Missing Data in Wind Farm Time Series: Properties and Effect on Forecasts

Rosemary Tawn and Jethro Browell
University of Strathclyde
Glasgow, UK
{rosemary.tawn; jethro.browell}@strath.ac.uk

Iain Dinwoodie
Natural Power
Stirling, UK
iaind@naturalpower.com

***Abstract*—Missing or corrupt data is common in real-world datasets; this affects the estimation and operation of analytical models where completeness is assumed or required. Statistical wind power forecasts utilise recent turbine data as model inputs, and must therefore be robust to missing data. We find that wind power data is 'missing not at random', with missing patterns also related to the forecast output. Approaches for dealing with this missing data in training and operation are proposed and evaluated through a case study, leading to a suggested forecasting methodology in the presence of missing data. In the training set, missing data was found to have significant negative impact on performance if simply omitted but this can be almost completely mitigated using multiple imputation. Greater increase in forecast errors is seen when input data are missing operationally, and re-training forecast models using the remaining inputs is found to be preferable to imputation.**

***Index Terms*—forecasting, missing data, time series, vector autoregression, wind power**

I. Introduction

Renewable electricity generation in Scotland has increased from 19.7% of total electricity generation in 2007 to 70.3% in 2017. This rise is due at least in part to measures aiming to limit the impact of climate change and is echoed across the European Union as a whole [1]. The variability of wind power production has a significant impact on power system operation [2], where accurate forecasts are required to aid decision making [3]. Very short-term (minutes to hours ahead) planning and operation is important for trading and to balance electricity supply and demand [4], and at these time horizons statistical methods typically outperform those based on numerical weather prediction [5]. These methods are driven by recent observations of wind speed and power from (potentially multiple) wind farms; however, this data is subject to information loss from communication errors or delays as well as maintenance operations and curtailments. This loss negatively impacts model estimation and operation, and therefore the predictive abilities, of forecasting models.

There has been little work to describe the general missing properties seen in wind farm operational data, and while some works have considered missing data in wind power time series for other applications, its impact on forecasts has not been assessed. The impact of missing data on monthly and annual average measurements was discussed for wind energy resource assessments [6] along with the corresponding impact on revenue [7]. Other applications include power curve estimation [8], wind farm control [9] and fatigue assessment [10].

In very short-term wind power forecasting studies, subsets of data with missing values are often simply omitted, which may bias model estimates and is not an option when producing operational forecasts. Recent works have focused on high dimensional modelling [11], dynamic models [12] and data sharing via privacy preserving algorithms [13], for example, but with the implicit assumption of data completeness.

Other research has presented methods for filling missing data in a wind time series; however, the simulated missing values are selected randomly throughout the time series [14] which does not reflect real patterns of missing data, and Lotfi [15] uses imputation by simple autoregressive or moving average models which are not suited to filling extended periods of missing data. The purpose of filling in a time series is generally to allow further analysis, for example to calculate energy yields or to detect sensor failures. By only reporting the accuracy of the imputation process itself, the financial or decision-making consequences of the proposed imputation methods are not addressed.

Fields that often utilise longitudinal studies, such as medical trials and political behaviour studies, have traditionally encountered significant levels of missing data [16] and as such have developed methods to quantify and account for its effects on study outcomes. Central to these methods is the classification of missing data into one of three types [17]: Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing Not at Random (MNAR), sometimes known as 'non-ignorable missing'. Data may be classified as MCAR when the probability of a data point being missing is completely independent of any variables in the dataset; in this case, the remaining complete data has the same distribution as the original population with no missing data and so there will be no bias introduced to the model estimation. MAR data occurs when 'missingness' in one variable is independent of its own value but does depend on the value of another. Finally, if the probability of a point being missing is dependent on the value it would have taken, then the missing pattern is classified as MNAR. In this case, the distribution of values in the remaining observed cases is not the same as that in the missing data points and so any model analysis ignoring this

discrepancy will produce biased results.

Several methods for dealing with missing data exist, with varying validity for the different missing data types. Simple methods include 'complete case analysis' where any data points with partial information are discarded [18]; this will produce biased results if data is not MCAR because the remaining data is no longer a random sample from the underlying distribution. This may be counteracted by weighting remaining data using missing probabilities [19], but this does not prevent reduction in size of the data set. Another popular ad-hoc method is imputation, where any missing values are filled with a simple substitute, such as the mean value [20], the preceding value, or the (weighted) nearest neighbour(s). While these may be acceptable methods for individual missing points and datasets with very low proportions of missing data, they are not suited to filling the long spans of missing data seen in wind power time series.

Final forecast errors should reflect the additional error from loss of information at missing values as well as inherent forecast uncertainty. Multiple imputation reconciles the need for complete data with an accurate estimation of final uncertainty by filling the data set several times according to a probability distribution for the missing values, producing multiple completed data sets leading to multiple forecast values, allowing calculation of the spread of forecast outcomes as well as a mean forecast. Care must be taken to correctly identify the distribution of missing values [21], although a multivariate normal distribution is also commonly used [22] as an approximation. It has been observed that results using multiple imputation are generally as good as those from more complex and rigorous alternatives [23].

