20th International Congress on Modelling and Simulation, Adelaide, Australia, 1–6 December 2013
www.mssanz.org.au/modsim2013

Multivariate Forecasting of Solar Energy

John Boland [a]

[a]*School of Information Technology and Mathematical Sciences and the Barbara Hardy Institute, University of South Australia, Mawson Lakes Boulevard, Mawson Lakes, SA, 5095, Australia*
Email: john.boland@unisa.edu.au

Abstract: When methods for forecasting solar radiation time series were first developed, the principal applications were for estimating performance of rooftop photovoltaic or hot water systems. If there were significant errors in the forecast, the consequences were not severe. In recent times there has been increasing development of larger solar installations, both large scale photovoltaic and also concentrated solar thermal. In order to first influence financial backers to participate in their development, and also to potentially compete in the electricity markets, better forecasting models are required than simple Box-Jenkins models, such as those outlined in Boland (2008). In Huang et al (2013), we developed a combination model linking a standard autoregressive approach with a resonating model borrowed from work on dynamical systems, and also an additional component that greatly enhances forecasting ability. This model was developed for a solar radiation series at a single site.

In this article I give an introduction to the tools needed for the multivariate forecasting of solar radiation. The modelling was developed for three sites in Guadeloupe, approximately 20 km. jointly from each other. One would expect significant cross correlation between the sites since they are affected by a common climate influence, Les Alizes, the Trade Winds. Thus, cloud bands inevitably pass over the sites at regular intervals. I demonstrate the techniques required to pre whiten the data (as far as possible) for a single site. The next step involved checking the cross correlation of the noise between sites, finding significant correlation between the sites at time t and also between the values at time t and time $t-1$. A subsequent one lag multivariate time autoregressive model was estimated. It was hoped that the three noise variables resulting from this modelling would be iid. However, this was not to be the case and all three noise series exhibited conditional heteroscedasticity. In this case, ARCH models sufficed to describe this behaviour.

Keywords: *Time series forecasting, multivariate series, ARCH model, CARDS model*

1 INTRODUCTION

We will describe the multivariate forecasting of solar radiation using three sites at Guadeloupe, in the French West Indies. The goal is to see how much forecasting skill we can attain, when you have data from three partially correlated sites. For the first analysis, we will concentrate on hourly solar radiation data, with the time for the sites being coincident. For this reason, we use the eleven months February to December 2011. We will follow the systematic procedure outlined in (3). When analysing a time series data set, the first step is to consider whether it contains a trend, or seasonality, or both. Following Boland (1), (2), we construct the Power Spectrum which gives the power in the series at frequencies 1 to 731 cycles per year. We illustrate this for the site of Desirade, latitude 16.32^o in Figure 1. There are a number of interesting features to this power spectrum, particularly if you compare it with the power spectrum for a site at latitude -34.22^o, Mildura, Australia, shown in Figure 2. For Mildura, the annual cycle is more pronounced than for Desirade, and also there are two prominent spikes at 364 cycles/year and 366 cycles/year. As explained in (3), these are called either beat frequencies or sidebands. They describe the amplitude modulation, the change in the amplitude of the daily cycle to suit the time of year. Their relative absence for Desirade shows that the amplitude of the daily cycle does not change significantly during the year. As well as this, the lower power at the annual cycle shows that there is not as great a difference over the year in the mean daily solar radiation. These two conclusions are well illustrated by comparing the daily mean radiation over the year for Desirade Figure 3 and Mildura Figure 4.

