.6.

Using the same MTMs for the generation of data in both locations would lead to very similar distributions of sub-hourly values. Since the distributions obtained with MTMs corresponding to the two locations are very different, the effects of the MTMs cannot be ignored.

Generalizing the procedure using the same MTMs for locations with very different meteorological conditions would lead to a significant loss of accuracy in the data. In conclusion, these plots show that it is fundamental to have different MTMs for different locations with completely different meteorological variables.

Another option for the generalization of the procedure could be to cluster more than one location according to their common climatic characteristics and create the group of MTMs in function of the number of these sets.

An important characteristic to classify the locations into some sets is the probability distribution function of the clearness index. In this work [8], it has been done a comparison between different probabilities density distribution functions of global solar irradiance in different regions. In [8], one-minute global horizontal solar irradiance distributions were studied at sites in five different climatic regions. For this purpose, a normalized clearness index on the air mass, K_t', has been used. The one-minute distributions of K_t' found have different properties, they are either unimodal or bimodal, depending on the location. These distributions are different for each of the locations analyzed in function of the climatic area where the locations are. These results point to the importance of local distribution and type of clouds variability in high-resolution irradiance distributions, and highlight the role of clearness index in differentiating these distributions.

The climate classification map proposed by Wladimir Köppen [1] has been used. A huge number of climate studies adopted this or a former release of the Köppen-Geiger map. In [1] is presented a new digital Köppen-Geiger world map on climate classification for the second half of the 20th century. This new map is based on recent datasets from the Climatic Research Unit [58] (CRU) of the University of East Anglia and the Global Precipitation Climatology Centre [59] (GPCC) at the German Weather Service. The classification used for this thesis methodology is only the one based on the "Main climates". According to this classification the whole world is classified into five Zones or Areas, without considering the more detailed classification on "Precipitations" and "Temperature". The five areas are the following and are reported in Figure 5.5:

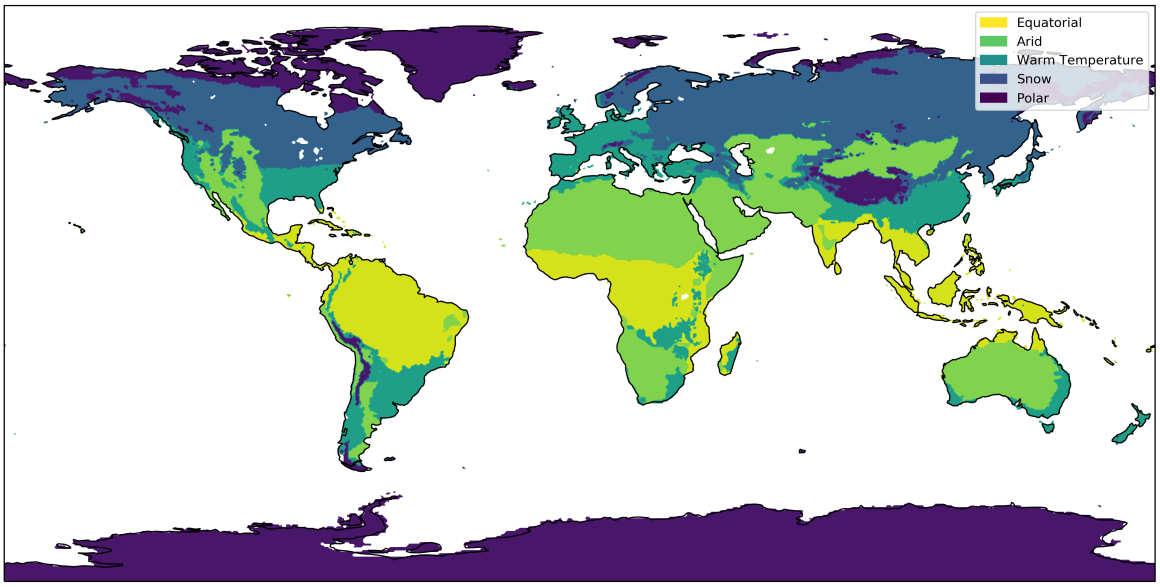


Figure 5.5: World map of climatic areas as proposed by Köppen-Geiger.

- A: equatorial
- B: arid
- C: warm temperature
- D: snow
- E: polar

The observation reported in [8] supports the hypothesis that generating different sets of MTMs for each geographical area could reduce the approximation that the generalization implies. Using this classification five different groups, one for each area, of MTMs can be created and used based on the area location under study belongs to. So the idea is to have five different MTMs, one for each zone, and within each zone five more MTMs that correspond to the initial classes in which the daily clearness index was classified.

Before proceeding with the classification and the creation of the MTMs for each area, it is important to verify if locations that belongs to different climatic areas have different probability distribution functions (PDF). This analysis has been performed for two different locations, Amsterdam and Ngarenanyuki, that belong to two different climatic zone, Zone C and Zone A, respectively. In Figure 5.6, the probability density function of the observed data for the two locations has been reported. The PDF of the two locations corresponds to

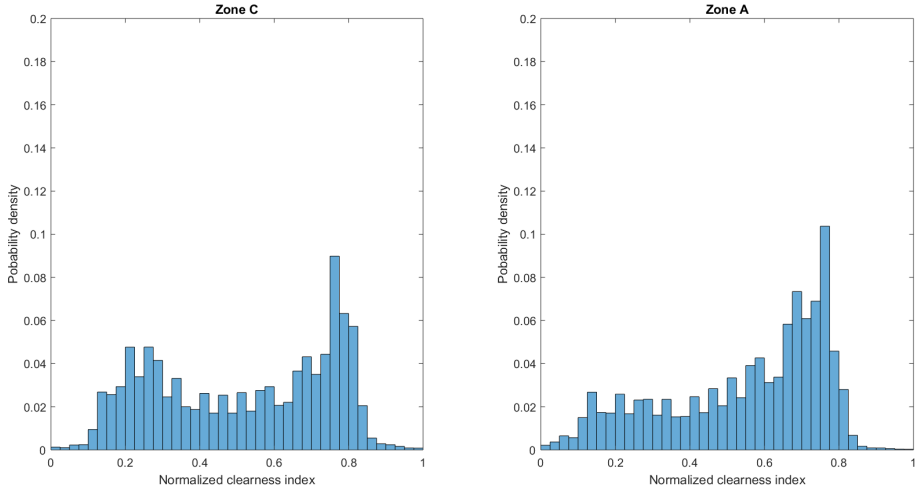


Figure 5.6: Probability Density Function for observed data for two different Climatic Areas.

the distribution found in [8].

To proceed with this approach and the creation of the generalized MTMs, it is necessary to create detailed matrices, with enough data for each group. To obtain more data, it has been decided to collect the data for the three years, from 2004 to 2006, for ten total locations that belong to the same area. This means to have in average 315,360.0 data for each Matrix, that is a consistent number.

In some cases the locations chosen to describe an area correspond to some capital and relevant cities, in other cases the locations are determined by matching a specific latitude and longitude, determined to be relevant for the full description of an area.

In Table 5.7 the locations used to built the MTMs for the different areas are reported with the corresponding coordinates. For the Polar area, there are no data available and the MTMs have not been built.

Area	**Locations**	**Coordinates** (ϕ, λ, STZ)
A	Ngarenanyuki	$-3.13°, +36.89°, 3$
	Loc_1	$-3.13°, +33.89°,$
	Loc_2	$+0.13°, +36.89°,$
	Loc_3	$-3.13°, +39.89°,$
	Dar es Salam	$-6.82°, +39.28°,$
	Maputo	$-25.89°, +32.60°,$
	Libreville	$+0.41°, +9.46°,$
	Lagos	$+6.52°, +3.37°,$
	Bangui	$+4.39°, +18.55°,$
	Accra	$+5.60°, -0.18°,$
B	Niger	$+19.68°, +8.07°,$
	$Algerian_1$	$+29.75°, +2.34°,$
	$Algerian_2$	$+27.28°, -3.97°,$
	$Algerian_3$	$+25.40°, +6.13°,$
	Mauritania	$+20.51°, -7.21°,$
	Ciad	$+18.20°, +18.44°,$
	Egypt	$+25.06°, +27.23°,$
	Sudan	$+19.20°, +27.94°,$
	Libya	$+25.24°, +15.62°,$
	Tamanrasset	$+22.78°, +5.52°,$
C	Milan	$+45.46°, +9.18°, 1$
	Barcelona	$+41.38°, +2.17°, 1$
	Lisbon	$+38.72°, -9.10°, 0$
	Paris	$+48.85°, +2.35°, 1$
	Berlin	$+52.52°, +13.40°,$
	Amsterdam	$+52.37°, +4.89°,$
	Loc_1	$+52.37°, +1.89°,$
	Loc_2	$+55.37°, +4.89°,$
	Loc_3	$+52.37°, +7.89°,$
	Loc_4	$+49.37°, +4.89°,$
D	Norway	$+64.77°, +11.76°,$
	$Sweden_1$	$+60.64°, +15.28°,$
	$Sweden_2$	$+65.26°, +16.69°,$
	Finland	$+63.69°, +28.95°,$
	Russia	$+62.57°, +37.50°,$
	Tallin	$+59.43°, +24.75°,$
	Helsinky	$+60.15°, +24.73°,$
	Moscow	$+55.75°, +37.61°,$
	Kiev	$+50.45°, +30.52°,$
	Oulu	$+65.01°, +25.46°,$
E	No data	No data

Table 5.7: Locations used to built the MTMs.

Once the MTMs have been built, for each locations around the world it is possible to synthetically generate data from the Markov process.

The final tool requires as input the coordinates of the location under study, the coordinates are the latitude, longitude and the Standard Time Zone. The tool can determine the climatic area 'X' at which the location belongs to.

Then, using the MTMs of the determined climatic area, it is possible to compute one year of high resolution clearness index values from the daily clearness index values.

The inclusion of the matrices dataset for the Climatic Areas in the tool, allowing to speed up the procedure without loosing the resolution of data and avoiding to look for a vast amount of high resolution data, difficult to find.

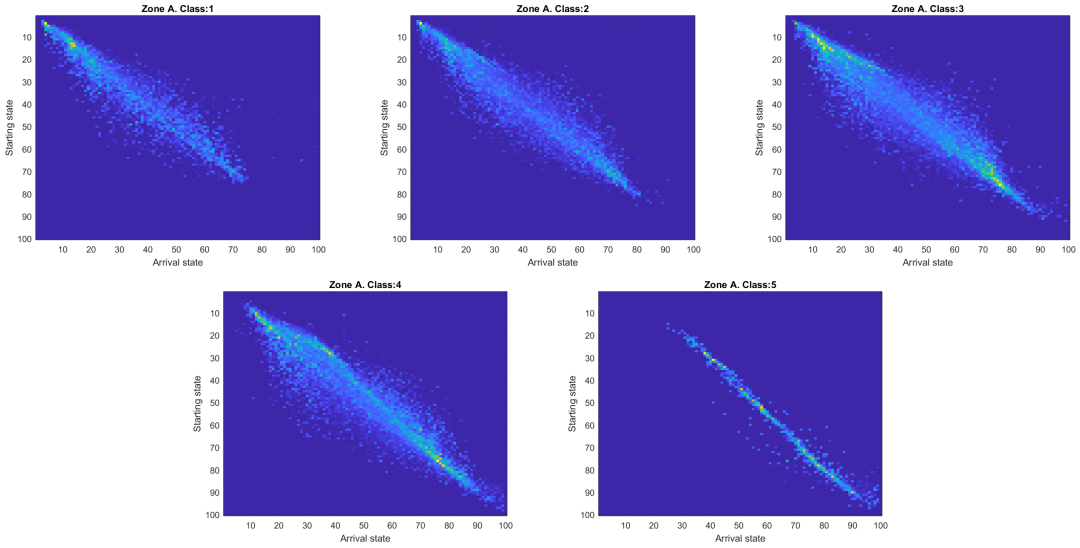
Analyzing the differences between the MTMs, it is possible to visualize some trends inside them. In Figure 5.7, the Markov Transition Matrices for two different Zones are reported. The location of Amsterdam belongs to Zone C, instead Ngarenanyuki belongs to Zone A.

Looking at Figure 5.7, it is possible to observe some similar trends. Since a second order Markov model has been used the MTMs of each class are tri-dimensional $100 \times 100 \times 100$. In order to show them in Figure 5.7 the sum across the third dimension has been performed. Indeed, the first and the last class of matrix, that correspond to values of k_t lower than 0.25 and higher than 0.75 respectively, contain a relatively low amount of data, because as discussed before these two classes are rare for both climatic areas. For what concerns Zone C (Amsterdam), the matrices of the class two is the most densely populated. This means that the probability to find values 0.25 and 0.4 are really high. Concerning Zone A, the most densely populated matrices are thethird and the fourth, that correspond to values between 0.4 and 0.75.

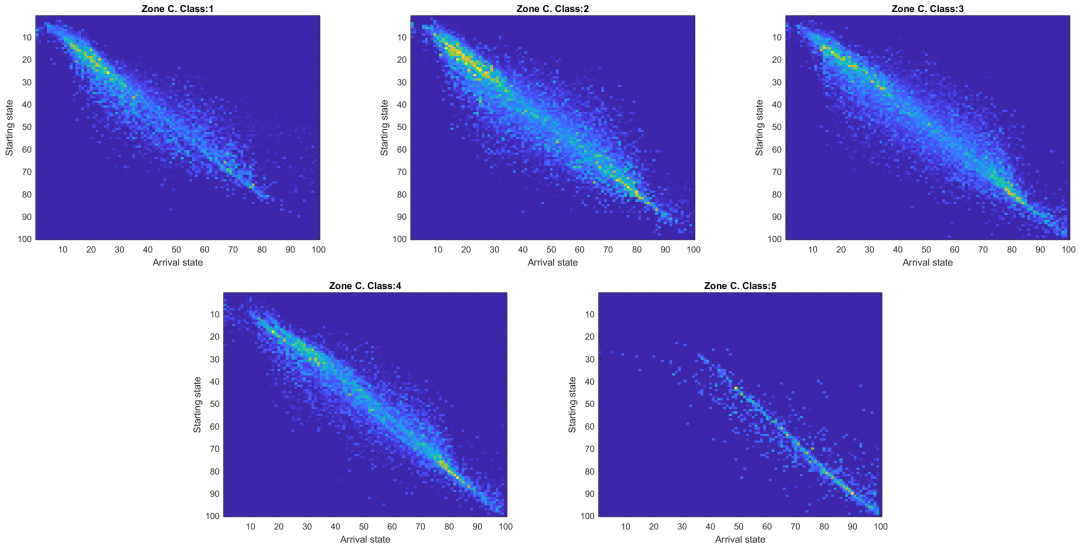
To confirm the importance of the climatic areas, it is relevant to look at the probability density distribution functions of the Synthetic generated data. It is possible to check if the PDF for Zone A, verified with the input data from Ngarenanyuki, and for Zone C, verified with input data from Amsterdam, correspond to the PDF proposed in [8]. Then it has to be verified if these probability densities correspond to the observed distribution, as shown in Figure 5.6 for the two different zones.

In [8], a bimodal PDF is proposed for the case of Area C. Indeed, the bimodal character observed is attributed to the presence of cloudy (peak at low K_t') and cloudless sky (peak at high K_t') conditions.

Dispersion and amplitude of the peaks vary for each location, but the form of the PDF



(a) Zone A



(b) Zone C

Figure 5.7: Markov Transition Matrix for each class.

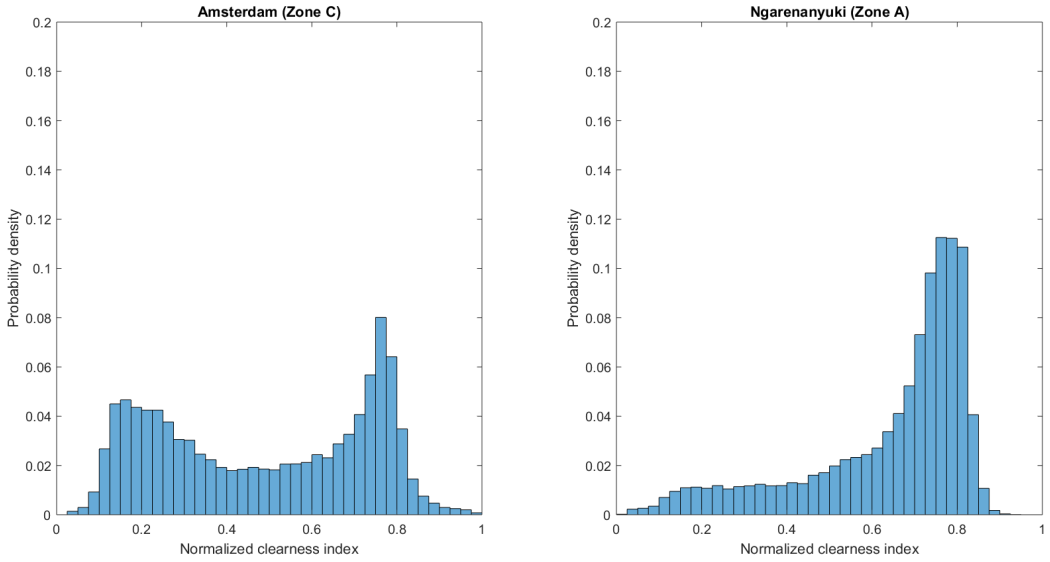


Figure 5.8: Probability Density Function for Synthetically Generated data for two different Climatic Areas.

for the zone, as can be seen in Figure 5.8, is well represented, and reflects different local atmospheric conditions.

Thanks to the results obtained with this analysis it is possible to highlight the importance the type of climatic conditions in high resolution irradiance distributions. The clearness index coefficient provides an important role in differentiating these distributions. The PDFs for high resolution as 10-min, 5-min and 1-min, clearness indexes could be either unimodal or bimodal, depending on the location. In addition, these distributions can be different even for sites in the same climate zone, for what concern the amplitude and relative width of the peaks.

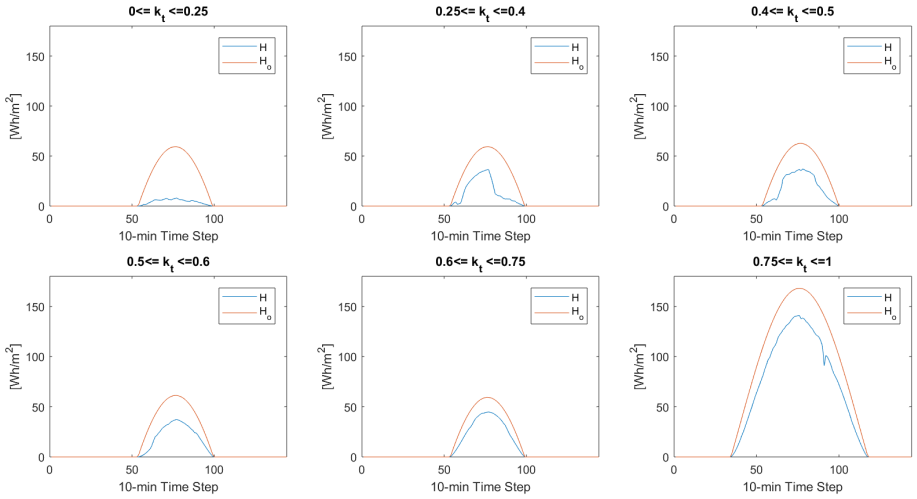
5.3 Final Configuration

In the end, the generalized procedure, that include the determination of the climatic zone is really fast and it allows to generate the data with a high accuracy. The higher result is the fact that to achieve this high resolution it is not necessary to look for a vast amount of data having all the step and model already generalized.

