Box-Jenkins Models

John Boland

Abstract *Time Series analysis is concerned with data that are not independent, but serially correlated, and where the relationships between consecutive observations are of interest.* O. E. Anderson, Time Series Analysis and Forecasting

1 Introduction

After we have identified and removed trends and seasonality, we assume that the residuals thus formed are a stationary time series. We then utilise particular methods to ascertain whether this residual series is the realisation of a purely random process, or if it contains serial correlation of some nature. We are going to examine the Autoregressive Integrated Moving Average Process. Another name for the processes that we will undertake is the Box-Jenkins (BJ) Methodology, which describes an iterative process for identifying a model and then using that model for forecasting. The Box-Jenkins methodology comprises four steps:

- Identification of process;
- Estimation of parameters;
- Verification, and;
- Forecasting.

Note that those four features define the original Box-Jenkins approach. We will add discussions about probabilistic forecasting and also synthetic generation of data sets.

We will be using the weather data Mildura Daily Solar Radiation and Mildura Daily Temperature for illustration, among other data sets. The first file comprises

Name of First Author
Name, Address of Institute, e-mail: name@email.address

Name of Second Author
Name, Address of Institute e-mail: name@email.address

twelve years of daily total global solar radiation on a horizontal plane, and the second is the same period of daily average temperature. Global solar radiation means the aggregate of the solar energy hitting a site directly (direct solar radiation) plus the solar energy reflected to that site off particles in the atmosphere (diffuse solar radiation).

2 Identification of Processes

There is more than one form of a Box-Jenkins model, and we are going to investigate the Non-Seasonal Autoregressive Moving Average Model. This model is so called because it combines both the concept of a moving average model and an autoregressive model. To proceed we need some definitions and terminology. Assume we have a stationary time series X_t, ie. no trend, seasonality, periodicity, and as well we assume homocedascity (no variation in time of the variance). We will relax this subsequently and discuss the various ways the variance may change, that is either systematically or stochastically.

We thus first deal with a series that we can assume is weakly stationary.

2.1 Weak Stationarity

A series is weakly stationary if $\forall t$,

$$\begin{aligned} E[X_t] &= \mu \\ Cov[X_t, X_{t-k}] &= \gamma_k. \end{aligned}$$

2.2 Strong Stationarity

A series is said to be strongly stationary if

$$(X_1, X_2, \ldots, X_m) =^d (X_{1+h}, X_{2+h}, \ldots, X_{m+h}) \tag{1}$$

Here, $=^d$, means equal in distribution, or all moments are equal, not just mean and variance as with weak stationarity.

3 Autoregressive Moving Average (ARMA) Models

The general form of these models is

$$X_t - \alpha_1 X_{t-1} - \alpha_2 X_{t-2} - \ldots - \alpha_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + \ldots + \theta_q Z_{t-q} \quad (2)$$

In this the X_t are identically distributed random variables $\sim (0, \sigma_X^2)$, and the Z_t are independent and identically distributed random variables or White Noise, $\sim (0, \sigma_Z^2)$. Note that there is often the misconception that the Z_t must be Normally distributed, but for many applications, they need not be even symmetric. We shall see later when a lack of Normality is critical.

α_p and θ_q are the coefficients of p and q order polynomials satisfying

$$\begin{aligned} \alpha(y) &= 1 - \alpha_1 y - \alpha_2 y^2 - \ldots - \alpha_p y^p \\ \theta(y) &= 1 + \theta_1 y + \theta_2 y^2 + \ldots + \theta_q y^q, \end{aligned}$$

and $\alpha(y)$ and θ_q are the autoregressive and moving average polynomials respectively.

Let B be the backwards shift operator, defined by $B^j X_t = X_{t-j}$, $j = 0, \pm 1, \pm 2, \ldots$. Then we can write Equation 2 in the form

$$\alpha(B) X_t = \theta(B) Z_t, \quad (3)$$

which is known as an $ARMA(p,q)$ process.

If $\alpha(z) = 1$, then we have a pure moving average process $MA(q)$,

$$X_t = \theta(B) Z_t. \quad (4)$$

Alternatively, if $\theta(z) = 1$, we have a pure autoregressive process $AR(p)$,

$$\alpha(B) X_t = Z_t. \quad (5)$$

4 The Sample Autocorrelation and Partial Autocorrelation Functions

The Autocorrelation and Partial Autocorrelation Functions provide a useful measure of the degree of dependence between values of a time series at specific interval of separation and thus play an important role in prediction of future values of a time series.

4.1 Digression

To fully understand these processes we need some concepts defined. We will briefly describe the concept of covariance. Suppose two variables X and Y have means μ_X , μ_Y respectively. Then the covariance of X and Y is defined to be

$$Cov(X,Y) = E(X-\mu_X)(Y-\mu_Y). \tag{6}$$

If X and Y are independent, then

$$Cov(X,Y) = E(X-\mu_X)(Y-\mu_Y) = E(X-\mu_X)E(Y-\mu_Y) = 0. \tag{7}$$

If X and Y are not independent, then the covariance may be positive or negative, depending on whether "high" values of X tend to go with "high" or "low" values of Y . It is usual to standardise the covariance by dividing by the product of their respective standard deviations to give a quantity called the correlation coefficient. If X and Y are random variables for the same stochastic process at different times, then the covariance coefficient is called an autocovariance coefficient, and the correlation coefficient is called an autocorrelation coefficient. If the process is stationary, the standard deviations of X and Y will be the same and their product will be the variance of either.

Let X_t be a stationary time series. The autocovariance function (ACVF) of X_t is given by

$$\gamma_X(h) = Cov(X_{t+h}, X_t), \tag{8}$$

and the autocorrelation function (ACF) of X_t is

$$\rho_X(h) = \frac{\gamma_X(h)}{\gamma_X(0)}. \tag{9}$$

4.2 The Sample Functions

The autocovariance and autocorrelation functions can be estimated from observations of $X_1, X_2, ..., X_n$ to give the Sample Autocovariance Function (SAF) and the Sample Autocorrelation Function (SACF), denoted by

$$r_k = \frac{\sum_{t=1}^{n-k}(x_t-\bar{x})(x_{t+k}-\bar{x})}{\sum_{t=1}^{n}(x_t-\bar{x})^2} \tag{10}$$

where k is the **lag**. Thus, the SACF is a measure of the linear relationship between time series observations separated by some time period, denoted the lag k. Similar to the correlation coefficient of linear regression, r_k will take a value between $+1$ and -1, and the closer to ± 1 the value is, the stronger the relationship. If the value is positive then the relationship is positive and vice versa.

