A Study of Principal Component Analysis Applied to Spatially Distributed Wind Power

Daniel J. Burke, *Student Member*, and Mark J. O'Malley, *Fellow.*

***Abstract*—Multivariate dimension reduction schemes could be very useful in limiting the number of random statistical variables needed to represent distributed wind power spatial diversity in transmission integration studies. In this paper, principal component analysis (PCA) is applied to the covariance matrix of distributed wind power data from existing Irish wind farms, with the eigenvector/eigenvalue analysis generating a lower number of uncorrelated alternative variables. It is shown that though uncorrelated, these wind components may not necessarily be statistically independent however. A sample application of PCA combined with multivariate probability discretisation is also outlined in detail. In that case study, the capability of PCA to reduce the number and prioritise the order of the alternative statistical variables is key to potential wind power production costing simulation efficiency gains, when compared to exhaustive multi-year time series load flow investigations.**

***Index Terms*-- power transmission, principal component analysis, statistics, time series, wind energy.**

I. Nomenclature

i - multivariate component index

t - number of multivariate historical wind power data observations

n - number of original wind power random variables

$\boldsymbol{X}$ - matrix of multivariate historical wind power data observations

$\boldsymbol{C_x}$ - covariance matrix of $\boldsymbol{X}$

$\boldsymbol{L}$ - PCA eigenvalue matrix with diagonal values l_i

$\boldsymbol{E}$ - PCA eigenvector matrix with columns e_i

σ - chosen number of retained components

$\boldsymbol{Q}$ - principal component time series matrix

$\overline{X}$ - vector of original wind variable capacity factor values

$\boldsymbol{X_{REC}}$ - reconstructed original variables for $\sigma<n$

$\boldsymbol{R}$ - reconstructed original variable residual errors

This work was conducted in the Electricity Research Centre, University College Dublin, Ireland, which is supported by the Commission for Energy Regulation, Bord Gáis Energy, Bord na Móna Energy, Cylon Controls, EirGrid, ESB Energy International, ESB Energy Solutions, ESB Networks, Gaelectric, Siemens, SSE Renewables, SWS Energy and Viridian Power & Energy. The work of D. J. Burke was supported in part by the Sustainable Energy Authority of Ireland through a postgraduate research scholarship from the Irish Research Council for Science Engineering and Technology.

The authors are with the School of Electrical, Electronic and Mechanical Engineering, University College Dublin, Dublin 4, Rep. of Ireland (e-mail daniel.burke@ucd.ie, mark.omalley@ucd.ie; telephone +353 1 716 1857; web http://erc.ucd.ie).

II. Introduction

Due to common weather patterns, wind power will exhibit reasonably strong statistical dependency within a small regional area. For more geographically distinct locations, this dependency reduces with increasing separation by distance [1]. While there will be many instances when wind power production at such distinct sites will be similar, it could be a relatively common occurrence that one wind farm might be at maximum power outpu0t with simultaneously little or no wind production at the others. Thus while generation expansion studies can justifiably use uni-variate statistical models for total system wind power production [2], in contrast the spatially distributed context of the transmission network planning problem requires an effective representation of multivariate wind statistical interdependency. Maintaining a distinct statistical representation for each potential wind farm location will result in a very large number of random variables [3]. If a more approximate model can capture the salient features of each location's uni-variate marginal distribution (i.e. the statistical behaviour of each wind farm's power output considered alone), as well as the spatial multivariate interdependency (i.e. the statistical relationship between power production at different sites), then considerable computational efficiency and/or model specification simplicity can result in wind power transmission study applications.

Principal component analysis (PCA) is a multivariate dimension reduction technique applicable to large statistical datasets – for example PCA was used to investigate dependency in transmission network flows in [4]. By performing eigenvector/eigenvalue analysis of the multivariate wind power covariance matrix, an arbitrarily lower number of alternative statistical variables (the 'principal components') could be determined [5]. Interestingly, the resultant variables are uncorrelated, and the sum of the eigenvalues corresponding to the retained principal components relates to how much of the original wind variables' variance is explained by the lower dimension transformation.

In this paper, the application of PCA to the distributed wind power statistical dimension reduction problem is analyzed in detail. The utility of PCA combined with multivariate probability discretisation to the production costing simulation of yearly wind power behaviour is also highlighted. It will be shown for this sample application how the capability of PCA to reduce the number and prioritise the order of the

transformed statistical variables is key to potential simulation efficiency gains.

Section III outlines the relevant PCA theory and related investigations for the distributed wind power case. Section IV describes the real historical wind power production data from multiple locations on the Irish power system that is used along with a simple test power system for study – some applications of PCA illustrating the computational accuracy/complexity trade-offs of dimension reduction and multivariate wind probability discretisation are outlined. Section V gives results for the PCA study and applications, with discussions and conclusions outlined in Sections VI and VII respectively.

III. Wind Power Component Analysis Study

A. Principal Component Analysis of Wind Power Data

For a given $t \times n$ matrix $\boldsymbol{X}$ of t observations from n distributed wind power random variables, the wind power covariance matrix $\boldsymbol{C_x}$ is a symmetric $n \times n$ matrix. Any symmetric nonsingular matrix such as $\boldsymbol{C_x}$ can be transformed to a diagonal matrix $\boldsymbol{L}$ through pre-multiplication and post-multiplication by a given orthonormal matrix $\boldsymbol{E}$ as in (1). $l_1, l_2, .. l_i .. l_n$, the diagonal values of $\boldsymbol{L}$, are the eigenvalues of the $\boldsymbol{C_x}$ matrix, and can be determined by solution of the classic characteristic equation (2), where I is the identity matrix of size n. The eigenvectors $e_1, e_2, .. e_i .. e_n$ comprising the columns of $\boldsymbol{E}$ are given by solutions of equations (3a) and (3b). For very large values of n, the eigenvalues and eigenvectors are usually determined by iterative numerical techniques [5].