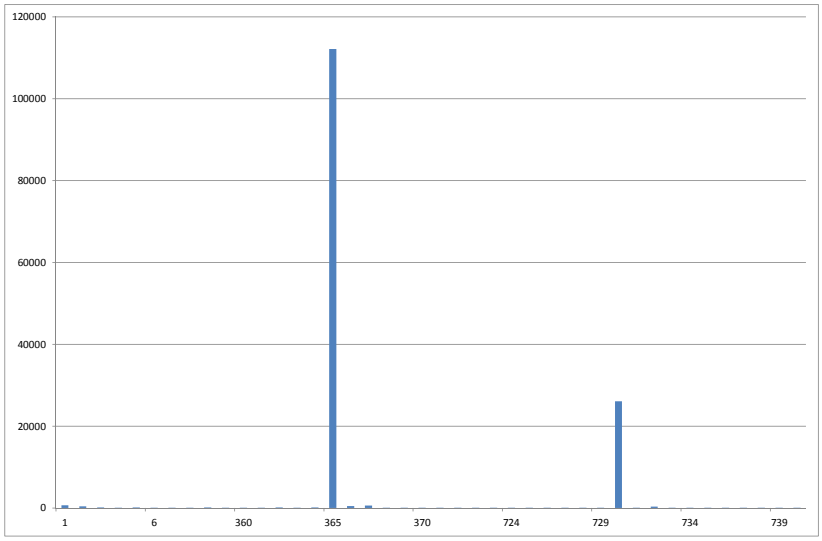

Figure 1. Power spectrum for hourly solar radiation for Desirade 2011 data

2 FOURIER SERIES MODEL

The power spectrum identifies which frequencies are significant contributors to what we will term the seasonality of the data. This seasonality is then well represented by a suitable Fourier series. Even though there are significant differences between the power spectra for Desirade compared to that of the Australian site of Mildura, we will still use the same formulation as in (3). Any insignificant frequencies will have a contribution to the series not far removed from zero. Equation 1 gives the Fourier series:

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

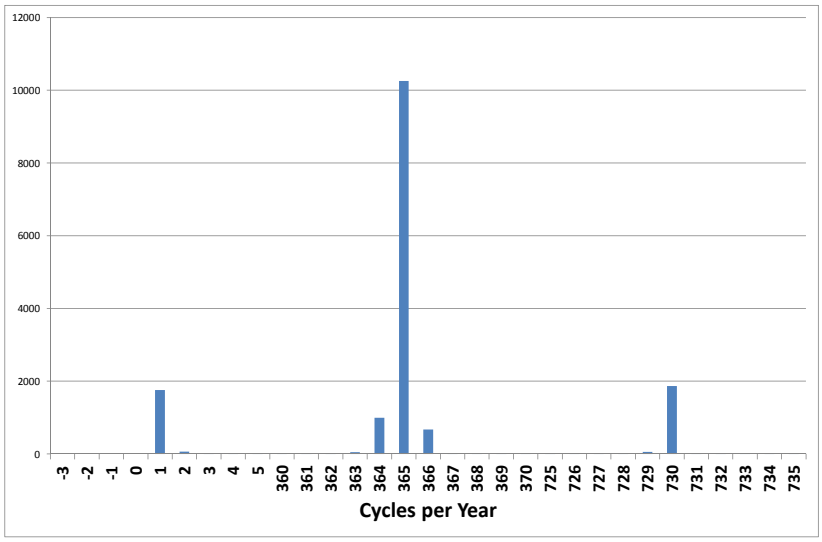

Figure 2. Power spectrum for hourly solar radiation for Mildura data

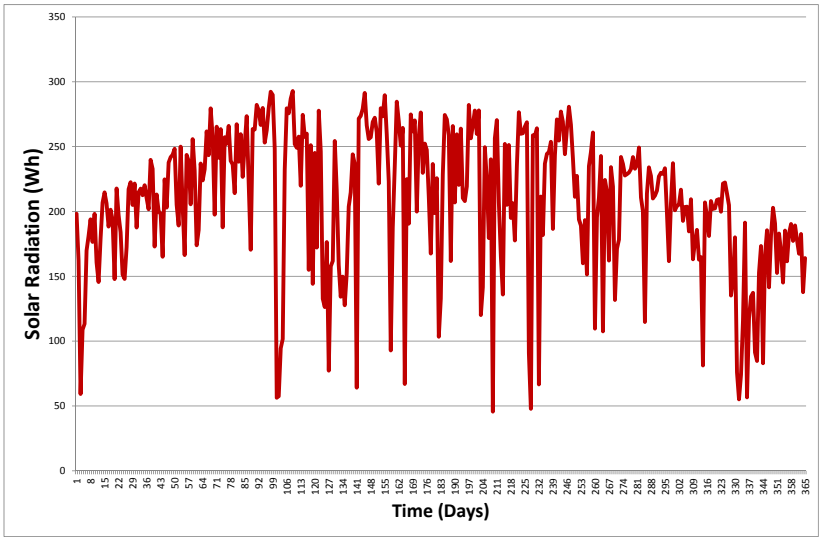

Figure 3. Daily mean solar radiation for Desirade

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 5 shows an illustration of the Fourier series model of the data.

When the Fourier series contribution is subtracted from the data series, the residual series is then modelled with the coupled autoregressive and dynamical system approach, details of which are given in (3). We now present illustrative results (with the seasonality added back in) for two separate days, one clear and one with clouds passing on a number of occasions. See Figures 6 and 7 for these two situations respectively. It should be noted that the normalised root mean square error for this in sample forecast was 20.8%, as compared to 18.5% for the same type of analysis for the Mildura data as reported in (3). This is a similar error but would point to a higher incidence of passing clouds. We suggest that in general the climate of Mildura would indicate that it is in general clearer than even the clearest part of Guadeloupe, that of Desirade.

The following table gives a summary of error measures for the three sites. The error measures are median absolute percentage error(MeAPE), mean bias error (MBE) and normalised root mean square error (nRMSE).