What relationship are we exactly talking about? Let's consider a lag of 1. A value close to 1 means that there is a strong correlation between x_t and x_{t-1}, x_{t-1} and x_{t-2} and so on down to the last observation. Thus, there is a strong correlation between x_t and some x_d, where x_d is an observation quite a number of time units away from x_t. Although the correlation is strong, it does not represent the individual correlations between the variables.

The Sample Partial Autocorrelation Function describes the correlation between observations at some time period, the lag, with the influence of the serial correlation removed. The autocorrelation function gives you the connection or correlation between the data values a certain number of time steps (lags) apart. However, if the value at time t is correlated with the value at time $t-1$ and the value at time $t-1$ is correlated with the value at time $t-2$, then there will be a significant correlation between the value at time t and the value at time $t-2$ because of the interconnection. The partial autocorrelation function strips away the interconnection and gives only "pure" correlation. We can calculate the PACF coefficients knowing the ACF ones. This requires use of the Yule-Walker equations which are

$$\begin{bmatrix} 1 & \rho_1 & \rho_2 & \dots & \rho_{k-2} & \rho_{k-1} \\ \rho_1 & 1 & \rho_1 & \dots & \rho_{k-3} & \rho_{k-2} \\ . & . & . & \dots & . & . \\ . & . & . & \dots & . & . \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \dots & \rho_1 & 1 \end{bmatrix} \begin{bmatrix} \alpha_{k1} \\ \alpha_{k2} \\ . \\ . \\ \alpha_{kk} \end{bmatrix} = \begin{bmatrix} \rho_1 \\ \rho_2 \\ . \\ . \\ \rho_k \end{bmatrix}$$

The partial autocorrelation is the α_{kk} term and they are solved progressively using

$$\alpha_{11} = \rho_1$$

$$\begin{bmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{bmatrix} \begin{bmatrix} \alpha_{21} \\ \alpha_{22} \end{bmatrix} = \begin{bmatrix} \rho_1 \\ \rho_2 \end{bmatrix}$$

$$\begin{bmatrix} 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_1 \\ \rho_2 & \rho_1 & 1 \end{bmatrix} \begin{bmatrix} \alpha_{31} \\ \alpha_{32} \\ \alpha_{33} \end{bmatrix} = \begin{bmatrix} \rho_1 \\ \rho_2 \\ \rho_3 \end{bmatrix}$$

In fact, a recursive formula due to Durbin is more useful in estimating the coefficients.

$$\begin{aligned}\hat{\alpha}_{p+1,j} &= \hat{\alpha}_{p,j} - \hat{\alpha}_{p+1,p+1}\hat{\alpha}_{p,p-j+1} \\ \hat{\alpha}_{p+1,p+1} &= \frac{r_{p+1} - \sum_{j=1}^{p} \hat{\alpha}_{p,j} r_{p+1-j}}{1 - \sum_{j=1}^{p} \hat{\alpha}_{p,j} r_j} \\ \hat{\alpha}_{1,1} &= r_1\end{aligned}$$

4.3 Examples

1. For an $AR(1)$ process, $\alpha_{11} = \alpha_1 = \rho_1$, or $\hat{\alpha}_{11} = \hat{\alpha}_1 = r_1$.
2. For an $AR(2)$ process, $\begin{bmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{bmatrix} \begin{bmatrix} \alpha_{21} \\ \alpha_{22} \end{bmatrix} = \begin{bmatrix} \rho_1 \\ \rho_2 \end{bmatrix}$
   $\Rightarrow \alpha_{22} = \alpha_2 = \dfrac{\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & \rho_2 \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{vmatrix}} = \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2}$.
3. Purely Random Process. A discrete process X_t is called a purely random process if the X_t are mutually independent, identically distributed random variables and

$$\gamma_k = Cov(X_t, X_{t+k}) = 0 \tag{11}$$

   The process is strictly stationary and the ACF is given by

$$\rho_k = \begin{cases} 1, & k = 0 \\ 0, & k = \pm 1, \pm 2, \ldots \end{cases} \tag{12}$$

4. Random Walk
   Suppose Z_t is white noise with mean $\mu = 0$ and variance σ_Z^2. A process X_t is denoted a random walk if

$$X_t = X_{t-1} + Z_t. \tag{13}$$

   The process is customarily started at zero when $t = 0$ and so

$$\begin{aligned} X_1 &= Z_1 \\ X_t &= \sum_{i=1}^{t} Z_t. \end{aligned}$$

   Thus, we find that $E(Z_t) = t\mu$ and $Var(X_t) = t\sigma_Z^2$, implying a nonstationary process. However, if we difference a random walk, we get

$$X_t - X_{t-1} = Z_t, \tag{14}$$

   a purely random process.

Once we have a stationary time series we can use these functions to determine whether to fit

1. A moving average process or;
2. An autoregressive process or;
3. An autoregressive integrated moving average process

We use (1) when the SACF has spikes at lags $1,2,,q$ and the SPACF dies down gradually.

We use (2) when the SACF dies down gradually and the SPACF has spikes at lags $1,2,,p$. If neither occurs then we may be faced with a non-stationary times series or a requirement to use a combination of both, which is option (3). If we were to refer to the theoretical sample autocorrelation function (TACF) and theoretical partial autocorrelation function (TPACF), it could be shown that for the case of the $MA(1)$ process that the TACF has a non-zero correlation at lag 1 and zero autocorrelations thereafter (ie., cuts off), while the TPACF dies down in a steady manner which can be shown to be dominated by damped exponential decay. For the $AR(1)process$, it can be shown that the TACF dies down in a damped exponential fashion while the TPACF has a non zero partial autocorrelation at lag 1, with zeros thereafter, that is, the TPAC cuts off. The graphs produced from the SACF are known as correlograms.

5 Moving Average Process

A moving average process $MA(q)$(of order q) has the present value of the series written as a weighted sum of or regression on past random shocks,

$$X_t = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} = \ldots + \theta_q Z_{t-q}, \tag{15}$$

where $Z_t \sim (0, \sigma_Z^2)$.

We find that

$$\begin{aligned} E(X_t) &= 0 \\ \sigma_X^2 &= \sigma_Z^2(1 + \sum_{i=1}^{q} \theta_i^2) \end{aligned}$$

6 Autoregressive Processes

The general form of an autoregressive process of order p, or $AR(p)$, is given by

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \ldots + \alpha_p X_{t-p} + Z_t. \tag{16}$$

This is like a multiple regression but X_t is regressed on past values of X_t, not on other predictor variables, hence the term autoregressive. Remember that another way to write this process is (using the backward shift operator B):

$$\alpha(B)X_t = Z_t, \tag{17}$$

where $\alpha(B) = 1 - \alpha_1 B - \alpha_2 B^2 - \ldots - \alpha_p B^2$ is called the $AR(p)$ operator.