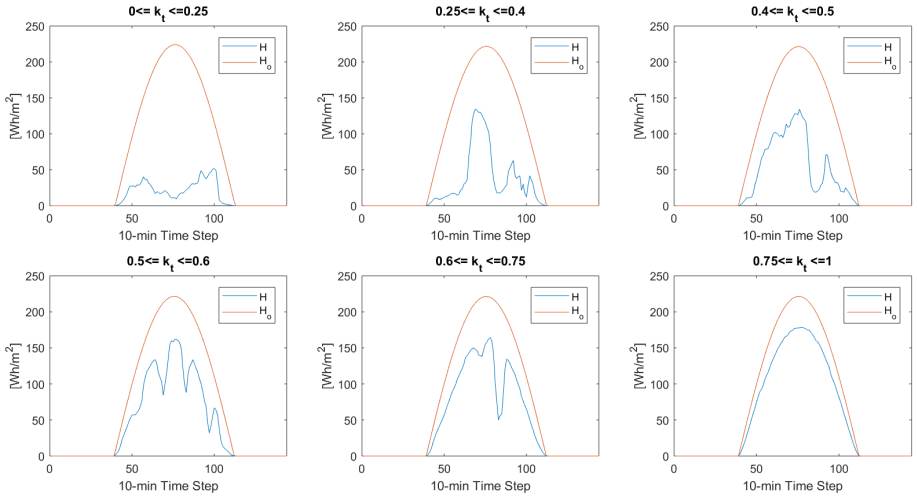
The knowledge of the statistical behavior of short-term variability of solar irradiance will facilitate more precise evaluation of the uncertainty in the long-term annual energy production of solar power plants.

In Figures 5.9 and 5.10 some representations of what is the output of the procedure.

In Figure 5.9 the 10-min global and extraterrestrial solar irradiation are plotted in function of the time of the day for different type of days, for both locations. The classification is based



(a) Amsterdam



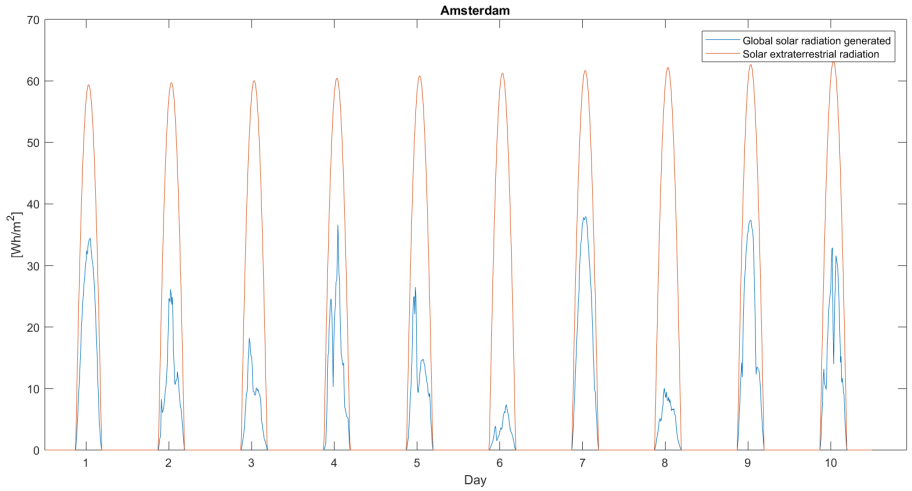
(b) Ngarenanyuki

Figure 5.9: Illustrative examples of solar irradiation plots for days with different values of K_t^d.

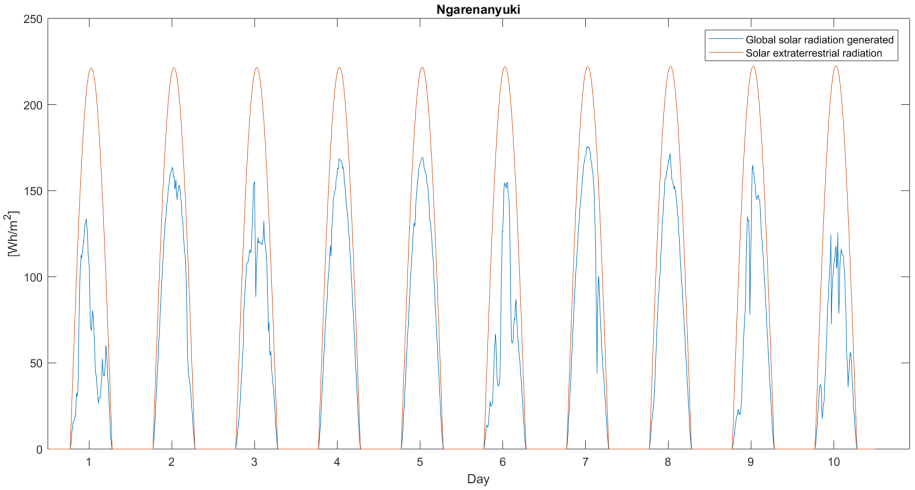
on the daily clearness index.

In Figure 5.10, 10-min global and extraterrestrial solar irradiation values for ten consecutive days are reported.

The flow chart in Figure 5.11 describes the entire final procedure of the tool and takes into account all the generalizations made to the models.



(a) Amsterdam



(b) Ngarenanyuki

Figure 5.10: Illustrative examples of solar irradiation plots for consecutive days.

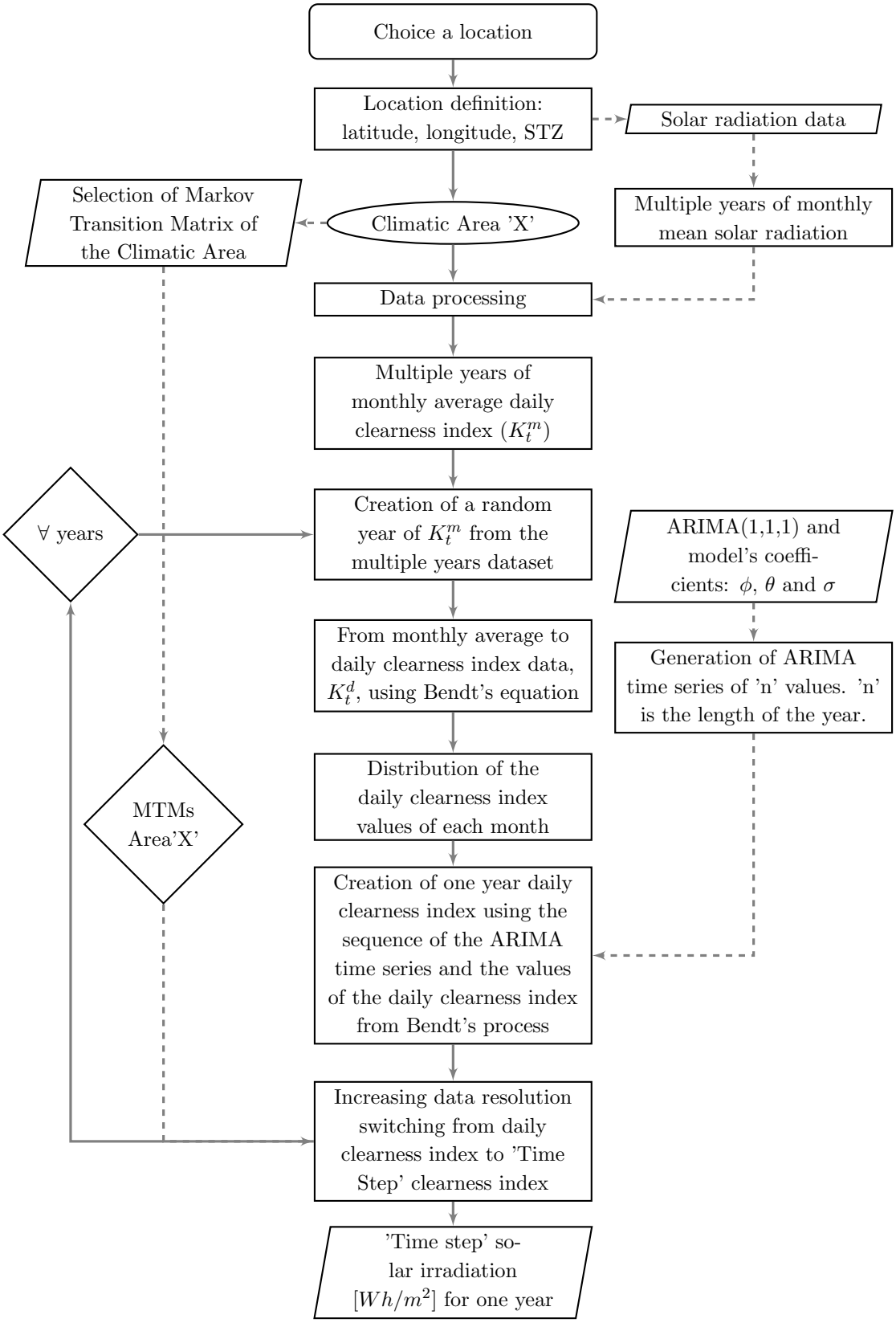


Figure 5.11: Flow chart of the final procedure that includes all the generalizations applied.

Chapter 6

Results and Validation

6.1 Applications

To calibrate and validate the tool proposed, the procedure has been tested on data relevant to the two pilot locations, Amsterdam and Ngarenanyuki. These belong to two different climatic areas, warm temperature and equatorial, respectively.

Ngarenanyuki has been chosen because in a rural area close by Arusha a mini-grid has been installed and includes a PV system, a diesel and a storage. From this system it is possible to collect data of the energy produced by the PV panels, i.e. to have a term of comparison for the performances evaluation.

Amsterdam has been chosen to have a locality in a different climatic area from the equatorial one. Futhermore, Amsterdam has really cloudy conditions all along the year and really different meteorological characteristics than Ngarenanyuki.

The idea to have two different locations with different meteorological conditions is an added value for the generalization of the procedure.

6.1.1 Data Collection

Freely available solar radiation data for long temporal range, with high temporal resolution and large spatial resolution, are scarce and difficult to find.

For the tests performed in the thesis project, it has been decided to use the data derived from the SoDa database – Solar Radiation Data Service [4]. SoDa provides access to a large set of information related to solar radiation and its use collecting information from a list of services and webservices.

Depending on the resolution needed for the input of the model and on the temporal

window, the data have been obtained from one of the three different datasets available on SoDa. Each solar radiation database collects data with different tools and with different resolutions. Below a description of the three different databases has been reported.

NASA-SSE The Surface meteorology and Solar Energy service provides time series of daily Global Horizontal Irradiation values. The series are available worldwide, covering a spatial resolution of 1° (approx. 100 km). Time coverage of data is from July 1983 to June 2005.

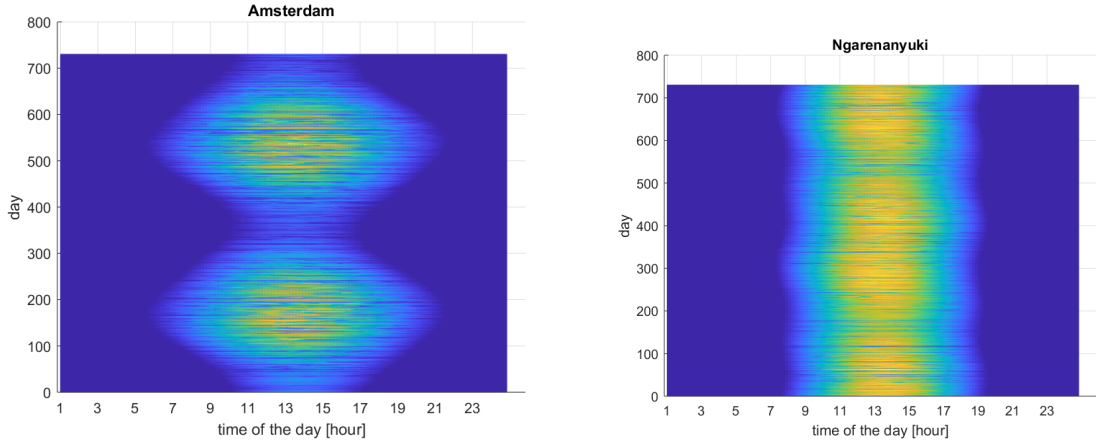
HelioClim-1 This service provides HelioClim-1 time series for monthly, weekly and daily Global Solar Irradiation values over a horizontal plane, as well as monthly irradiation received by a plane normal to sun rays. The geographical coverage corresponds to the Meteosat satellite field of view, i.e. covers Europe, Africa, the Atlantic Ocean, Middle East. The spatial resolution is approximately 20 km. The time coverage of data is from 1985 to 2005.

HelioClim-3 This services provides time series of the radiation components over a horizontal, fix-tilted and normal (to beam radiation) plane for the actual weather conditions as well as clear-sky conditions. Geographical coverage corresponds to the Meteosat satellite field of view, i.e. covers Europe, Africa, the Atlantic Ocean, Middle East. Spatial resolution is 3 km at Nadir, and approx. 4-5 km at 45° of latitude. Data are available with a time step ranging from 10 min to 1 month. The time coverage is from February 2004 up to current day-2. Free data are only available until December 2006.

The data with the highest resolution come from HelioClim-3, that provides a satellite-derived solar radiation database.

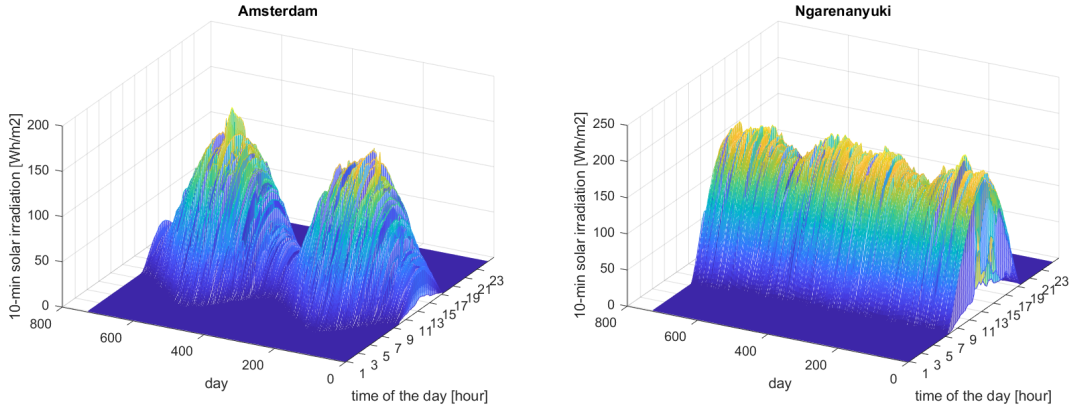
The procedure has been applied to the three years of freely available data – from 2004 to 2006. Furthermore, the calibration of the model and the daily resolution validations have been performed using also the data from the other two databases, that have a lower resolution but provide a larger temporal coverage.

The resolution chosen for the implementation of the thesis work is of 10 minutes. This resolution allows to have a higher level of detail for the simulations than the most commonly used in literature and in commercial tools (commonly set to hourly time-step). The choice not to further decrease the time-step is due to the limited availability of data at such resolutions.



(a) Amsterdam. (b) Ngarenanyuki.

Figure 6.1: Two years of solar irradiation.



(a) Amsterdam. (b) Ngarenanyuki.

Figure 6.2: Two years of solar irradiation.

6.1.2 Data Exploration

In Figures 6.1 are reported plots for two consecutive years of synthetic generated data from Amsterdam and Ngarenanyuki.

In Figure 6.2 are reported the same plots of Figure 6.1 but seen in 3 dimensions, having in the y-label the value of the 10-min solar irradiation.

6.2 Validation Methodology

The procedure allows the synthetic generation of multiple years of solar radiation values, while increasing their resolution compared to the monthly input data.

Figures 6.3 and 6.4 show multiple years of synthetically generated data for the two locations under study, Amsterdam and Ngarenanyuki respectively. For the specific case of the

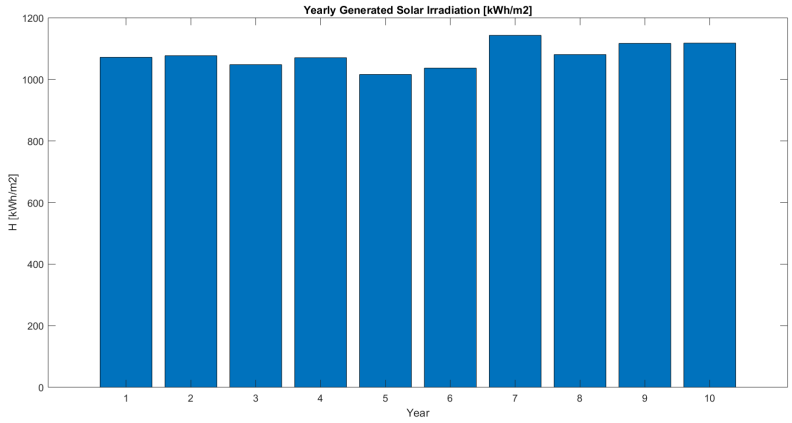


Figure 6.3: Ten years of yearly sum of 10-min global solar irradiation for Amsterdam.

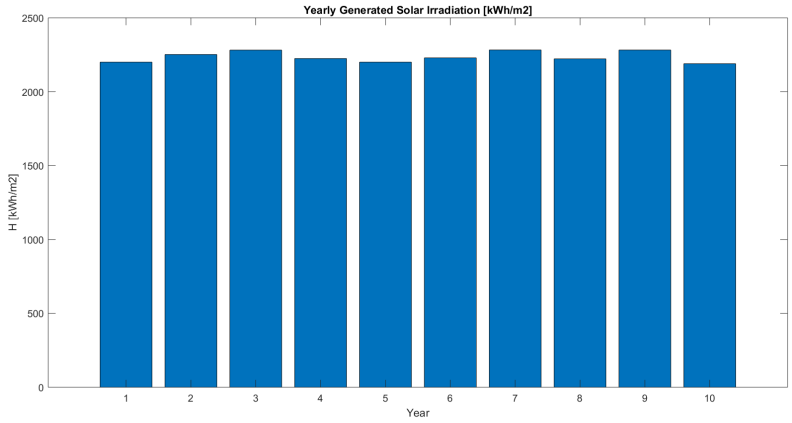


Figure 6.4: Ten years of yearly sum of 10-min global solar irradiation for Ngarenanyuki.

thesis, the tool has been used to obtain 10 years of power produced by a PV panel.

The entire procedure has been validated step by step, challenging all the most relevant assumptions.

Two different types of validation have been used. The first set of validation procedures are *internal*, and have been applied to each model and correlation used as part of the procedure. These validations are necessary for the calibration of the model.

The second type of validation is *global*. It allows to validate the entire tool, comparing the results obtained with the actual time series.