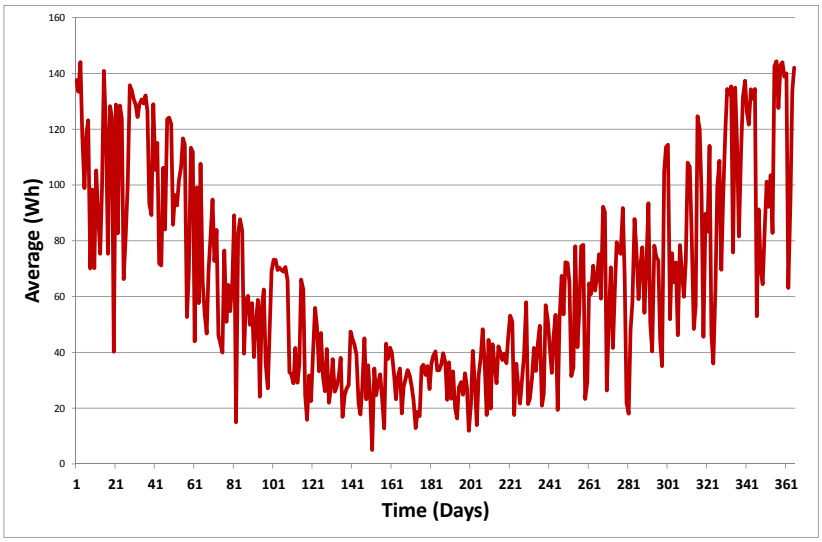

Figure 4. Daily mean solar radiation for Mildura

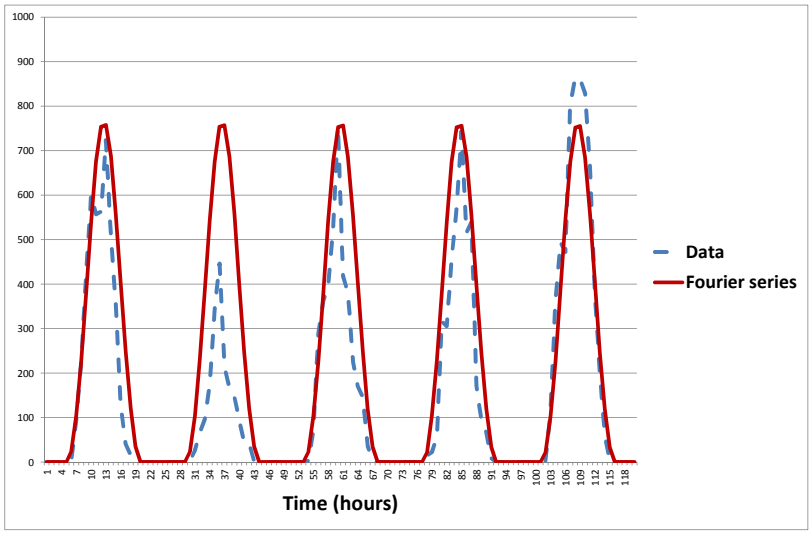

Figure 5. Five days solar radiation and corresponding Fourier series representation

There are two things worth mentioning with these measures. We use the median APE instead of the mean APE since the divisor is at times quite small in this calculation, resulting in an abnormally large error when in fact the absolute error itself is quite small. Also, some authors are suggesting that for normalising the RMSE, one should use a clear sky value in the denominator rather than the mean of the data. We shall in this discussion use the mean, but reserve the right to think again in future as to the preferred approach. It should be noted that the trend toward using a clear sky value mirrors the approach often taken in error analysis for wind farms, using the installed capacity rather than mean output as the normalising constant.

	Desirade	Fouillole	Petit-Canal
MeAPE	11.9%	16.6%	12.1%
MBE	2.99	1.55	3.35
nRMSE	20.8%	26.0%	21.4%

