

# Empirical Orthogonal Function Analysis of wind farm power output

Lynne McArthur <sup>a</sup>

<sup>a</sup>*School of Mathematical and Geospatial Sciences, RMIT University, Melbourne, Australia*  
Email: [Lynne.McArthur@rmit.edu.au](mailto:Lynne.McArthur@rmit.edu.au)

**Abstract:** One of the major challenges to supporting, facilitating and developing wind generated power is matching supply and demand. Wind generated power is obviously subject to fluctuations due to variation in the wind. There is also a shutdown mechanism which is employed when the wind becomes very strong which prevents damage to the turbines. Thus, when the wind is light or very strong, there is no power generated. Predicting the output from the turbines is currently an important research topic .

In the past, time series analysis and other methods have been employed in order to better understand the nature of the data. These include ARMA and GARCH models as well as relatively new methods of detrended fluctuation analysis, (Ward et al., 2009; Boland et al., 2009; Magnano and Boland, 2007; Kavasseri and Nagarajan, 2004). The behaviour of complex time series, as is the wind farm power output, and as is seen in the financial sector has been and is currently well researched. Some of these time series have been found to be non-linear, stochastic and chaotic, and are notoriously difficult to model. There is evidence however, that there could be some scaling behaviour apparent in these series (Weron and Przybylowicz, 2000).

The particular objective of this project is to reduce the error in predicting the power supply generated by wind farms (or the individual turbines) five minutes into the future so that the power company is able to guarantee the promised power. Currently this is not possible due to the financial risks involved in not delivering the guaranteed amount. Apart from climatic research, there is not much in the way of research into spatial correlations where records are available of similar variables at different locations. Since the temporal nature of the time series is currently being thoroughly investigated by many researchers, this study explores the possible spatial correlations between wind farms.

This paper describes the investigation into the spatial correlation between five South Australian wind farms using empirical orthogonal function analysis, (EOF). If there is spatial correlation between farms, this can be used to better predict what might happen at one farm based upon what has happened at a different farm at some earlier time. There has also been speculation among experts that power output of sites along a weather front behave in a similar way.

One of the main difficulties with determining the power output of wind turbines is that wind speed is not a useful predictor. The wind speed isn't actually measured at the height of the blades and if instruments were installed to measure the wind speed immediately in front of the turbine blades, then the likelihood of correct measurements would be compromised by their existence.

This study has focused on the spatial correlations between farms and the initial results from this study indicate that there is evidence of some spatial correlation between some of the farms.

**Keywords:** Empirical Orthogonal Function Analysis, wind power prediction, electricity generation, spatial correlation

## 1 INTRODUCTION

Eigenanalysis, also known as Empirical Orthogonal Function (EOF) analysis, is a technique used for determining the dominant modes of variation in data sets (Hannachi, 2004; Aubrey and Emery, 1983), and distinguishing between spatial and temporal variation. It is closely related to principal component analysis (PCA) and Singular Value Decomposition (SVD), although there are subtle differences (Hannachi, 2004). Given a spatial and temporal set of data, as is the wind farm power output data, the method of eigenanalysis can be used to isolate the temporal variation from the spatial variation in the data.

EOF analysis was initially applied in the social and behavioural sciences but since 1951 it has been widely used for climate and weather prediction, and to other areas of climatology such as sea level change, sea surface temperatures, climatology models and for long-term coastal profile evolution models (Craddock, 1975; Aubrey and Emery, 1983; Cahalan et al., 1997; Larson et al., 1999; Chiera, 1999; Giese and Carton, 1999; Uddstrom and Oien, 1999; Kim and North, 1999; Kim and Wu, 1999; Tseng et al., 2000; Robertson et al., 2000; Walland et al., 2000; Ji et al., 2000; Hannachi, 2004). It has also been used to separate the temporal and spatial variation in catch and effort data of South Australian commercial fisheries, (McArthur, 2003).

