Journal of Atmospheric and Solar-Terrestrial Physics 107 (2014) 42–47

Contents lists available at ScienceDirect

Journal of Atmospheric and Solar-Terrestrial Physics

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

Markov processes and Zipf's law in daily solar irradiation at earth's surface

J.M. Vindel*, J. Polo

Renewable Energy Division (CIEMAT), Avenida Complutense 40, 28040 Madrid, Spain

ARTICLE INFO

Article history:
Received 10 April 2013
Received in revised form
16 October 2013
Accepted 28 October 2013
Available online 11 November 2013

Keywords:
Markov process
Zifp's law
Daily solar irradiation

ABSTRACT

Sequences of two consecutive days of solar irradiation (global horizontal and direct normal) have been studied here by different approaches. The frequency vs. rank relationships have been analyzed as an attempt to explore whether the Zifp's law is fulfilled, yielding to a partial fulfillment and observing that a good logarithmic fit can be applied to the data in the whole range. In addition, the pdfs of increments in two consecutive daily irradiation values are also studied, showing a relationship between persistence and the coefficients of the logarithmic fit. Finally, it has been shown that a Markov process can represent properly sequences of two consecutive daily irradiation values, for both global horizontal and direct normal components. Thus, synthetic series can be generated by Markov chains for characterizing daily global and direct irradiation.

1. Introduction

According to Zipf scaling law (Zipf, 1949), whose origin is in the Pareto distribution (Pareto, 1986), the frequency of occurrence of a word behaves as a power-law with its rank. The rank to each word is assigned in a decreasing order of its frequency, so the most frequent word has a rank of 1, the second most frequent has a rank of 2, and so on. More precisely, if words are ordered according to their frequencies, then frequency of a word with a rank of r fulfils the following law:

$$f(r) \sim 1/r^p \tag{1}$$

where p is the Zipf exponent which characterizes the structure of the language. Thus, this law can be used for characterizing the distribution of words within sequences. Even though this methodology was first studied in natural languages, it has been later applied in many other domains, including geography (Simon, 1955; Krugman, 1996; Gabaix, 1999), biology (Mantegna et al., 1994; Suzuki et al., 2005), Internet traffic (Adamic and Huberman, 2000), climatology (Primo et al., 2007). This methodology of analysis has been used here to explore the relationship between frequency and rank in sequences of daily solar irradiation and to study whether these sequences can be generated from a Markov stochastic process. Some studies of Markov processes applied to global irradiation have been conducted in the past using a different approach (Amato et al., 1986; Aguiar et al., 1988; Mellit et al., 2005). The Markov process analysis of solar radiation is extended here to the direct normal component of solar radiation. The possibility of characterizing daily solar irradiation sequences by using Markov chains allows to generate synthetic time series that can be used to fill gaps in long term sequences of measurements of both global horizontal and direct normal daily irradiances. In addition, another interesting potential application of this methodology is the application to short term forecasting of daily irradiation. The goodness of the synthesized series generated assuming a Markov process will also be checked using a Kolmogorov–Smirnov test (Chakravarti et al., 1967; Justel et al., 1997).

2. Markov process fundamentals

A Markov process is a stochastic process that fulfils the following condition (Markov, 1907; Cox and Miller, 1965; Feller, 1968):

$$\begin{aligned} P(X_n = x_n / X_1 = x_1; \quad X_2 = x_2, \ldots; \quad X_{n-1} = x_{n-1}) \\ = P(X_n = x_n / X_{n-1} = x_{n-1}) \end{aligned} \tag{2}$$

That is to say, the probability that variable X, in the step n, X_n, is in the state x_n, having been in the states x_1, x_2,…, x_{n-1} in the corresponding former steps, only depends on its state in the step $n-1$. Thus, a Markov process can be thought of as 'memoryless', as what the state corresponding to a certain step only depends on the state in the current step, and not on the former steps. In this type of processes, the probabilities of moving from one state to another can be described from a transition matrix, A. The element corresponding to the i row, and to j the column of this matrix, p_{ij}, represents the probability that variable X moves to state x_j, being in the state x_i in the previous step. If the set of states

* Corresponding author. Tel.: +34 913466043.
E-mail address: josemaria.vindel@ciemat.es (J.M. Vindel).

1364-6826/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jastp.2013.10.017

Table 1
Ground stations details.

Name	Latitude (°N)	Longitude (°E)	Height (m)	Records for training GHI/ DNI (days)	Records for testing GHI/ DNI (days)
Carpentras	**44.03**	**5.06**	**100**	**365/730**	**4748/4383**
Tamanrasset	**22.79**	**5.53**	**1377**	**365/365**	**3652/3652**
Sede Boqer	**30.85**	**34.78**	**480**	**365/731**	**2557/2191**
Bondville	**40.05**	**−88.37**	**212**	**366/731**	**2922/2557**
Darwin	**−12.42**	**130.89**	**29**	**365/365**	**2922/2922**

GHI: Global Horizontal Irradiation, DNI: Direct Normal Irradiation.

corresponding to the variable is the vector $X=\{x_1, x_2,\ldots, x_r\}$, then the set of states after n steps is the vector $X_n=X\cdot A^n$.