The $n \times t$ principal wind power components matrix $\boldsymbol{Q}$ is determined by the linear orthonormal transform (4) – the original variable set $\boldsymbol{X}$ is first centered by subtracting the column mean vector of capacity factor values $\overline{X}$, and its transpose then pre-multiplied by $\boldsymbol{E}^T$. This essentially corresponds to a translation and rotation of the original co-ordinate axes, with each new axis chosen to explain as much of the variance in the original wind power dataset as possible. The rows of $\boldsymbol{Q}$ correspond to the resultant uncorrelated principal components.

$$E^T.C_x.E = L \tag{1}$$

$$\det(C_x - lI) = 0 \tag{2}$$

$$[C_x - l.I]k_i = 0 \tag{3a}$$

$$e_i = \frac{k_i}{\sqrt{k_i{}^T.k_i}} \tag{3b}$$

$$Q = E^T.(X - \overline{X})^T \tag{4}$$

It can be shown that the trace of matrix $\boldsymbol{L}$, $Tr(\boldsymbol{L})$ is equal to $Tr(\boldsymbol{C_x})$. Each eigenvalue, corresponding to the variance of one principal wind component, therefore relates to how much of the variance in the original multivariate wind power data which that component explains. For highly correlated wind datasets, the first few principal components will explain the majority of the original variance. The eigenvalues can be plotted in order of decreasing size – this approach is sometimes used in PCA as an approximate visual technique to determine how many principal components should be retained. The approximate 'scree' or 'broken-stick' tests propose the decreasing eigenvalue plot corner-point discontinuities as the natural number of components to retain, though more sophisticated statistical tests have also been proposed [5]. Some PCA implementations advise the use of the correlation matrix instead of the covariance matrix for $\boldsymbol{C_x}$, giving different resultant principal components [5]. The correlation matrix is most useful in situations where the original variables have different units of measurement, or when there are other significant differences in their variances. For the wind PCA study of this paper however, the consistent scaling of the nominal 1MW wind power time series in Section IV allows the use of the covariance matrix.

B. Wind Power Data Reconstruction Residual Effects

For wind power and transmission system load flow studies using any chosen number of retained principal components σ, a reconstructed estimation of the original wind power time series $\boldsymbol{X_{REC}}$ can be determined in (5) by inverting the PCA transform of (4), using only the relevant truncated E and Q matrix rows and columns. If less than the overall n principal components are used, this will result in a non-zero residual error matrix $\boldsymbol{R}$ with respect to the original variables of $\boldsymbol{X}$, as in (6).

$$X_{REC} = (E^T)^{-1}{}_{trunc}.Q_{trunc} + \overline{X} \tag{5}$$

$$R = X - X_{REC} \tag{6}$$

C. Multivariate Wind Probability Discretisation

Probability discretisation procedures attempt to reduce the computational burden of multivariate statistical model simulation by grouping similar historical data observations for a given set of random variables. A simple example of this concept is illustrated for two random statistical variables in Fig.1, where one representative case in the centre of each square 'bin' is probability-weighted by the number of original data-point observations contained within it. In this way, the ~ 130 original random data samples implicitly describing the statistical dependency between the two variables are effectively represented by a much lower number of cases (4×4=16), each then with a modified probability-weighting.

Note however that if the range of each variable had been evenly binned into 5 regions instead of 4, then the positions of the representative cases would perhaps map the original data-sample spread with slightly more precision (i.e. less error associated with discretising what is in truth a continuous probability distribution). On the other hand there would have been a less efficient reduction in case dimensionality however (i.e. 5×5=25). Equally critical is the fact that if there was a third random statistical variable (and therefore a three-dimensional statistical dependency space), then the efficiency of the discretisation procedure would furthermore be considerably less (i.e. $4 \times 4 \times 4$=64). The benefit of a probability discretisation approach to reducing the cardinality of samples used to represent a distribution is therefore conditional on the number of statistical variables present, as

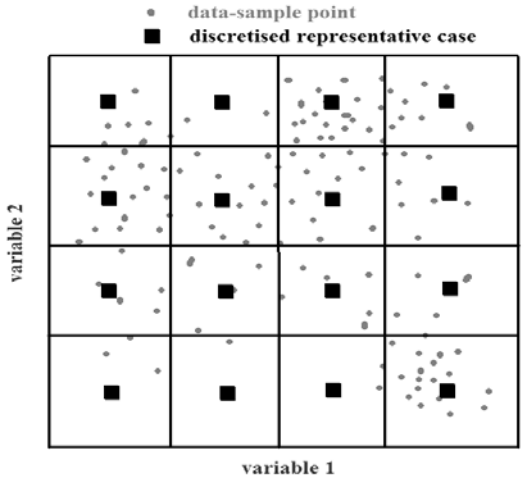

Fig.1 – Conceptual example of probability discretisation for two random variables.

well as the respective density of discrete bins used to categorise the range of their values.

This type of discretisation procedure was studied for the multivariate wind power data case in [6] - individual wind power and load demand historical time series samples were grouped into uniformly sized multi-dimensional histogram bins (analogous to Fig.1). A multidimensional wind volume is generally quite 'full' with samples however, as even medium-scale geographical separation of wind locations on the transmission network results in little grouping around the main multidimensional diagonal. The efficiency/accuracy trade-off of such probability discretisation approaches directly applied to raw multivariate wind power data is therefore low in most cases - e.g. ~ 20% reduction in model sample size for the seven wind farm case study in [6]. Discretisation approaches applied directly to the wind data may have greater justification for distribution system applications, as later investigated in [7], where wind farms are much more closely located than in the transmission system case.

A more refined approach might alternatively try to reduce the number of variables prior to probability discretisation with an intelligent dimension reduction scheme such as PCA. Through analysis of the PCA eigenvalues, the selection of which principal components should be retained is decided. Equally important however is the fact that in many cases, the eigenvalues of even the retained components may have different relative importance. Therefore for a subsequent probability discretisation in the transformed principal component domain, the binning density of each retained component can be tailored with respect to its relative importance. This will have a significant influence on the overall efficiency of the probability discretisation approach, and is one illustration of the potential of multivariate component analysis procedures.