7 Simple Time Series Example

We will look at 98 years of the level of Lake Huron, one of the Great Lakes between Canada and USA. The level in the series is measured once a year on the same time. See Figure 1 for the series over time. Before performing any ARMA modelling, we need to remove the trend from the data. In this case, we first tried a linear trend but that did not seem potentially as suitable as a curvilinear trend. We fit the most simple curvilinear model, a quadratic. Note that this is still a linear model as the coefficients to be estimated still appear linearly. When this is done, it is noted that the coefficient of determination, or R^2, the amount of variance explained by the model, increases from 26.5% to 39.6%. One important point to note that in reality perhaps something like a decaying exponential might be more appropriate, implying that the level is decreasing towards an asymptote, rather than rising at the end as in Figure 2.

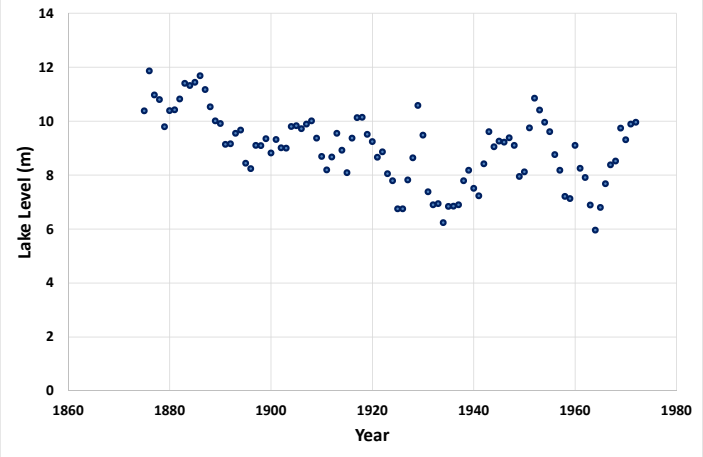

Fig. 1 Lake Huron level in meters

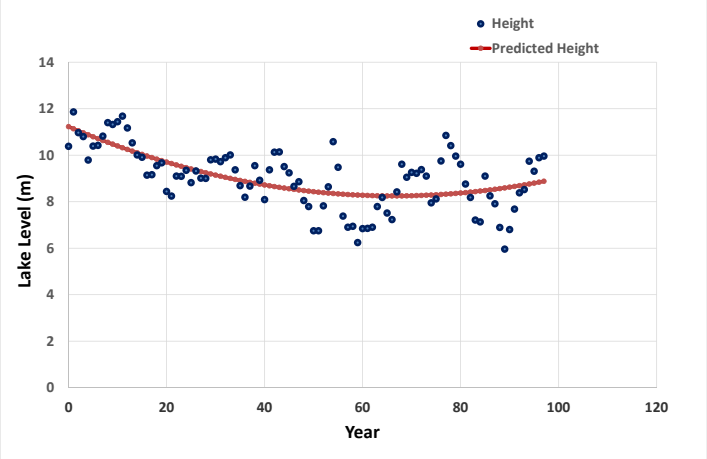

Fig. 2 Lake Huron level with quadratic fit

7.1 The data with trend removed

If we define the original data as $L(t)$, and the quadratic trend as $T(t)$, the next step after estimating the trend equation is to subtract it from the original data to form the residual series $R(t) = L(t) - T(t)$. The question then is whether there is any serial correlative structure in $R(t)$. We follow the procedure described above for identification of the type of ARMA process. For this we use the sample autocorrelation function (SACF) in Figure 3 and sample partial autocorrelation function (SPACF) in Figure 4.

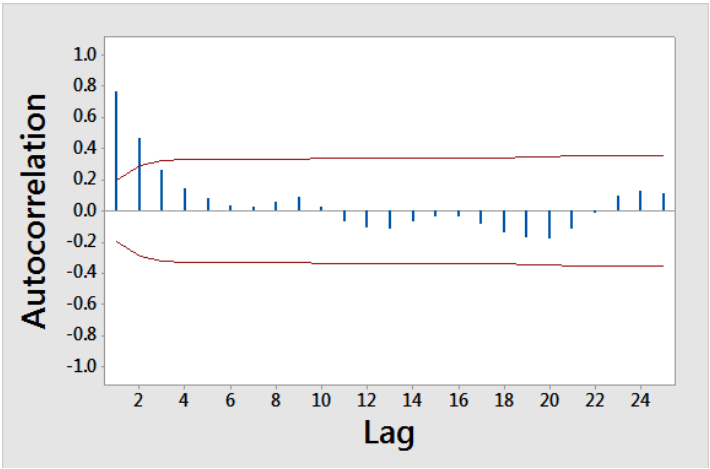

Fig. 3 SACF for the Lake Huron detrended data

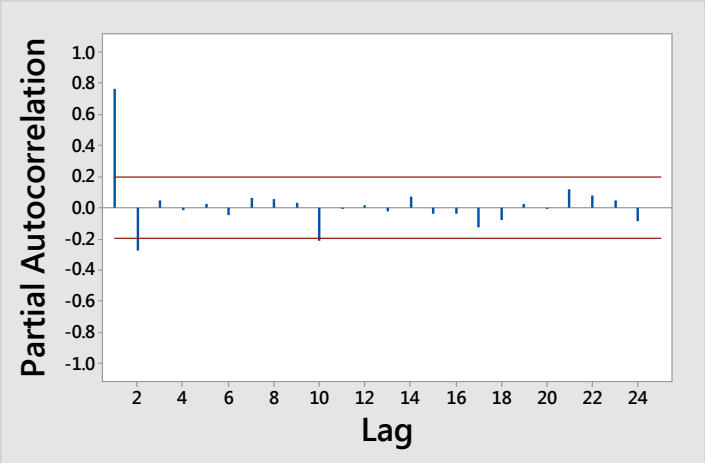

Fig. 4 SPACF for the Lake Huron detrended data

This fits the simplest situation possible with the SACF decreasing gradually with increasing lags into the past and the SPACF stopping abruptly after two lags. This would imply that an $AR(2)$ model might well fit the series. The next step is to see if we can verify this conjecture or if we need to alter it. To do so, we first try overfitting, that is, we try to fit an $AR(3)$ model with a constant term included in the specification. Minitab, or any other statistics package, will give the output shown in Table 1. How you read this is as a test of the Null Hypothesis for any autoregressive coefficient or the constant. In the table, the p-value for the first two AR coefficients is < 0.05, implying that they are significantly different from zero, whereas for the thrid and the constant, the opposite os true.