The method involves calculating the eigenvectors (functions) and associated eigenvalues of the covariance matrix of a given spatial and temporal data set. The eigenvectors are termed ‘empirical orthogonal functions’, (EOFs) and the set of corresponding eigenvalues are the variance spectra. The first eigenfunction best describes the data in a least squares sense; the second eigenfunction best describes the residual of the data in the least squares sense, and so on. In addition, all eigenfunctions are orthonormal, (Aubrey and Emery, 1983; McArthur, 2003).

The EOFs correspond to a statistically optimal description of the data with respect to how the variance is concentrated in the functions. The functions are ranked according to decreasing eigenvalues and the variance typically drops off very quickly, thus only a limited number of the lowest ranked functions are required to explain most of the variance in the data. This property sometimes motivates the use of EOF analysis as a data reduction technique or to separate signal from noise. Thus, for an  $n \times m$  data matrix, the first  $k$  dominant functions, where  $k \ll \min(n, m)$ , can be considered to represent the deterministic part of the data and the remaining functions to represent the noise.

Broadly, the objectives of eigenanalysis are to

- a) reduce the dimensionality of the problem by replacing the measured and intercorrelated variables by a smaller number of uncorrelated variables, (the EOFs), particularly in meteorological applications,
- b) to interpret these uncorrelated variables in terms of recognisable physical processes, (Craddock, 1975)
- c) separate the spatial and temporal variance in the data.

The process can reduce the dimensionality of a given data set by a factor of ten and the first three or four functions can generally be used to capture most of the variation in the data. There is no set rule for selecting the truncation point, although in practice the scree plot can inform the decision.

One of the main problems with this analysis is that it can be difficult to identify the underlying physical processes that the EOFs represent. However, reducing the dimensionality simultaneously results in removal of the ‘noise’ from the data set, and provides a good basis for regression analysis, assuming the material is amenable to linear treatment (Craddock, 1975), and that there are predictor variables available.

The technique is applied to five wind farms from South Australia, to determine if spatial correlation is evident.

## 2 METHOD

Given a set of data composed of both temporal and spatial information, let  $D$  be the matrix with  $n$  rows of temporal data and  $m$  columns of spatial data. That is, for each location, there is a column of measurements taken at consecutive time steps. In order to apply the EOF analysis, we need a square positive definite matrix, so firstly the covariance matrix of  $D$  is determined. Given that  $D$  is an  $n \times m$  array, two covariance

matrices can be formed: one will have dimensions  $n \times n$  and the other  $m \times m$ . These will be the temporal and spatial covariances matrices respectively.

## 2.1 Background Theory to EOF

Let  $\mathbf{A}$  be a positive definite symmetric  $n \times n$  square matrix. The associated  $\lambda_i$  for  $i = 1, 2, \dots, n$  are positive real eigenvalues and the  $\mathbf{v}_i$  are orthonormal eigenvectors of  $\mathbf{A}$ . From spectral decomposition theory, the following equations hold:

$$\begin{aligned} \mathbf{v}_i \cdot \mathbf{v}_i &= 1 \\ \mathbf{v}_i \cdot \mathbf{v}_j &= 0 \quad \text{for } i \neq j \\ \mathbf{A} &= \lambda_1 \mathbf{v}_1 \mathbf{v}_1^T + \lambda_2 \mathbf{v}_2 \mathbf{v}_2^T + \dots + \lambda_n \mathbf{v}_n \mathbf{v}_n^T \end{aligned} \quad (1)$$

$$\mathbf{A} \mathbf{v}_i = \left( \sum_{j=1}^n \lambda_j \mathbf{v}_j \mathbf{v}_j^T \right) \mathbf{v}_i = \lambda_i \mathbf{v}_i \quad (2)$$

since  $\mathbf{v}_j \mathbf{v}_j^T = \mathbf{I}$ , where  $\mathbf{v}_i^T$  is the transpose of  $\mathbf{v}_i$ .