IV. Multivariate Data and Test System Analyses

A. Irish Regional Wind Power Time Series Data

Fig.2 illustrates the locations of the Irish wind farms used as the database for this study. Geographically adjacent wind

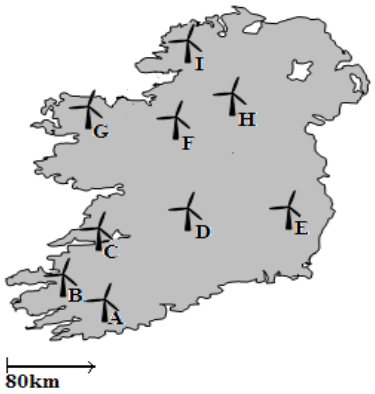

Fig.2 - Irish wind power zones used in this study (each with 2-5 wind farms).

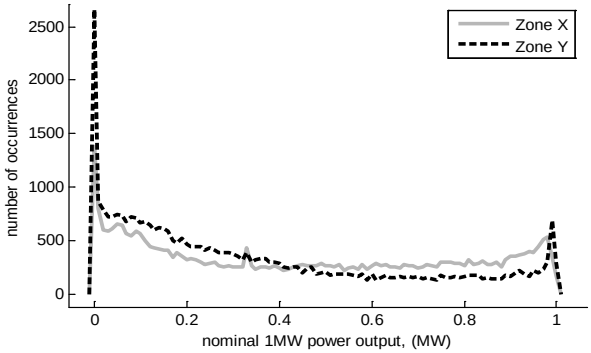

Fig.3 – Histograms of marginal probability densities for Zones X, Y.

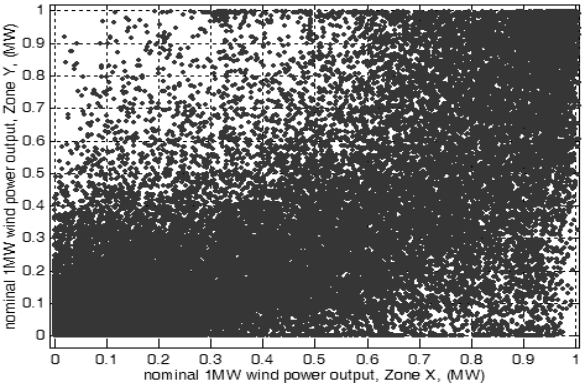

Fig.4 – Scatter plot of nominal wind power output for Zones X, Y.

farms were arbitrarily grouped into nine wind clusters Zone A, B, C, D, E, F, G, H, and I, with each zone cluster based on real data from 2 to 5 close-by individual existing wind farms. Each wind zone was modeled by summing the respective wind farm power time series, and then rescaling by the zone's total capacity to give consistent 1-MW nominal wind power time series models for each region. Synchronously recorded historical power output data from the year 2008 was used (inherently representing the relevant marginal statistics and multivariate dependency), arbitrarily taken at 2-hourly intervals from the original 15-min recordings, thus giving 4392 multivariate samples overall.

The marginal probability density functions (i.e. the probability density function for a single site) of power output at two typical Irish wind regions, Zones 'X' and 'Y', are illustrated in Fig.3 (X and Y are not linked explicitly to Fig.2 for commercial sensitivity reasons). Clearly the individual wind power output patterns over an extended timeframe

correspond to non-parametric statistical distributions (resulting from passing the Weibull wind speed distribution through the non-linear turbine power curves). It should be noted that such distributions cannot be described by first- and second-order statistical moment derived mean and variance information alone. The scatter plot of their joint power production, as illustrated in Fig.4, furthermore outlines their non-parametric bi-variate statistical dependency. Fig.4 also emphasizes the unsuitability of naive probability discretisation approaches directly applied to a multivariate wind power dataset as in [6]. If each of the n wind zones were modeled by d discrete binning density, then the spread of the scatter plot in Fig.4 (each of the $(10\times10=100)$ bins contains at least 1 data sample) would suggest that d^n, the maximum number of multidimensional probability-weighted discrete cases, would be intolerably large if more than 3 or 4 wind zones are studied i.e. the 'curse of dimensionality'.

B. Test Power System Information

The test power system used for the economic dispatch and power flow studies of this paper is illustrated in Fig.5. This has a 35-bus, 54-line network, denoted as 'Area 1' (based on a very simplified model of the Irish 'All-Island' 220/275/400KV high-voltage transmission system). It contains a mixture of base-load and mid-merit fossil-fuel (coal and peat) steam turbine generation, combined-heat-and-power gas plants (CHP), combined-cycle gas turbines (CCGTs), higher-efficiency aero-derivative gas turbines (ADGTs), lower-efficiency open-cycle gas turbines (OCGTs), as well as a few gas/oil-distillate 'peaking' units, amounting to 10.4GW conventional plant capacity overall. 500MW of HVDC interconnection capacity to a much larger separate power system denoted as 'Area 2' (based on an approximate model of the Great Britain generation portfolio) is available at both buses 12 and 34, with these interconnectors denoted as 'IC-1' and 'IC-2' in Fig.5. Conventional plants in Area 2 are grouped approximately into multiple generation capacity blocks of similar plant-type, all connected at a single transmission node. Conventional plant performance data, fuel prices, load profile, load magnitude (accounting for projected load growth to an Area 1 maximum peak value of 9.61GW), and the assumed load geographic distribution are mainly consistent with [8]. Coincident load profile information for Area 2 was sourced from [9]. Additional information on the test network branch reactance and thermal capacity parameters, the assumed system geographical load spread, and the conventional generation portfolio network locations as applied in this investigation are given in the Appendix section of [10]. The wind power collective Zones A, B, C, D, E, F, G, H, and I are modeled in Fig.5 as connected to network buses 3, 4, 9, 17, 12, 25, 15, 28, and 30 respectively. Wind capacity installation in Area 2 was assumed zero – the performance of statistical component analysis for wind power output in Area 1 is of primary interest. All model development for this paper was carried out in MATLAB [11] and GAMS [12], using the MATLAB/GAMS interface available at [13].