$$\begin{aligned} H_0 &: \rho = 0 \\ H_a &: \rho \neq 0 \\ \alpha &= 0.05 \end{aligned}$$

Table 1 Final Estimates of Parameters

Type	Coefficient	SE Coef	T	p
AR 1	1.0345	0.1032	10.03	0.000
AR 2	-0.3641	0.1438	-2.53	0.013
AR 3	0.0667	0.1051	0.63	0.527
Constant	0.00744	0.06982	0.11	0.915
Mean	0.0283	0.2656		

Therefore, one then re-estimates the $AR(2)$ coefficients to obtain the model, as in Eqn 18.

$$X_t = 1.016X_{t-1} - 0.299X_{t-2} + Z_t \tag{18}$$

This completes the first two steps of the Box-Jenkins process, that of Identification and Estimation. The third step is to Verify that this is the best model possible. There are two related aspects to this. One is the Ljung-Box test for residual autocorrelation between 1 and 12 lags, 13 and 24, and so on. Results of this test are given in Table 2. Essentially these are hypothesis tests as well and if the p-values in the last row are all > 0.05, then there is no residual autocorrelation. Aligned with this is the SACF of the final noise series, the $Z(t)$. One hopes that all the spikes are within the confidence intervals for non significant correlation. This is shown in Figure 5. It is readily seen that both criteria are satisfied.

Table 2 Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag	12	24	36	48
Chi-Square	5.3	10.3	16.8	25.8
DF	10	22	34	46
P-Value	0.872	0.983	0.994	0.993

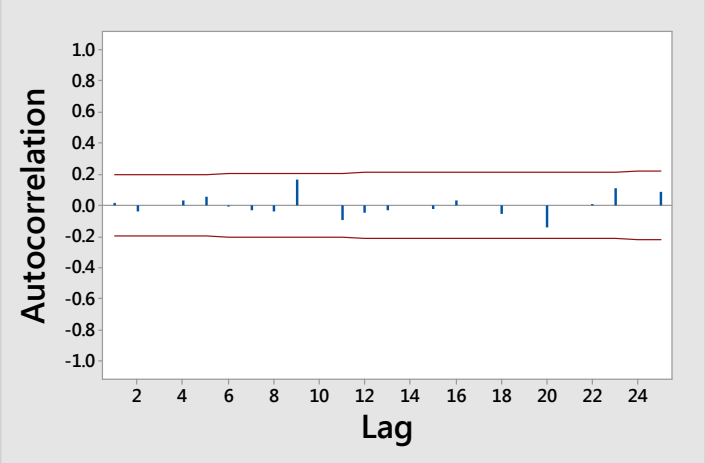

Fig. 5 SACF for the Lake Huron noise series

There is still the question of what characteristics does the noise series Z_t possess. Z_t is supposed to be white noise with mean $\mu = 0$ and variance σ_Z^2. White noise means that the series is independent in time, and identically distributed. The results of the Ljung-Box test and the SACF of the noise series indicate lack of serial

correlation. We will see in the Chapter on conditional heteroscedasticity (changing variance) that a series may be serially uncorrelated but still dependent. The variance may have serial correlation. The squared noise terms are a proxy for the variance, and one can test the SACF of them to check for changing variance. In this example, the SACF of the squared noise terms shows no residual autocorrelation. This also implies identically distributed, at least in the first two moments. In this instance let's examine the distribution of the noise terms, noting that in a perfect situation we would be able to conclude they are Normally distributed. Figure 6 gives the histogram of the noise with a Normal curve overlaid, for the Normal distribution with the same mean and variance. Once again one could say this is the descriptive statistics version of implying that the noise is normally distributed. However, a Q-Q plot shows the correspondence or not for a distribution to be Normal - see Figure 7. If the dots are close the the line, then one can infer Normality. There is also the results of a test for the Null Hypothesis of Normality and the p-value in the graphic is > 0.05, and so we can conclude that $Z\ N(0, 0.679^2)$.

Note that this example behaves exactly as one might hope and is essentially a special case, since this concordance is almost never true for real data.

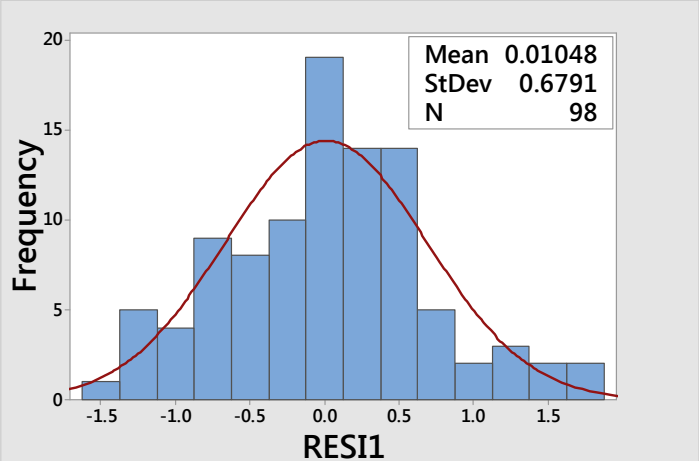

Fig. 6 Histogram of noise with Normal curve overlaid

8 Climate Variables

In this section, we will examine the Box-Jenkins approach for both hourly solar and wind energy series, and ambient temperature, plus their attributes on a five minute time scale. The time scales from five minute to one to two hours is the realm of statistical forecasting models. There is a lot of work in the literature about the use

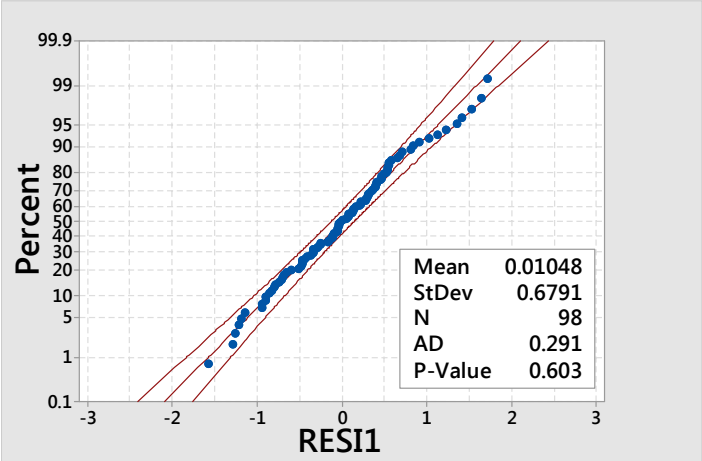

Fig. 7 Probability plot for the noise

of Artificial Neural Networks in this arena - see for instance [1] and [2]. I suggest that this results from an approach that is not necessarily trying to find the most parsimonious model for the phenomenon, but instead a conservative approach in that it attempts to take care of any possible influence. The Box-Jenkins approach, if successful in typifying the characteristics of the series, provides a model that offers a physical interpretation of the process. Thus I suggest it is a better place to begin. It is fascinating to examine the differences not only on the time scale, but also between solar, wind and temperature series.