The  $n \times m$  data matrix contains five minute power output values for each farm. The positive definite symmetric covariance matrices are calculated using the following equations (Aubrey and Emery, 1983), where  $\mathbf{X}$  is the data set with the arithmetic mean for each farm removed.

The spatial mean is removed from the data and the covariance matrices (equation 3) calculated. The eigenequations are then solved and the dominant functions are identified. A scree plot is constructed to obtain a visual indication of the degeneracy of the eigenvalues, and the cumulative contribution to the variance is calculated to determine the number of functions required to reconstruct the data sets, with the noise removed.

Let  $\mathbf{C}_t$ , be a positive definite symmetric  $n \times n$  temporal covariance matrix. Then,  $\mathbf{C}_x$  will be the associated  $m \times m$  spatial covariance matrix.

$$\begin{aligned} \mathbf{C}_x &= \frac{1}{m n} \mathbf{X} \mathbf{X}^T \\ \mathbf{C}_t &= \frac{1}{m n} \mathbf{X}^T \mathbf{X}. \end{aligned} \quad (3)$$

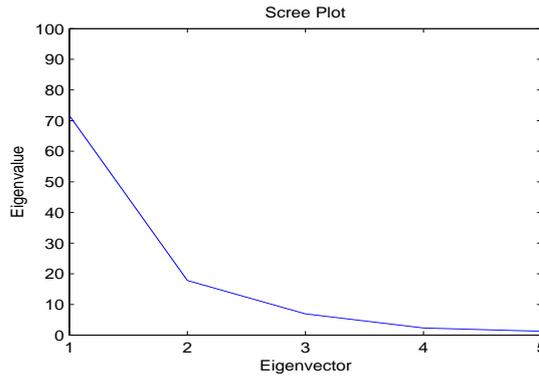
The spatial and temporal eigenfunctions,  $\mathbf{v}_k$  and  $\mathbf{u}_k$  respectively, for  $k = 1, 2, \dots, N$  where  $N = \min(n, m)$ , are determined by solving the following two equations for the covariance matrices  $\mathbf{C}_x$  and  $\mathbf{C}_t$ :

$$\begin{aligned} \mathbf{C}_x \mathbf{v}_k &= \lambda_k \mathbf{v}_k & \text{and} & \\ \mathbf{C}_t \mathbf{u}_k &= \lambda_k \mathbf{u}_k. \end{aligned} \quad (4)$$

The original data set can be reconstructed using equation (6), by limiting the reconstruction to the first  $l < N$  functions. In general the first three to five eigenfunctions explain 95% of the variability, and in so doing, one can effectively eliminate the noise from the data. For this analysis, the number of eigenfunctions required was determined by the degeneracy of the eigenvalues (degenerate eigenvalues are those which are not distinguishable) and it was determined that the first three functions ( $l = 3$ ) were sufficient to explain  $> 95\%$  of the variation. Let  $X_{x,t}$  be the matrix containing the reconstructed data;

$$\mathbf{X}_{x,t} = \sum_{j=1}^l \left( \sqrt{(\lambda_j m n)} \mathbf{u}_j \mathbf{v}_j^T \right). \quad (6)$$

The spatial mean is removed from the data and the covariance matrices (equation 3) calculated. The eigenequations are then solved and the dominant functions are identified. A scree plot is constructed to



**Figure 1.** Scree plot illustrating the degeneracy of the eigenvalues.

obtain a visual indication of the degeneracy of the eigenvalues, and the cumulative contribution to the variance is calculated to determine the number of functions required to reconstruct the data sets, with the noise removed.

## 2.2 Data

Data has been provided by the Australian Energy Market Operator, AEMO, and consists of 219 days of 5 minute power output data (approximately 63,000 data points) from five separate farms in South Australia. Confidentiality prohibits the association of the data to the farm, hence the farms are identified by number.