6.3 PV Power Output Results

Using the synthetically generated solar radiation data, it is possible to compute the incident radiation on the PV panel H_T, using the set of equations introduced in section 2.6. Using

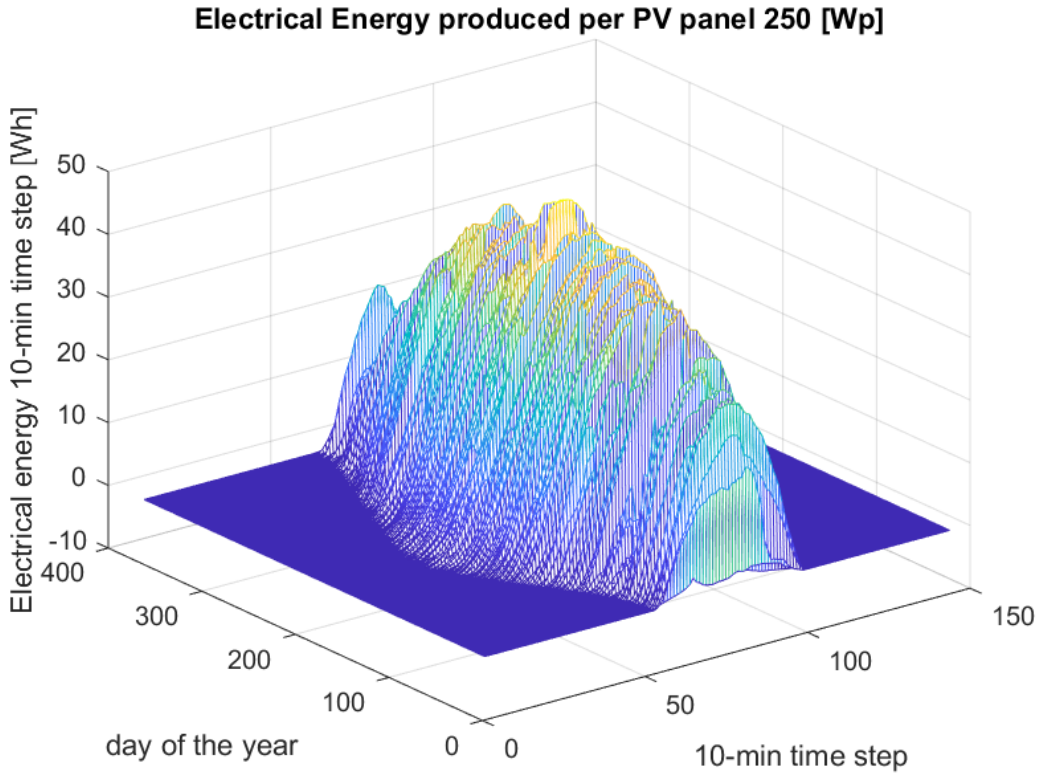
Ground	**Reflectivity**
Water	7%
Coniferous forest	7%
Bituminous and gravel roof	13%
Dry bare ground	20%
Weathered concrete	22%
Green grass	26%
Dry grassland	20–30%
Desert sand	40%
Light building surfaces	60%

Table 6.1: Table of some characteristic values for ground reflectivity.

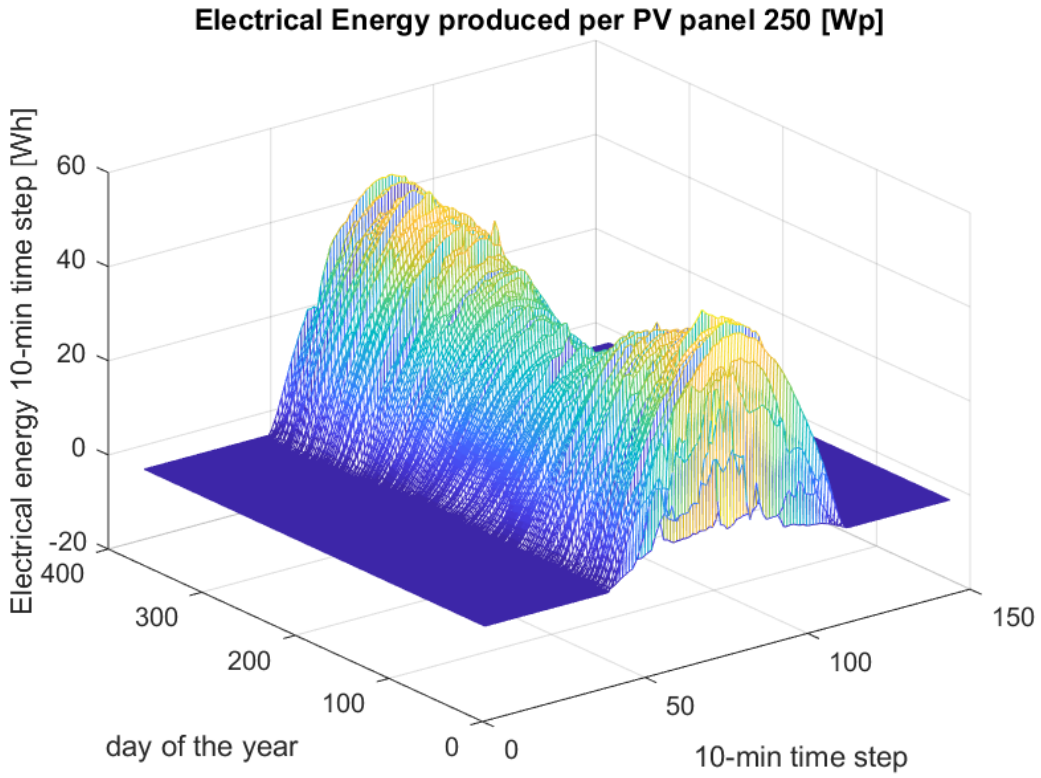
equations 2.31 and 2.32, it is possible to compute the yearly electrical energy produced based on the incident radiation on the PV panel. Then, the calculated power output is compared with the electricity production of the real system.

The electrical energy produced by one PV panel for a typical year is reported in Figure 6.5 for the two specific cases of Amsterdam and Ngarenanyuki. The yearly production for a 250 Wp PV panel is 260 kWh for Amsterdam and 405 kWh for Ngarenanyuki. These values are affected by the characteristics of the PV system. In general, these parameters can be set by the user of the tool. These inputs are listed below and the values used in the calculations are reported in Table 6.2.

- Reflectivity of the ground, ρ. In Table 6.1 some characteristic values are reported.
- Time step of the system.
- Tilt angle of the PV panel, β_{PV}.
- Azimuth angle of the PV panel, ψ_{PV}.
- Data design that depends on the manufactures of the PV panel: nominal power, $P_{PV,nom}$ $[W]$, the maximum power point efficiency $\eta_{mp,STC}$, the coefficient of temperature-power, γ_{PV}, the derating factor, d_{PV}, and the balance of system efficiency, η_{BOS}.
- Ambient temperature, T_{amb}. In the current state of the tool the ambient temperature data for the two locations used for the calibration, Amsterdam and Ngarenanyuki, are taken from the same dataset used for the solar radiation, SoDa. From the website it is possible to download three years of free data with a resolution of ten minutes.



(a) Amsterdam.



(b) Ngarenanyuki.

Figure 6.5: Yearly plot of the electrical energy produced by a PV panel.

Parameter	Value	Unit
ρ	0.5	$[-]$
β_{PV}	30	$°C$
ψ_{PV}	0	$°C$
$P_{PV,nom}$	250	$[W]$
$\eta_{mp,STC}$	15	%
γ_{PV}	0.24	%
d_{PV}	2	%
η_{BOS}	80	%

Table 6.2: Values of some parameters and characteristics used for the derivation of the panel's output, yearly electrical energy.

The entire procedure, starting from the n-years of global solar radiation synthetically generated to n-years of output electrical energy from the PV panel, is reported in Figure 6.6.

In Table 6.2 some important input values used for the derivation of the final results of the validation procedure are reported.

If the characteristics of the system are unknown, the efficiency $\eta_{mp,STC}$ can be calculated as follows:

$$\eta_{mp,STC} = \frac{P_{PV,nom}}{A_{PV} \cdot G_{STC}} = \frac{250[Wp]}{1.6[m^2] \cdot 1[kW]} = 15.6\% \tag{6.1}$$

In Figures 6.7a and 6.7b, ten years of yearly electric energy output $[Wh/y]$ from a single PV panel, obtained through the synthetically generated data for Amsterdam and Ngarenanyuki, are represented. It is possible to observe the decreasing trend due to the degradation of the panels, modelled through the derating factor d_{PV}.

6.4 Internal Validation

In each part of the work, after using some models or correlations to build the procedure, internal validations have been executed.

6.4.1 Bendt's Correlation Results

The Bendt's correlation aims at reproducing the daily clearness index distribution based on its monthly average value. To determine if this method is suitable for the application of interest, its results can be compared to the observed daily values of the same month. In Figures 6.8a and 6.8b, the results for three different months are plotted for both locations.

In order to determine how well the calculated values fit the observed values, the Root

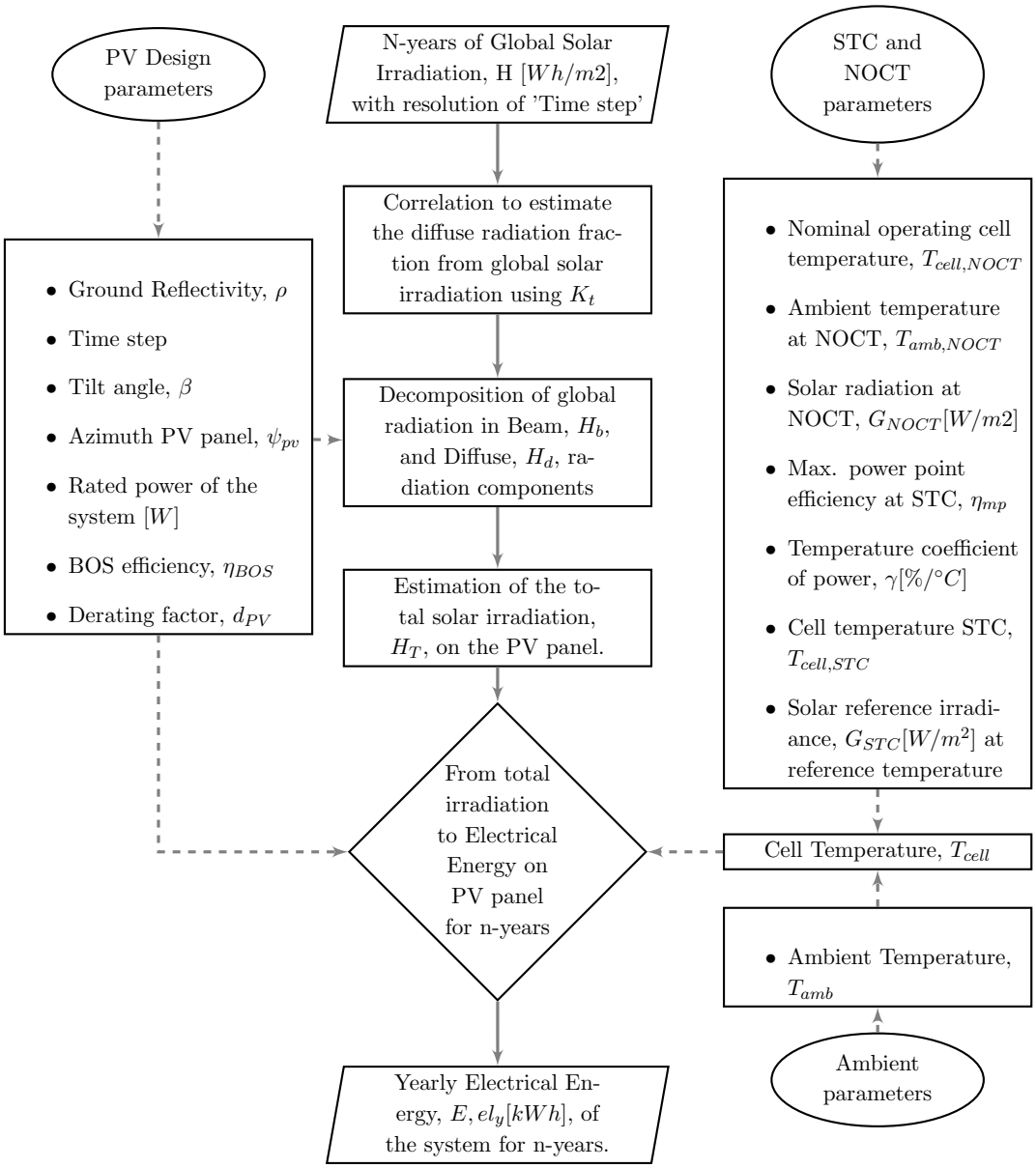
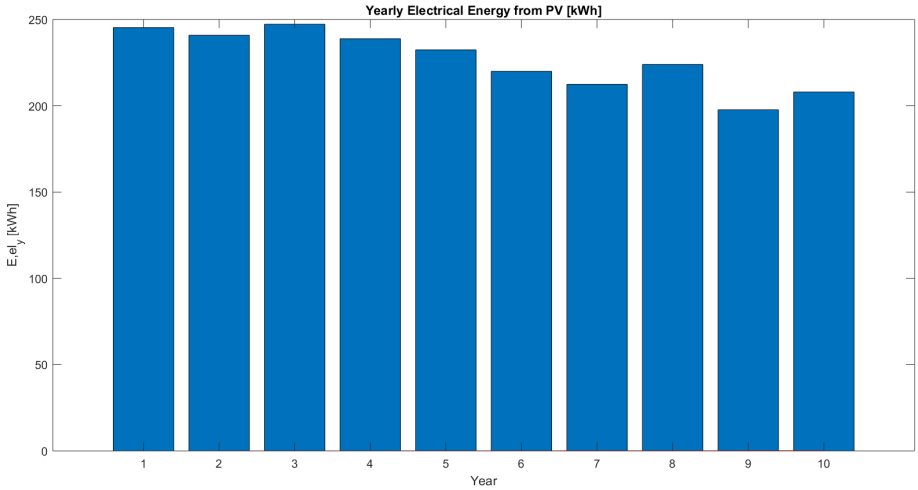
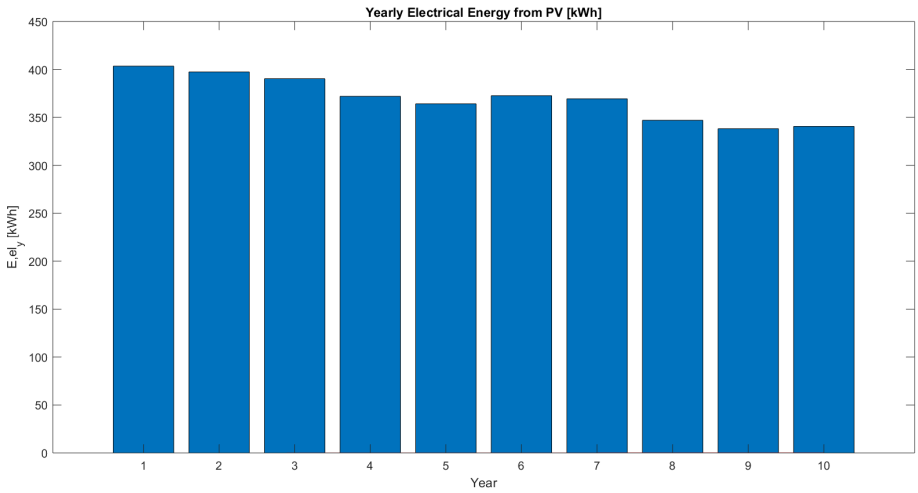


Figure 6.6: Flow chart that represents the process to determine the PV power.

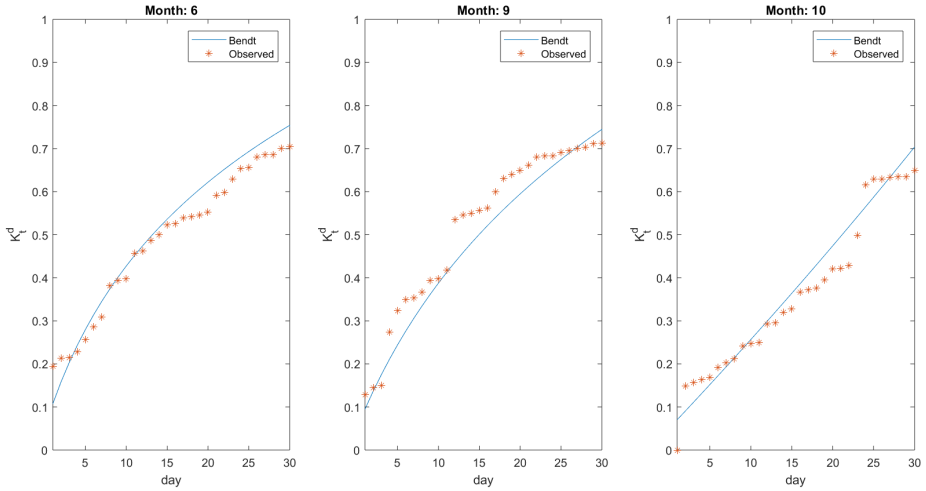


(a) Amsterdam.

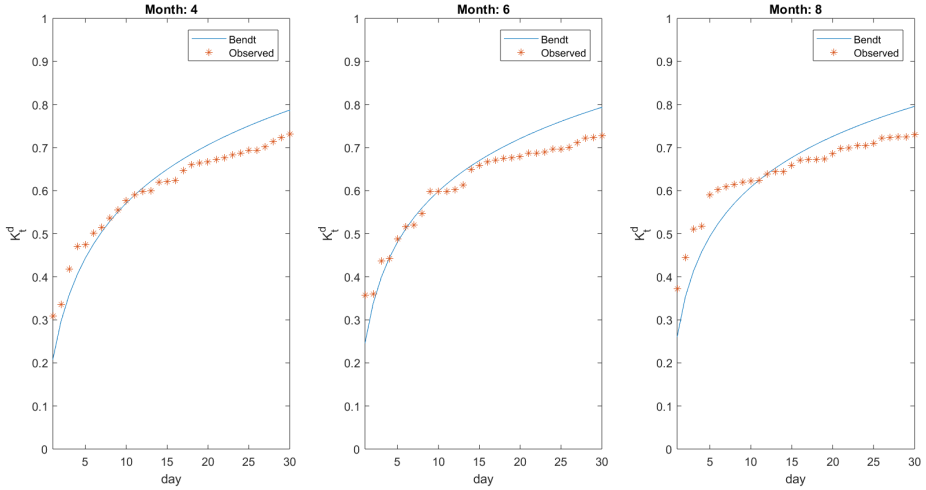


(b) Ngarenanyuki.

Figure 6.7: Ten years of yearly electrical energy produced by the PV panel after taking into account the derating factor.



(a) Amsterdam.



(b) Ngarenanyuki.

Figure 6.8: Correlation of Bendt for observed data for different months.

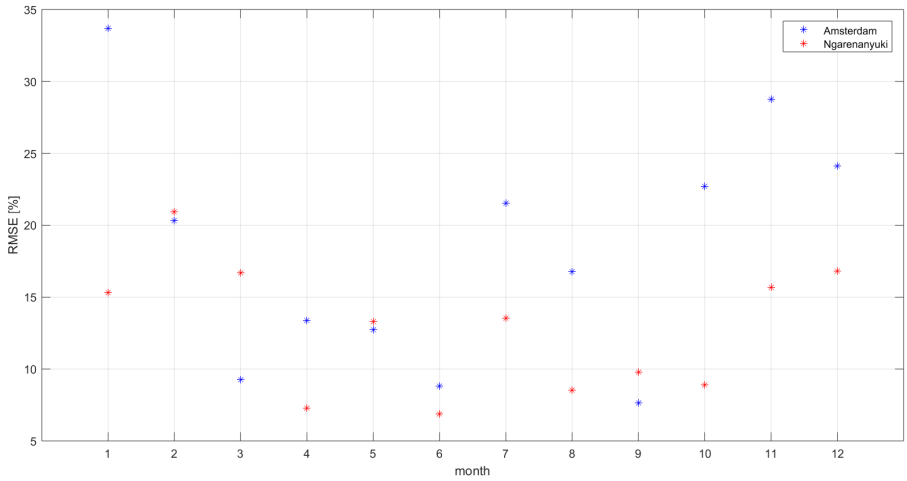


Figure 6.9: Root Square Mean Error between generated and observed, one year.

Square Mean Error (RSME) and the Mean Absolute Percentage Error (MAPE) for the period of one year are analysed. In Figure 6.9, the RMSE for each month for both locations is plotted. The resulting RSME for Amsterdam is around 15% and for Ngarenanyuki is around 12%. In Figure 6.10, the estimated monthly MAPE for both locations is shown. 15.5% is the mean values observed for Amsterdam and 7.7% is the mean value observed for Ngarenanyuki.

The RMSE and MAPE values have been calculated comparing a single synthetically generated year with a single observed year. As explained in section 3.3, the monthly average value used for each month is randomly sampled from all the available observed years. For this reason, a given month of the single observed and generated year might have a different average value. Therefore, the RMSE and MAPE include this error component, and are a slight overestimate of the error in the distribution of each month.

6.4.2 ARIMA Model Results

In Figures 6.11a and 6.11b, the Autocorrelation and Partial Correlation Correlograms of daily data generated by ARIMA(1,1,1) model are represented.

