

Renewable Resource Dataset Generators

Gruffudd A. Edwards and Rod W. Dunn, *Member, IEEE*
Department of Electronic and Electrical Engineering, University of Bath
Email: gae20@bath.ac.uk

Abstract—Designing flexible networks to cope with the broad range of plausible scenarios for the future of the electricity system in Great Britain (GB) demands adequate and appropriate renewable resource data. This paper reports on research aimed at developing algorithms capable of producing synthetic time series datasets of arbitrary length to represent the variable renewable resources available to generators. Attention to date has been on the wind resource, as this is the dominant technology. The algorithms will produce the datasets using time series models – building upon an existing ‘Bath Wind Model’ methodology, but modelling the resources as seasonal long memory processes. The datasets will be suitable for a range of studies, but particularly system adequacy studies using Monte-Carlo simulation.

Index Terms—Monte-Carlo Simulation, seasonal long-memory process, solar energy, wind energy, wind modelling.

I. INTRODUCTION

There exists considerable uncertainty regarding the way in which electricity transmission and distribution networks in Great Britain (GB) will evolve up to 2050, the only certainty being radical changes of some kind. Accurate and appropriate renewable resource time series datasets are essential for the designing of flexible networks capable of coping with the broad range of plausible scenarios for the future of the GB electricity system.

This paper presents the initial results of a PhD project concerned with the development of novel algorithms to produce arbitrary-length synthetic time series datasets, representing the variable and non-dispatchable renewable resources available to generators connected to the GB networks. The algorithms will produce the data using statistical time series models, with no explicit physical basis. Attention to date has been on representing wind, due to the current domination of wind generation, but it is hoped that extension to include the solar resource will be possible before the end of the project.

The primary application foreseen for the datasets will be in facilitating more accurate assessment of system security indices for future scenarios with high penetrations of renewable generation - particularly wind. Such assessments require sequential Monte-Carlo simulations with resource time series used as trial inputs. While some simulations are only concerned with the differences between total available generation and total demands, accurate assessment must include transmission constraints. This makes information on the spatial distribution of the resource essential. Such spatial information will also be of great use for a wide variety of economic studies – e.g. enabling exploration of the utilisation patterns of a proposed network extension.

The time series models must therefore be multivariate, and the datasets will present simultaneous hourly averaged resource availabilities for 20 GB zones. Spatial and temporal correlation structures will ensure that the resources’ chronological characteristics and spatial distribution patterns, as found in historical data, are accurately reproduced. Care is taken to ensure that the data realistically reflect changes in the resource occurring on all timescales – from turbulent fluctuations to the continuous ‘climate’ changes between decades.

A starting point for developing the algorithms is provided by the previously developed ‘Bath Wind Model’ methodology [1], but with the added sophistication, including modelling the resource availability as seasonal long memory process.

II. THE BATH MODEL

The Bath Wind Model methodology is based on the idea that the vector of GB zonal wind speeds at time t , $\mathbf{U}_t = [U_{1t}, U_{2t}, \dots, U_{nt}]^T$ is a 4th order autoregressive, or AR(4), process. Each zone is assumed to ‘see’ a uniform wind speed throughout it and was represented for any given historical hour by the recorded wind speed at a selected Met Office weather recording station. In forming the vector \mathbf{U}_t all component wind speeds have their mean values removed so that we have a zero mean stochastic process satisfying the equation

$$\mathbf{U}_t = \mathbf{a}_1 \mathbf{U}_{t-1} + \mathbf{a}_2 \mathbf{U}_{t-2} + \mathbf{a}_3 \mathbf{U}_{t-3} + \mathbf{a}_4 \mathbf{U}_{t-4} + \mathbf{Z}_t, \quad (1)$$

where the $\mathbf{a}_1, \dots, \mathbf{a}_4$ are autoregression coefficient matrices which correlate the values of each variable in \mathbf{U}_t to their own past values and also the other variables’ past values. The vector \mathbf{Z}_t is a composed of independent and identically distributed Gaussian white noise terms. The model parameters are estimated by using Matlab’s System Identification Toolbox [2] to find the best fit of this model to the historical data. Once fitted, Matlab can use the model to generate artificial time-series of any desired length. The model clearly captures the correlation between different sites, along with the autocorrelation functions of the vector components – although not perfectly, as explained below.