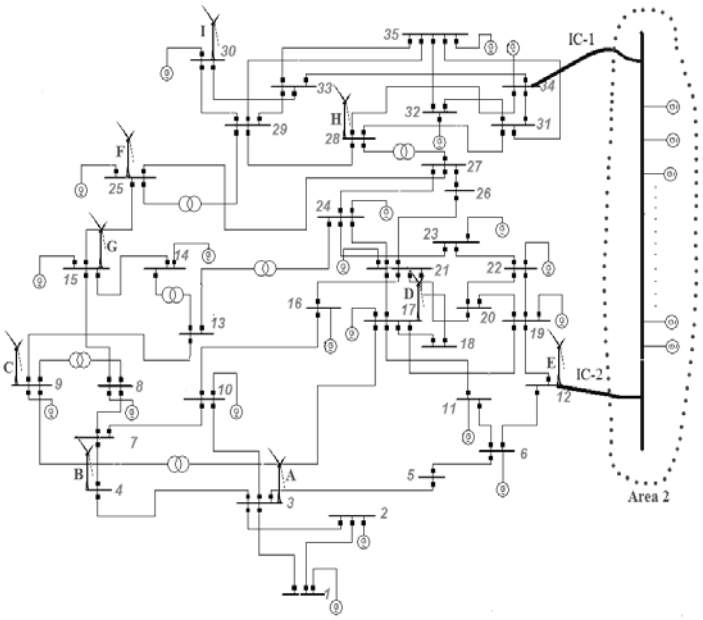

Fig.5 - Test power system network schematic and wind zone locations.

TABLE-I
ZONAL WIND CAPACITY ALLOCATION USED FOR SCOPF ANALYSIS, (MW)

Zone	A	B	C	D	E	F	G	H	I
Capacity	1068	422	467	460	1255	1158	0	770	399

C. Residual Error System Power Flow Case Studies

The residual error (6) impact of discarding increasing numbers of statistical wind components on transmission network power flow modeling accuracy was investigated using a linear DC time series power flow model. The original wind power output time series (i.e. with all 4392 historical samples included and no probability discretisation applied) were re-constructed from different numbers of retained principal components as in (5). Two case study investigations were implemented with the zonal wind capacity allocations as given in Table-I, which result from an optimal non-firm wind capacity connection algorithm (6GW total wind capacity, approx. 33% wind energy penetration) as described in [14]:

- Case-I – A network-unconstrained economic dispatch model illustrating the total power flow requirements.
- Case-II – An 'N-1' security constrained optimal power flow model (SCOPF) analysis of network congestion indices such as wind farm annual energy curtailment.

D. Probability Discretisation Application Study

The SCOPF analyses of Section IV-C above are carried out with the full 4392 historical data samples, for completeness, to investigate the PCA residual error impacts alone. A further study is proposed in this subsection where the probability-discretisation procedure is applied to the lower numbers of retained principal components, to investigate if a potentially much smaller number of probability-weighted SCOPF cases may be solved for similar resulting accuracy. Again, as outlined in Section III-C, the benefit of discretising the principal components instead of the original wind variables (as was applied in [6]) is that fewer variables may need to be

discretised (depending on what dimension reduction is appropriate), and equally importantly the binning-density of the retained components can be tailored with respect to the relative importance of their associated eigenvalues.

For example, if the 9 original wind zones and the customer demand profile were equally and arbitrarily discretised with 10-bin marginal variable binning density, then there would be a maximum of 10^{10} possible discrete cases in the multi-dimensional probability dependency space (though of course most of these bins would be un-occupied as there is a much lesser length of time series data). If on the other hand 3 principal components are deemed appropriate to be retained, and the first principal component and load demand variable are binned with 10-bin density while all other subsequent components are binned with 5-bin density (for example), then a maximum of $10\times10\times5\times5$ = 2500 possible probability-weighted discrete cases would be relevant. The objective of the studies in this sub-section is thus to establish the trade-off between any computational-time advantage (i.e. comparing the lower number of probability-weighted SCOPF solver implementations required) with any accuracy degradations from such PCA discretisation models. To determine spatial wind power inputs to the SCOPF model, the inverse transform of (5) is applied to the representative set of discretised principal component cases.

V. Results

A. Principal Components and Residual Error Investigation

The correlation matrix for the nine wind zones is given in Table-II. Clearly the inter-zonal correlation reduces from Zone A to Zone I, as might be suggested by the geographical separation in Fig.2. As Fig.4 suggests, these linear-dependency metrics explain only part of the overall dependency structure however. The covariances of the PCA results are given in Table-III. As the covariance matrix is diagonal, clearly the resultant principal components have been de-correlated.

A plot of the eigenvalues given in decreasing order, as per the diagonal of Table-III, is illustrated in Fig.6. Clearly there is a rapid reduction in the amount of the original wind variance explained by retention of more than 3 or 4 principal components. Also the first principal component and its associated eigenvalue is significantly more important than the others, due to the relatively high linear correlations in Table-II. On the basis of the arbitrary 'scree' test alone, retaining principal components 6-9 would seem to add very little value. The first principal component corresponds to the common power production patterns in the original multivariate set. The other components do not have such a tangible explanation.