In this paper, complete data from 10 UK wind farms is used to quantify the increase in error in wind power forecasts from different types of missing data through a vector autoregressive (VAR) model. Forecast skill reflects the loss of information from missing data as it involves fitting the model to the time series directly. In addition, forecasts are used by many groups of people ranging from wind farm operators to electricity traders and power system managers. Indeed, while this work uses wind turbine data, the analysis and approaches used may be applied to other sensor-based time series such as smart meters where similar issues with missing data are also seen. Section II introduces the analyses of real data including a test for MNAR missingness as well as outlining the missing data cases considered in the case study. Case study model setup, assumptions and creation of missing data are laid out in Section III. Properties of missing data and its effects on forecast error are presented and discussed in Section IV, while Section V contains the conclusions.

II. Methodology

Linear regression is a powerful and flexible approach widely used in time series forecasting and central to several of the methods in this work. The target variable to be predicted, y, is modelled as a weighted sum of input variables $x_1, x_2, ...x_N$, linear in the weights b_i,

$$y = b_0 + b_1 x_1 + b_2 x_2 + ...b_N x_N + \varepsilon \quad , \tag{1}$$

where ε is an error term to account for any random noise component in the model. This may be written in vector form as

$$y = X\beta + \varepsilon \quad , \tag{2}$$

where β contains all the parameters; these are estimated using sets of known inputs and outputs, typically to minimise some function of ε. The dimensions of y and X are then increased by stacking input/output rows in X and y. Now each row in X represents a single set of inputs corresponding to the output in the same row in y (labelled $s_1, s_2..s_n$), and each column in X contains all the instances of the same variable (e.g. x_1)

$$X = \begin{pmatrix} 1 & x_1^{s_1} & \dots & x_N^{s_1} \\ 1 & x_1^{s_2} & & \\ \vdots & \ddots & & \\ 1 & x_1^{s_n} & \dots & x_N^{s_n} \end{pmatrix}, \beta = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_N \end{pmatrix}, y = \begin{pmatrix} y^{s_1} \\ y^{s_2} \\ \vdots \\ y^{s_n} \end{pmatrix} .$$

In time series analysis, inputs may be lagged versions of the output variable so that, for example, $x_1 = y_{t-1}$: this is known as autoregression. Vector autoregression extends the model to include multiple outputs (forecasts at multiple sites) and allows for dependencies between sites within the forecast model. For k forecast outputs, β becomes an $(N \times k)$ matrix and the dimensions of y become $(n \times k)$.

A. Classification of Missing Data

The general properties of missing data seen in wind power time series are presented, with the proportion of missing data broken down by origin. The distributions of lengths of missing periods are also found. Although a definitive classification of data as MNAR is generally not possible [24], physical contextual justification along with the missing indicator method [25] is used to test for the likelihood of MNAR data. This method involves adding a binary indicator variable into a regression model to encode whether the original variable is present or missing for every time step. A missing indicator is included for every input and a linear regression using one lag at each site is performed. The model is then formulated as

$$y = X\beta_X + Z\beta_Z \ , \tag{3}$$

where Z is the matrix of indicators, with one for each element in X. Missing elements in X are filled with zeros. Coefficients in β_Z that are significantly different from zero imply that the mean of the missing data points is not equal to that of the non-missing data. Therefore, the observed and unobserved data follow different distributions and the data for this variable can be said to be MNAR at the chosen significance level. The proportion of site forecasts for which the indicator variable is significant then gives a suggestion of the likelihood that variable displays MNAR missing data. As well as indicating that missing and non-missing cases follow different distributions, a significant coefficient for the

indicator variable suggests the output y has some dependence on whether the value is missing, which is a significant factor in the performance of different mitigation methods [26].

B. Missing Data Scenarios

Data may be missing from the training dataset in both the X and Y matrices, compromising parameter estimation in the model training process, or input data required to generate a forecast may be missing, compromising the production of operational forecasts. In addition to the ideal case with no missing data, the following missing data scenarios are simulated to evaluate the effect of different types of missing data on forecast performance:

- Missing training data: Five levels of missing training data are synthesised, mimicking patterns seen in real datasets
- Missing spatial forecast inputs at a single time instance, i.e. the columns in X for multiple sites with the same time lag are missing
- Missing temporal forecast inputs from a single location: the columns in X for multiple lags from a given target site are missing

The structure of the forecasting method, including the missing data mitigation methods detailed below, is shown in Fig. 1. The following methods for dealing with missing data are by no means exhaustive, but represent a range of approaches.

Missing Training Data: In addition to complete case analysis (where rows with any missing data in the training set are dropped [18]), three other approaches for managing missing training data are tested. The first alternative is the application of inverse probability weights to correct the bias in complete case estimation. Two other imputation-based methods are tested; mean imputation is a simple approach while the more involved multiple imputation technique takes relationships with other variables in the data into account.

Inverse probability weighting estimates the probability of a given row in the training data being complete using logistic regression [19]. Due to the importance of variable selection for this model, principal component analysis was used to ensure no linear dependencies between inputs. The non-missing data are then weighted in the final model estimation by the inverse of their probabilities of being complete. This ensures instances (rows in X) with a low probability of being complete are given a higher weighting, so that rows which are more likely to be similar to the missing rows contribute more to the model fitting, with the intention of correcting the bias introduced by retaining only complete rows. However, significant information loss, especially at high missing data proportions, is still associated with this method.