To perform a straightforward comparison, it is possible to compare the ACF and PACF for observed and generated daily values. For both locations, the year observed used for the comparison is 2005. In Figures 6.12a and 6.12b, it is possible to see that for both cases the model follows the measured data (similar values).

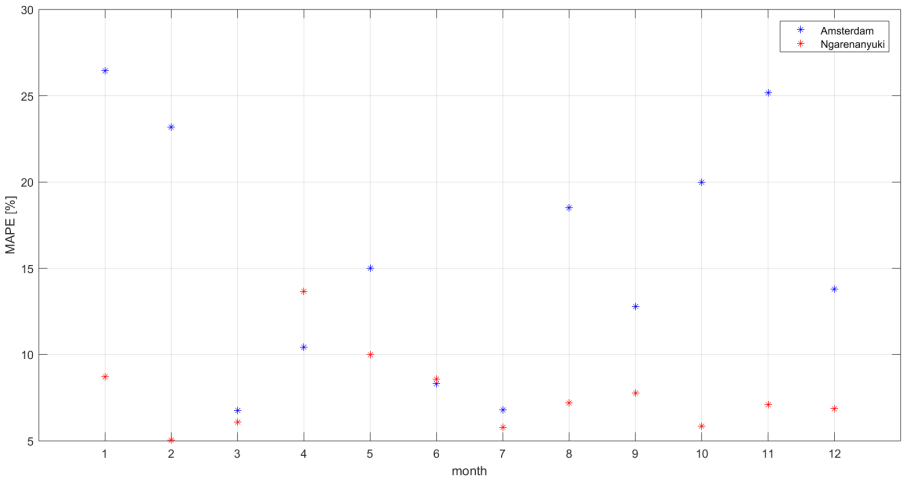


Figure 6.10: MAPE estimated monthly values for one year, both locations.

6.4.3 Markov Model Results

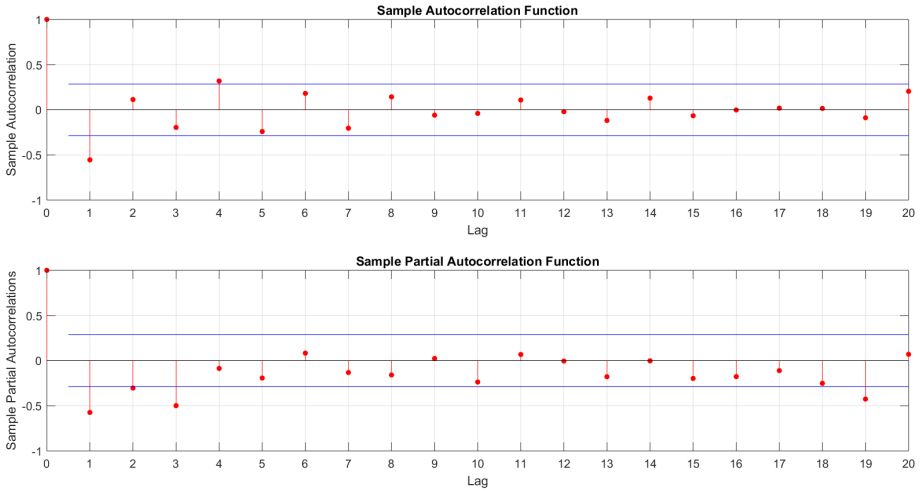
In order to validate Markov and observe its quality inside the model, it is useful to look at and compare the distributions of days with an equal average daily clearness index.

The ACF plots for the observed normalized clearness index and synthetically generated data sets shown in Figures 6.13a and 6.13b are generally similar for the two locations and for the three different days. Most of the autocorrelation functions are gradually reducing with no clear periodicities. The biggest difference is seen in the ACF plots for $K_t^d = 0.44$ for Ngarenanyuki, reported in Figure 6.13b. In this plot the ACF for the observed data is higher than that of the synthetically generated data. This means that consecutive observations in the normalized clearness index observed data have higher correlations. This high correlation produces fewer fluctuations in the radiation plot compared to the synthetically generated data.

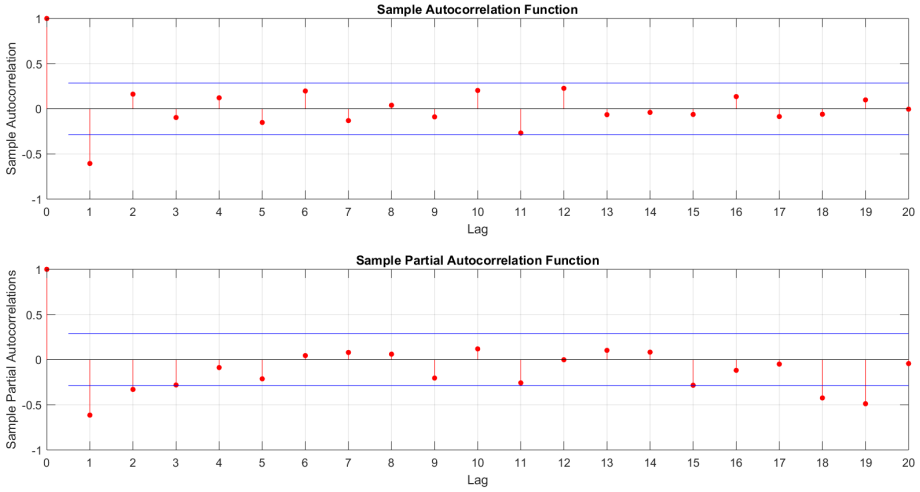
It is important to notice that the ACF is calculated for two random days from the observed and generated data. Since the goal of the methodology is to generate synthetic years rather than forecasting, it is expected that two single days can differ significantly. For the sake of the validation, it is interesting to note that the behaviour of observed and generated data within each class is consistent and differs significantly across different classes.

6.5 Global Results for Amsterdam and Ngarenanyuki

Since the methodology is aimed at the synthetic generation of data rather than at making forecasts, the evaluation of the results cannot be performed through a point-to-point compar-

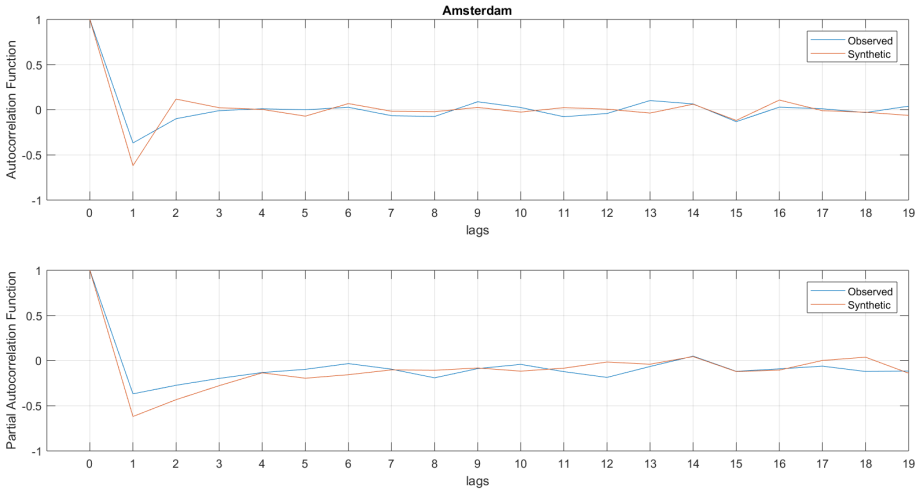


(a) Amsterdam.

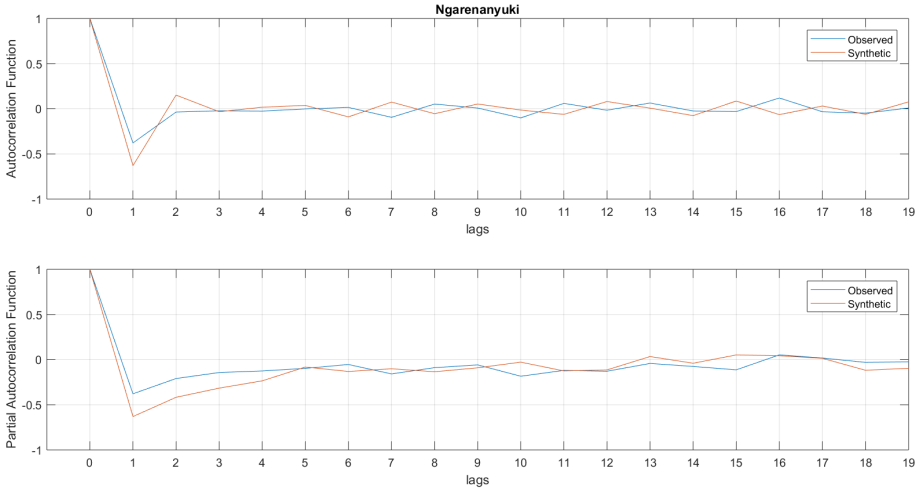


(b) Ngarenanyuki.

Figure 6.11: Correlograms of Autocorrelation (ACF) and Partial Correlation (PACF) of the model daily data for one year.

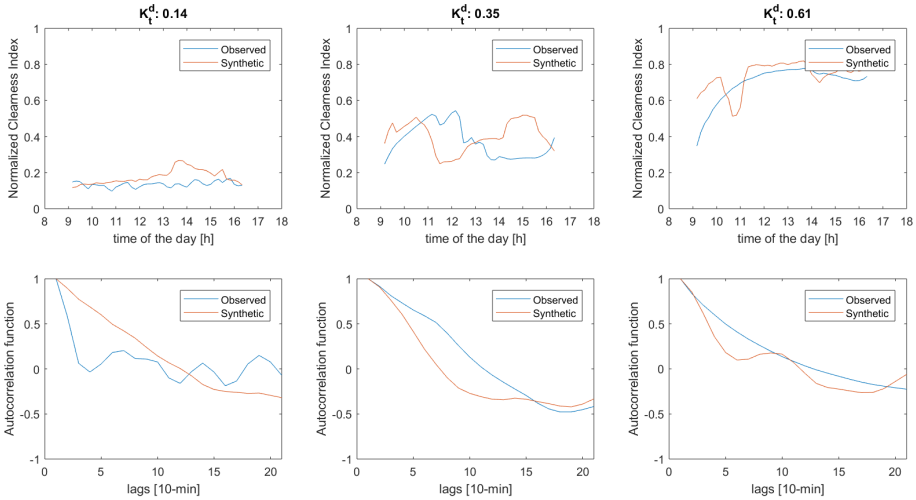


(a) Amsterdam.

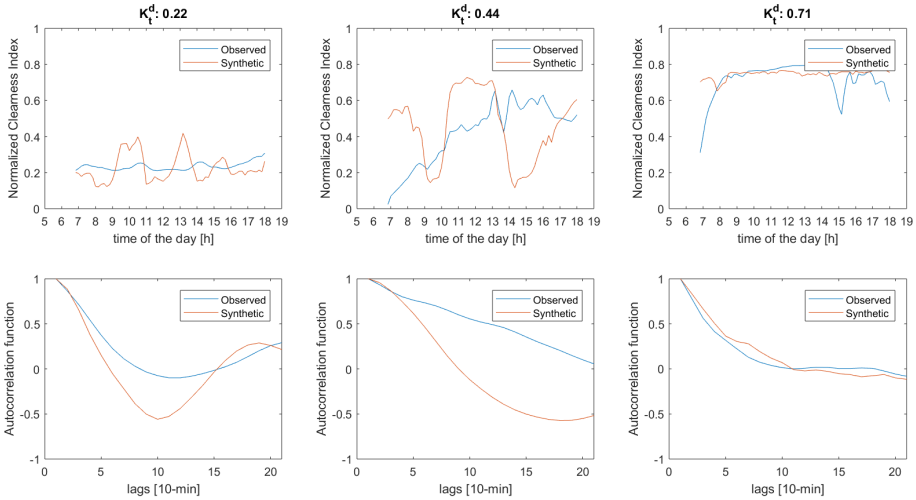


(b) Ngarenanyuki.

Figure 6.12: Corelograms comparison of Autocorrelation (ACF) and Partial Correlation (PACF) of the observed and model daily data for one year.



(a) Amsterdam.



(b) Ngarenanyuki.

Figure 6.13: Plots of normalized clearness index and corresponding autocorrelation functions.

ison. Indeed, the purpose is to generate a realistic time series, rather than determining the exact value of each time step. Therefore, the quality of the results is qualitatively measured by looking at the fluctuations of the output values and checking if the behaviour is similar to the behaviour of the observed data.

In Figures 6.14a and 6.14b, some examples of daily profiles for 10-min irradiation generated and observed are reported for different meteorological conditions. Each subplot represents a different sky condition, determined in function of the daily clearness index K_t^d.

The generated data seem to preserve the general behavior of the observed ones. Indeed, for the Amsterdam case the data have low fluctuations during the day for observed and generated data. The first and the last-but-one subplot have a slightly anomalous behavior. This is due to the fact that, despite the similar K_t^d, the extraterrestrial solar irradiation is completely different. This indicates that the days could belong to two different seasons. For Ngarenanyuki's case, the data have really high fluctuations for K_t^d lower than 0.7, and the generated data seem to reproduce well the behaviour.

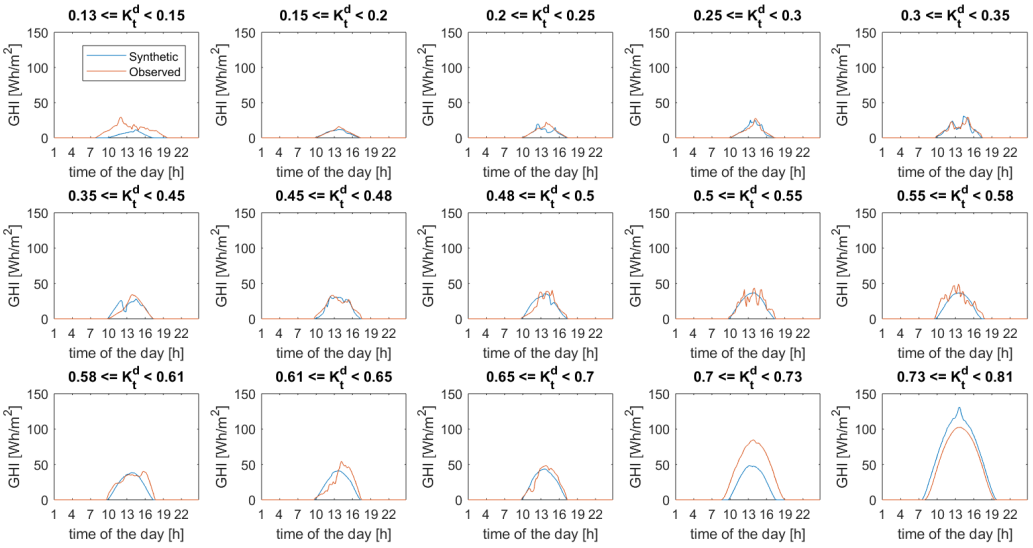
6.5.1 Climatic Areas

The validation of the Climatic Area has already been descripted in subsection 5.2.4. The validation is performed by looking at the probability density functions and making a comparison between the observed and the synthetically generated *pdf* for each area.

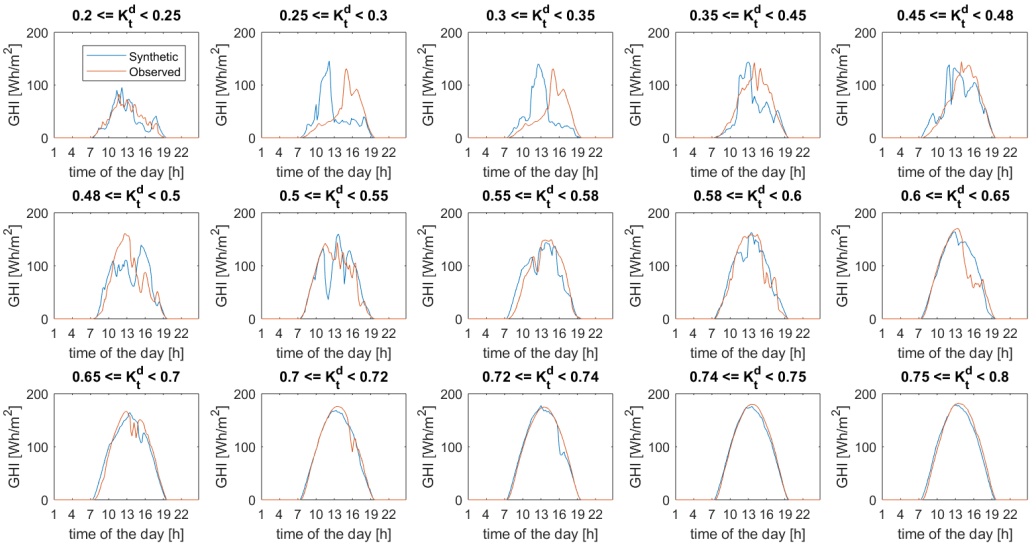
Figure 6.15 shows the similarity between the *pdf* of the observed data and of the synthetically generated data. The similarities between the *pdf*s can bee seen not only on the shape, unimodal and bimodal, of the curves but also on the amplitudes and values of the peaks. Indeed the peaks in Zone A, equatorial, are higher than the peaks for Zone C, warm temperature.

It is interesting to observe the effect of changing the MTMs for the area of interest. Figure 6.16b shows how the *pdf*s change using the MTMs of equatorial area for Amsterdam and the MTMs for warm temperature area for Ngarenanyuki.

In the new *pdf* of the synthetic generated data computed using the wrong MTMs, some changes can be observed for both locations. For Ngarenanyuki, the *pdf* of clearness index values is increased for low values, and another little hump is present. For what concerns Amsterdam, the *pdf* is higher for high values of clearness index.

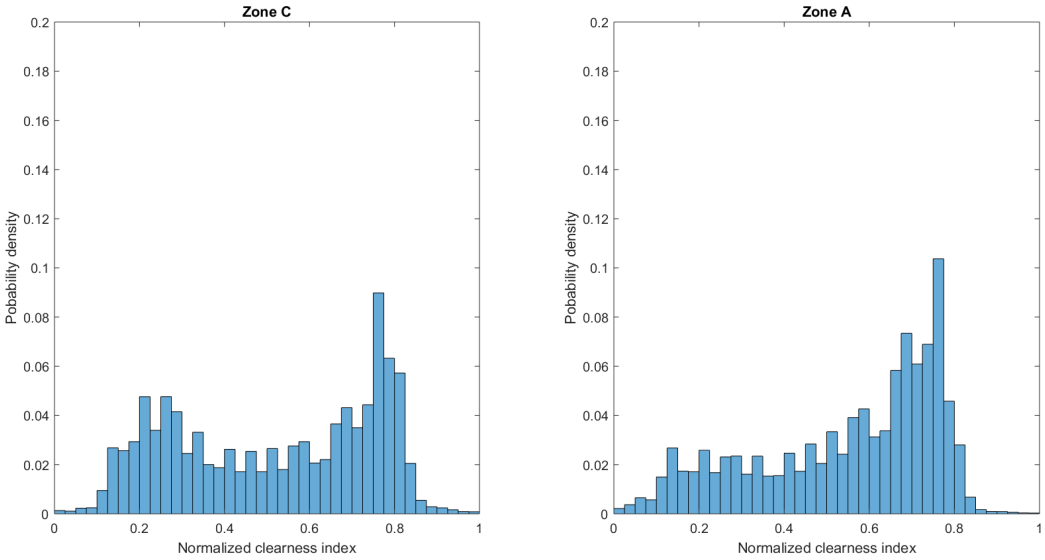


(a) Amsterdam.