A time series plot of the nine principal components is given in Fig.7. The residual error analysis of (6) for wind Zone A is illustrated in Fig.8 with different numbers of retained principal components. Retaining a few components alone leads to significant residual error in this statistical variable. Retaining more principal components results in a better estimation of the original Zone A variable, but as Fig.6 might suggest, the

TABLE-II
ZONAL CORRELATION COEFFICIENTS BEFORE PCA TRANSFORMATION

ZONE	A	B	C	D	E	F	G	H	I
A	1.00	0.89	0.84	0.81	0.65	0.64	0.64	0.65	0.55
B	0.89	1.00	0.88	0.86	0.67	0.70	0.71	0.70	0.62
C	0.84	0.88	1.00	0.88	0.69	0.69	0.73	0.69	0.63
D	0.81	0.86	0.88	1.00	0.75	0.74	0.76	0.74	0.68
E	0.65	0.67	0.69	0.75	1.00	0.65	0.63	0.70	0.63
F	0.64	0.70	0.69	0.74	0.65	1.00	0.84	0.82	0.85
G	0.64	0.71	0.73	0.76	0.63	0.84	1.00	0.76	0.82
H	0.65	0.70	0.69	0.74	0.70	0.82	0.76	1.00	0.78
I	0.55	0.62	0.63	0.68	0.63	0.85	0.82	0.78	1.00

TABLE-III
PRINCIPAL COMPONENT COVARIANCE AFTER PCA TRANSFORMATION

PC	1	2	3	4	5	6	7	8	9
1	0.61	0	0	0	0	0	0	0	0
2	0	0.072	0	0	0	0	0	0	0
3	0	0	0.041	0	0	0	0	0	0
4	0	0	0	0.022	0	0	0	0	0
5	0	0	0	0	0.015	0	0	0	0
6	0	0	0	0	0	0.013	0	0	0
7	0	0	0	0	0	0	0.011	0	0
8	0	0	0	0	0	0	0	0.01	0
9	0	0	0	0	0	0	0	0	0.009

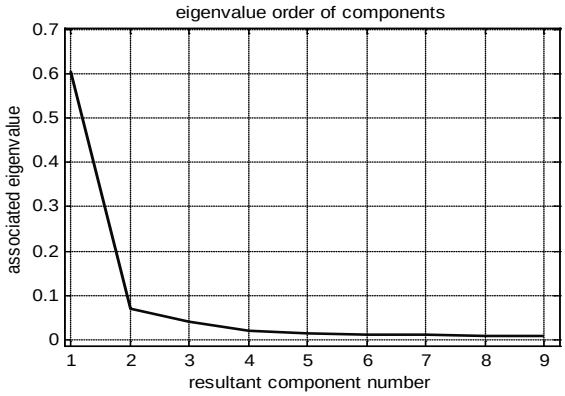

Fig.6 – Ordered eigenvalue plot associated with each principal component.

incremental benefit reduces somewhat for the higher-number principal components. The reduction in the root-mean-square (rms) of the residual error in (6) for a selection of the wind zones is given in Fig.9. Clearly, retaining additional principal components gives a much greater reduction in the rms error for some variables more than others – see the disproportionate reduction in the rms error for Zone E with the retention of principal component 3. This illustrates an occasionally observed effect in PCA, where the discarded principal components such as in Fig.6 can sometimes correspond to specific individual variables in the original multivariate set, rather than being shared amongst all. Thus careful residual error analysis must always be performed for $\boldsymbol{X_{REC}}$ when using PCA to ensure that no single wind zone is overwhelmingly impacted by the dimension reduction process.

The impact of residual errors on the network power flow modeling of the test system in Fig.5 is illustrated in Fig.10 and Fig.11. As a typical example, histograms of the Case-I dispatch model power flows in network branch 15-25 over the extended reconstructed multivariate time series are given in Fig.10. Wind capacity at zone F is connected to bus 25. Clearly there are some non-negligible power flow model differences associated with the multivariate reconstruction residual error (6) for different numbers of retained

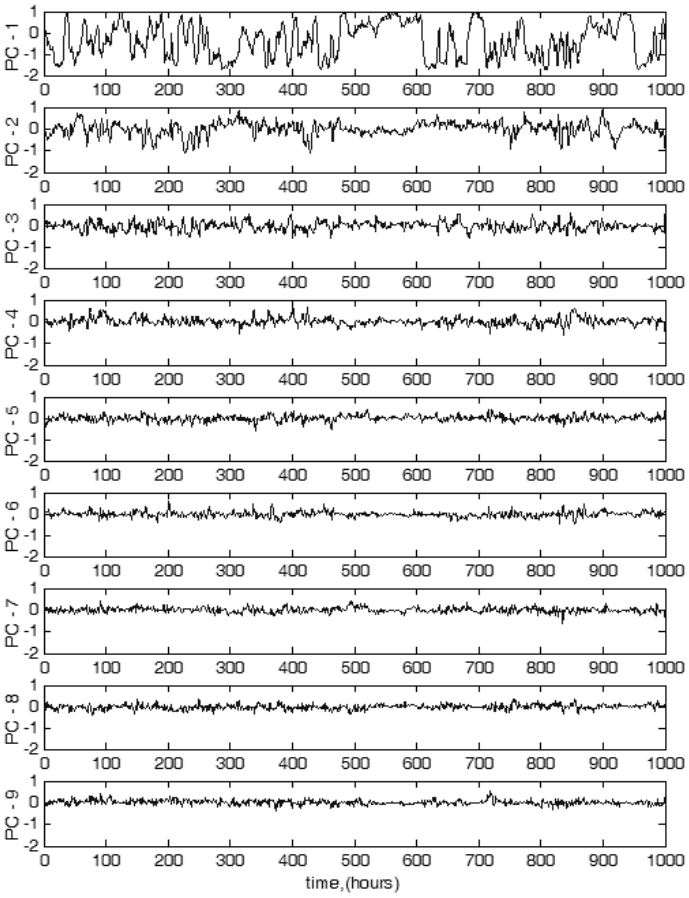

Fig.7 – The resulting principal component time series.