Mean imputation involves simply replacing all missing values with the mean of that variable. In the case of normalised data where all columns have zero mean and unit variance, this is equivalent to filling all missing values with zeros. This method is simple to implement and preserves all of the available information, but artificially reduces the standard errors associated with imputed variables as well as altering the inter-variable dependencies.

Multiple imputation may be achieved by any method that allows for selection of the missing value from a probability distribution or even a group of possible values. Here the Markov Chain Monte Carlo method of Schunk [22] is followed. All missing values are first initialised with the mean of that variable before columns in X are iterated over, estimating new values for the missing points in each column in turn. For each iteration, a linear model taking all other columns as inputs is used to estimate the conditional mean and variance of the missing data. The new imputed variable is then a random draw from a gaussian distribution with the given mean and variance, with censoring where necessary: for power variables the distribution is censored at zero and rated power for example. The iterative process is repeated until the average distance between imputations stabilises. Imputations are repeatedly generated to create multiple separate pseudo-complete datasets. All further analysis (in this case the fitting and evaluation of a forecasting model) is carried out separately on all the imputed datasets, with the results from each combined to give a final result which accounts for the additional variation due to uncertainty in the missing values.

Missing Input Data: Missing input data occurs when incoming data feeds with the most recent information are down, or latencies mean data from site are not yet available by the time a new forecast must be issued: these latencies may vary with time so the exact combination of available forecast inputs also changes. A new forecast cannot be generated without adapting the model in some way; for each different case of missing data, two approaches to deal with missing inputs are considered. In the first approach, alternative models are fit which do not require the missing value(s) and in the second, missing data are filled with estimates.

In the first approach, named the *'re-train'* method, the linear model is re-configured and re-trained without the missing forecast input(s) (columns are dropped from X). The models used are computationally efficient enough that re-training a model whenever any forecast inputs are missing is a feasible approach. While the re-trained model loses some (potentially informative) input variables, it does not rely on estimated input values. In the second approach (the *'impute'* method), a regression model is fitted to predict the missing value(s) using the remaining available forecast inputs. The original forecasting model with all forecast inputs is then used. This requires an extra precursor model in addition to the main forecasting model which will change depending on the specific combination of inputs missing at a given time, but again this is only a modest computational burden. This additional model can never exactly replicate the values of the missing forecast inputs - they will always be estimates - so it is expected that this process will decrease the skill of the final forecast.

III. Data and Case Study

Dataset A comprised wind speed and power time series from 30 European wind farms, with missing data, with a mean site capacity of 41.8 MW and a range of 129.8 MW. Two years of 10-minute resolution data was re-sampled to 30-minute

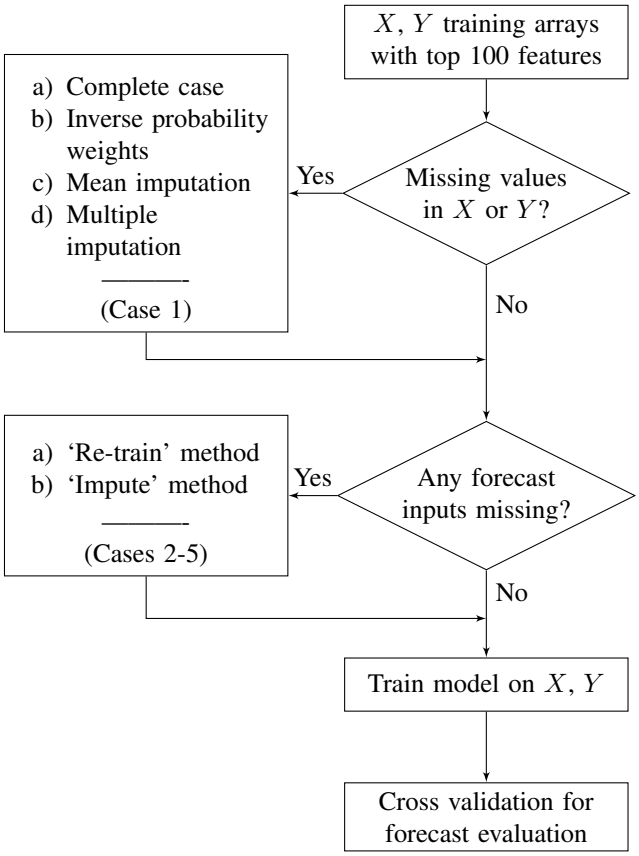


Fig. 1. Flowchart of the process for identifying and dealing with missing data within forecasts. Multiple methods a), b), c) were applied to find the optimal approach for each stage. Cases refer to the figures and discussion in section IV-B.

resolution in line with the time resolution used in the case study. This data was used to analyse levels and patterns of missing data seen in wind time series, allowing replication of these patterns in dataset B.

The case study is based on dataset B, which includes ten sites with a complete dataset at half-hourly resolution. A VAR model was used to test the effect of different types of missing data, with errors evaluated through 5-fold cross-validation. Input variables comprised 60 lags (representing 30 hours) of both power and wind speed, two-hourly mean and standard deviation measures, and monthly and diurnal dummy variables to account for seasonality and day/night variations. LASSO regularisation was used to perform feature selection as a precursor step to the model fitting [27], with the top 100 features kept as inputs for the final model. The optimum regularisation parameter value for each fold combination is found through nested cross validation prior to forecast training and testing, using only the most significant features.

