

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/331606775>

Analysis of wind missing data for wind farms in Isthmus of Tehuantepec

Conference Paper · November 2018

DOI: 10.1109/ROPEC.2018.8661457

CITATIONS

2

READS

306

5 authors, including:



Claudia N. Sánchez

Universidad Panamericana

26 PUBLICATIONS 70 CITATIONS

[SEE PROFILE](#)



Josué Enriquez_Zárate

AP-ENGINEERING INNOVACIÓN TECNOLÓGICA EN ENERGÍAS S.A DE C.V.

28 PUBLICATIONS 141 CITATIONS

[SEE PROFILE](#)



Ramiro Velázquez

Universidad Panamericana

184 PUBLICATIONS 1,451 CITATIONS

[SEE PROFILE](#)



Mario Graff

Consejo Nacional de Ciencia y Tecnología

104 PUBLICATIONS 797 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:



Data mining [View project](#)



Condition Monitoring of Damaged Bearings [View project](#)

Analysis of wind missing data for wind farms in Isthmus of Tehuantepec

Claudia N. Sánchez ^{*‡}, J. Enríquez-Zárate [†] Ramiro Velázquez* Mario Graff [‡] and S. Sassi [§]

^{*} Universidad Panamericana. Facultad de Ingeniería.

Aguascalientes, Aguascalientes, 20290, México. Email: {cnsanchez, rvelazquez}@up.edu.mx

[†] Universidad de los Andes. Facultad de Ingeniería y Ciencias Aplicadas.

Las Condes, Santiago, Chile. Email: jenriquezza@gmail.com

[‡]INFOTEC - Centro de Investigación e Innovación en Tecnologías de la Información y Comunicación.

Aguascalientes, Aguascalientes, México. Email: mario.graff@infotec.mx

[§]Department of Mechanical and Industrial Engineering, College of Engineering,
Qatar University. Al Tarfa, Doha 2713, Qatar. Email: sadok.sassi@qu.edu.qa

Abstract—The availability of reliable data related to the behavior of the wind in a wind farm is very useful to determinate with accuracy some aspects as power curve of a wind energy turbine and the wind's speed in an interval of times. The precision in the prediction process of wind behavior is useful for reducing structural efforts in the wind-turbine rotor system, and even in the tower section. In practice, sensors are used to acquire data for monitoring wind farms which occasionally may tend to fail causing an incomplete information from the sensor. This work is focused on the imputation of missing data based on a combination of interpolation and regression models. Our experiments show that this approach is useful for correlated time series, considering the direction and speed wind for 20 and 40 meters of height in wind farms located in the Isthmus of Tehuantepec in Oaxaca state in Mexico. Finally, the approach of combination methods is effective to solve the problem of missing data in the database of wind in wind farms.

Keywords—Wind missing data, imputation on time series, correlated time series, machine learning.

I. INTRODUCTION

Recently, the generation of electricity from wind energy around the world has a pronounced interest for low power wind turbine as well as for high-power wind turbines. The quality and wind speed establish the capacity of power's extraction in a geography zone, used to determine the maximum efficiency of a wind-turbine from its power output curve. An unfortunate monitoring process causes a wind prediction problem, which limits the possibility of obtaining the maximum production of energy of the wind farm. For different reasons, sometimes the data provided by sensors that measure the wind velocity and direction have a lot of missing data. Imputation method involves the replacement of missing values with some values that have been estimated based on data mining of available information in the data set [6]. However, several studies have shown that the performance of missing values imputation algorithms is significantly affected by factors such as correlation structure in the data, the missing data mechanism, the distribution of missing entries in the data, and the percentage

of missing values in the data. Selecting the right algorithm may significantly boost the accuracy of the imputation results since there is no single imputation algorithm that performs the best in every situation [4].

There are several techniques or algorithms that perform the imputation in time series. Norazian *et al.* used interpolation and mean imputation techniques to calculate missing values in air pollution [12]. Data Interpolating Empirical Orthogonal Functions (DINEOF) [2] was used in a multivariate approach to reconstruct missing data in sea surface temperature, chlorophyll, and wind satellite fields.

The method of wind shear coefficient (WSC) [15] is an approach to interpolate the measured wind data of wind farm. The equation of WSC is defined as $V_2 = V_1 * (h_2/h_1)^\alpha$, where h_1 and h_2 are the installed height of two sensors of the wind mast, V_1 and V_2 are wind speeds of the two installed heights, and α is wind shear coefficient, it is obtained as $\alpha = \log(V_2/V_1)/\log(h_2/h_1)$. However, in real work, influenced by environment and measured error, wind shear coefficient α is not a constant. This brings a great impact on the practical engineering application of the above approach [15]. Given the success of the application of Machine Learning (ML) techniques, some of them have been applied to the reconstruction of missing data. Lui *et al.* [9] used Gaussian process regression models to calculate missing data for wind power prediction. In [10] a new Measure Correlate Predict (MCP) method was used to establish a missing wind speed data using temporal interpolation and extrapolation method, considering all mixed uncertainties, based on granular computing theory by adopting the cloud model method; support vector regression method, artificial neural networks, genetic algorithm and fuzzy c-means clustering algorithm as tools. The maintenance of a wind turbine is a necessary service to keep it working and producing electrical energy. The ML for detection and diagnostic of delamination in Wind Turbine Blades (WTB) is presented in [3] using Autoregressive Yule-Walker model, which is employed for feature extraction using time series analysis and predictive models for fault detection