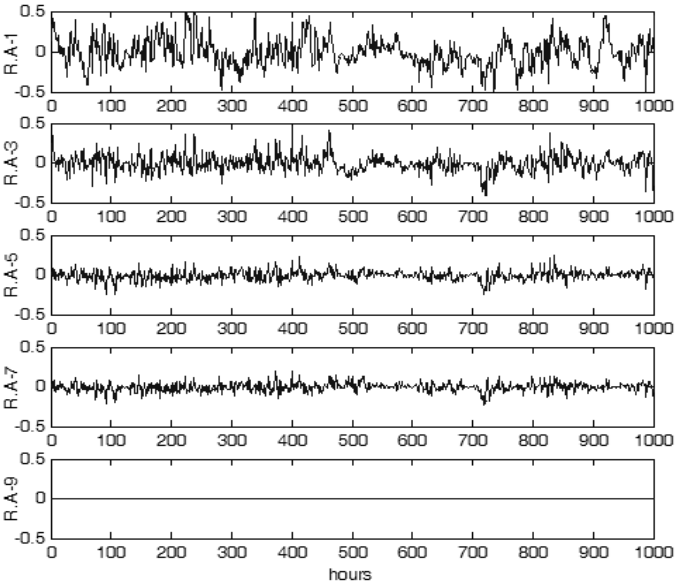

Fig.8 – Zone A residual error for 1, 3, 5, 7, 9 retained principal components.

components. Interestingly however, viewing the results in Fig.11 for the SCOPF analysis applied in Case-II, the selected wind zone energy curtailments due to network congestion were not as significantly influenced as might be initially expected from the magnitude of the residual errors observed. One possible reason is that the types of residual errors in Fig.8 are at least somewhat symmetric, and thus there may be as many hours when the curtailment is overestimated incorrectly as

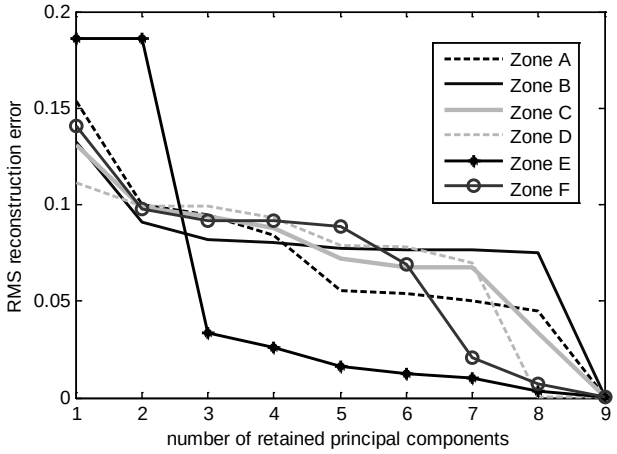

Fig.9 – Reduction in zonal rms reconstruction error for additional components

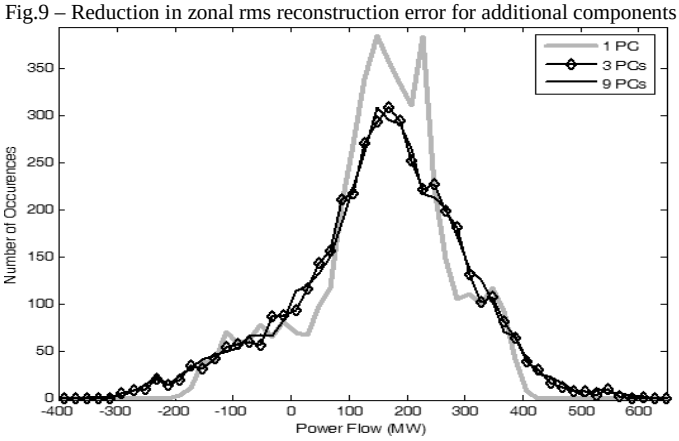

Fig.10 – Case-I : Transmission network power flow modeling residual error (line 15-25).

underestimated incorrectly, giving approximately the same net overall yearly percentage value.

A scatter plot illustrating the dependency between the first two principal components resulting from the wind data PCA study is given in Fig.12. While Table-III does prove these components to be de-correlated, one can clearly see that there is still some general statistical dependency present. For example choosing samples of principal component 1 at both the lower and upper values of its domain will limit the range of values given by principal component 2 relative to samples chosen in the middle of the domain of principal component 1. This emphasizes the drawbacks of wind component analysis based on second order moment statistics information (i.e. covariance) alone.

B. Probability Discretisation Application Results

The results of the principal component discretisation study, as outlined in Section IV-D, are presented in Fig.13 and Table-IV. Fig.13 illustrates the number of discrete probability-weighted representative cases that need a separate SCOPF analysis (i.e. the number of multidimensional bins with more than one sample), for different numbers of discretised components retained. Clearly, by applying the principal component analysis first (i.e. reducing multivariate dimension), to be subsequently followed by the discretisation step (with binning density tailored by eigenvalue size – 10 bins

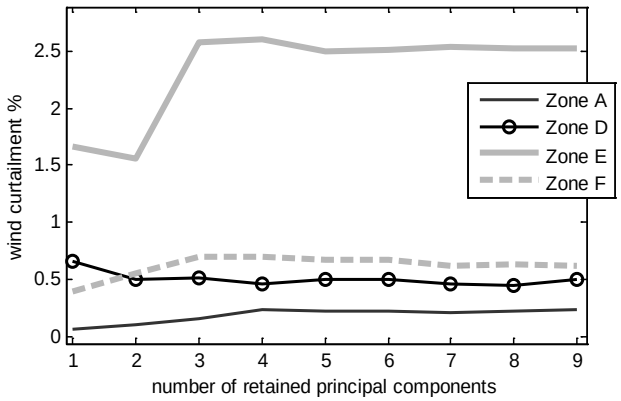

Fig.11 – Case-II : Wind energy curtailment % with respect to number of retained components' residual errors (for 4392 time series samples).