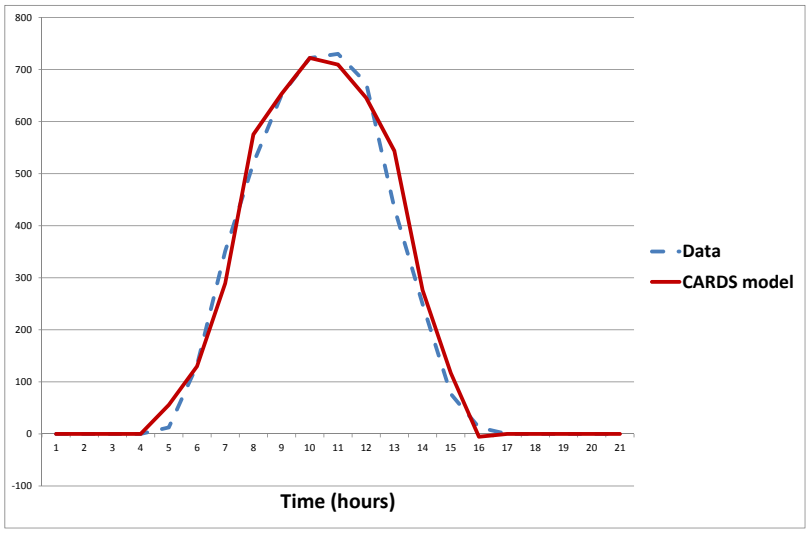

Figure 6. The CARDS model on a clear day

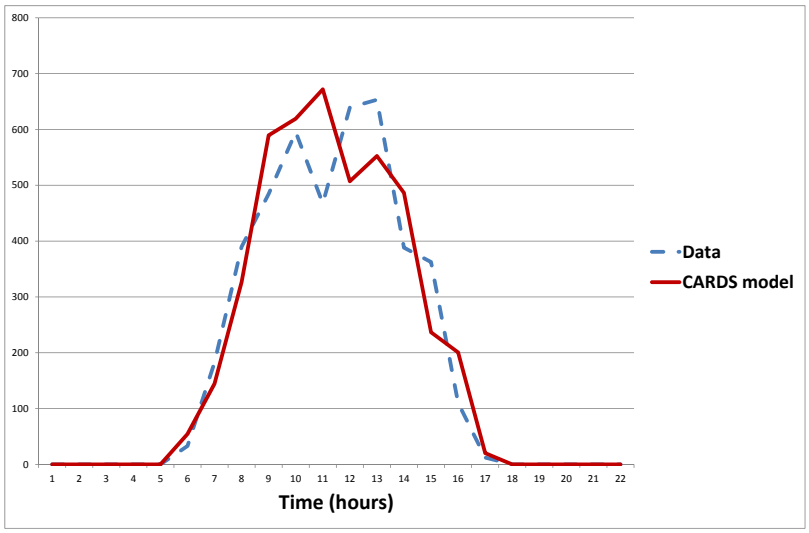

Figure 7. The CARDS model on a cloudy day

2.1 Multivariate model

The preceding modelling has pre-whitened the three series, in other words, identified, analysed and removed any structure that depends on the individual series. What is left are three residual time series, $X_i, \quad i = 1, 2, 3$. What we are going to do now is to determine if there is any cross correlation between these three series. Here, we designate X_1 as the series corresponding to Desirade, X_2 Fouillole, and X_3 Petit-Canal. When we perform a simple correlation at time t, we find that all three pairwise correlations are small but significant. They are

	Desirade	Fouillole	Petit-Canal
Desirade	1	0.177	0.243
Fouillole	0.177	1	0.278
Petit-Canal	0.243	0.278	1

We next checked the correlations for each site with the other sites at one hour time lag. In each case, all proved to be significant, although once again positive but not large. They are given in the following tables.

	Fouillole	Petit-Canal
Desirade	0.126	0.117

	Desirade	Petit-Canal
Fouillole	0.150	0.130

	Fouillole	Desirade
Petit-Canal	0.232	0.199

This allows us to write down equations for the X_i, in the following form, similar to the vector autoregressive (VAR) model.

$$\begin{aligned} X_{1,t} &= \alpha_{11} + \alpha_{12}X_{2,t} + \alpha_{13}X_{3,t} + \beta_{12}X_{2,t-1} + \beta_{13}X_{3,t-1} + Z_{1,t} \\ X_{2,t} &= \alpha_{22} + \alpha_{21}X_{1,t} + \alpha_{23}X_{3,t} + \beta_{21}X_{1,t-1} + \beta_{23}X_{3,t-1} + Z_{2,t} \\ X_{3,t} &= \alpha_{33} + \alpha_{31}X_{1,t} + \alpha_{32}X_{2,t} + \beta_{31}X_{1,t-1} + \beta_{32}X_{2,t-1} + Z_{3,t} \end{aligned}$$

After performing the regression, this becomes

$$\begin{aligned} X_{1,t} &= -3.11 + 0.0824X_{2,t} + 0.1830X_{3,t} + 0.0498X_{2,t-1} + 0.0819X_{3,t-1} + Z_{1,t} \\ X_{2,t} &= 0.2766X_{1,t} + 0.1398X_{3,t} + 0.1214X_{1,t-1} + 0.1005X_{3,t-1} + Z_{2,t} \\ X_{3,t} &= 0.2093X_{1,t} + 0.1798X_{2,t} + 0.2024X_{1,t-1} + 0.1212X_{2,t-1} + Z_{3,t} \end{aligned}$$

The expectation is that the Z_i will each be independent and identically distributed on an individual basis. If that is so, then for each series, one will be able to use the principle of ergodicity to estimate the variance of the forecast for each series at each time t, by calculating the three sample ensemble variances. To check this condition, we squared each Z_i and then looked at the sample autocorrelation and partial autocorrelation functions (SACF, SPACF). The SACF for each series decayed slowly with a sinusoidal variation due to the series all being zero at night. An example SPACF representation for Desirade is given in Figure 8.