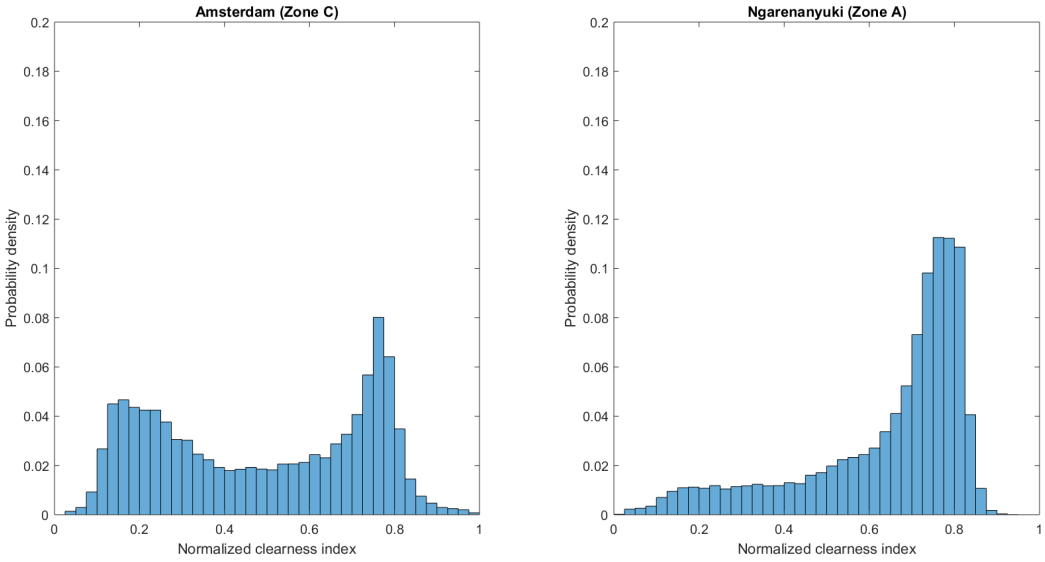


(b) Ngarenanyuki.

Figure 6.14: Daily examples of the results of the synthetic generation method compared to the observed time series for global horizontal irradiation.

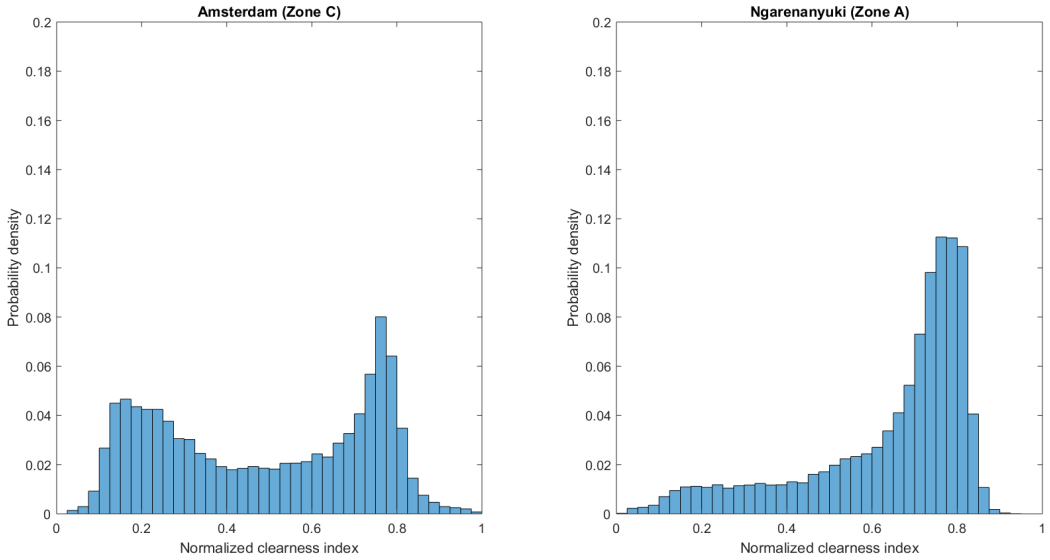


(a) Observed Zone C(left) and Zone A(right).

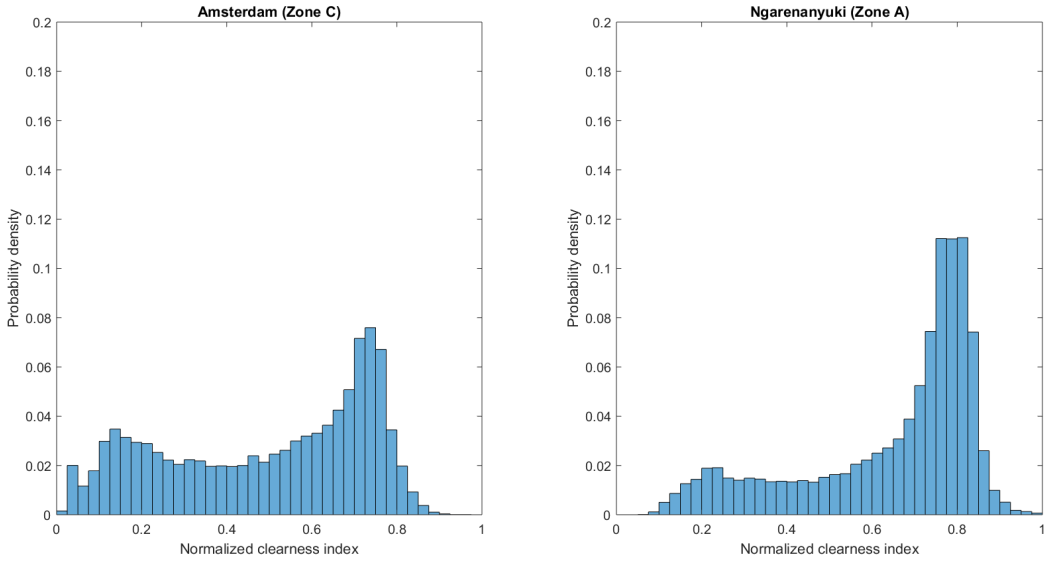


(b) Synthetic generated Zone C(left) and Zone A(right).

Figure 6.15: Probability Density Function for two different Climatic Areas.



(a) Synthetic generated with their MTMs. Zone C(left) and Zone A(right).



(b) Synthetic generated with MTMs from another zone. Zone C(left) and Zone A(right).

Figure 6.16: Probability Density Function for two different Climatic Areas.

6.5.2 Probability density function

The first thing to check is if the data, inside each Markov class, suffer of some alterations.

A way to check these alterations is to compare the probability distribution function of the observed data and synthetically generated data for each class, for both locations.

It is important to keep in mind that the two locations belong to different climatic areas. Amsterdam belongs to the Warm Temperature Area and Ngarenanyuki belongs to the Equatorial Area. Figures 6.17 and 6.18 show the *pdf*s of the normalized clearness index, $K'_{t,10min}$ for the five Markov classes. A close agreement between the *pdf* of the observed data and generated data can be observed. The synthetically generated data preserve the characteristics of the observed data plots. The synthetically generated data distribution is slightly smoother than the distribution of observed data because more years are generated through the synthetic generation.

It is interesting to note that the distributions are approximated so closely that the small differences between the distributions in the observe data for the two locations are preserved in the synthetically generated data. For instance, in the plot for fifth class, the plots for Zone A (Ngarenanyuki), both observed and generated, have higher peaks than the pdf plots for Zone C (Amsterdam) data. This demonstrates a good ability of the tool in simulating particular climatic conditions.

6.5.3 Statistical parameters

Figures 6.19a and 6.19b show a comparison between some statistical parameters obtained from the synthetically generated and the normalized observed data. The data under study are the normalized, 10-minute resolution clearness index time series.

The plots of the synthetic data seem to reproduce well the trend of the observed time series, showing a good match between the statistical properties. However, slight differences are observed in the plots of standard deviation. Generally, for days with K_t^d comprised between 0.4 and 0.6, the values of standard deviation obtained in Zone C are higher than the values obtained from Zone A. For days with daily clearness index lower than 0.4 – cloudy days – the standard deviation obtained from Zone C is lower than the values obtained from Zone A. These factors can be seen in both the plots, for generated and observed data.

The differences discussed above are due to the different climatic characteristics of the two locations. A lower value of standard deviation suggests that the type of clouds found in an area are more uniform in nature and do not produce as many large fluctuations in solar

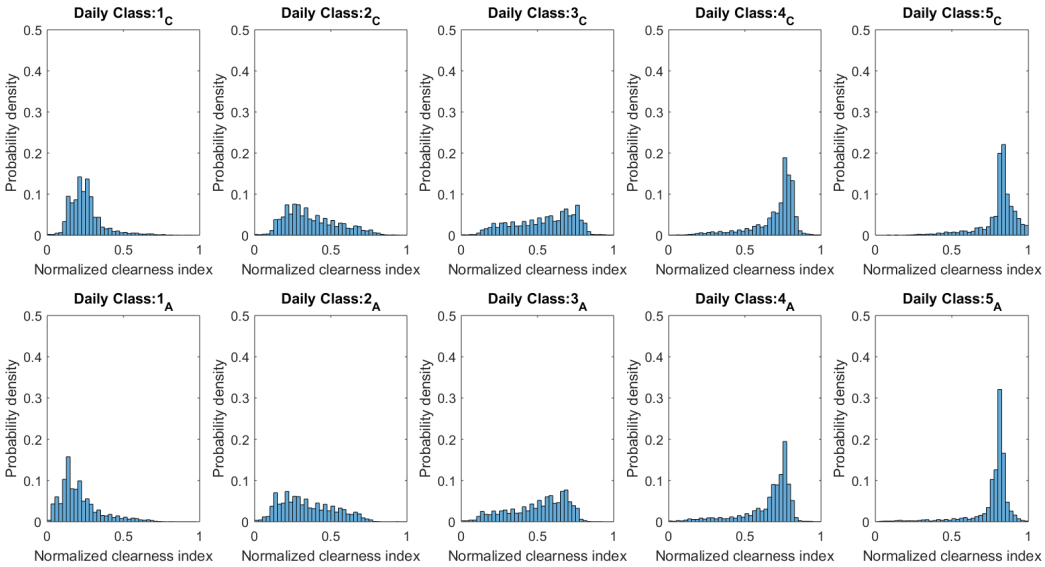


Figure 6.17: Probability Density Function for Observed data for Zone C(top) and Zone A(bottom).

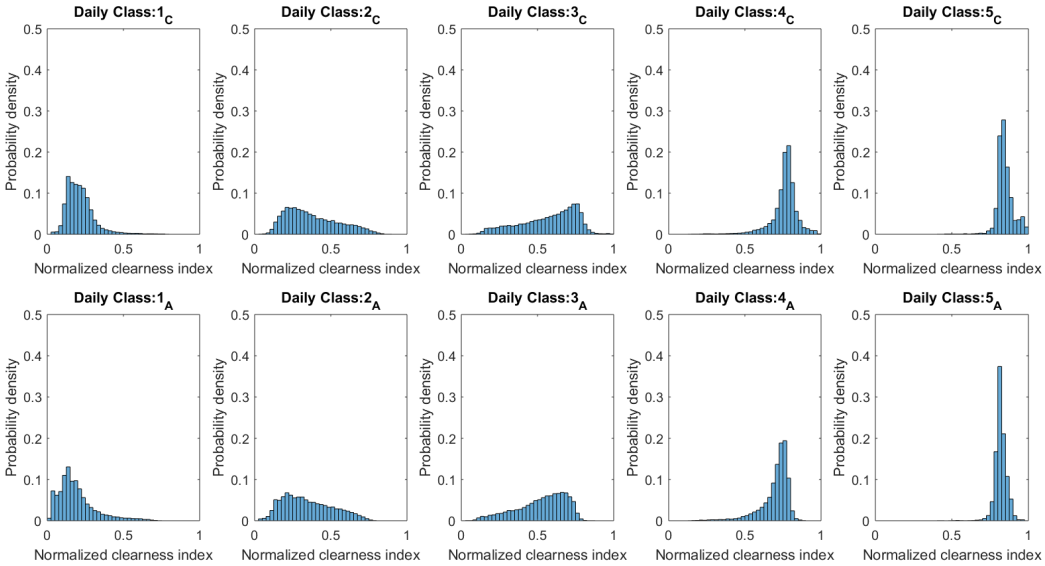


Figure 6.18: Probability Density Function for Synthetically Generated data for Zone C (top) and Zone A (bottom).

radiation.

From the plots, it can be deduced that the fluctuations in solar radiation in cloudy days are lower in Amsterdam than in Ngarenanyuki. On the other hand, for variable and sunny days in Amsterdam the fluctuations in solar radiation are really high. This provides a further confirmation of the dependence between the nature of clouds on location. The general trend among all the cases is similar: for low and high K_t^d, the standard deviation is low. Instead, for intermediate values the standard deviation is higher. This confirms higher fluctuations of solar radiation in these days.

Figures 6.19a and 6.19b show also skewness and kurtosis of the distributions, as a function of daily clearness index. These plots also show some differences derived from the different locations and some similarities in the general trend.

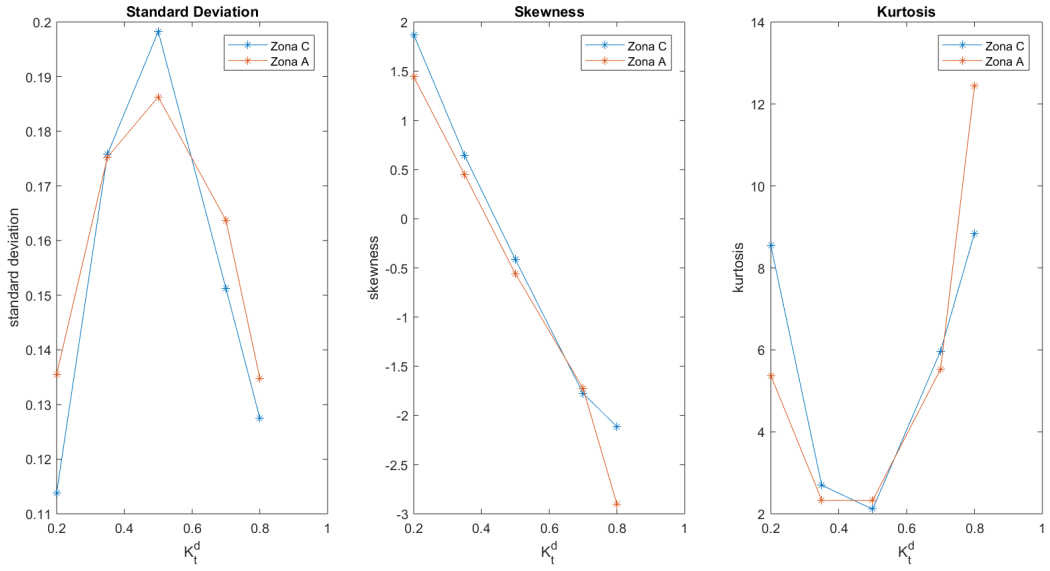
The skewness of the *pdf*s for all cases, observed and generated, for both locations, changes sign, from positive to negative, decreasing down to negative values for K_t^d equal to or higher than 0.4. The skewness plot, from positive to negative values, shows that the distributions are skewed to the right and to the left for low and high values of K_t^d, respectively. This means that for lower values of K_t^d, the distributions are asymmetric to the left, and for higher K_t^d the distributions are asymmetric to the right. This confirms what is shown in the pdf plots.

Kurtosis characterizes the peakiness or flatness of the distribution compared to the normal distribution. Positive kurtosis (leptokurtic distribution) shows a relatively peaky distribution, whereas negative kurtosis values (platykurtic distribution) show a relatively flat distribution.

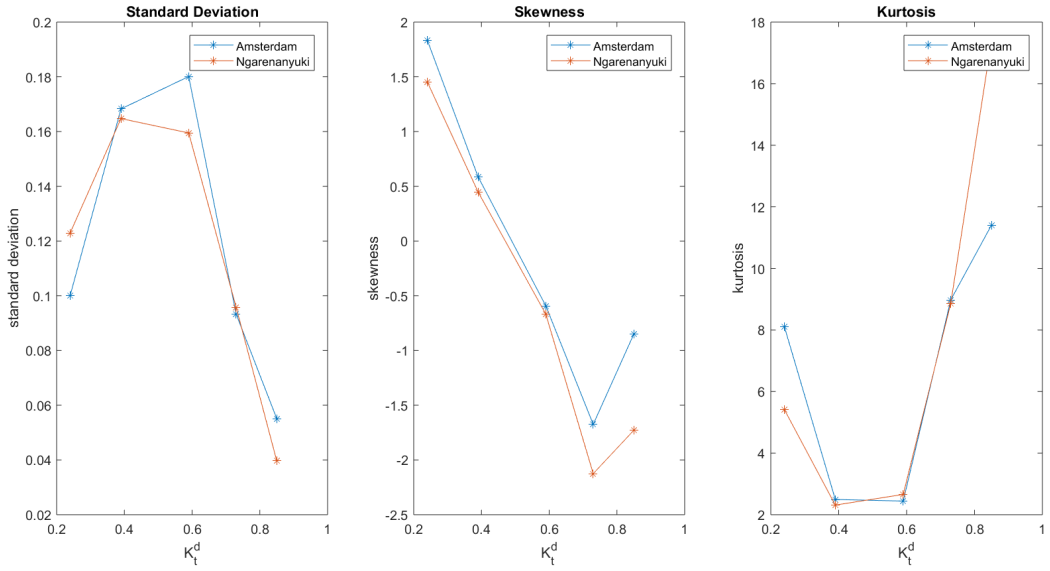
The kurtosis of these pdfs shows a maximum for kth $K_t^d = 0.8$, and low kurtosis for K_t^d between 0.2 and 0.6. The distribution are confirmed by the analysis on the kurtosis values, the kurtosis of the *pdf*s is higher for extreme K_t^d analyzed than for central values. For high valued of daily clearness index, the pdf of the Equatorial Area presents a really high kurtosis values, higher than the Warm Temperature Area. These distributions therefore have very sharp peaks and are representative of very homogeneous clear sky situations for Zone A.

6.5.4 Cumulative distribution function

The synthetically generated time series have been examined to determine their ability to preserve the properties of the observed clearness index time series. A satisfactory accordance has been noted between the observed and the generated data from different perspectives. In order to validate the model, it is important to verify that some statistical characteristics of the synthetic and observed data sets have a good agreement.



(a) Observed data.



(b) Synthetic generated data.

Figure 6.19: Comparison of general statistical parameters for sets of K_t' data.

One of the main characteristics that can be compared is the cumulative distribution of the 10-minutes and daily values. The cumulative distribution functions of the synthetically generated and observed solar irradiation data with 10 minutes resolution are comparable and show similar behaviours, as shown in Figures 6.20a and 6.20b.

Furthermore, despite the procedure is not intended to strictly reproduced the observed data point by point, the trends and the dynamic of the fluctuations should be comparable between observed and generated data. In particular, the RMSE (Root Mean Squared Error) computed on the distribution between the synthetic data and the observed values of 10 minutes solar global irradiation is 4.5% for Amsterdam and 3.2% for Ngarenanyuki.

It is also possible to compare the cumulative distribution functions for the daily solar irradiation values as shown in Figures 6.21a and 6.21b. Analyzing three years of observed and generated data, the RMSE is 6% for Amsterdam and 8.5% for Ngarenanyuki.

6.6 Qualitative Assessment

6.6.1 Yearly comparison

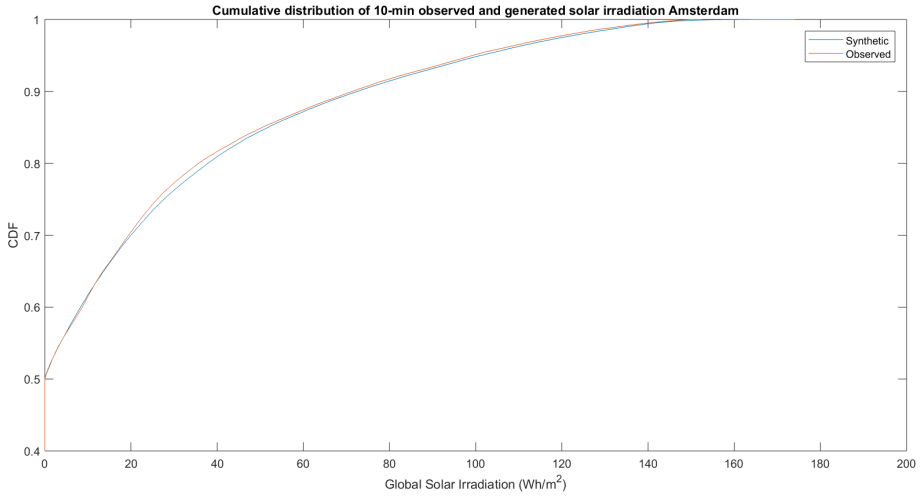
The principal intent of the method of synthetic generation of 10-min irradiation is to generate data that preserve as much as possible some general physical characteristics, such as total energy. The idea is to generate data statistically representative of the observed ones. Indeed, some generated solar energy quantities, such as the monthly average or the daily sum, should be comparable and should remain quite constant when compared to the same quantities from the observed time series.