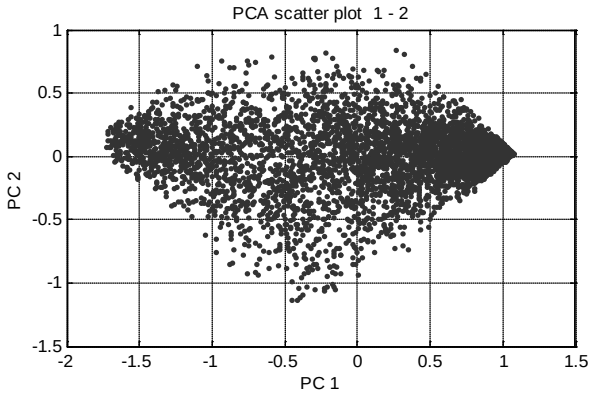

Fig.12 – scatter plot of retained principal components 1 and 2.

for the load demand and principal component 1 random variables, with 5 bins for each additional retained component), will significantly reduce the number of SCOPF implementations from the original 4392 time series samples. If the multi-year high-frequency 15-minute data had initially been proposed instead of the reduced 4392 samples in Section IV-A, then the relative value of the discretisation efficiency would obviously be correspondingly multiplied.

The accuracy degradation associated with component discretisation would appear to be relatively mild for this test system data-set, as suggested by the wind energy curtailment estimates in Table-IV – note the last row of Table-IV corresponds to the 'correct' wind energy curtailment value given by the full 4392 time series samples and all 9 components retained (i.e. neither PCA residual nor multivariate discretisation error). Rows 1-8 of Table-IV are influenced by both residual error (as in Fig.11) as well as any additional error related to probability discretisation. Even for relatively few components, and less than 1000 SCOPF runs, there is only small deviation from the correct values suggested by the original 4392-sample model.

VI. Discussion

Using a sample wind time series SCOPF application, this paper has illustrated the utility of PCA to wind power transmission models based on multivariate probability discretisation. By performing a component analysis and

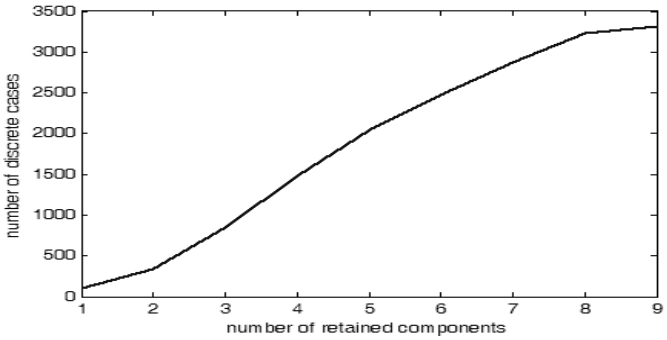

Fig.13 – The number of probability-weighted discrete cases for different numbers of retained components.

TABLE-IV
Impact of Probability Discretisation - Wind Energy Curtailment Estimates, (%)

No. of Discretised Components	Zone A	Zone D	Zone E	Zone F
1	0.055	0.696	1.823	0.431
2	0.195	0.455	1.653	0.614
3	0.312	0.408	2.655	0.794
4	0.298	0.439	2.703	0.728
5	0.311	0.497	2.601	0.718
6	0.316	0.457	2.627	0.821
7	0.268	0.386	2.628	0.772
8	0.307	0.373	2.643	0.835
9	0.261	0.524	2.690	0.757
'Correct' Value	0.227	0.495	2.526	0.617

dimension reduction prior to the multivariate discretisation step, the design of which can subsequently be tailored with respect to the importance of the respective component eigenvalues, a significantly fewer number of probability-weighted discrete cases approximated the spatial wind power diversity with a reasonably good degree of end-result accuracy. This type of modeling strategy may have relevance for future wind/transmission optimization models solved through decomposition strategies, where a large number of sub-problem implementations based on time-series samples would otherwise be required to model different wind power flow conditions [14]. A very desirable characteristic of the PCA technique is that it is performed on zero-mean centered data (4), and therefore the precise wind zone capacity factor values can be preserved for the original variable reconstructions in (5). Combined wind/transmission optimization problem solutions are quite sensitive to wind resource quality differences in the power system geographical area [14]. Preservation of the capacity factor characteristics, in addition to the reasonably stable wind energy curtailment results in Fig.11 (relative to the impact of other general modeling uncertainties [15]), could indicate that optimization solution quality with a few retained components describing the wind variations might still be reasonably accurate, though further study is required.

The data used in this study was sourced from wind farms on the Irish power system, though generality to other similar wind regimes should apply. Note that the real Irish 'All-Island' power system has approximately 100 wind farms at present, which on the basis of the results in this paper might be reasonably accurately represented by many fewer random

statistical variables. It must be acknowledged however that the efficiency of the discretisation study reported here in part stems from the fact that the wind power pattern in a reasonably small-to-medium sized geographical area such as the test-system in Section IV has a strong degree of correlation, and thus the first few principal components were far more significant than the others. Furthermore, using a probability discretisation approach to combine separate time series samples with no regard for their sequential dependencies might not be appropriate for certain power systems with a significant amount of short-term energy storage capacity [10].

PCA, which is based on the principle of de-correlation, will be of greatest relevance for multivariate Gaussian datasets, as the linear PCA transformation will then still give Gaussian principal components, and uncorrelated multivariate Gaussian random variables also have the desirable property of statistical independence. As was illustrated in Fig.12 however, this property does not necessarily extend to other more generally dependent multivariate distributions such as in the distributed wind power case. For more general multivariate distributions, the covariance which is determined by second-order statistical moment information is but one measure statistical dependency only, and zero correlation does not necessarily imply independence [16]. Higher-order statistical moments are required to fully describe the non-parametric marginal statistical and multivariate dependency distributions that apply to the distributed wind power generation case. A different approach may therefore be required to find fully independent wind components.

Further investigation of alternative multivariate component analysis techniques for the distributed wind power case was also carried out as part of this paper's study. The classic 'independent component analysis' (ICA) model, based on the assumption of a linear mixing-matrix [17] was applied to the multivariate wind data-set of Section IV for example. This 'Fast-ICA' algorithm [18],[19] is a negentropy-maximization based optimization technique, with the PCA de-correlation step of this paper as an initialization stage. It was found however that the application of this rather advanced statistical source separation technique had little additional impact on reducing the retained dependence in the results of the PCA stage, and thus it is not reported in detail in this paper.