8.1 Hourly Solar Energy Series

We model the seasonality using Fourier Series. Equation 19 gives the Fourier series:

$$FS_t = \alpha_0 + \alpha_1 \cdot \cos\frac{2\pi t}{8760} + \beta_1 \cdot \sin\frac{2\pi t}{8760} + \alpha_2 \cdot \cos\frac{4\pi t}{8760} + \beta_2 \cdot \sin\frac{4\pi t}{8760} + \sum_{n=1}^{2}\sum_{m=-1}^{1}\left(\alpha_{nm} \cdot \cos\frac{2\pi(356n+m)t}{8760} + \beta_{nm} \cdot \sin\frac{2\pi(365n+m)t}{8760}\right) \quad (19)$$

Here, α_0 is the mean of the data, α_1, β_1 are coefficients of the yearly cycle, α_2, β_2 of twice yearly and α_{nm}, β_{nm} are coefficients of the daily cycle and its harmonics and associated beat frequencies. An inspection of the power spectrum would show that we need to include the harmonics of the daily cycle (n = 1, 2) and also the beat frequencies (m = -1, 1). Figure 18 shows an illustration of the Fourier series model of the data.

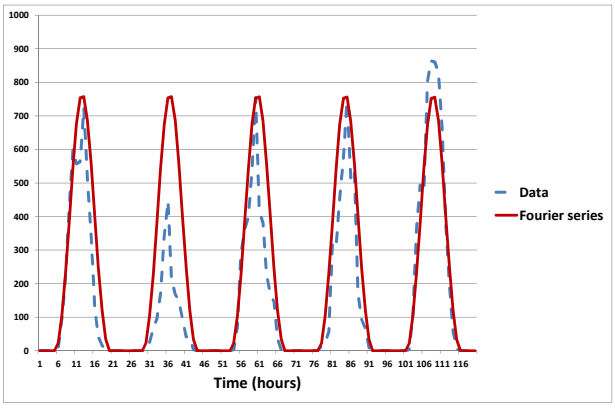

Fig. 8 Five days solar radiation and corresponding Fourier series representation

When the Fourier series contribution is subtracted from the data series, the residual series can be modelled with the coupled autoregressive and dynamical system approach, details of which are given in [6]. In this instance, we use a simpler but similarly effective procedure, that of a short lag autoregressive model. This simplification of the ARMA process is possible since in this instance, the moving average (MA) contribution is not significant. This will be shown in subsequent analysis, using the SACF and SPACF.

Let the original series be denoted by S_t and then the residual series is given by $X_t = S - t - FS_t$.

The general form of an autoregressive process of order p, or $AR(p)$, is given by

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \ldots + \alpha_p X_{t-p} + Z_t. \tag{20}$$

This is like a multiple regression but X_t is regressed on past values of X_t, not on other predictor variables, hence the term autoregressive. In this, $Z(t)$ is white noise with variance σ_Z^2.

8.2 Selection Criteria

We start by examining the sample autocorrelation and partial autocorrelation functions.

Taken together, these diagrams lead us to try and model the solar residual series using an $AR(p)$ configuration, with $p \leq 5$. Note that Minitab restricts p to be five or less for reasons of parsimony. One could use and $ARMA(p,q)$ but the diagrams indicate that an purely autoregressive model should suffice. The good thing about this is that it is easily explainable physically. How do we decide how many lags to use? We can do it roughly by examining the output from a couple of models. Let's start with $AR(5)$. Since the p-values are greater than 0.05 for both the fifth lag and

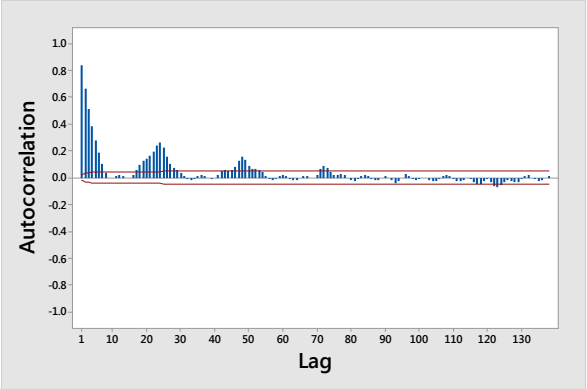

Fig. 9 Solar SACF

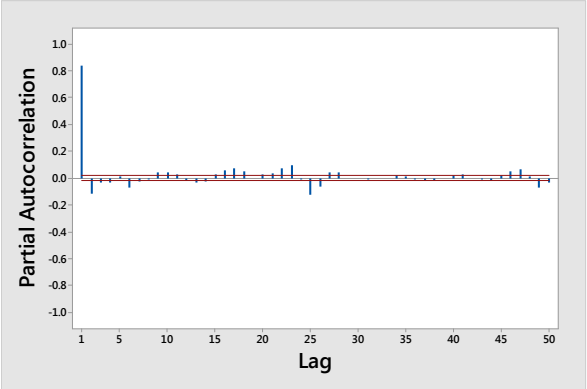

Fig. 10 Solar SPACF

the constant, they will not be needed so we can try an $AR(4)$, with zero constant. This seems appropriate, except of course the Ljung-Box tests fail but we will examine this further - in most real instances this will happen, but will not detract from the process. One thing we notice though is that the third lag is also not significant and the coefficient for the fourth lag is not very arge compared to the first and second. Perhaps we could get by with only an $AR(2)$ model, and this is what we show next.