Forecasts are evaluated using normalised Mean Absolute Error (NMAE), where MAE is divided by site capacity in order to compare sites of different sizes. For all the missing data cases, a 2.5-hour ahead horizon is used as statistical forecasts tend to outperform both persistence and numerical weather model based methods on this time scale.

The missing data patterns observed in dataset A were replicated in dataset B in the case study to allow comparison to the complete data case. Although the creation of MNAR missing data patterns has been studied [28], the methods focus on datasets with a small number of variables or where the 'rules' for missingness can be simply simulated. The availability of a 'real' dataset from which to replicate missing patterns allowed for a nearest neighbours approach. The two pairs of most correlated sites between datasets A and B were found using the R^2 correlation coefficient and then used to calculate the Euclidean distance between power values in Y in datasets A and B. For each row in Y in dataset B, the most similar row (nearest neighbour) for the two most correlated sites in Y in dataset A was found. The missing data pattern from the corresponding row in X in dataset A was then reproduced in that input/output pair of dataset B to give the 'closest' reproduction of missing data, labelled 'medium'. Datasets with deliberately higher and lower levels of missing data were created following the same procedure but using a different number of nearest neighbours, picking the highest or lowest missing data pattern within this subset as the one to replicate (Table I). An approach using the probability of missing data in a certain variable given the output power was also tested but resulted in all input/output pairs containing missing data, leaving no training data for the complete case analysis. This is perhaps due to the lack of dependency between missing data across variables in this method.

TABLE I
MISSING DATA CREATED IN THE COMPLETE DATASET. 'KNN' GIVES THE NUMBER OF NEAREST NEIGHBOURS SELECTED FROM AND '% ROWS WITH MISSING DATA' INDICATES THE REDUCTION IN SIZE OF THE AVAILABLE DATASET WHEN USING COMPLETE CASE ANALYSIS.

	KNN	Missing Data %	% Rows With Missing Data
Low	3	1.36%	42%
Low-Medium	2	2.48%	56%
Medium	1	6.15%	76%
Medium-High	2	9.57%	94%
High	3	11.65%	99%

A. Model assumptions

We employ a preliminary step to select out the most informative 100 features which allows fast calculation of the final step of model training and testing with a set value of the regularisation parameter. However, re-running only the final step of model training and testing when different inputs are missing requires two assumptions:

1) Missing a small number of the top 100 features makes negligible difference to the final forecast error
2) the optimum regularisation parameter is the same.

When the complete model process was re-iterated with missing data present, an improvement in error of 0.02% was seen compared to re-running only the final step, with a

significantly longer computation time. As such, retraining only the final model without repeating the feature selection process is justified in the presence of missing data. Forecasts with a longer horizon display very little difference in error (0.001%) between a model trained specifically for that horizon and one using the features and regularisation parameter optimised for a 1-step-ahead forecast. Both scenarios result in a change in the optimum regularisation parameter of more than 20% but in both cases this clearly has a negligible impact on final forecast error. The coefficients affected most by regularisation are those with small magnitude, which by definition will also be the ones with the smallest impact on the final forecast.

IV. Results and Discussion

The patterns and levels of missing data seen in real time series are presented before a set of missing data scenarios are evaluated in the case study, including testing methods to mitigate information loss.

A. Analysis of real missing data

For wind turbines, any measurements taken during non-routine operation may be considered invalid or missing as they are not representative of the unconstrained behaviour that data analysis is generally aiming to capture, i.e. power production may not match what the wind farm or turbine is normally capable of in those wind conditions. As such, three main sources of missing data were identified: data missing in the raw time series due to sensor measurement, recording or communications failures; missing periods due to site-wide maintenance works; and curtailments (when controller action is taken to limit power output). The shutdown of individual turbines may be compensated for by renormalising site power production and so is not considered as missing data. The proportion of time points in the series affected by each of these missing sources was found separately (Table II) in addition to the combined effect. Fig. 2 shows the levels of missing data seen at the sites analysed. Although the majority of sites displayed low levels of missing data with medians of 2.70% and 1.57% for power and wind speed respectively, it can be seen that a number of sites have far higher levels of over 30% missing data.

TABLE II
Mean missing data proportions by type. The total proportion of missing data is less than the sum of each individual type due to overlap in missing time points.

	Power	Wind Speed
Raw Data	5.32%	4.86%
Maintenance	0.57%	0%
Curtailments	2.89%	0%
Overall	5.71%	4.86%

The maximum observed length of missing data extended to 29 days for both wind speed and power. The mean length of a period of missing values is 3.0 hours, with a mean length of non missing periods (i.e. mean time between instances of missing data) of 48 hours. These long sections of missing data

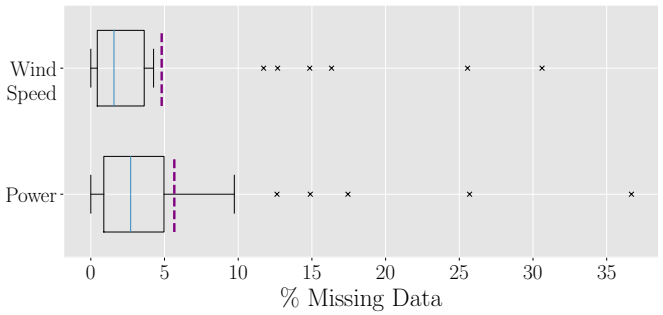


Fig. 2. Percentages of missing data seen overall in power and wind speed variables for a group of 30 wind farms. The dashed purple line represents the mean value.