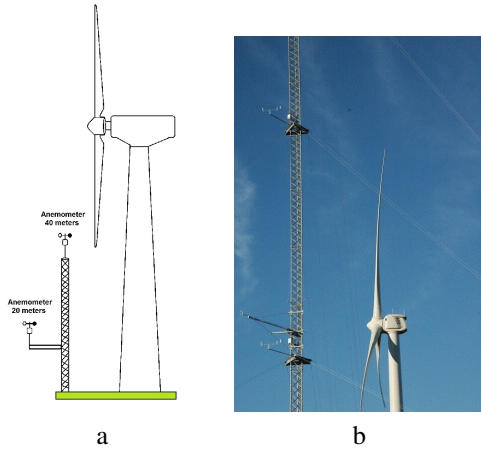


Fig. 1. Instrumented meteorological tower in a wind farm: a) Schematic diagram, b) Clipper wind turbine located in the Isthmus of Tehuantepec, Oaxaca, Mexico.

using Artificial Neural Network (ANN). The monitoring of power curves can be used for both wind turbines and wind farms with the intention of extracting the maximum power. The ML methodology is used in [11] as an approach for detection of an anomalous functioning condition in a wind farm, which can affect the power extraction. The authors herein present the use and comparison of three different ML models such as a self-supervised neural network called GMR (Generalized Mapping Regressor), a feed-forward Multi-Layer Perceptron (MLP) and a General Regression Neural Network (GRNN).

This paper is organized as follows. Section II provides the data description and exploration. The techniques that are applied to predict the missing values are shown in Section III. Experiments and results are described in Section IV. Finally, conclusions are given in Section V.

II. DATA DESCRIPTION AND EXPLORATION

The public data was obtained from the Instituto Nacional de Electricidad y Energías Limpias (INEEL). It was stored in EXCEL files named with the key DP-XXZZ-MM-YYYY, where DP means public data, XX are representative letters of the Mexican Republic state where the station is installed, ZZ the station number, MM the month and YYYY the year. For this paper, the file DP-LV01-06-2000.xls was used. It corresponds to June 2000 of one station installed in the Isthmus of Tehuantepec, Oaxaca, México. Figure 1 shows an instrumented meteorological tower.

The data contains 9 columns:

- Julian day, the day number of the year.
- Hour and minute in format hhmm.
- Average wind velocity at 20 meters high (m/s).
- Wind direction in grades (from 0° to 359°) that are measured from the north clockwise.
- Standard deviation of the wind velocity at 20 meters high (m/s).
- Maximum wind velocity at 20 meters high (m/s).

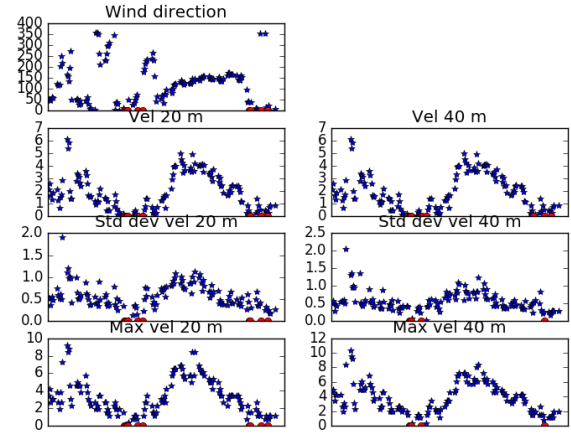


Fig. 2. Data corresponding to the 167th day

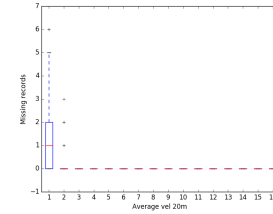


Fig. 3. Velocity at 20m Vs Missing value number

- Average wind velocity at 40 meters high (m/s).
- Standard deviation of the wind velocity at 40 meters high (m/s).
- Maximum wind velocity at 40 meters high (m/s).