3. Experimental data

Daily sums of global horizontal and direct normal irradiation corresponding to five stations of the BSRN (Baseline Surface Radiation Network) have been used in this work. The five stations and theirs study periods have been the following: Carpentras (1998–2011), Bondville (2000–2008), Tamanrasset (2001–2011), Sede Boqer (2003–2010) and Darwin (2003–2011). These stations have been chosen to cover a wide range of climatic regions. Detailed information on geographic coordinates and number of days used for training and checking the models of the ground stations are shown in Table 1.

Quality check procedures according to MESoR project and BSRN recommendations have been used to ensure the quality of the ground data used (McArthur, 1998; Hoyer-Klick et al., 2008).

4. Methodology

Complete sequences of daily irradiation values for both global horizontal and direct normal have been built for this analysis. Gaps in the sequences have been filled by persistence or monthly mean substitution. It is worth to remark that gaps represented less than 1% of the daily records in all the stations, with the exception of Carpentras where it represented less than 4% of the whole data; thus, the observations and resusequences of daily irradlts of the work should not be affected by the gap filling procedure. The iation measurements have been converted into daily normalizing indexes by dividing daily values by the corresponding value at the top of the atmosphere (TOA). Thus, global horizontal irradiation was normalized to daily clearness index, *kt*, (resulting from the ratio of the daily global horizontal to the daily global horizontal at TOA) and direct normal was normalized to daily transmittance index, *kb*, (resulting from dividing daily direct normal irradiation by the solar constant integrated along the daytime). The use of normalized indexes like the clearness index or the transmittance is widely used in the statistical studies of solar radiation time series to remove the deterministic components due to astronomic influences on the solar radiation reaching the Earth at a particular location (Iqbal, 1983).

Next, clearness and transmittances indexes have been divided in several categories, in order to discrete the irradiation. In particular nine categories, whose limits are the percentiles 11, 22, 33, 44, 55, 66, 77 and 88, have been assigned to a letter (from A to I, respectively). The percentiles values were chosen to range homogeneously the whole distribution. That is, the use of percentiles of irradiation distribution function ensures values of irradiation in all the categories, avoiding local singularities in the solar radiation behavior. Therefore, each category corresponds to one state of the process; next, sequences of two consecutive letters have been considered, which means that the series are formed by groups of 2 consecutive days. So, each sequence represents a word and thus, the number of different words is 81. The rank and the frequency corresponding to each of these words have been obtained, in order to check the relation frequency vs. rank.

Assuming a Markov process, a transition matrix has been estimated considering the frequencies of moving from one state to another (these empirical frequencies are considered as the corresponding transition probabilities), and with this matrix, a synthesized series has been able to estimate employing the following procedure (Aguiar et al., 1988): firstly, the states kb_i *and* kt_i, corresponding to the most probable clearness and transmittance index, respectively, have been chosen; next, a number between 0 and 1 is randomly selected, and the corresponding elements to i-th row of the transition matrix, p_{i1}, p_{i2}, …, p_{ij} are added until their sum is greater than the random number chosen before; then, the states kb_j and kt_j are the states corresponding to second day: By this way, sequences of two letters are built: the first letter corresponds to states kb_i *and* kt_i, and the second to kb_j and kt_j. The Markov transition matrices have been obtained for each station considering 1 or 2 years of data (for Tamanrasset the year 2001; for Darwin, 2003; for Sede Boqer, 2003 in the case of the global radiation and 2003–2004 in the case of the direct radiation; for Carpentras, 1998 for global and 1998–1999 for direct; for Bondville, 2000 for global and 2000–2001 for direct). Finally, using the Markov transition matrices obtained a new synthetic series have been produced and compared to empirical sequences, obtained considering the rest of the years of study, using for this purpose the relation frequency vs. rank. Thus, the use of a period of data for development the Markov model and a different period of data for validation provides a solid argument in order to characterize the behavior of the sequences by Markov chains.

Moreover, the goodness of the synthesized series built from a Markov process has been checked by means of a Kolmogorov–Smirnov test. Kolmogorov–Smirnov test is a general non-parametric method for comparing two samples (Massey, 1951). This statistical test is based on estimating the statistical defined by the maximum value of the absolute difference between two cumulative distribution functions. However, the use of this test in this work is slightly different. Instead of the maximum value, here the differences corresponding to each category are plotted and all of them are compared to the threshold value for passing the test.

5. Results

In Fig. 1 probabilities (normalized frequencies) of empirical series of clearness and transmittance indexes vs. corresponding ranks have been plotted. Logarithmic fits have been included showing, by means of the determination coefficients, the goodness of these fits for whole ranks set. The Zipf's law (a power relation between frequencies and ranks) is only fulfilled for sections of ranks (Fig. 3). In any case, the shape of the curve implies the most usual sequences of 2 days are much more frequent than the least usual sequences. In this sense, the coefficients of the logarithmic terms allow to characterize these variations between the most frequent sequences and the least frequent ones: the higher this coefficient (in absolute value), the higher the difference between the frequencies of the extremes. According to the figure, in the case of direct normal radiation, Darwin has the highest coefficient, though very close to Tamanrasset and to Sede Boqer; next, Carpentras and Bondville. With respect to global radiation, the