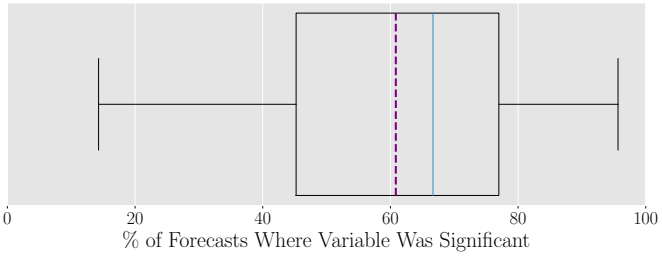


Fig. 3. Percentage of forecasts where the original variable is classed as MNAR by the missing indicator test. Dashed purple line represents the mean.

suggest the mechanism causing missing data is not completely random. The distributions of missing data for power and wind speed are also similar as the dominant cause of missing data is raw points missing, which affects both variables.

Planned maintenance activities are often scheduled for times with lower wind speeds and any work on a turbine will have an associated maximum safe wind speed over which activities will be cancelled; this suggests a correlation between times of missing data due to planned maintenance and the value of missing variables, making the data missing not at random. Wind farm sites may be more likely to be curtailed close to rated power from grid restraints limiting power flows; again this would cause an MNAR data pattern from curtailments. In addition to the physical justifications given above, an MNAR pattern was tested for using the missing indicator method for each site wind speed and power variable. The percentage of the time the corresponding missing indicator variable was classed as significant was found (Fig. 3). On average across all variables, the missing indicator was significant in the forecast 60.8% of the time. This suggests it is likely the MNAR data pattern applies to both wind speed and power measurements across the majority of sites studied. Any model that ignores MNAR missing data will under-represent behaviour seen under missing data scenarios in the training dataset and therefore might be expected to perform worse under these conditions.

B. Case study: effect of missing data on forecast error

As a benchmark to compare worsening in performance due to missing data, the forecast model with no missing inputs was evaluated and compared to a simple persistence model.

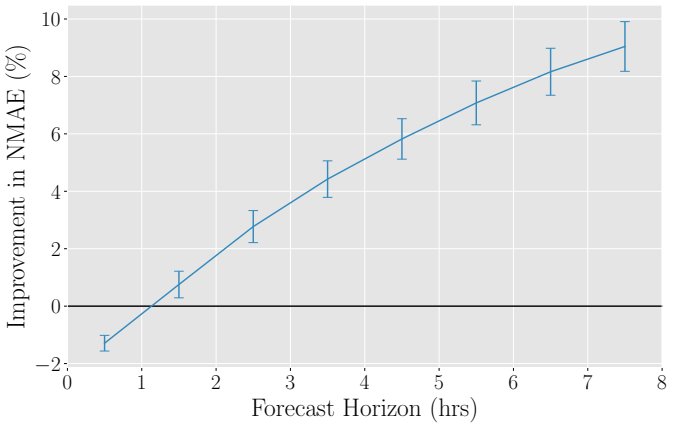


Fig. 4. Model error relative to the persistence model, at varying forecast horizons, when no data is missing. Improvement is calculated as ($NMAE_{persistence}$ - $NMAE_{VAR}$)/$NMAE_{persistence}$ × 100 so that a lower VAR error results in a higher (positive) improvement value. Error bars represent the variation over different sites.

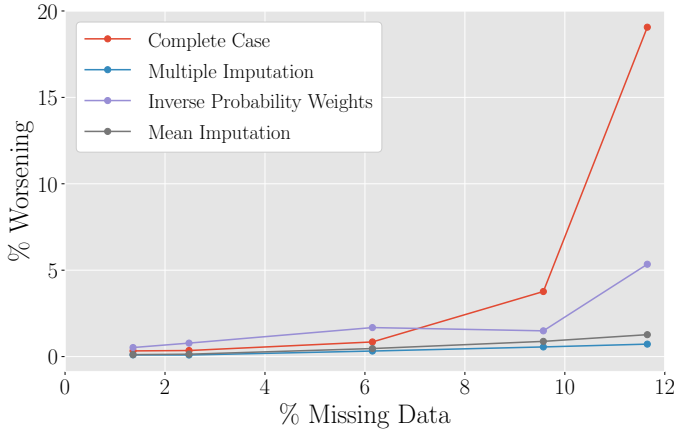


Fig. 5. Case 1: Complete case analysis used on missing data in the training set. Worsening is measured as the percentage increase in NMAE, compared to the case with a full training dataset. Error bars represent variation across different sites.

The VAR model outperforms persistence at every site for all forecast horizons 90 minutes ahead or greater, with improved relative performance at longer forecast horizons (Fig. 4).