## 3 RESULTS

In this study we are primarily interested in spatial correlations; this section describes the results of the spatial analysis for the five farms.

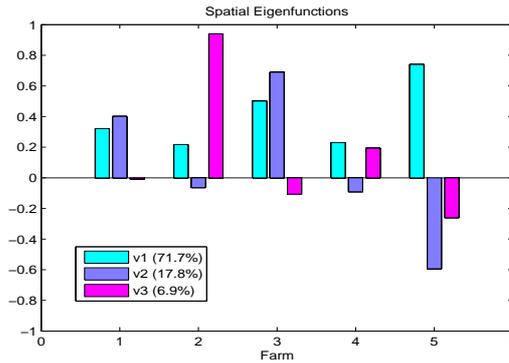
The eigenvalues ( $\lambda_k$ ), ( $k = 1, 2, \dots, 5$ ) are identical for equally ranked spatial and temporal functions and each  $\lambda_k$  represents the amount of variance in the direction of each of the orthogonal EOFs. Thus the contribution to the variability of the data by each of the EOFs is given by the eigenvalues. The scree plot, Figure 1, displays the size of the eigenvalues in order of their rank, clearly showing the degeneracy which indicates the number of functions required to reconstruct the ‘noiseless’ data.

The contribution of each of the first three eigenfunctions to the variability of the data is given by the first three eigenvalues which are:

$$\lambda_1 = 71.7\%, \quad \lambda_2 = 17.8\%, \quad \lambda_3 = 6.9\%.$$

The first three eigenfunctions are shown in Figure 2. It can be seen that the first eigenfunctions are all positive for each of the farms. This indicates some spatial similarities, although it isn’t necessarily possible to identify the corresponding physical attributes. The second eigenvalues are strong and positive for farms one and three, and negative for the others. This indicates that a secondary physical property, which explains approximately 18% of the variation in the data, is different for farms one and three compared to the other three farms. The third eigenfunction indicates that farms two and four have similar properties as do three and five. However, since these functions explain less than 8% of the variation, not too much importance can be placed upon this.

In summary, there is evidence that farms one and three have correlated data, and that the others also have some correlation, although not as strong. Once the farms have been identified, factors such as wind speed and direction at farm A, and the position and range of the weather front can be used to help predict the power outlet at farm B. This information is valuable in enabling predictive capabilities from one farm to another and reducing the error in predicting power outlet.



**Figure 2.** The first three spatial eigenfunctions for the five farms.

It is probable that a multiplicity of physical factors cause the variation in the observed data, and for this reason it is often difficult to isolate and identify individual factors, and thus assign physical properties to the EOFs. However, the results indicate that there is some correlation between the sites and therefore this information can be used to enhance the predictability of the power outlet from one farm based on another with some degree of confidence.

#### 4 CONCLUSIONS

Eigenanalysis of the wind farm data has shown that there are spatial correlations between the five farms, which is possibly related to their physical location. There has been suggestion (pers. comm.) of similar wind patterns occurring along lines of weather fronts, so this could be related to the correlation indicated between farms one and three by the second eigenfunction. The eigenvalues indicate how much variation in the data is explained by each of the orthogonal eigenfunctions, which are approximately 72%, 17% and 7% respectively.

The analysis has shown that there is a high probability that the measured power output from one farm can be used as an indication of the power generated by a second farm. In addition, this technique can be used to remove noise from the data, and to separate the spatial and temporal variation in the data set. This itself is a valuable process, as the spatial correlation can be identified between farms.

#### ACKNOWLEDGEMENT

This study was supported by the University of South Australia, School of Mathematics and Statistics.