The big question is how do we decide if we only need two lags compared to four? An aid in identifying the appropriate model is the the *Akaike Information Criterion (AIC)*. It is defined by

```
Final Estimates of Parameters

Type         Coef  SE Coef      T      P
AR   1     0.9328   0.0107  87.28  0.000
AR   2    -0.0934   0.0146  -6.40  0.000
AR   3    -0.0005   0.0146  -0.03  0.972
AR   4    -0.0421   0.0146  -2.88  0.004
AR   5     0.0093   0.0107   0.87  0.383
Constant  -0.0047   0.5951  -0.01  0.994
Mean       -0.024    3.069


Number of observations:  8760
Residuals:    SS =  27160072 (backforecasts excluded)
              MS =  3103  DF = 8754


Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag            12     24     36     48
Chi-Square  104.9  518.3  675.6  989.6
DF              6     18     30     42
P-Value     0.000  0.000  0.000  0.000
```

Fig. 11 AR(5) with constant

```
Final Estimates of Parameters

Type       Coef  SE Coef      T      P
AR   1   0.9325   0.0107  87.30  0.000
AR   2  -0.0934   0.0146  -6.40  0.000
AR   3  -0.0014   0.0146  -0.09  0.925
AR   4  -0.0334   0.0107  -3.13  0.002


Number of observations:  8760
Residuals:    SS =  27162435 (backforecasts excluded)
              MS =  3102  DF = 8756


Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag            12     24     36     48
Chi-Square  107.7  522.4  681.5  997.6
DF              8     20     32     44
P-Value     0.000  0.000  0.000  0.000
```

Fig. 12 AR(4) with no constant

$$\begin{aligned} AIC(k) &= \frac{2}{N}[-ln(L)+k] \\ &= ln(\tilde{\alpha}_k^2)+2\frac{k}{N} \end{aligned} \tag{21}$$

Here, L is the likelihood and k is the number of parameters in the model. $\tilde{\alpha}_k^2$ is the maximum likelihood estimate of the residual variance. Note that this form is only

```
Final Estimates of Parameters

Type       Coef  SE Coef       T      P
AR   1   0.9375   0.0106   88.38  0.000
AR   2  -0.1209   0.0106  -11.40  0.000


Number of observations:  8760
Residuals:    SS =  27221645 (backforecasts excluded)
              MS =  3108  DF = 8758


Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag            12     24     36     48
Chi-Square  121.6  507.7  647.9  923.6
DF             10     22     34     46
P-Value     0.000  0.000  0.000  0.000
```

Fig. 13 AR(2) with no constant

valid for near Gaussian errors, but it should be said that since it is relative, that is comparing one model versus another, it should still be usable.

The goal is to pick the model that minimises the criterion.

The first term measures the goodness of fit of the model and the second term is called the penalty function of the criterion because it penalises a candidate model by the number of parameters. This is indicative of the concept of parsimony, the least complicated model is best. Another common criterion is the *(Schwarz) Bayesian information criterion (BIC)*, given by

$$BIC(k) = ln(\tilde{\alpha}_k^2) + \frac{kln(N)}{N}. \tag{22}$$

Note that the BIC is a harsher penalty, so to add more parameters, it must be a substantial improvement. Note that in multiple linear regression, there is a similar criterion, called the partial F test.

For this example of solar radiation, we find that $BIC(4) = 8.0439$, while $BIC(2) = 8.0438$. The values are very close and even if the precedence were to be reversed, one could still argue for using the $AR(2)$ model, where you are getting the same fit for less work.

Therefore, the final model for the residual series is $R_t = 0.938R_{t-1} - 0.121R_{t-2} + Z_t$.

Note for later: it is assumed in the theory that Z_t is independent and identically distributed. However we will see that neither is perfectly true for solar radiation series.

One final check is to see, even if it does not imply independence, that you have at least removed most of the autocorrelation. To check this, you plot the SACF of Z_t. Note that there are significant spikes around lags $24, 48, 72\ldots$. This is the effect of including night in the series. Other than that there is little of any significance. Most

researchers would stop here, saying we have independent errors, but we have only shown lack of correlation, which we will see later is not the same as independence.

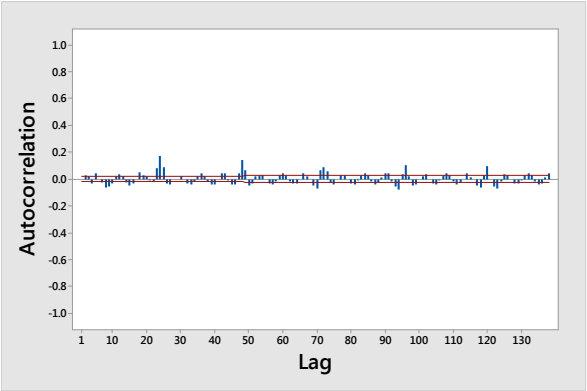

Fig. 14 SACF of the error series

9 Error analysis

Error analysis is a very important step for data analysis and forecasting. It is a tool for distinguishing which model is better. In Hoff et al.'s paper, error analyses are classified into two classes. One class consists of absolute dispersion errors which are root mean square error (RMSE) and mean absolute error (MAE). The other class represents relative percentage errors. Relative percentage errors use absolute dispersion error, such as RMSE or MAE divided by the mean of the real data to produce a normalized value. In this section several of the most commonly used error value measures will be introduced. These are median absolute percentage error (MeAPE), mean bias error (MBE), normalised root mean square error (NRMSE), normalised mean absolute error (NMAE), and Kolmogorov-Smirnov test integral (KSI). MeAPE captures the size of the errors and avoids distorting the results for solar and wind energy forecasting. If one uses mean absolute percentage error (MAPE), then with the variables under consideration, wind and solar, there are certain points in time that can produce large errors, and the distribution of errors becomes skewed. So, using MeAPE instead of MAPE for renewable energy forecasting can provide a more accurate perspective. In turn, MBE is used to determine whether any particular model is more biased than another. NRMSE measures overall model quality related to the regression fit. What this means is that is how far the data deviates from the model. What is more informative is, in essence, how

far the regression line is from the line $Y = X$, where the $y's$ are the predicted values from the model, and $x's$ are the data values. Interestingly, Willmott and Matsuura produce convincing arguments as to why the mean absolute error (MAE) is a superior error measure to the RMSE. They argue that the RMSE is a function of three characteristics of a set of errors.

It varies with the variability within the distribution of error magnitudes and with the square root of the number of errors ($n^{1/2}$), as well as with the average-error magnitude (MAE).

KSI is a new model validation measure based on the Kolmogorov-Smirnov (KS) test which has the advantage of being non-parametric. KS test is a nonparametric test for the equality of continuous, one dimensional probability distributions. It can be used to compare a sample with a reference probability distribution, in which case it is named a one sample KS test, or to compare two samples, when it becomes a two sample KS test. The KSI measure was proposed in Espinar et al. to assess the similarity of the cumulative distribution functions (CDFs) of actual and modelled data over the whole range of observed values.

Definitions of all the measures are as follows:

1. Median Absolute Percentage Error

$$MeAPE = MEDIAN\left(\left|\frac{\hat{y}_i - y_i}{y_i}\right| \times 100\right) \tag{23}$$

2. Mean Bias Error

$$MBE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i) \tag{24}$$

3. Normalised Root Mean Squared Error

$$NRMSE = \frac{\sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}}{\bar{y}} \tag{25}$$

where $\hat{y}_i$ are predicted values, y_i are measured values and $\bar{y}$ is the average of measured values.