The 20 ‘Bath Zones’ are derived from 17 zones created by National Grid - based the boundaries between them intersecting the major interconnectors of interest on the transmission network. The major difference introduced is the addition of three zones in northern Scotland, reflecting the area’s increasing importance.

Once generated, the wind speeds must be scaled-up to match those experienced at a height of 80m (assumed turbine hub height) using the simplified relationship

$$V_1/V_2 = (H_1/H_2)^{1.7}. \quad (2)$$

These wind speeds are then transformed into per-unit zonal power outputs using a manufacturer’s wind power curve. Converting these into power outputs and using them as inputs into a Monte-Carlo simulation requires a scenario including zonal generation capacities and demand distributions. The availability of each generator for each trial is modelled as an independent Bernoulli variable (i.e. randomly on or off).

A more sophisticated version of this methodology was developed for a project commissioned by National Grid [3]. The main difference is the introduction of four different types of windfarm location: coastal, lowland, upland and offshore. A scenario must specify the generation capacity of each type in each zone - although not all zones have each type, particularly land-locked ones. A second difference is to differentiate between transmission and distribution connected capacity – for each location type within each zone. Offshore projects are all assumed transmission connected.

The next difference is that account is taken of the fact that the wind climates at windfarm locations are generally different from those at the Met Office stations. This is achieved through speed up ratios (for constant heights of 10m) – a different one for each location type in each zone. Different hub height speed-up ratios are also introduced for each location type, although in this case they do not vary between zones. A final difference is that two slightly different turbine power curves are used – one being more suitable in situations where the resource is only moderate. Offshore turbines are assumed to all have a capacity of 3MW, while onshore ones are all 2MW.

This methodology is powerful, fast and has been shown to reproduce several statistical properties of the historical data – such as average variance and the correlation between zones. However it has several shortcomings and these are discussed in the following sections, along with ways in which they may be overcome.

III. CAPACITY FACTORS

The capacity factors of synthetic datasets produced by the model are too high, compared to empirical data. This is discussed in [3], the main suggested reason being that the availabilities attributed to wind generators in the Bernoulli simulation are too high. This is likely to be true to some extent: the assigned availabilities are 97% for onshore wind and 90% for offshore, which is larger than the values assumed by National Grid in a recent study [4] of 95% and 85%, respectively. Clearly the difference is not significant.

The main reason for the high capacity factors is that the model produces data with a Gaussian distribution, while it is

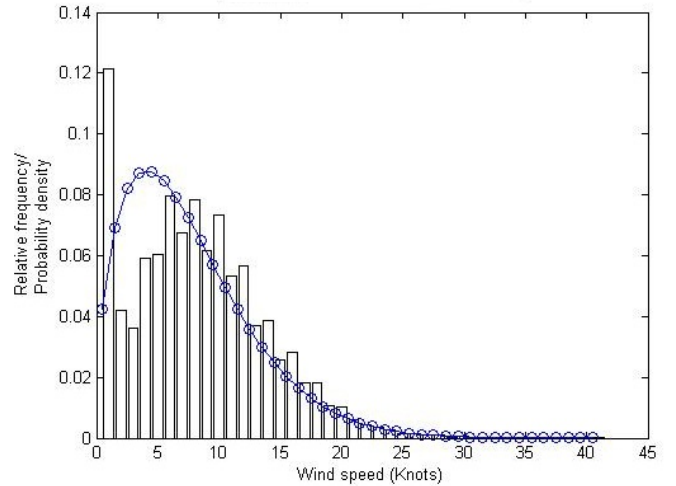


Fig.1. Comparison of a normalised histogram for a Met Office station’s data (3 year sample) and the best-fit Weibull probability density function for that data.

an established fact that wind speeds are well approximated by Weibull distributions, as can be seen in Fig. 1.

The model fitting ensures that the simulated data has the same mean as the historical data. However wind speed distributions are asymmetrical, and their mean may be considerably greater than their median. Since the synthetic data have a Gaussian distributions the median will be the same as the mean and artificially high, resulting in artificially high capacity factors. This problem can be overcome by transforming the historical data so that it has an approximately Gaussian distribution before model fitting takes place, then reversing the transformation for the simulated data. This can be done through the use of a Box-Cox transformation, as demonstrated in [5].