Statistical transformation of the individual (marginal) random wind variables to Gaussian marginal densities, using their marginal cumulative distribution followed by the inverse Gaussian cumulative distribution, was reported in [20] with the aim of designing a multivariate wind time series synthesis model. Such transforms could be relevant in order to generate independent wind components in the transformed domain using a subsequent PCA de-correlation step. The implicit assumption of multivariate Gaussianity in the transformed domain (only the marginals are transformed in [20], with no specific treatment of any multivariate dependency information) would need to be carefully tested before assuming such PCA results are completely independent though (as a set of marginal Gaussian variables is not necessarily multivariate Gaussian). The significant non-linearity of the combined statistical transform in [20] could also be analytically unwieldy for some applications. Analysis of multivariate dependency in distributed wind speed data as opposed to wind power data may also be worthy of investigation, though retention of the piecewise non-linear wind turbine speed/power curves in power system study formulations could be a significant practical drawback. Investigating such alternative approaches to the distributed multivariate wind data dimension and dependence reduction problem is an interesting future research topic.

VII. Conclusions

This paper presents a multivariate dimension reduction study as applied to spatially distributed wind power historical data. Using principal component analysis, it is shown that the strong dependency in the distributed wind power data allows a reasonably significant dimension reduction to retain a large proportion of the original statistical behaviour. Careful residual analysis should always be performed for the discarded components though, as they may sometimes correspond directly to specific wind zones only. A lower number of statistical variables combined with an effective probability discretisation approach could also reduce model dimensionality for computationally efficient power system study applications. Even though they are shown to be de-correlated, the retained wind power components from a linear mixing model cannot be guaranteed to have the stronger statistical property of independence however.

VIII. References

[1] T. Ackermann, (Editor) *Wind Power in Power Systems*, Wiley, 2005.

[2] R. Doherty, H. Outhred, M.J. O'Malley 'Establishing the Role That Wind Generation May Have in Future Generation Portfolios', *IEEE Transactions on Power Systems*, Vol. 21, No.3, August 2006.

[3] G. Papaefthymiou, A. Tsanakas, D. Kurowicka, P. H. Schavemaker, and L van der Sluis, "Probabilistic Power Flow Methodology for the Modeling of Horizontally-Operated Power Systems," International Conference on Future Power Systems, Amsterdam, Nov. 2005.

[4] S. Deladreue, F. Brouaye, P. Bastard, L. Peligry 'Using Two Multivariate Methods for Line Congestion Study in Power Systems Under Uncertainty', *IEEE Trans. Power Systems*, Vol. 18, No.1, February 2003.

[5] J.E. Jackson, *A Users Guide to Principal Component Analysis*, Wiley New York 1991.

[6] D. Burke and M.J. O'Malley 'Optimal Wind Power Location on Transmission Systems – A Probabilistic Approach', presented at the IEEE PMAPS Conference, Puerto Rico, May 2008.

[7] L.F. Ochoa, C.J. Dent, G.P. Harrison 'Distribution Network Capacity Assessment: Variable DG and Active Networks', *IEEE Trans. Power Systems*, Vol. 25, No.1, February 2010.

[8] 'All Island Grid Study, Workstream 4 – Analysis of Impacts and Benefits', Irish Government Department of Communications, Energy and Natural Resources/United Kingdom Department of Enterprise, Trade and Investment, Jan. 2008. Available online - http://www.dcenr.gov.ie/Energy/North-South+Co-operation+in+the+Energy+Sector/All+Island+Electricity+Grid+Study.htm

[9] http://www.nationalgrid.com/uk/Electricity/Data/

[10] D.J. Burke 'Accommodating Wind Energy Characteristics in Power Transmission Planning Applications', PhD Thesis, University College Dublin, Ireland, 2010.

[11] MATLAB, available at http://www.mathworks.com/
[12] General Algebraic Modeling System, GAMS – available online http://www.gams.com/
[13] 'Matlab and GAMS – Interfacing Optimization and Visualization Software', by M.C. Ferris – available online at http://www.cs.wisc.edu/math-prog/matlab.html
[14] D.J. Burke, M.J. O'Malley 'A Study of Optimal Non-Firm Wind Capacity Connection to Congested Transmission Systems', *IEEE Transactions on Sustainable Energy* (Accepted, In Press 2010).
[15] D.J. Burke, M.J. O'Malley 'Factors Influencing Wind Energy Curtailment', *IEEE Transactions on Sustainable Energy* (Accepted, In Press 2010).
[16] G. Papaefthymiou, D. Kurowicka 'Using Copulas for Modeling Stochastic Dependence in Power System Uncertainty Analysis', *IEEE Trans. Power Systems*, Vol. 24, No.4, Feb. 2009.
[17] A. Hyvarinen J. Karhunen and Erkki Oja, *Independent Component Analysis*, Wiley New York 2001.
[18] A. Hyvarinen 'Fast and Robust Fixed-Point Algorithms for Independent Component Analysis', *IEEE Trans. Neural Networks*, Vol. 10, No.3, May 1999.
[19] http://www.cis.hut.fi/projects/ica/fastica/
[20] B. Klockl 'Multivariate Time Series Models Applied to the Assessment of Energy Storage in Power Systems', presented at the IEEE PMAPS Conference, Puerto Rico, May 2008.

I. BIOGRAPHIES

Daniel Burke (M 2007) graduated from University College Cork, Ireland with a BEEE in Electrical and Electronic Engineering in 2006. He is currently conducting research for a PhD in power systems at the Electricity Research Centre in University College Dublin, Dublin, Ireland. He is a postgraduate student member of the IEEE.

Mark O'Malley (F'07) received B.E. and Ph. D. degrees from University College Dublin in 1983 and 1987, respectively. He is the professor of Electrical Engineering in University College Dublin and is director of the Electricity Research Centre with research interests in power systems, grid integration of renewable energy, control theory and biomedical engineering. He is a fellow of the IEEE.