There is an associated measure, called the Skill Score (SS). It is defined as

$$SS = 1 - \frac{RMSE_{Forecast}}{RMSE_{Reference}} \tag{26}$$

where $RMSE_{Reference}$ is the root mean squared error of some reference or benchmark model. This is usually one of two formulations, the naive persistence forecast, $X_{t+1} = X_t$, or the so-called smart persistence forecast. This is used when the variable under consideration has trend or seasonality embedded in it. An example is the case of hourly solar radiation. In the literature what is usually done is to

model the seasonality by defining the clear sky index (CSI). A precursor to this is a clear sky model, a physical model that attempts to give radiation values for a perfectly clear sky for a particular location, for each hour of the year. The CSI is then obtained by dividing the global horizontal radiation (GHI) measured for that hour by the clear sky model for that hour. It is a vexed measure, since Ineichen (2016), a respected researcher in the field, feels the need to check the validity of models - in this paper seven models often used are tested, and three suggested as good. Also, the smart persistence forecast will then include a seasonality model so the results for the Skill Score could well depend on the model chosen.

4. Normalised Mean Absolute Error

$$NMAE = \frac{\frac{1}{n}\sum_{i=1}^{n}(|\hat{y}_i - y_i|)}{\bar{y}} \tag{27}$$

5. Kolmogorov-Smirnov Integral

$$KSI(\%) = 100 \times \frac{\int_{x_{min}}^{x_{max}} D_n dx}{\alpha_{critical}} \tag{28}$$

where x_{max} and x_{min} are the extreme values of the independent variable, and $\alpha_{critical}$ is calculated as $\alpha_{critical} = V_c \times (x_{max} - x_{min})$. The critical value V_c depends on population size N and is calculated for a 99% level of confidence as $V_c = 1.63/\sqrt{N}, N \geq 35$. The D_n are the differences between the cumulative distribution functions (CDF) for each interval. The higher the KSI value, the worse the fit of the model to data.

It is worth noting that in assessing a model against the actual data, NRMSE measures how close the points are clustered around the regression line for the relationship between the observed and predicted values, while KSI and MBE assess the distribution of points around the unit line, $\hat{y} = y$. Thus by considering a set of diverse measures the aim is to allow for a more complete comparison of the proposed models. For example, additional information on the CDFs carried by KSI and MBE can be used to distinguish between models with similar MeAPE or NRMSE values.

Even though these measures are mainly used for comparison of models, we will report the results for this final model here. Note that this is for the final model, consisting of the $AR(2)$ part for the deseasoned data plus the seasonal Fourier Series model, as compared to the original series. This is depicted in Figure 15 and Table 3.

Measure	Percentage
NMBE	0.51
NMAD	10.73
NRMSE	15.51

Table 3 Error measures for solar model

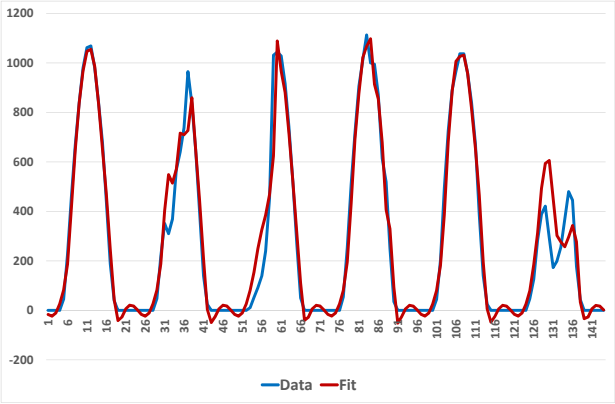

Fig. 15 Solar data and fitted model

9.1 Hourly Temperature Series

We model the seasonality using Fourier Series in the same way.

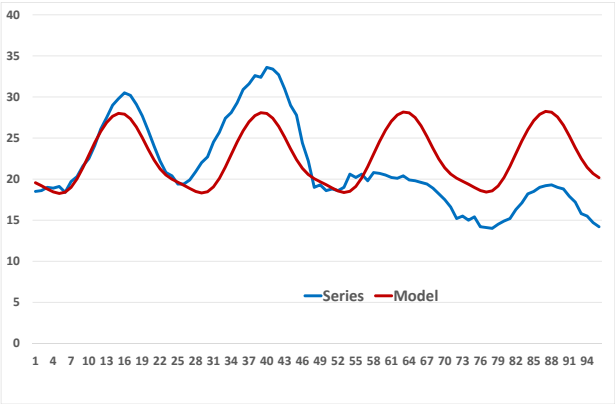

Fig. 16 Four days temperature and corresponding Fourier series representation

When the Fourier series contribution is subtracted from the data series, the residual series can be modelled with the coupled autoregressive and dynamical system approach, details of which are given in [6]. In this instance, we use a simpler but similarly effective procedure, that of a short lag ARMA process is possible since in this instance. This will be shown in subsequent analysis, using the SACF and SPACF.

Let the original series be denoted by T_t and then the residual series is given by $X_t = T - t - FS_t$.

The general form of an ARMA process of order (p,q), or $AR(p,q)$, is given by

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \ldots + \alpha_p X_{t-p} + Z_t + \beta_1 Z_{t-1} + \beta_2 Z_{t-2} + \ldots + \beta_q Z_{t-q}. \quad (29)$$

In this, $Z(t)$ is white noise with variance σ_Z^2.

9.2 Selection Criteria

We start by examining the sample autocorrelation and partial autocorrelation functions.

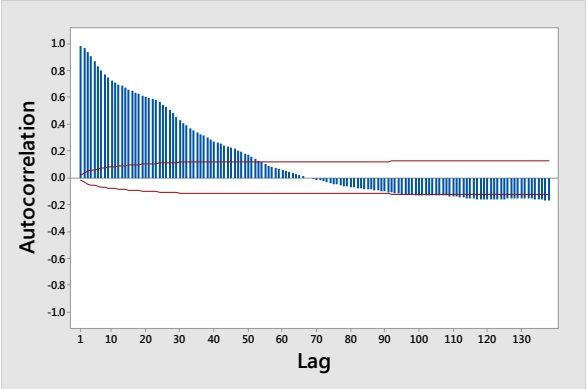

Fig. 17 Temperature SACF

This set of Figures sets the scene for modelling the temperature residuals with an $ARMA(p,q)$ process. I tried an $ARMA(5,2)$ and $ARMA(2,2)$ model and compared them. Note that the usual method is to start from below and work up. Interestingly, even though when adding more predictor variables, one lowers the mean square error of the residuals, the model in this case actually had a lower error for the final residuals. One conjectures that this is because the lag two autoregressive coefficient is not significantly different from zero in the $ARMA(5,2)$ but is predominant in $ARMA(2,2)$. There are very good error metrics for temperature, Table 4, and the fit is shown in Figure 19.