A problem that can be clearly seen in Fig. 1 is that there is a very significant excess of zero wind speed recordings in the recorded data, at the expense of other low wind speeds. This is due to cup anemometer friction, and the Author’s attempts to eliminate genuine errors in a rigorous way proved futile. The zeros have a warping effect on the Weibull parameter estimation process – although mathematical methods for overcoming this will be investigated.

It is likely that the wind climates at windfarm locations within zones differ from those at the corresponding Met Office stations not only in terms of the average wind speed, but also in terms of their Weibull distribution shapes – defined by shape parameter k . In order to reflect these differences in the synthetic data, slightly different reverse Box-Cox transformations can be used, making use of the property that if U has a Weibull distribution with shape parameter k , then U^m is also Weibull distributed with parameter k/m . The way in which the shape parameter changes with location type is an area to be researched.

IV. SEASONAL CHARACTERISTICS

The wind resource displays clear seasonal patterns of availability – on both annual and diurnal time scales. As

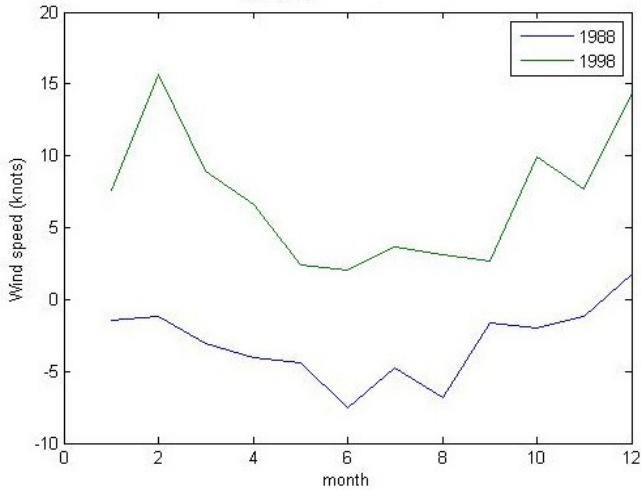


Fig.2. Monthly averaged wind speeds for a single location, in 1988 and 1998, minus the 20 year average.

reported in [6], the wind resource is considerably greater in winter and, on average greater during the day than at night.

These seasonal trends are dealt with by most wind modellers through subtraction of a deterministic function from historical data before model fitting, simulation of a non-seasonal process and then addition of the deterministic component back onto the simulated data. The function may be for every hour of the year, or a separate model for each month may be created, each with a fixed mean, and a different diurnal profile is subtracted from each. Additionally, modellers may construct a deterministic function for the standard deviation and divide the historical data by it, since the variance of wind speeds is not stable, as is required for AR modelling.

This method has limitations: developing separate models causes discontinuity at the transition between months; and as Fig. 2 shows, although every year has some features in common, the annual pattern is far from deterministic. This indicates that in order to model this aspect correctly, the annual seasonality must be incorporated into the stochastic modelling.

This is in contrast to diurnal seasonality, which is well represented by deterministic profiles, as Fig. 3 demonstrates.

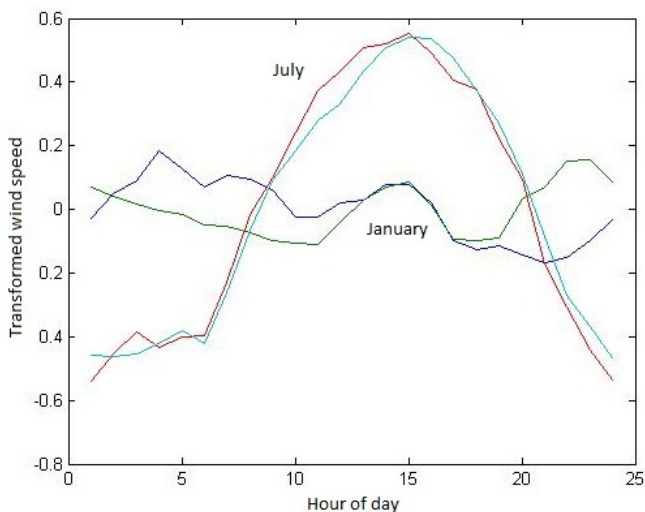


Fig.3. Mean diurnal profiles for Box-Cox transformed Met Office data: two decade-long samples for a single station, for January and July – with annual seasonality removed.