In Figures 6.22a and 6.22b, three years of yearly 10-min global solar irradiation are reported. The comparison is between the generated and the observed data, from three different databases, HelioClim-3, HelioClim-1 and NASA. HelioClim-3 is the one from all the data have been taken. For Amsterdam the distance in percentage between the observed data from HelioClim-3 and the generated ones for the yearly values averaged on three different years of calculations is of 3% yearly. For Ngarenanyuki, the yearly averaged distance in percentage is 2.4% yearly.

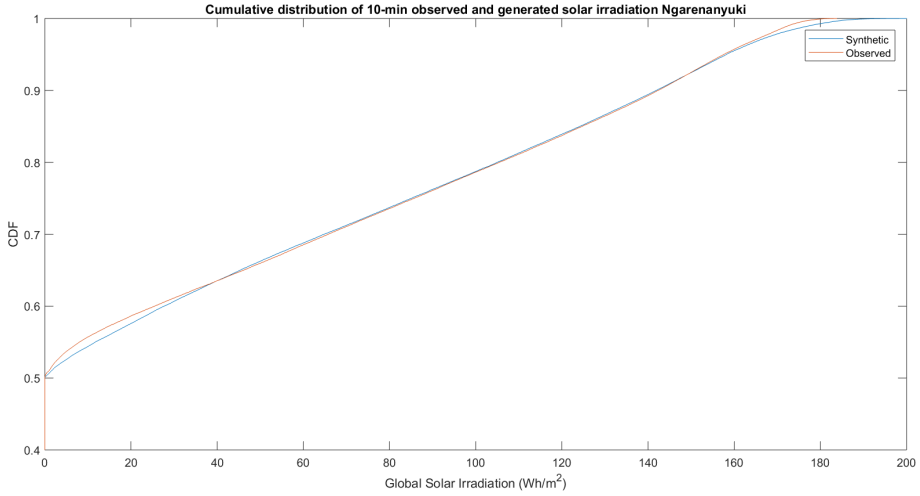
The error between the observed and generated data has been computed taking into account the yearly values with the following formulation (6.2):

$$error = \frac{(H_{y,obs} - H_{y,sg})}{H_{y,obs}} \tag{6.2}$$

Where $H_{y,obs}$ is the amount of yearly energy derived from observed data and $H_{y,sg}$ is the

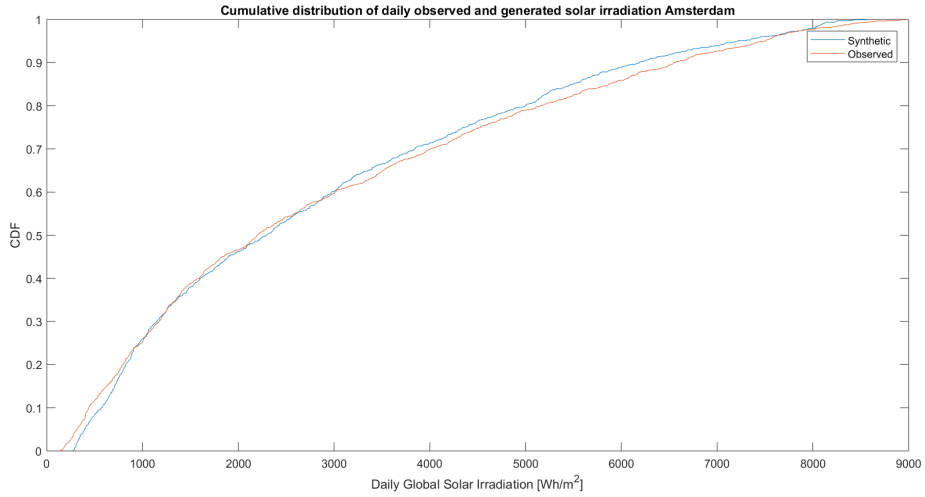


(a) Amsterdam.

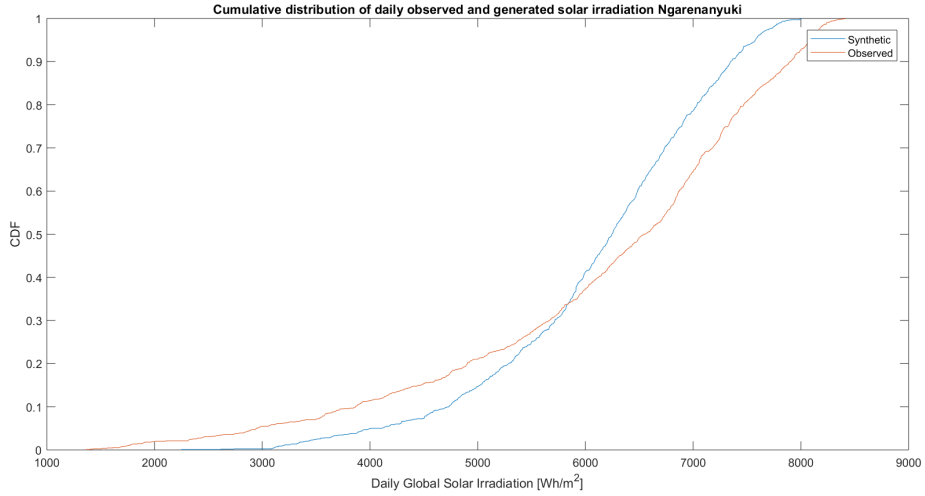


(b) Ngarenanyuki.

Figure 6.20: Comparison of the cumulative distribution functions of observed and generated 10-min solar radiation.

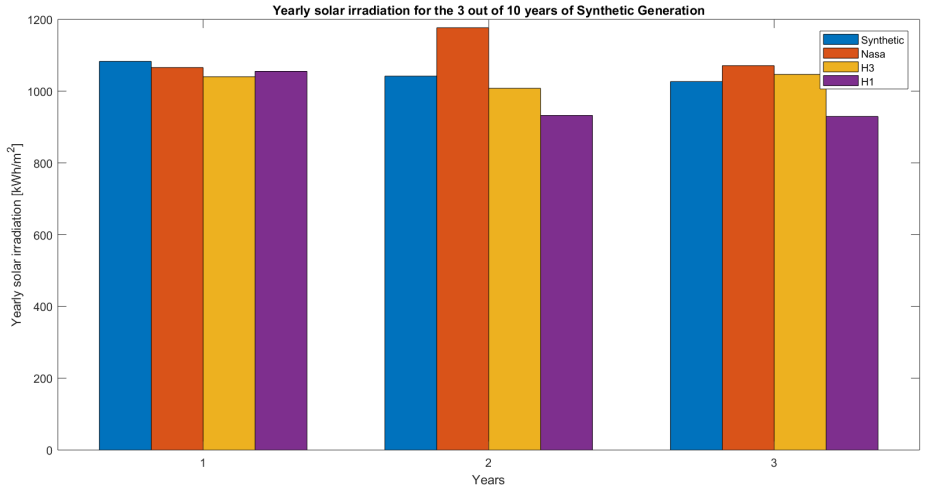


(a) Amsterdam.

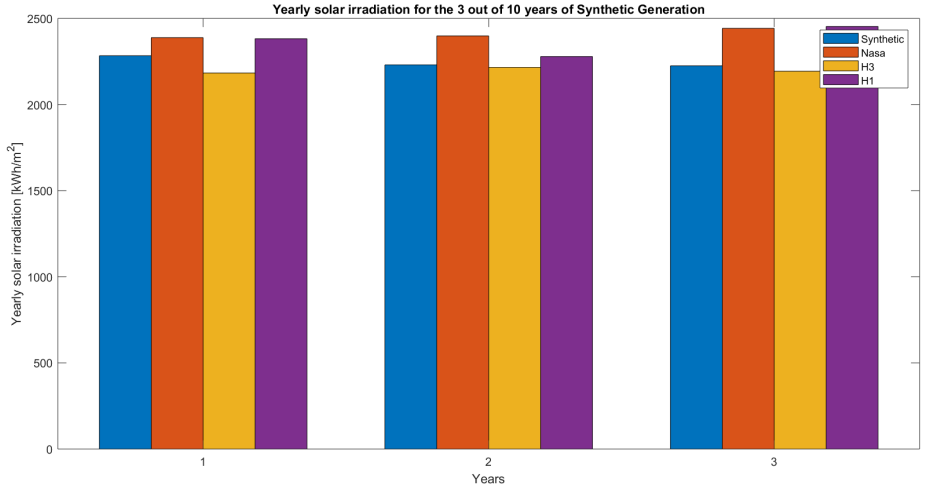


(b) Ngarenanyuki.

Figure 6.21: Comparison of the cumulative distribution functions of observed and generated daily solar radiation.

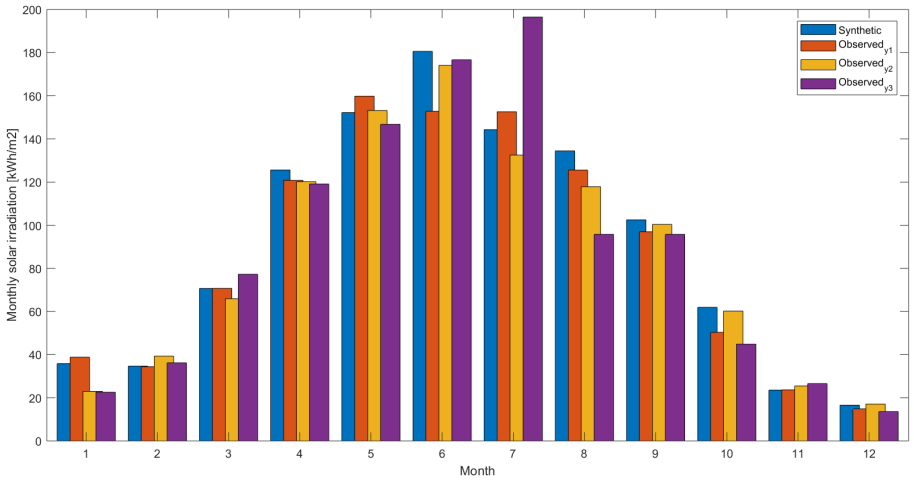


(a) Amsterdam.

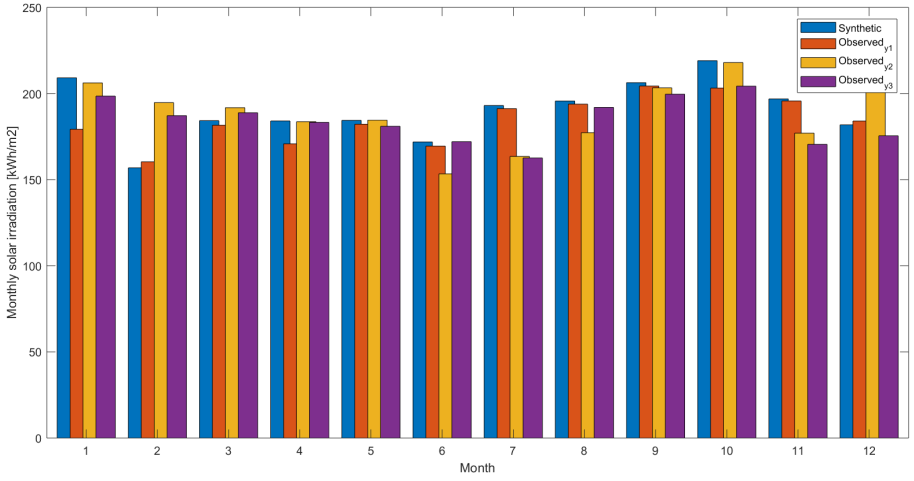


(b) Ngarenanyuki.

Figure 6.22: Comparison among different database and generated yearly data for three different years.



(a) Amsterdam.



(b) Ngarenanyuki.

Figure 6.23: Monthly solar irradiance observed, for three different years, and generated.

yearly energy computed from synthetic generation values.

6.6.2 Monthly comparison

Figures 6.23a and 6.23b show a comparison between the monthly total solar irradiance of synthetic generated data compared with three different observed year. The RMSE for Ngarenanyuki of year one is 8.8%, year two is 5% and year three is 3.7%, for a mean of 5-6%. For Amsterdam it is around 10%.

Figures 6.24a and 6.24b are reported box plot for Amsterdam and for Ngarenanyuki of the observed data in comparison with the average value for each month of several synthetic generated data. For both locations the monthly averages of the generated years falls between

the 25-th and 75-th percentile of the observed data.

6.6.3 Daily comparison

Figure 6.25 shows a comparison between observed and synthetically generated daily solar irradiation values. In order to highlight the seasonal trend, the moving average of the data is shown. The rolling mean has been calculated with a window of thirty days.

It can be seen that the synthetic data follow the seasonal trend and the years are significantly different from each other. The range of observed data is based on three sampled years for both the testing areas.

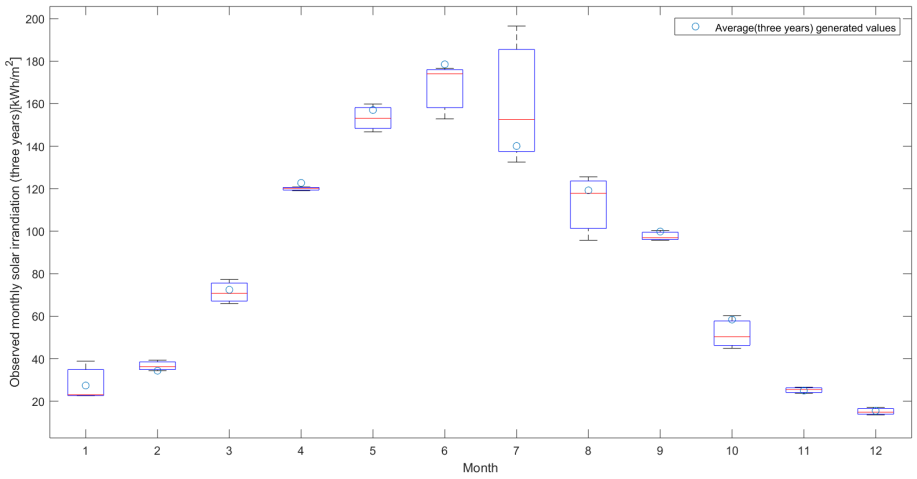
A comparison of the distribution of observed and generated daily global irradiation values is shown in Figure 6.26. The data used to create the plot are three years of observed daily values and ten years of generated daily values for each test location. The y-axis of the plot is the value of the daily global irradiation, while the x-axis shows the frequency of each value. The two shapes show the probability distribution of observed and generated daily values. The dashed lines represent the 25th, 50th and 75th percentiles of each distribution.

It can be seen that there is a slight difference between the observed and generated distributions, especially for Ngarenanyuki. The difference is due to the use of the Bendt's correlation, which is deterministic and is not adapted to the specific area.

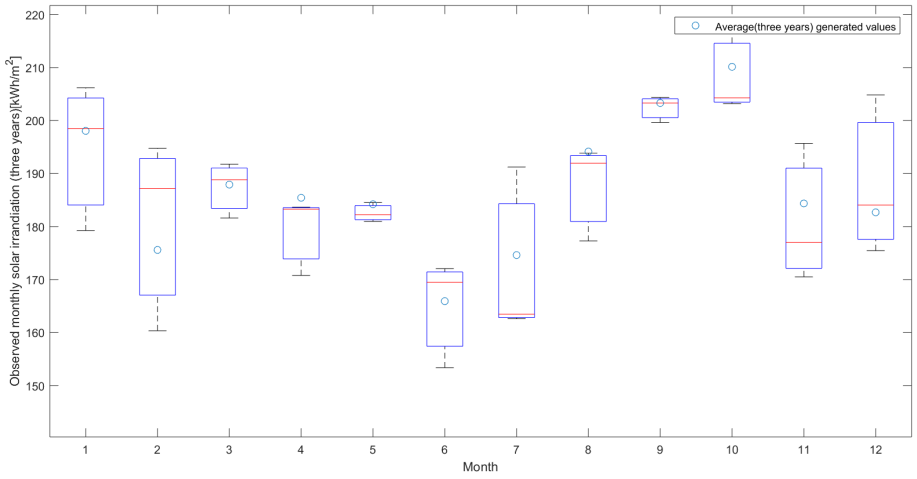
6.6.4 Sub-hourly comparison

Figure 6.27 shows a comparison between the generated and observed distributions of 10-minutes global solar irradiation values. The x-axis of the plot represents the probability distribution of the 10-minutes values, split into generated and observed data. It can be observed that the distributions of observed and generated values are very similar, and in particular that the difference between the daily distributions shown in Figure 6.26 are greatly reduced when passing to 10-minutes values. This shows that the stochastic component of the ARIMA and Markov increase the reliability of the results, in addition to leading to independent profiles each time they are used. It can be noticed that the distributions for the two locations are significantly different, with the median value for Ngarenanyuki being larger than the median in Amsterdam.

Finally, a breakdown by month of the 10-minutes values distributions is reported in Figure 6.28. Although the procedure is aimed at the generation rather than forecasting of solar profiles, meaning some differences between specific months are expected, the distributions

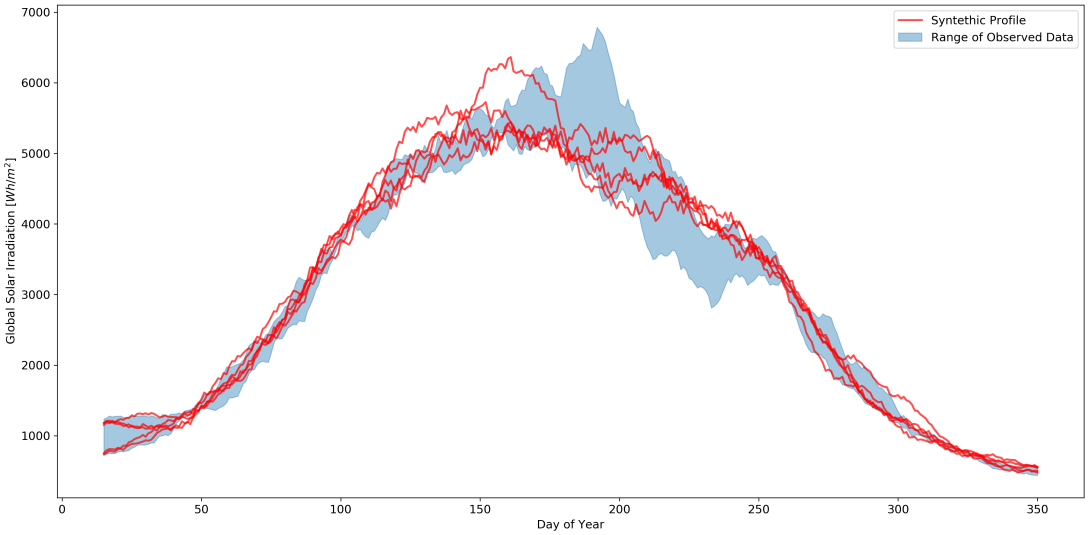


(a) Amsterdam.