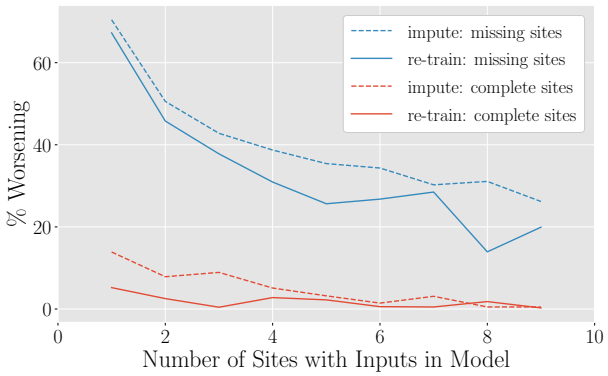
Five missing data scenarios were tested with multiple approaches to missing data applied to each. Fig. 1 shows the forecasting process with missing data methods. The forecasting performance of each method was compared to that of the model where no data is missing by calculating the 'worsening', that is, the percentage increase in NMAE of the model with missing data relative to that of the model with no missing data.

Multiple simulations of each case were run, with each site in turn (or a different combination of sites, in the case of multiple missing sites) missing data. Overall worsening plotted is the average worsening across all simulations.

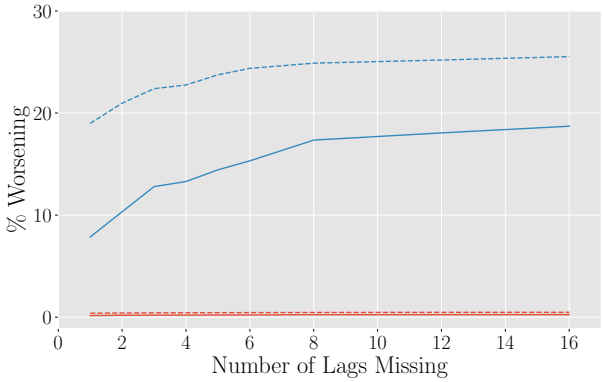
In the first case, data missing in the training set was considered; Fig. 5 shows the effect of the level of missing data on forecast performance. Complete case analysis, inverse probability weightings, mean imputation and multiple imputation were tested as mitigation methods. A greater proportion of missing data dramatically reduces the number of complete rows remaining in the training dataset as shown in Table I, decreasing the accuracy of the model fit for complete case and inverse probability weighting methods, where incomplete rows are discarded. At 11.65% missing data, forecasts using a complete case approach are 19% worse than when the training dataset is complete. Correcting the bias of the complete case approach through inverse probability weightings improved forecasts at missing data levels of 9% or more, although care must be taken to choose a suitable number of components in the principal components analysis used for this. Mean imputation displays perhaps surprisingly good performance, given the data is MNAR: for the highest level tested of 11.65% missing data, forecasts using mean imputation had NMAE 1.27% higher than the model with no missing data. Multiple imputation shows the best performance, with worsening of 0.72% across all missing data levels. The benefit of multiple imputation is most pronounced at high levels of missing data, although even modest improvements in forecast skill at lower levels may still be valuable in some applications. Multiple imputation also results in more consistent forecast errors across the set of sites analysed (no individual sites showed significantly more worsening under this method than others). However, multiple imputation is a more complex method to implement than mean imputation, both in the imputation process itself and in the combining of results from the different imputations in the final analysis. Of the methods tested, those that utilise all the available data clearly outperform those where incomplete cases are ignored. The increased skill of multiple imputation likely comes from the modelling of inter-variable relationships in the imputation process.

Cases 2-5 represent scenarios where some combination of forecast inputs is missing, seen for example during a long communications failure with a site. Two mitigation methods are tested: firstly, the 're-train' method involves deleting missing forecast input variables from the training matrix X and training a new model from this. While this reduces the number of forecast inputs in the model, no estimation of missing measurements is needed. In contrast, the 'impute' method fits a separate, precursor model using the available forecast inputs to predict the missing one(s). This allows the original model to be used with all forecast inputs but adds an extra modelling step, introducing an extra source of uncertainty. Results are plotted separately for complete sites (where all forecast inputs from that site are available) and missing sites, where some or all of the forecast inputs are missing.

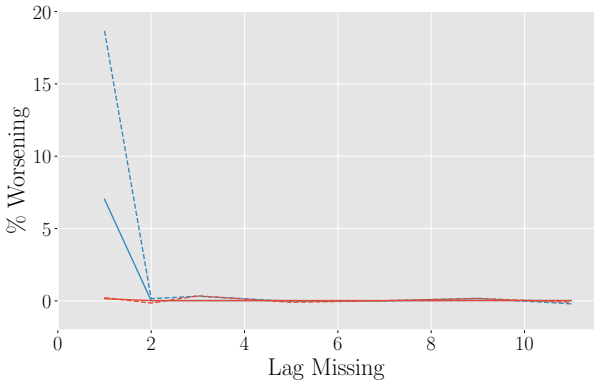
Case 2 examines the scenario where all information for a number of sites is missing; forecast performance at a missing site clearly benefits from other sites with complete inputs. Increasing the number of sites in the model may increase the probability of a complete site similar to the missing site, therefore improving forecast errors, illustrated in Fig. 6a. This demonstrates the advantage of multivariate models such as VAR in providing greater robustness to missing data through