To produce the profiles seen in the figure, a moving average was calculated for each wind speed datum of the 24 hour period for which it is at the centre (i.e. 13 hours either side of it), and this average was subtracted from the original datum. For this new dataset of differences, monthly means were calculated for each hour of the day. To avoid discontinuities at the transition between months, interpolation methods may be used to produce a matrix containing the diurnal trend for each hour of the year - a column for each hour of the day and a row for each day of the year. (Leap years had a separate matrix.) It is of course possible to work in terms of days rather than months from the outset, but this would involve sample sizes of only 10 for each hour. The fact that the profiles for the two decades are in good agreement, particularly for July, indicates that the deterministic assumption is valid.

V. CLIMATE CHANGES AND LONG MEMORY

Another problem with regular autoregressive or, to generalise, autoregressive moving average (ARMA) models is that they produce stationary data – i.e. the data shows no significant long term variations between years or decades. Such climate changes are however a characteristic of the wind resource [7]. One type of process which does behave in this way are called long-memory processes. It has been established [8] that wind should indeed be modelled as such a process, which is also true of many aspects of the climate. Such processes are defined by the fact that their spectral densities become unbounded for some frequency in the range $[0, \pi]$.

The simplest type of long memory process, known as AFRIMA processes, has 0Hz as the unbounded frequency, and this is the type of process assumed in a widely cited paper [8]. However it is shown in [9] that this is not the case – due to the seasonal nature of the wind resource, the spectral density in fact peaks at the annual frequency. Such a process is known as a seasonal persistent process (SPP). This has been confirmed during this research, as seen in Fig. 4, for a

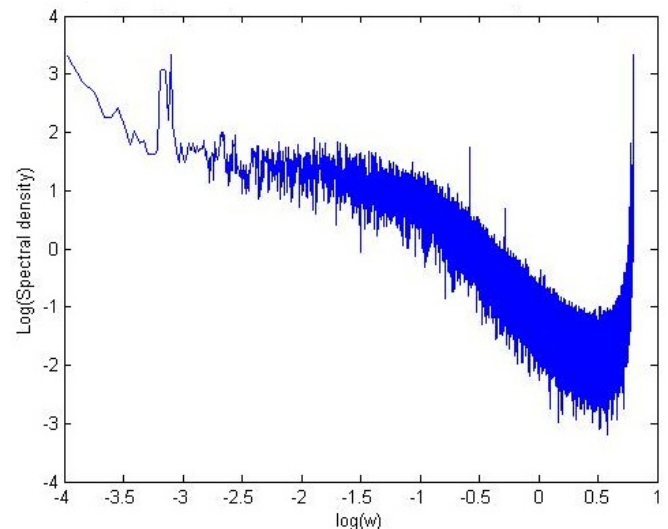


Fig.4. Log-log discrete periodogram for a Box-Cox transformed 20 year Met Office station sample, smoothed with a simple linear filter.

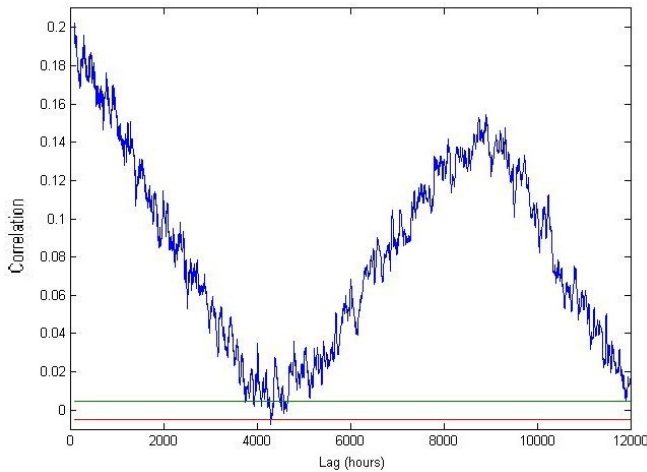


Fig.5. Auto-correlogram for a diurnally de-trended, Box-Cox transformed 20 year sample, for lags of 100 – 12,000 hours, with white noise bounds included.