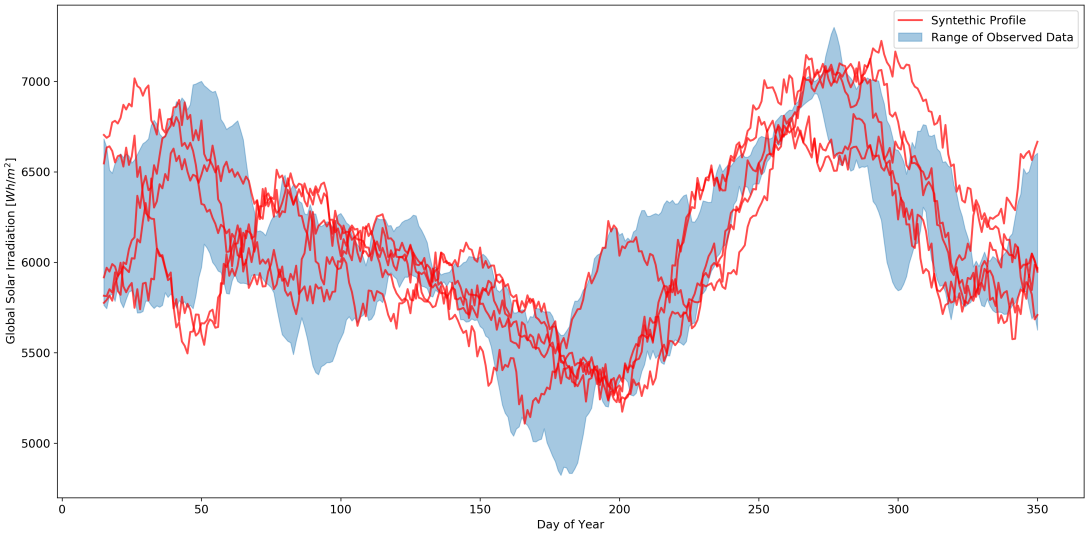


(b) Ngarenanyuki.

Figure 6.24: Box plot of monthly solar irradiance observed, for three different years, and average monthly generated values.

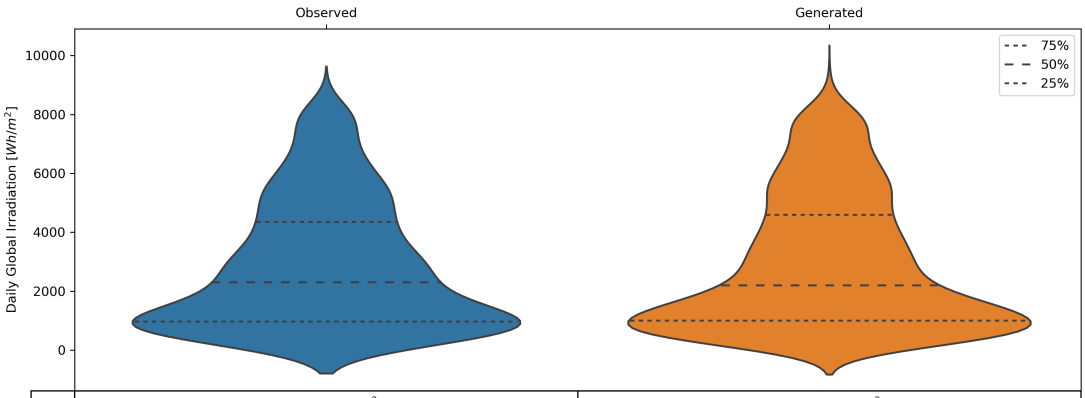


(a) Amsterdam.



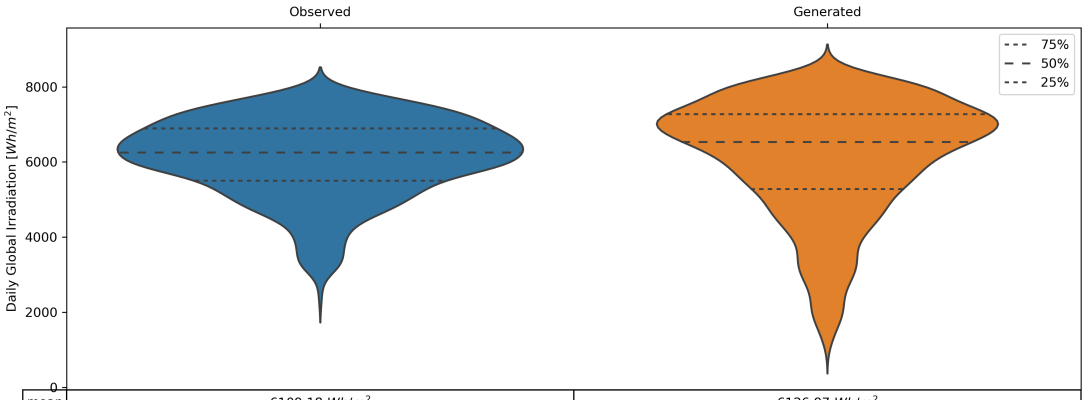
(b) Ngarenanyuki.

Figure 6.25: Comparison of observed and synthetic daily Irradiation.



	Observed	Generated
mean	2851.59 Wh/m^2	2913.37 Wh/m^2
std	2171.15 Wh/m^2	2271.94 Wh/m^2
25%	965.05 Wh/m^2	998.53 Wh/m^2
50%	2299.70 Wh/m^2	2195.00 Wh/m^2
75%	4354.04 Wh/m^2	4590.68 Wh/m^2

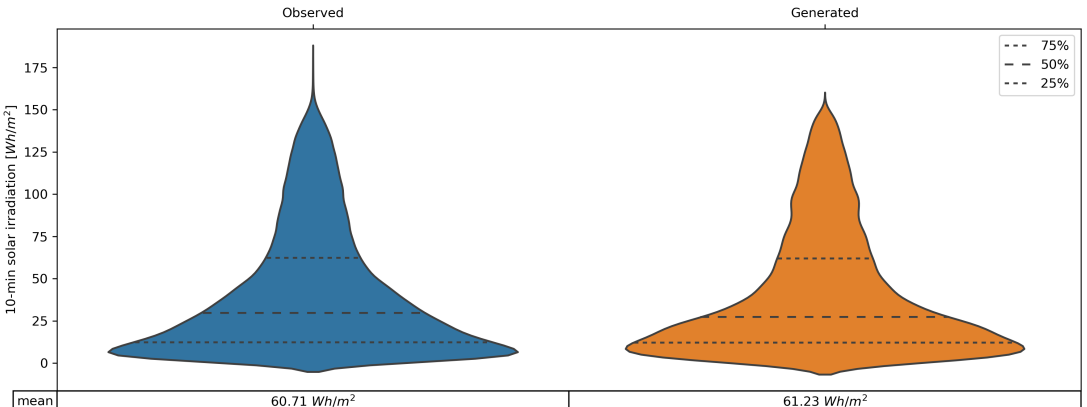
(a) Amsterdam.



	Observed	Generated
mean	6109.18 Wh/m^2	6126.97 Wh/m^2
std	1046.03 Wh/m^2	1556.66 Wh/m^2
25%	5499.55 Wh/m^2	5282.77 Wh/m^2
50%	6246.40 Wh/m^2	6531.35 Wh/m^2
75%	6892.35 Wh/m^2	7274.62 Wh/m^2

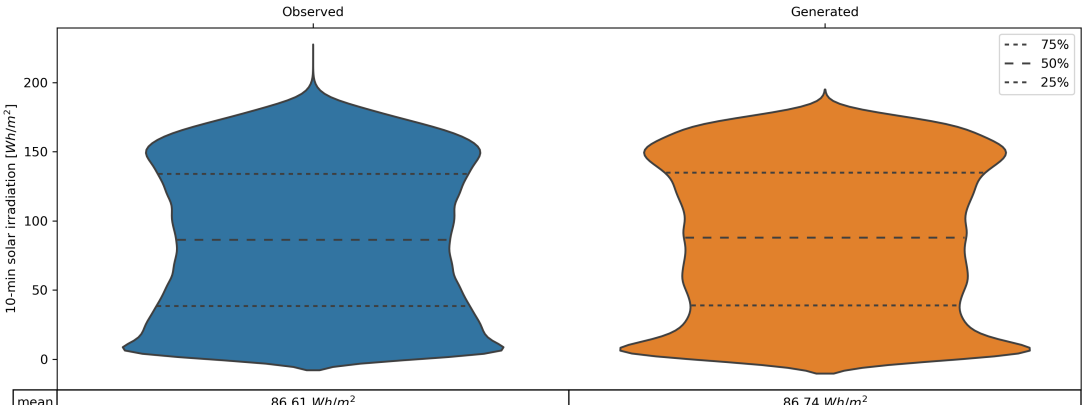
(b) Ngarenanyuki.

Figure 6.26: Comparison of observed and synthetic daily Irradiation distributions.



	Observed	Generated
mean	60.71 Wh/m^2	61.23 Wh/m^2
std	34.18 Wh/m^2	34.36 Wh/m^2
25%	31.12 Wh/m^2	32.43 Wh/m^2
50%	51.23 Wh/m^2	50.79 Wh/m^2
75%	85.00 Wh/m^2	85.01 Wh/m^2

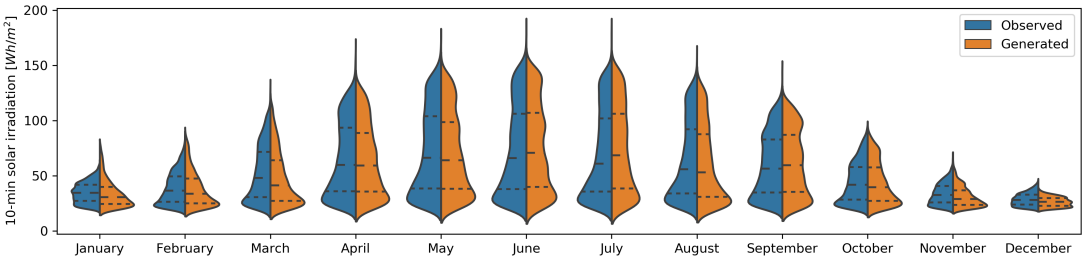
(a) Amsterdam.



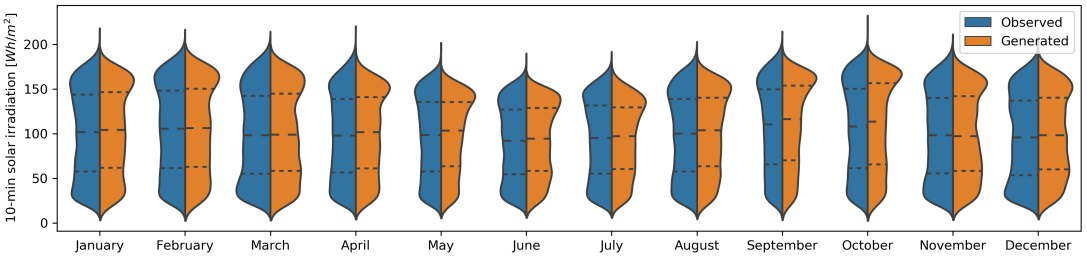
	Observed	Generated
mean	86.61 Wh/m^2	86.74 Wh/m^2
std	53.38 Wh/m^2	53.37 Wh/m^2
25%	38.99 Wh/m^2	38.52 Wh/m^2
50%	87.81 Wh/m^2	86.38 Wh/m^2
75%	134.85 Wh/m^2	133.98 Wh/m^2

(b) Ngarenanyuki.

Figure 6.27: Comparison of observed and synthetic 10-min Irradiation distributions.



(a) Amsterdam.



(b) Ngarenanyuki.

Figure 6.28: Comparison of observed and synthetic 10-min Irradiation distributions by month.

are very well reproduced and follow the same behaviour for both the observed and generated data. It is easy to see the stark difference in the seasonal trends for the two locations.

6.7 Final Validation

For the final validation of the model it is useful to compare the results obtained from the tool with the measurements obtained from the real system in Ngarenanyuki.

Out of one year, 118 days are missing in the production measured data. Therefore, it is not possible to draw a detailed comparison of the yearly production. A more specific comparison is presented for monthly, daily and sub-hourly values. This lack of data is a significant example of the challenges faced in collecting solar measurements in rural areas of developing countries, even with an existing system in place. In addition to the missing data, the measurements are affected by a number of factors that are not considered in the model. These factors include curtailment, wind, dust and the difficulty in regular maintenance operations in these areas.

In Figure 6.29 it is possible to see the daily production along the year of a PV panel of 250 Watt installed for the secondary school in Ngarenanyuki, reported in kWh/kWp.

Based on the measurements, the yearly average production is approximately estimated to be 335 kWh for the third year of life of the PV panel. In Figure 6.7b, that represents the PV

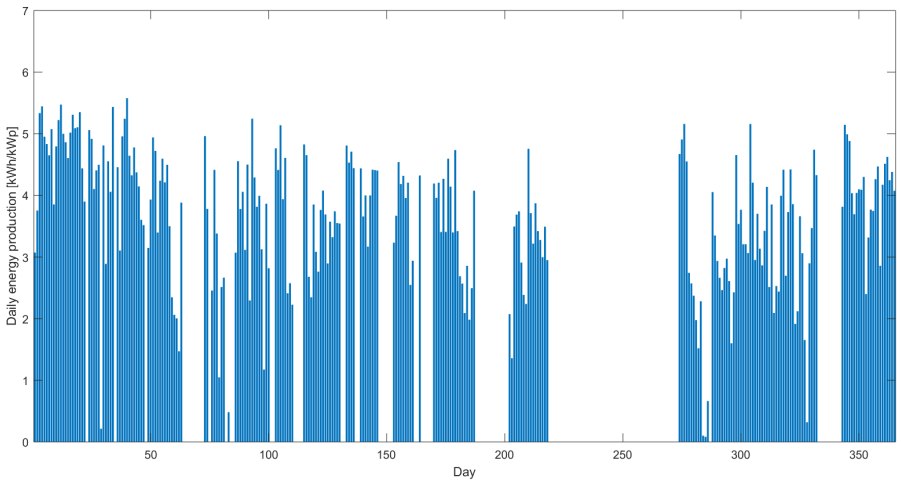


Figure 6.29: Daily energy production of the PV system installed in Ngarenanyuki.

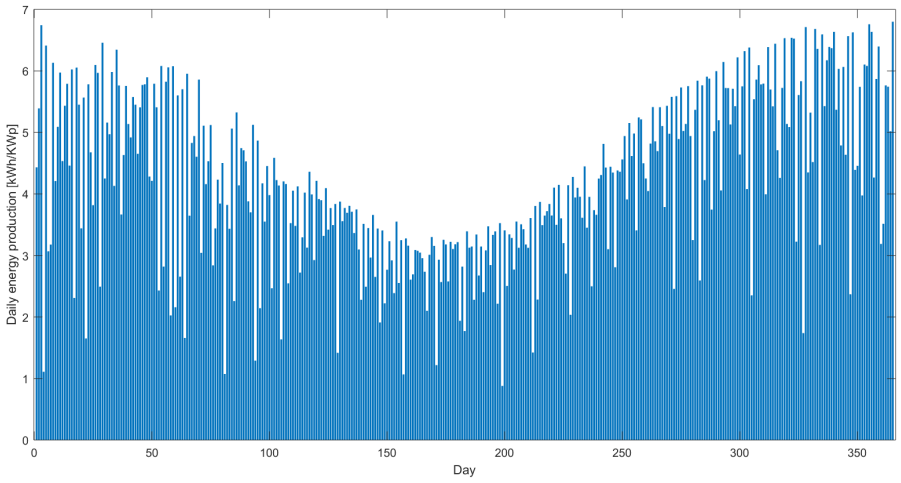


Figure 6.30: Daily energy production of the PV system with the data generated by the procedure.

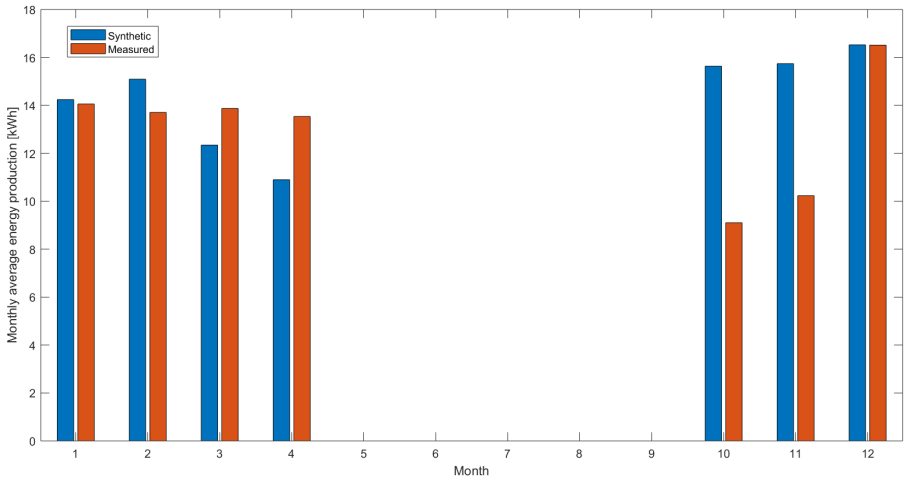


Figure 6.31: Monthly average energy produced by the PV panel with synthetic generated data (blue) and monthly average energy produced by the PV panel installed (red) in Ngarenanyuki.

production from synthetically generated input, the third year of life has a power production of 375 kWh. Taking into consideration that the PV production in Ngarenanyuki is curtailed when the batteries are full, the two yearly values are quite similar. Furthermore, the system is located in a rural area in a developing country and most of the time the local conditions make difficult the maintenance of the system. It should also be taken into account that the measured data have been manipulated to obtain the yearly value, leading to a further approximation.

Three different analyses have been developed on the available data:

- comparison between monthly average energy produced;
- comparison between the cumulative distributions of 10-minutes energy produced;
- comparison between 10-minutes resolution daily profiles.

In Figure 6.31 is shown the comparison between the measured and synthetically generated values of the monthly average energy produced. The data are only shown for months with at least 25 days of measured values. The production is calculated for a 3 kW system, which is installed in Ngarenanyuki. It is possible to observe a general correspondence in the monthly values. Two months are particularly low in the measured data, this could be due to several local conditions.

Figure 6.32 represents the cumulative distribution function of the values with a resolution of 10-min for real data and results obtained from the synthetically generated data for a 250

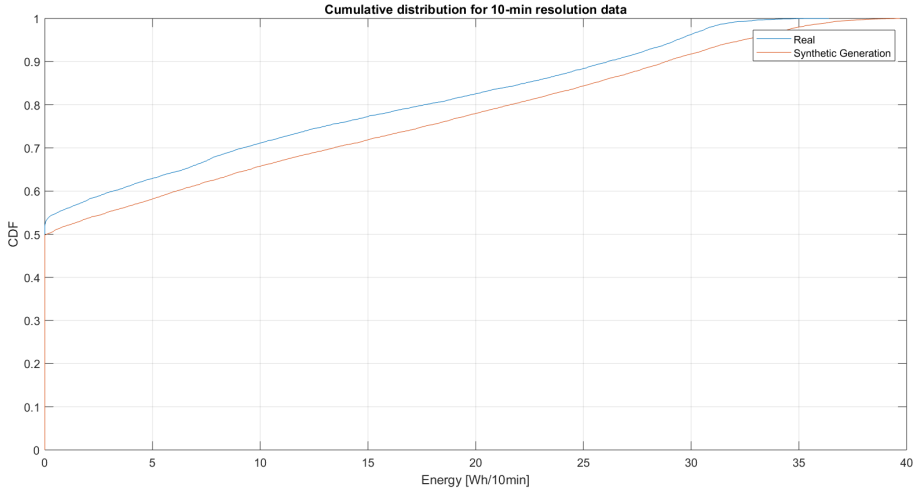


Figure 6.32: Cumulative distribution function 10-min resolution for a 250 Wp PV panel.

Wp PV panel. The cumulative distributions are built based on the first three months of the year, due to the significant gaps in measured data from the other months. The curves are shifted with the real CDF being lower than the synthetically generated CDF. This means that the real values are lower, and could be due to the curtailment of PV production when the batteries are full or other reasons mentioned above. Despite this shift, the shape of the distribution is extremely well reproduced, meaning the synthetically generated values closely represent the real data.