What this means is that we need to fit an Autoregressive Conditional Heteroscedastic (ARCH) model to the squared residuals, using them as the best estimator for the variance of the series (4). What this means is that when forecast the value of the series for time $t + 1$, when we have the history up to time t, we also need to forecast the variance at time $t + 1$ using this ARCH model and knowledge of the squared final residuals up to time t. We can use this forecasted variance to construct error bounds on the forecasts. After estimating the parameters for the ARCH models, we have

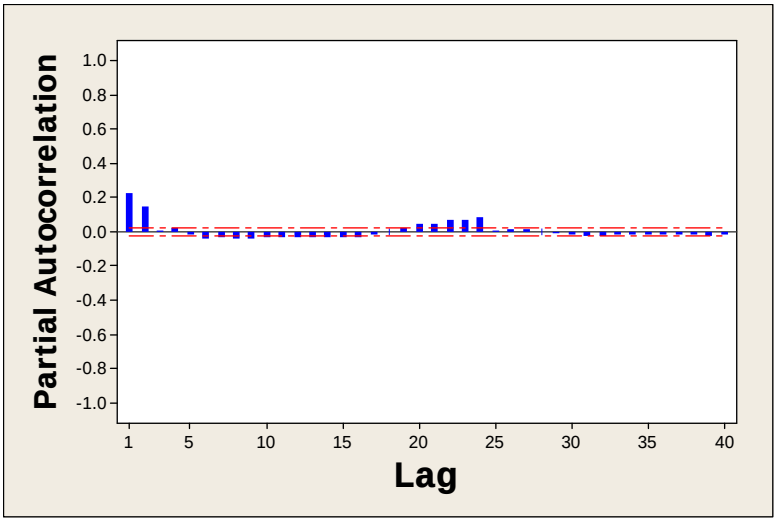

Figure 8. SPACF for squared residuals for Desirade

$$
\begin{aligned}
\sigma^2_{X_{1,t}} &= 0.2259Z^2_{1,t-1} + 0.1816Z^2_{1,t-2} \\
\sigma^2_{X_{2,t}} &= 0.2458Z^2_{2,t-1} + 0.1583Z^2_{2,t-2} + 0.0514Z^2_{2,t-3} + 0.0633Z^2_{2,t-4} \\
\sigma^2_{X_{3,t}} &= 0.2677Z^2_{3,t-1} + 0.1344Z^2_{3,t-2} + 0.0569Z^2_{3,t-3}
\end{aligned}
$$

3 CONCLUSION

This article gives an introduction to the tools needed for the multivariate forecasting of solar radiation. I have demonstrated the techniques required to pre whiten the data (as far as possible) for a single site. The next step involved checking the cross correlation of the noise between sites, finding significant correlation between the sites at time t and also between the values at time t and time $t-1$. A subsequent one lag multivariate time autoregressive model was estimated. It was hoped that the three noise variables resulting from this modelling would be iid. However, this was not to be the case and all three noise series exhibited conditional heteroscedasticity. In this case, ARCH models sufficed to describe this behaviour. Two immediate tasks come to mind. One is to utilise the models for the levels of the three series and the corresponding ARCH models for the conditional variances to demonstrate the probabilistic forecasting of multivariate solar time series. The other issue is to check for all these effects for time series of solar radiation for multiple sites in other locations.

ACKNOWLEDGEMENT

This work has been performed with funding from the Australian Renewable Energy Agency.

REFERENCES

[1] Boland, J. (1995) Time Series Analysis of Climatic Variables, *Solar Energy*, Vol. 55, No. 5, pp. 377-388.

[2] Boland J. (2008) Time series and statistical modelling of solar radiation, *Recent Advances in Solar Radiation Modelling*, Viorel Badescu (Ed.), Springer-Verlag, pp. 283-312.

[3] Jing Huang, Malgorzata Korolkiewicz, Manju Agrawal and John Boland, (2013) Forecasting solar radiation on an hourly time scale using a coupled autoregressive and dynamical system (CARDS) model, *Solar Energy*, 87, pp. 136-149.

[4] Ruey Tsay, (2005) *Analysis of Financial Time Series*, Second Edition, Wiley Series in Probability and Statistics, Wiley-Interscience, Hoboken, New Jersey.