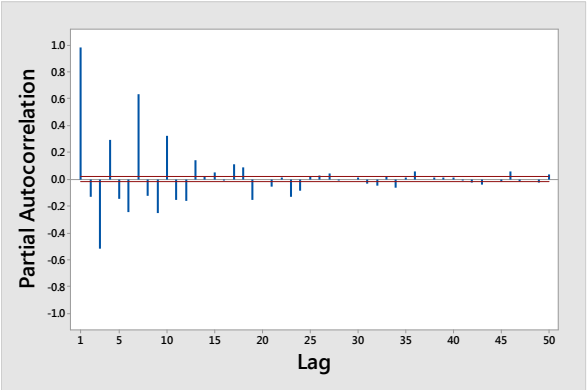

Fig. 18 Temperature SPACF

Measure	Percentage
NMBE	-5.52^{-5}
NMAD	2.04
NRMSE	2.73

Table 4 Error measures for temperature model

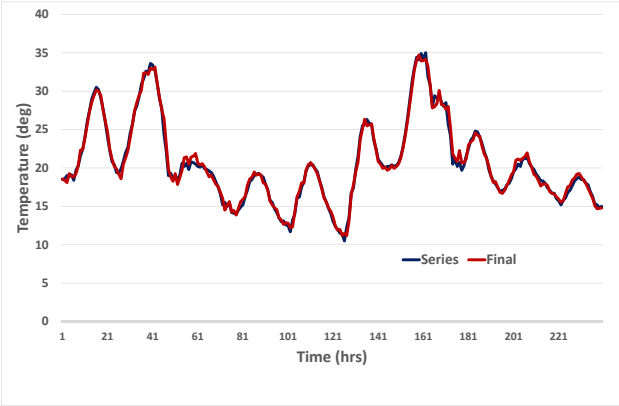

Fig. 19 Temperature data and fitted model

References

1. Maimouna Diagne, Mathieu David, Philippe Lauret , John Boland , Nicolas Schmutz. Review of solar irradiance forecasting methods and a proposition for small-scale insular grids,

Renewable and Sustainable Energy Reviews, **2013**, **27**, pp. 65-76.
2. Rich H. Inman, Hugo T.C. Pedro, Carlos F.M. Coimbra, Solar forecasting methods for renewable energy integration, *Progress in Energy and Combustion Science*, **2013**, *39*, pp. 535-576.
3. Lauret, Philippe, Voyant, Cyril, Soubdhan, Ted, David, Mathieu and Poggi, Philippe, A benchmarking of machine learning techniques for solar radiation forecasting in an insular context, *Solar Energy*, **2015**, *112*, pp. 446-457.
4. Peng, Zhenzhou, Yoo, Shinjae, Yu, Dantong, Huang, Dong, Kalb, Paul and Heiser, John, 3D Cloud Detection and Tracking for Solar Forecast Using Multiple Sky Imagers, *Proceedings of the 29th Annual ACM Symposium on Applied Computing, SAC '14*, **2014**, *Gyeongju, Republic of Korea*, pp. 512-517.
5. Lauret, Philippe, Diagne, Maïmouna and David, Mathieu, A Neural Network Post-processing Approach to Improving NWP Solar Radiation Forecasts, *Energy Procedia*, **2014**, *57*, pp. 1044-1052.
6. Jing Huang, Malgorzata Korolkiewicz, Manju Agrawal and John Boland. Forecasting solar radiation on an hourly time scale using a coupled autoregressive and dynamical system (CARDS) model. *Solar Energy*, **2013** 87, pp. 136-149.
7. Bird R. E., and Hulstrom R. L. Simplified Clear Sky Model for Direct and Diffuse Insolation on Horizontal Surfaces. **1981**. Technical Report No. SERI/TR-642-761, Golden, CO: Solar Energy Research Institute.
8. Boland J. Time series and statistical modelling of solar radiation, *Recent Advances in Solar Radiation Modelling*, Viorel Badescu (Ed.), Springer-Verlag, **2008** pp. 283-312.
9. Boland, J. Time Series Analysis of Climatic Variables, *Solar Energy*, **1995**, **55**, No. 5, pp. 377-388.
10. John Boland and Ted Soubdhan. Spatial-temporal forecasting of solar radiation, *Renewable Energy*, **2015 75** pp. 607-616.
11. Pierre Ineichen. Validation of models that estimate the clear sky global and beam solar irradiance *Solar Energy*, **2016**, *132*, pp. 332-344.
12. Badescu, V., Gueymard, C., Cheval, S., Oprea, C., Baciu, M., Dumitrescu, A., Iacobescu, F., Milos, I. and Rada, C., Accuracy analysis for fifty-four clear-sky solar radiation models using routine hourly global irradiance measurements in Romania, *Renewable Energy*, **2013**, *55*, pp. 85-103.
13. AERONET, **2014**, $http://aeronet.gsfc.nasa.gov/$
14. Ozone, World Ozone Monitoring Mapping, **2014**, $http://es-ee.tor.ec.gc.ca/e/ozone/ozoneworld.htm$.
15. Hornik, K., Stinchcombe, M., White, H. Multilayer feedforward networks are universal approximators. *Neural Networks* **1989**, *2(5)*, 359-366.
16. Bishop, C.M. Neural networks for pattern recognition. **1995**. Oxford University Press, Oxford.
17. Lauret, P., Fock, E. , Mara, T.A.. A node pruning algorithm based on the Fourier amplitude sensitivity test method. *IEEE Trans Neural Networks* **2006**, *17(2)*, 273-293.
18. MacKay, D.J.C. A practical Bayesian framework for back-propagation networks. *Neural Computation* **1992**, *4(3)*, pp. 448-472.
19. Lauret, P., Fock, E., Randrianarivony, R.N., Manicom-Ramsamy, J.F. Bayesian neural network approach to short time load forecasting. *Energy Conversion and Management*. **2008**, *49(5)*, 1156- 1166.
20. Chatfield, C.Time series analysis, an introduction. Chapman and Hall/CRC. **2004**.
21. C Voyant, M Muselli, C Paoli, ML Nivet, Numerical Weather Prediction (NWP) and hybrid ARMA/ANN model to predict global radiation, *Energy*, **2012**, *39(1)*, pp. 341-355.
22. Zagouras, Athanassios, Pedro, Hugo T.C. and Coimbra, Carlos F.M., Clustering the solar resource for grid management in island mode, *Solar Energy*, **2014**, *110*, pp. 507-518.