In Figure 6.33 are reported some examples of daily PV production profiles with data resolution of 10-min. In the left plot are reported some examples for the real PV production data from the Ngarenanyuki system, on the right are reported data derived from the synthetically generated input.

Although the data available for the PV system in Ngarenanyuki are more than the data available in most rural areas in developing countries, they still present significant amounts of missing data. Therefore only partial comparisons are possible. This is a very good example of the issues faced when sizing distributed energy systems in rural areas of developing countries, where it is more challenging and expensive to ship, install, maintain and monitor measurement systems.

It should also be considered that the actual measurements collected in Ngarenanyuki are significantly affected by several factors that are not included in the model, such as wind and dust. This notwithstanding, the synthetically generated profiles represent well the observed data. In particular, the distribution of 10-minutes values is very well approximated. In

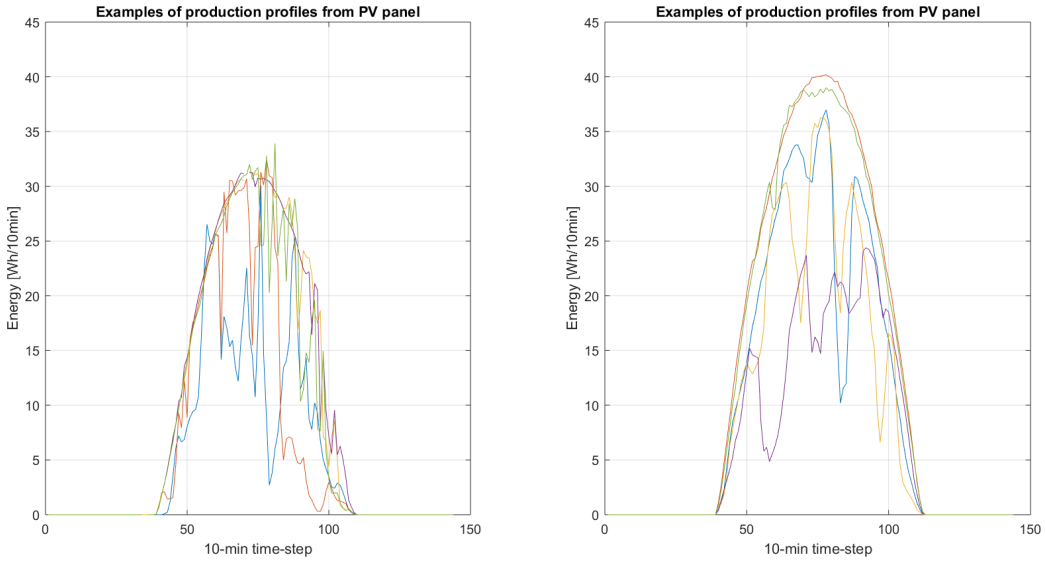


Figure 6.33: Example of PV production. Left: Real data. Right: Generated data.

addition to this, the behaviour of the fluctuations of 10-minutes values within each day is well reproduced. As stated in section 2.4, the distribution and fluctuations of sub-hourly values are the most important factors in the synthetic generation of solar radiation profiles for distributed energy systems.

Chapter 7

Conclusion

7.1 Conclusions

The increasing development of Distributed Energy Systems and the deployment of PV in these systems requires a specific attention. Especially in developing countries, this technology is becoming more and more important not only due to the abundant resources many of these countries have access to, but also due to its modularity and its independency on National electric grids. This is particularly relevant in rural areas, where grid expansion costs are very high.

The variability of solar irradiation and the impossibility to dispatch PV power production make it challenging to dimension distributed energy systems. The extensive simulations required to model such systems lead to the need of large amounts of input data. This requirement is especially critical in rural areas of developing countries, where it is more difficult to deploy measurement systems and investment costs are often a major barrier.

A possible solution to mitigate this problem is using tools that, through stochastic models, can compensate this gap enabling a more robust approach to the dimensioning of distributed energy systems. Several tools exist that already optimize and dimension distributed systems. The goal of this thesis is to elaborate a methodology that requires in input a low amount of low-resolution data, but can generate multiple years of high resolution solar profiles.

The proposed methodology is capable of synthetically generating several years of sub-hourly data based on monthly average input data. It is a general procedure that can be applied to any location. In its most data-demanding steps, the proposed methodology provides the flexibility to use data from other locations if local measurements are not available. This enables detailed simulations of distributed energy systems in areas where solar irradiation

data are missing, while giving to the user the freedom to balance accuracy and requirements in terms of input data.

The methodology is based on several statistical models and some deterministic correlations, and uses the clearness index as climatic variable to model solar radiation. The clearness index values are turned from monthly to a distribution of daily values using Bendt's correlation. Subsequently, the sequence of the days within each month is determined through an ARIMA(1,1,1) model. Finally, the clearness index values with sub-hourly resolution are generated using a second order Markov model.

The ARIMA model calibration has been generalized, allowing the use of data from other locations. In addition to this, it has been shown that the Markov Matrix building process can be based on measurements obtained in other locations within the same Climatic Area.

The methodology has been implemented in MATLAB and designed to be integrated with PoliNRG, but can be used to provide the input to any energy systems simulation software.

The models used have been validated step by step and calibrated with the observed data of two specific locations, Amsterdam and Ngarenanyuki (Tanzania). The tool is not aimed at forecasting exactly solar radiation data, but the final objective is to generate realistic solar profiles that preserve the general characteristics and behaviour of solar radiation for a specific area. The performance of the model and its reliability have been tested for several years and the results are satisfactory.

7.2 Limitations of the methodology

The proposed methodology presents some limitations and approximations that reduce the accuracy of the results.

The main limitation of the model is that the resolution of the generated data depends on the resolution of the input data, due to the need to create the Markov Transition Matrices. To compensate this strict correlation, a generalized calibration procedure based on data from other locations within the same climatic area has been proposed and validated.

In general, the quality of the results is affected by the amount and resolution of the input data. In particular, the availability of a higher number of years of input data would increase the accuracy of the generated data.

The second weakness is due to the use of the Bendt's correlation to generate the distribution of daily values from the monthly average values. This deterministic correlation does not take into account any stochastic characteristic of the data. Although this deterministic step

is followed by several stochastic transformations (ARIMA and Markov model) that reduce its impact on the final results, the overall methodology could be improved by replacing this step with a stochastic procedure that always generates individual results.

Furthermore, this first part of the procedure could be improved by determining a specific correlation for each climatic area as it has been done for the MTMs.

Another aspect that could be improved is the trade-off between the determination of the Markov classes and the amount of data needed to build the Matrices. There is not any available bibliography reference that describes exactly how to determine the number and range of the Markov classes. This phase of the methodology could be improved and could be linked to the climatic areas, creating different ranges and classes for each area.

For what concerns the PV panel simulation, two limitations are present in the tool. The first limitation depends on the fact that the split between diffuse and direct solar irradiation is performed with a correlation demonstrated for a hourly time step. As discussed in 2.6.1, other existing formulations are either very data-intensive or have not been validated for higher resolutions.

The second limitation in the PV panel simulation concerns the atmospheric temperature. This data is used to compute the actual cell temperature to determine the power temperature losses. To implement perfectly also this aspect, this temperature should be synthetically generated for all the years of the simulation. To further improve this aspect, the correlation between ambient temperature and solar irradiation should be considered.

7.3 Future Developments

Synthetic generation procedures could be adopted also to generate other meteorological variables, such as wind speed or atmospheric temperature.

A future development could take into account the inclusion of the synthetic generation of wind speed profiles. This would allow the modelling and optimization of combined wind and solar energy systems.

Bibliography

[1] Köppen-Geiger. *World map of the Köppen-Geiger climate classification update*. http://koeppen-geiger.vu-wien.ac.at/present.htm. Accessed: 2017-03-11.

[2] National Renewable Energy Lab NREL. *National Renewable Energy Lab*. https://www.nrel.gov/research/publications.html. Accessed: 2017-03-11.

[3] International Renewable Energy Agency IRENA. *Renewable Cost Database*. http://www.irena.org/publications/2018/Jan/Renewable-power-generation-costs-in-2017. Accessed: 2017-03-11.

[4] Solar Radiation data SoDa. *Solar Radiation data*. http://www.soda-pro.com/. Accessed: 2017-03-11.

[5] J. M. Santos, J. M. Pinazo, and J. Canada. "Methodology for generating daily clearness index index values Kt starting from the monthly average daily value Kt. Determining the daily sequence using stochastic models". In: *Renewable Energy* 28.10 (2003), pp. 1523–1544. ISSN: 09601481. DOI: 10.1016/S0960-1481(02)00217-3.

[6] B. O. Ngoko, H. Sugihara, and T. Funaki. "Synthetic generation of high temporal resolution solar radiation data using Markov models". In: *Solar Energy* 103 (2014), pp. 160–170. ISSN: 0038092X. DOI: 10.1016/j.solener.2014.02.026. URL: http://dx.doi.org/10.1016/j.solener.2014.02.026.

[7] M. Jurado, J.M. Caridad, and V. Ruiz. "Statistical distribution of the clearness index with radiation data integrated over five minute intervals." In: (1995).

[8] Carlos M. Fernández-Peruchena and Ana Bernardos. "A comparison of one-minute probability density distributions of global horizontal solar irradiance conditioned to the optical air mass and hourly averages in different climate zones". In: *Solar Energy* 112.February 2015 (2015), pp. 425–436. ISSN: 0038092X. DOI: 10.1016/j.solener.2014.11.030.

[9] D. G. Erbs, S A Klein, and J. A. Duffie. "Estimation of the diffuse radiation fraction for hourly, daily and monthly-average global radiation". In: *Solar Energy* 28.4 (1982), pp. 293–302. ISSN: 0038092X. DOI: 10.1016/0038-092X(82)90302-4.

[10] Annette Hammer et al. "Solar energy assessment using remote sensing technologies". In: *Remote Sensing of Environment* 86.3 (2003), pp. 423–432. ISSN: 00344257. DOI: 10.1016/S0034-4257(03)00083-X.

[11] REN21. *Renewables 2017 global status report 2017*. 2017. ISBN: 9783981810769.

[12] SDGs. *SDGs*. https://sustainabledevelopment.un.org. Accessed: 2017-03-11.

[13] Benjamin Pillot, Marc Muselli, and Philippe Poggi. "On the impact of the global energy policy framework on the development and sustainability of renewable power systems in Sub-Saharan Africa: the case of solar PV". In: (2017), pp. 1–25. arXiv: arXiv:1704.01480v1.

[14] SDG7. *SDG7*. https://sustainabledevelopment.un.org/sdg7. Accessed: 2017-03-11.

[15] Khalid Malik. *Human Development Report 2014, Reducing Vulnerabilities and Building Resilience*. 2014. ISBN: 9789211263688.

[16] I.E.A. "CO2emissions from fuel combustion – highlights 2014." In: (2014).

[17] SE4ALL. *Sustainable Energy for All*. https://www.seforall.org/about-us. Accessed: 2017-03-11.

[18] M. Moner-Girona, S. SzabóS, and S. Rolland. "Finance Mechanisms and Incentives for Photovoltaic Technologies in Developing Countries". In: (2012).

[19] A. Chaurey and T. C. Kandpal. "Assessment and evaluation of PV based decentralized rural electrification: An overview." In: (2010).

[20] Chukwuma Leonard et al. "Electricity for development : Mini-grid solution for rural electrification in South Africa". In: *ENERGY CONVERSION AND MANAGEMENT* 110 (2016), pp. 268–277. ISSN: 0196-8904. DOI: 10.1016/j.enconman.2015.12.015. URL: http://dx.doi.org/10.1016/j.enconman.2015.12.015.

[21] Department of Minerals DOE and Energy Pretoria. "Mini-grig viability and replication potential: The Hluleka and Lucingweni pilot projects". In: (2008).

[22] Douglas Banks et al. "Integrated Rural Energy Utilities A review of literature and opportunities for the Establishment of an IREU". In: (2008).

[23] A. Cherni et al. "Energy supply for sustainable rural livelihoods. A multi-criteria decision-support system." In: (2007).

[24] Juan C Rojas-zerpa and Jose M Yusta. "Energy for Sustainable Development Methodologies , technologies and applications for electric supply planning in rural remote areas". In: *Energy for Sustainable Development* 20 (2014), pp. 66–76. ISSN: 0973-0826. DOI: 10.1016/j.esd.2014.03.003. URL: http://dx.doi.org/10.1016/j.esd.2014.03.003.

[25] Subhes C Bhattacharyya and Debajit Palit. "Mini-grid based off-grid electri fi cation to enhance electricity access in developing countries : What policies may be required ?" In: *Energy Policy* 94 (2016), pp. 166–178. ISSN: 0301-4215. DOI: 10.1016/j.enpol.2016.04.010. URL: http://dx.doi.org/10.1016/j.enpol.2016.04.010.

[26] Stefano Mandelli et al. "Effect of load profile uncertainty on the optimum sizing of off-grid PV systems for rural electrification systems for rural electrification". In: *Sustainable Energy Technologies and Assessments* 18.October 2016 (2017), pp. 34–47. ISSN: 2213-1388. DOI: 10.1016/j.seta.2016.09.010. URL: http://dx.doi.org/10.1016/j.seta.2016.09.010.

[27] TRNSYS. *Transient System Simula-tion Tool Thermal Energy System Specialists LLC*. http://trnsys.com/. Accessed: 2017-03-11.

[28] Claudio Brivio et al. “A novel software package for the robust design of off-grid power systems”. In: *Journal of Cleaner Production* 166.August (2017), pp. 668–679. ISSN: 0959-6526. DOI: 10.1016/j.jclepro.2017.08.069. URL: http://dx.doi.org/10.1016/j.jclepro.2017.08.069.

[29] Openmod. *Openmod*. http://www.openmod-initiative.org/manifesto.html. Accessed: 2017-03-11.

[30] Stefano Mandelli, Marco Merlo, and Emanuela Colombo. “Energy for Sustainable Development Novel procedure to formulate load pro fi les for off-grid rural areas”. In: *Energy for Sustainable Development* 31 (2016), pp. 130–142. ISSN: 0973-0826. DOI: 10.1016/j.esd.2016.01.005. URL: http://dx.doi.org/10.1016/j.esd.2016.01.005.

[31] R Aguiar and M. Collares-Pereira. “Statistical properties of hourly global radiation”. In: *Solar Energy* 48.3 (1992), pp. 157–167. ISSN: 0038092X. DOI: 10.1016/0038-092X(92)90134-V.

[32] J. Boland. “Time-series analysis of climatic variables”. In: *Solar Energy* 55.5 (1995), pp. 377–388. ISSN: 0038092X. DOI: Doi10.1016/0038-092x(95)00059-Z. URL: http://www.sciencedirect.com/science/article/pii/0038092X9500059Z.

[33] Viorel Badescu. *Modeling Solar Radiation at the Earth ' s Surface*. 2008, p. 536. ISBN: 978-3-540-77454-9. DOI: 10.1007/978-3-540-77455-6. URL: http://link.springer.com/10.1007/978-3-540-77455-6.

[34] Jamal Hassan. “ARIMA and regression models for prediction of daily and monthly clearness index”. In: *Renewable Energy* 68 (2014), pp. 421–427. ISSN: 0960-1481. DOI: 10.1016/j.renene.2014.02.016. URL: http://dx.doi.org/10.1016/j.renene.2014.02.016.

[35] HOMER Energy Radiation Incident on th PV array. *Radiation Incident on th PV array*. https://www.homerenergy.com/support/docs/3.11/how_homer_calculates_the_radiation_incident_on_the_pv_array.html. Accessed: 2017-03-11.

[36] V. A. Graham and K. G. T. Hollands. “A method to generate synthetic hourly solar radiation globally”. In: *Solar Energy* 44.6 (1990), pp. 333–341. ISSN: 0038092X. DOI: 10.1016/0038-092X(90)90137-2.

[37] V. A. Graham, K. G. T. Hollands, and T. E. Unny. “A time series model for Kt with application to global synthetic weather generation”. In: *Solar Energy* 40.2 (1988), pp. 83–92. ISSN: 0038092X. DOI: 10.1016/0038-092X(88)90075-8.

[38] J. Polo et al. “A simple approach to the synthetic generation of solar irradiance time series with high temporal resolution”. In: *Solar Energy* 85.5 (2011), pp. 1164–1170. ISSN: 0038092X. DOI: 10.1016/j.solener.2011.03.011. URL: http://dx.doi.org/10.1016/j.solener.2011.03.011.

[39] J. M. Bright et al. “Stochastic generation of synthetic minutely irradiance time series derived from mean hourly weather observation data”. In: *Solar Energy* 115 (2015), pp. 229–242. ISSN: 0038092X. DOI: 10.1016/j.solener.2015.02.032. URL: http://dx.doi.org/10.1016/j.solener.2015.02.032.

[40] B.Y. Liu and R.C. Jordan. “The interrelationship and characteristic distribution of direct, diffuse and total solar radiation.” In: (1960).

[41] P. Bendt, M. Collares-Pereira, and A. Rabl. "The frequency distribution of daily insolation values." In: (1981).

[42] K.G. Hollands and R.G. Huget. "A probability density function for the clearness index, with applications." In: (1983).

[43] W. E. Knowles Middleton. "Bouguer, Lambert, and the Theory of Horizontal Visibility." In: (1960).

[44] S.H. Abdulla, S.A. Klein, and W.A. Beckman. "A new correlation for prediction of the frequency distribution of daily solar radiation." In: (2000).

[45] G. C. Tiao. "Time Series : ARIMA Methods". In: (1994).

[46] Statistic How To. *Unit Root*. http://www.statisticshowto.com/unit-root/. Accessed: 2017-03-11.

[47] Assess Stationarity of a Time Series MATLAB. *MATLAB*. https://it.mathworks.com/help/econ/assess-stationarity-of-a-time-series.html. Accessed: 2017-03-11.

[48] Radiation Incident on th PV array HOMER Energy. *HOMER*. https://www.homerenergy.com/support/docs/3.11/how_homer_calculates_the_radiation_incident_on_the_pv_array.html. Accessed: 2017-03-11.

[49] J.A. Duffie and W.A. Beckman. "Solar Engineering of Thermal Processes 2nd edition." In: (1991).

[50] PV Efficiency at Standard Test Conditions HOMER Energy. *HOMER*. https://www.homerenergy.com/support/docs/3.11/pv_efficiency_at_standard_test_conditions.html. Accessed: 2017-03-11.

[51] HOMER Energy. *HOMER Energy Published Solar Data*. https://www.homerenergy.com/support/docs/3.11/published_solar_data.html. Accessed: 2017-03-11.

[52] Gordon Reikard. "Predicting solar radiation at high resolutions : A comparison of time series forecasts". In: *Solar Energy* 83.3 (2009), pp. 342–349. ISSN: 0038-092X. DOI: 10.1016/j.solener.2008.08.007. URL: http://dx.doi.org/10.1016/j.solener.2008.08.007.

[53] Ilhami Colak et al. "Multi-period Prediction of Solar Radiation Using ARMA and ARIMA Models." In: (2016).

[54] A Shamshad et al. "First and second order Markov chain models for synthetic generation of wind speed time series". In: 30 (2005), pp. 693–708. DOI: 10.1016/j.energy.2004.05.026.

[55] R. Perez et al. "Making full use of the clearness index for parameterizing hourly insolation conditions." In: (1990).

[56] C. Chatfield. *The analysis of time series*. 1991.

[57] G.M. Box Jenkins. *Time series analysis: forecasting and control*. 1988.

[58] Climatic Research Unit CRU. *Climatic Research Unit*. http://http://www.cru.uea.ac.uk/. Accessed: 2017-03-11.

[59] Global Precipitation Climatology Center GPCC. *Global Precipitation Climatology Center*. https://www.dwd.de/EN/ourservices/gpcc/gpcc.html. Accessed: 2017-03-11.