#### REFERENCES

- Aubrey, D. and K. Emery (1983). Eigenanalysis of recent United States sea levels. *Continental Shelf Research* 2(1), 21–33.
- Boland, J., B. Ridley, and M. Agrawal (2009). Modelling the conditional variance of wind farm output utilising realised volatility. In *Proceedings of the ISES Solar World Congress 2009: Renewable Energy Shaping Our Future, Johannesburg, South Africa, 11 - 14 October 2009*, pp. 25–33.
- Cahalan, R. F., L. E. Wharton, and M. L. Wu (1997). Empirical orthogonal functions of monthly precipitation and temperature over the United States and homogeneous stochastic models. *Journal of Geophysical Research-Atmospheres* 101(D21), 26309–26318.
- Chiera, B. (1999). *Australian Sea Levels, ENSO and Greenhouse Forcing: Modelling and Predictability Analysis*. Ph. D. thesis, University of South Australia.
- Craddock, J. M. (1975). Problems and prospects for eigenvector analysis in meteorology. *The Statistician* 22(2), 133–145.

- Giese, B. S. and J. A. Carton (1999). Interannual and decadal variability in the tropical and midlatitude Pacific Ocean. *Journal of Climate* 12(2), 3402–3418.
- Hannachi, A. (2004). A Primer for EOF Analysis of Climate Data. Department of Meteorology, University of Reading Reading RG6 6BB, U.K., [Online: Accessed July 31, 2011]. URL: <http://www.met.rdg.ac.uk/han/Monitor/eofprimer.pdf>.
- Ji, M., R. W. Reynolds, and D. W. Behringer (2000). Use of TOPEX/Poseidon sea level data for ocean analyses and ENSO prediction: Some early results. *Journal of Climate* 13, 216–231.
- Kavasseri, R. G. and R. Nagarajan (2004, November 20,). Evidence of crossover phenomena in wind speed data. [Online: Accessed March 17, 2010] URL: <http://arxiv.org/abs/physics/0411194>,.
- Kim, Y. J. and G. R. North (1999). EOF-based linear prediction algorithm: Examples. *Journal of Climate* 12(7), 2076–2092.
- Kim, Y. J. and Q. G. Wu (1999). A comparison study of EOF techniques: Analysis of nonstationary data with periodic statistics. *Journal of Climate* 12, 185–199.
- Larson, M., H. Hanson, N. C. Kraus, and J. Newe (1999). Short- and long-term responses of beach fills determined by EOF analysis. *Journal of Waterway Port Coastal & Ocean Engineering-ASCE* 125(6), 285–293.
- Magnano, L. and J. Boland (2007). Generation of synthetic sequences of electricity demand: Application in south australia. *Energy* 32(11), 2230–2243.
- McArthur, L. C. (2003). *Quantification of the relationships between fish populations and seagrass*. PhD Thesis, University of South Australia, Adelaide.
- Robertson, A. W., C. R. Mechoso, and Y. J. Kim (2000). The influence of Atlantic sea surface temperature anomalies on the north Atlantic oscillation. *Journal of Climate* 13, 122–138.
- Tseng, C., C. Y. Lin, S. C. Chen, and C. Z. Shyu (2000). Temporal and spatial variations of sea surface temperature in the East China Sea. *Continental Shelf Research* 20, 373–387.
- Uddstrom, M. J. and N. A. Oien (1999). On the use of high-resolution satellite data to describe the spatial and temporal variability of sea surface temperatures in the New Zealand region. *Journal of Geophysical Research -Ocean* 104, 20729–20751.
- Walland, D. J., S. B. Power, and A. C. Hirst (2000). Decadal climate variability simulated in a coupled general circulation model. *Climate Dynamics* 16(2-3), 201–211.
- Ward, K., M. Korolkiewicz, and B. J. (2009). Modelling residual wind farm variability using HMMs. In R. Anderssen, R. Braddock, and L. Newham (Eds.), *18th World IMACS MODSIM Congress, Cairns, Australia 13-17 July, 2009*, pp. 4467–4473.
- Weron, R. and B. Przybylowicz (2000). Hurst analysis of electricity price dynamics. *Physica A* 283, 462–468.