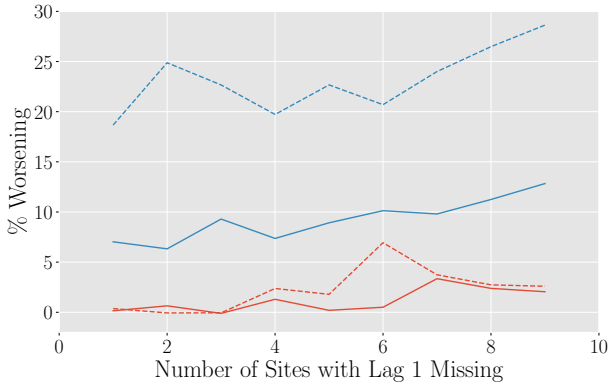
(a) Case 2: Impact of number of sites as inputs in VAR model on forecast error. Forecast error improves when more sites are included in the model, particularly for sites that are missing forecast input data.



(b) Case 3: Impact of cumulative number of lags missing from one site on forecast error (normalised to site capacity). Forecasts at missing sites are worse for longer missing periods.



(c) Case 4: Impact on forecast error of an individual lag missing from one site.



(d) Case 5: Impact on forecast error of number of sites with lag 1 missing. The more sites missing data simultaneously, the worse the error.

Fig. 6. Worsening is measured as the percentage increase in NMAE, compared to the case with no missing data. The legend in (a) applies to all subfigures. 'missing sites' show the average forecast error worsening at sites where some or all forecast inputs are missing, while 'complete sites' shows the worsening at sites whose forecast inputs are all available.

the inter-site relationships captured. The 're-train' method consistently outperforms the 'impute' method, possibly due to added error from an additional estimation step in the 'impute' method.

Case 3 examines the effect of the length of a missing period at a single site; as may be expected, removing the most recent lags makes the greatest difference to forecast error as these lags tend to carry the highest weight in the regression model (indicating they are the best predictors). Forecasts continue to worsen with increasing length of missing period, but the largest proportion of the loss of forecast skill comes from missing the most recent information, shown in Fig. 6b. The worsening seen at the complete sites increases very slightly with increased missing period but is small in comparison with that at missing sites.

Case 4 evaluates the impact of the loss of a single input variable. This reinforces the finding of case 3 that missing the first lag, corresponding to the most recent information, has the largest detrimental effect on the forecast. As might be expected, missing a single piece of information has a smaller effect than missing several points, shown by the lower levels of worsening in Fig. 6c compared to the other cases.

In Case 5 the effect of data missing simultaneously across sites is shown by evaluating forecast performance with lag 1 missing at a number of sites. As expected, an increased number of sites with the most recent information missing results in a worsening of forecast performance across all sites, but notably more so at missing sites, shown in Fig. 6d.

While these results stem from a VAR forecasting model, it is expected that similar relationships between missing variables and loss of forecast skill would be seen with other models. Univariate models where each site forecast comes from its own separate model would be expected to be less robust to missing data, as there are no inter-site dependencies available to (at least partially) compensate for missing values. The persistence model is an extreme case, where a long missing period at one site means a forecast will be based on a single piece of out-of-date information.

V. Conclusions

The properties of missing data in real SCADA time series have been found, before the effect of various missing data

scenarios on forecast skill were simulated through a case study. Real wind power data is shown to have typical median levels of missing data of 2.70% for the power variable and 1.57% for wind speed. However, some sites may display levels up to 36%, greatly reducing forecast skill. Data is missing not at random, meaning care must be taken to use an appropriate missing data technique. The impact of missing data on wind power forecasts in an autoregressive framework has been demonstrated, with the most appropriate mitigation methods identified. The key results are summarised:

- Missing training data can have a significant impact on results if not dealt with appropriately; multiple imputation is found to be the best of the methods considered here to compensate for this
- If inputs to an operational forecast model are missing, retraining the model without these inputs results in better performance than filling the missing values using a regression model based on available inputs
- Forecast error improves across all sites when more sites are included in the model, with particular improvement at sites that are missing forecast input data; therefore, spatio-temporal models including a greater number of sites are more robust to missing data
- Forecasts continue to worsen with increasing length of missing period, but the largest proportion of the loss of forecast skill comes from missing the most recent information

While these results are from a case study using a VAR forecasting model, future work could extend this to other models. It is expected the results would be similar for other models, as the change in forecast skill is related more to the loss of information from the missing variable(s) than the modelling framework itself. In summary, awareness of the properties of missing data, its potential impact on model performance and use of suitable mitigation techniques is essential to realise that model's full potential.

Acknowledgements

Rosemary Tawn is supported by The Data Lab Innovation Centre funding with co-funding from Natural Power Consultants Ltd. Thanks also to Ciaran Gilbert for reviewing the paper. Case study data was provided by Natural Power.

References

[1] P. Matthews and I. Scherr, *Annual Compendium of Scottish Energy Statistics*. Scottish Government, 2019.

[2] P. Pinson, "Wind energy: Forecasting challenges for its operational management," *Statistical Science*, vol. 28, no. 4, pp. 564–585, Nov. 2013.