Met Office station 20 year sample. It is not clear from this plot that the peak at the annual frequency ($w = 7.8 * 10^{-4}$) is greater than that for $w \rightarrow 0$, however removing the log from the y-axis shows that this is indeed the case. The other feature of note on Fig. 4 is the straight downward slope of the spectral density for the lowest frequency octaves – a characteristic of long memory processes.

Further evidence is displayed in Fig. 5, which shows that the autocorrelation function decays far more slowly than would be expected from a regular AR process, and where the annual seasonality is very clear. The data used for this plot had been diurnally de-trended as described above, to give a line that is considerably less noisy in appearance than without the de-trending.

As an SPP, the wind speed for a single site at time t , U_t , must satisfy

$$(1 - 2\nu B + B^2)^\delta \Phi(B)U_t = \Theta(B)Z_t, \quad (3)$$

where B is the backward shift operator, ν is a parameter dictated by the seasonal period, δ is a parameter that dictates the slope of the straight line on the periodogram and the width of the unbounded peak, and Φ and Θ are the autoregressive and moving average polynomials, respectively. Constraints are: $|\nu| < 1$ and $0 < \delta < 0.5$. In the multivariate case, U_t and Z_t are again vectors while Φ and Θ are matrix polynomials. For a purely AR process, which the Author believes is likely to be true, Θ is simply the identity matrix. Model fitting and simulation are much easier if we assume that ν and δ are constant for all sites, which seems physically reasonable, such that the fractional differencing operator $(1 - 2\nu B + B^2)^\delta$ is a scalar. Fitting the model (3) to 20 locations, one in each zone, is the main task of this research.

The rather counterintuitive concept of fractional differencing is made easier to comprehend by the fact that if the operator is moved to the right hand side of (3), it can be

expanded as a (Gegenbauer) polynomial, allowing (3) to be expressed in an infinite moving average form as

$$\Phi(B)U_t = \Theta(B)\sum_{k=0}^{\infty} C_k(\delta, \nu)Z_{t-k}, \quad (4)$$

where the Gegenbauer coefficients are given by

$$C_k(\delta, \nu) = \sum_{j=0}^{\lfloor k/2 \rfloor} (-1)^j \Gamma(\delta + k - j) (2\nu)^{k-2j} \dots / \Gamma(\delta)j!(k-2j)! \quad (5)$$

It is clear that the fitting and simulation of an SPP is highly involved. Various authors have developed a range of methodologies, perhaps the most promising of which involves the use of discrete wavelet transforms [10] – Matlab's Wavelet Toolbox should provide considerable assistance with this method. The novelty of this research lies in estimating and simulating wind as an SPP in the multivariate case.

VI. SPATIAL DIVERSITY

The last potential improvement to the Bath model to be attempted relates to its lack of spatial diversity. A single wind speed time series (for a zone), converted to a power output via a power curve for a single turbine, produces a wind speed time series with excess variance compared to a spatially diverse collection of wind farms within that zone. This statement is supported by empirical evidence, as reported in [11]. The overall issue would be alleviated somewhat by introducing a larger number of zones, but this would cause problems due to an unmanageable number of parameters to be estimated (the number of elements in each matrix varies with the square of the number of locations). Other techniques include applying a moving average filter, as discussed in [11] or time-slipping (i.e. overlaying a part of the previous hour's output on the current one) as discussed in [12]. Another subtlety presented in [11] is to use a 'softer' turbine power curve that reflects the output from a single windfarm, rather than a single turbine. The potential problem with the first two methods is that they may destroy some of the vital spatial information generated.

A method that will be investigated in this research is to create additional 'sub-locations' within each zone, in addition to core locations. All parameter values for location within each zone would be the identical, and regression would only be with the core locations in other zones. The difference between the locations within a zone would be that they experience a different, but highly correlated white noise input (e.g. $R^2 = 0.8$).

VII. CONCLUSIONS

After presenting reasons why accurate, multivariate renewable resource data is essential at this period of uncertainty regarding the future, a review was given of an existing method of generating such data - the Bath Wind Model. Despite its novelty, several aspects of this

methodology may be improved, and doing so is the task of the PhD project described here. The improvements include ensuring that the generated data have suitable Weibull distributions, ensuring they have the correct seasonal and diurnal characteristics, reproducing the natural climate changes which are constantly occurring and ensuring realistic spatial diversity.

Given the complexity of the enhanced methodology, the question naturally arises of why one needs to produce synthetic data, given that around 100 years of accurate historical data exists, courtesy of the UK Met Office. Data which also has a spatial resolution far greater than a 20 zone model. The reason is that although 100 years represents nearly 877,000 hours of data, researchers often need to examine only specific times – e.g. the current range for possible peak demand of 5.00pm – 7.00pm during December to February. Being that specific reduces the number of hours to little more than 18,000 hours. A rigorous Monte-Carlo simulation may require up to 100,000 trials, particularly if trying to establish the precise probability for a highly infrequent event, such as a significant loss of load. With a synthetic dataset generator a researcher can be as specific about times as they wish and conduct a study of any size. Additionally, several runs of the same simulation can be run with slightly different resource data – e.g. one with a windy decade another with a calm one, so that values such as loss of load expectation can be established with confidence intervals.

ACKNOWLEDGEMENTS

The authors would like to thank the British Atmospheric Data Centre for allowing access to the MIDAS database of synoptic records, making this research possible. This PhD project is part of Supergen FlexNet, a consortium of Universities and industrial stakeholders that aims to lay out the steps necessary to ensure future-proof flexible networks in the context of challenging carbon reduction targets. The research forms part of the ‘Future Shape and Size of the Network’ work-stream.

REFERENCES

- [1] Miranda, M.S. & Dunn, R.W., “Spatially correlated wind speed modelling for generation adequacy studies in the UK”. *Proceedings of the Power Engineering Society General Meeting*, Tampa, Florida, 2007. IEEE, pp. 24 -28, 2007.
- [2] Ljung, L. Matlab: *System Identification Toolbox 7: User’s Guide* [online]. Natick, MA: The Mathworks. Available at: http://www.mathworks.com/access/helpdesk/help/pdf_doc/ident/ident.pdf [Accessed 24 March 2010].
- [3] Li, F., Dunn, R.W., Miranda, M.S., Kuri, B., *Sufficiency of transmission capacity for a system with wind generation* (report commissioned by National Grid). Bath: University of Bath, 2006.
- [4] National Grid, *Operating the electricity transmission networks in 2020: initial consultation*. Warwick: National Grid Electricity Transmission plc, 2009.
- [5] Miranda, M.S. et al. “Bayesian inferencing for wind resource characterisation”. *9th International Conference on Probabilistic Methods Applied to Power Systems*, 2006, Stockholm. PMAPS, pp.1 - 6, 2006.

- [6] Sinden, G., “Characteristics of the UK wind resource: long-term patterns and relationship to electricity demand”. *Energy Policy*, 35(1), pp.112 -127, 2007.
- [7] Petersen, E.L. et al., *Wind power meteorology*. Roskilde: Riso National Laboratory, (Riso-I-1206(EN)), 1997.
- [8] Haslett, J. & Raftery, A.E., “Space-time modelling with long-memory dependence: assessing Ireland’s wind power resource”. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 38(1), pp.1-50, 1989.
- [9] Bouette, J.C. et al., “Wind in Ireland: long memory or seasonal effect?”, *Stochastic Environmental Risk Assessment*, 20(3), pp.141 - 151, 2006.
- [10] Whitcher, B., “Wavelet-based estimation for seasonal long-memory processes”. *Technometrics*, 46(2), pp.225 -238, 2004.
- [11] Holttinen, H., “Hourly wind power variations in the Nordic countries”. *Wind Energy*, 8, pp.173 -195, 2004.
- [12] Ilex & Strbac, G., “Quantifying the system costs of additional renewables in 2020”. A report to the Department of Trade and Industry. Oxford: Ilex Energy Consulting, 2002.