[3] J. Wang, Y. Song, F. Liu, and R. Hou, "Analysis and application of forecasting models in wind power integration: A review of multi-step-ahead wind speed forecasting models," *Renewable and Sustainable Energy Reviews*, vol. 60, pp. 960–981, Jul. 2016.

[4] P. Pinson, C. Chevallier, and G. N. Kariniotakis, "Trading wind generation from short-term probabilistic forecasts of wind power," *IEEE Transactions on Power Systems*, vol. 22, no. 3, pp. 1148–1156, Aug. 2007.

[5] G. Giebel, R. Brownsword, and G. Kariniotakis, "The state of the art in short-term prediction of wind power: A literature overview," Risø DTU, Tech. Rep. 2nd Edition, 2011.

[6] J. Salmon and P. Taylor, "Errors and uncertainties associated with missing wind data and short records: Uncertainties and missing data," *Wind Energy*, vol. 17, no. 7, pp. 1111–1118, Jul. 2014.

[7] A. Coville, A. Siddiqui, and K.-O. Vogstad, "The effect of missing data on wind resource estimation," *Energy*, vol. 36, no. 7, pp. 4505–4517, Jul. 2011.

[8] Y. Hu, Y. Qiao, J. Liu, and H. Zhu, "Adaptive confidence boundary modeling of wind turbine power curve using SCADA data and its application," *IEEE Transactions on Sustainable Energy*, vol. 10, no. 3, pp. 1330–1341, Jul. 2019.

[9] S. H. Hosseini, C. Y. Tang, and J. N. Jiang, "Calibration of a wind farm wind speed model with incomplete wind data," *IEEE Transactions on Sustainable Energy*, vol. 5, no. 1, pp. 343–350, Jan. 2014.

[10] M. Martinez-Luengo, M. Shafiee, and A. Kolios, "Data management for structural integrity assessment of offshore wind turbine support structures: data cleansing and missing data imputation," *Ocean Engineering*, vol. 173, pp. 867–883, Feb. 2019.

[11] J. W. Messner and P. Pinson, "Online adaptive lasso estimation in vector autoregressive models for high dimensional wind power forecasting," *International Journal of Forecasting*, Apr. 2018.

[12] J. Browell, D. R. Drew, and K. Philippopoulos, "Improved very short-term spatio-temporal wind forecasting using atmospheric regimes," *Wind Energy*, May 2018.

[13] Y. Zhang and J. Wang, "A distributed approach for wind power probabilistic forecasting considering spatio-temporal correlation without direct access to off-site information," *IEEE Transactions on Power Systems*, vol. 33, no. 5, pp. 5714–5726, Sep. 2018.

[14] Y. Mao and M. Jian, "Data completing of missing wind power data based on adaptive BP neural network," in *2016 International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*. Beijing: IEEE, Oct. 2016, pp. 1–6.

[15] B. Lotfi, M. Mourad, M. B. Najiba, and E. Mohamed, "Treatment methodology of erroneous and missing data in wind farm dataset," in *Conference on Systems, Signals & Devices*. Sousse: IEEE, Mar. 2011, pp. 1–6.

[16] J. G. Ibrahim and G. Molenberghs, "Missing data methods in longitudinal studies: a review," *TEST*, vol. 18, no. 1, pp. 1–43, May 2009.

[17] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976.

[18] R. Little, "Regression with missing X's: a review," *Journal of the American Statistical Association*, vol. 87, no. 420, 1992.

[19] S. R. Seaman and I. R. White, "Review of inverse probability weighting for dealing with missing data," *Statistical Methods in Medical Research*, vol. 22, no. 3, pp. 278–295, Jun. 2013.

[20] R. Little and D. B. Rubin, *Statistical analysis of missing data*, 2nd ed. Wiley, 2002.

[21] P. D. Allison, "Multiple imputation for missing data: a cautionary tale," *Sociological Methods & Research*, vol. 28, no. 3, 2000.

[22] D. Schunk, "A Markov chain Monte Carlo algorithm for multiple imputation in large surveys," *Advances in Statistical Analysis*, vol. 92, no. 1, pp. 101–114, Feb. 2008.

[23] J. Honaker and G. King, "What to do about missing values in time-series cross-section data," *American Journal of Political Science*, vol. 54, no. 2, pp. 561–581, Apr. 2010.

[24] N. J. Horton and K. P. Kleinman, "Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models," *The American Statistician*, vol. 61, no. 1, pp. 79–90, Feb. 2007.

[25] T. D. Pigott, "A review of methods for missing data," *Educational Research and Evaluation*, vol. 7, no. 4, pp. 353–383, Dec. 2001.

[26] Y. Ding and J. Simonoff, "An investigation of missing data methods for classification trees applied to binary response data," *Journal of Machine Learning Research*, vol. 11, 2010.

[27] G. James, D. Witten, T. Hastie, and R. Tibshirani, Eds., *An introduction to statistical learning: with applications in R*, ser. Springer texts in statistics. New York: Springer, 2013, no. 103, oCLC: ocn828488009.

[28] M. S. Santos, R. C. Pereira, A. F. Costa, J. P. Soares, J. Santos, and P. H. Abreu, "Generating synthetic missing data: A review by missing mechanism," *IEEE Access*, pp. 1–1, 2019.