The data can be seen as seven time series with a value every ten minutes. Figure 2 shows the time series of the 167th day. The red points are the missing data. There are 4319 records, 68 of them have missing data (1.5%). Two rules were found: (1) when a value of the velocity at 20m is missing, also the values of the wind direction, standard deviation at 20m and the maximum velocity at 20m are missing; and (2) when a value of the velocity at 40m is missing, also the values of standard deviation at 40m and the maximum velocity at 40m are missing (see Table I). There are 62 and 39 records where the values of the velocity at 20m and 40m are missing, respectively. In Figure 2, it can be seen that the missing values are in period of times where the wind velocity is low. The same pattern was observed in the remaining days. In order to analyze the relationship between the wind velocity and the number of missing data, for every two hours (12 records), the average of the velocity at 20m (the missing data not included) and the number of missing data are calculated. In Figure 3, it can be seen that the missing data is present where the wind velocity has low values. The majority of them are where the wind velocity is less or equals than 1 m/s, and some of them when it is between 1 and 2 m/s.

TABLE I
ORIGINAL DATA EXAMPLE

Julian day	HourMinute	Vel 20m	Wind direction	Std dev 20m	Max vel 20m	Vel 40m	Std dev 40m	Max vel 40m
160	1840	0.428	269.5	0.406	1.527	0.232	0.362	1.527
160	1850	0.312	244.1	0.461	1.527	0.204	0.368	1.527
160	1900	0.005	249.1	0.044	0.382	0	0	0
160	1910	0	0	0	0	0	0	0
160	1920	0.046	249.1	0.139	0.764	0	0	0
160	1930	0.151	249.1	0.222	0.764	0	0	0
160	1940	0.022	249.1	0.094	0.764	0	0	0
160	1950	0	0	0	0	0.003	0.031	0.382
160	2000	0	0	0	0	0.104	0.224	0.764
160	2010	0.605	311.8	0.467	1.527	0.627	0.26	1.527
160	2020	0	0	0	0	0.003	0.031	0.382
160	2030	0	0	0	0	0	0	0
160	2040	0	0	0	0	0	0	0
160	2050	0.018	328.5	0.097	0.764	0.075	0.196	0.764
160	2100	0	0	0	0	0	0	0
160	2110	0	0	0	0	0.033	0.116	0.764
160	2120	0.283	328.5	0.313	1.145	0.85	0.517	1.527
160	2130	0.78	328.4	0.21	1.145	0.904	0.287	1.527
160	2140	0.736	339	0.24	1.145	0.261	0.419	1.527
160	2150	1.487	354.9	0.468	2.291	1.616	0.382	2.673
160	2200	1.797	351.9	0.299	2.673	2.016	0.236	2.673

Figure 2 also shows a correlation between the wind velocity at 20 and 40m, standard deviation of the wind velocity at 20 and 40m, and finally maximum wind velocity at 20 and 40m. The correlation coefficient of these relations are 0.99, 0.97 and 0.99 respectively.

III. METHODOLOGY

The proposed methodology for prediction and measurement of imputation data is performed in the following steps: (1) remove some values, (2) interpolate, (3) predict values using regression models, and (4) combine interpolation and regression outputs. The overall framework is shown in the Figure 4.

A. Remove some values

In order to measure the performance of each imputation technique, it is necessary to remove values in valid records to compare the real values and imputed values.

The removing process is performed trying to preserve the essence of the missing values in the original data: the missing values appear in period of times where the wind velocity is low. Every 12 records, the wind velocity average is calculated, if there are missing data in those records, they are omitted. If the average wind velocity is less than 1.5 m/s, some values are removed. With 60% of probability, the values of wind direction, wind velocity at 20m, wind velocity standard deviation at 20m and maximum wind velocity at 20m are removed. On the other hand, with 40% of probability, the values of wind velocity at 40m, wind velocity standard deviation at 40m and maximum wind velocity at 40m are removed. Figure 5, left side, shows the values that were removed from the 166th day.

B. Interpolation

Interpolation is a technique capable of calculating the values of missing data within the range of given known values.

The linear interpolation fits a straight line between two given points a and b . If a point between those points a and b is given, the output can be calculated as the value in the line. The line equation is:

$$y = y_1 k(x + x_1) \quad (1)$$

where $k = (y_2 - y_1)/(x_2 - x_1)$. To predict the value of x , the constraint $x_1 < x < x_2$ needs to be respected.

The cubic interpolation fits a polynomial function given several points. The function equation is:

$$y = ax^3 + bx^2 + cx + d \quad (2)$$

where a, b, c and d are parameters that need to be optimized.

Both interpolation techniques, linear and cubic, were performed using the python library scipy [8].

C. Wind direction interpolation

The interpolation of the wind direction cannot be applied with the original data using the continuous range 0° to 359° . For example, if the linear interpolation is used and the points values for the times t and $t + 2$ are 354 and 2, the middle point in the time $t + 1$ calculated by the interpolation will be 178. However, the correct value can be obtained if the values -2 and 2 are used for the times t and $t + 2$. To avoid this problem, we transform all the values of the wind direction with the sinusoidal and cosine functions. Then, an interpolation for each transformation is performed. Finally, the wind direction values are obtained using the function atan2 with the interpolated values.

D. Regression

To take advantage of the strong correlation between the variables measured at 20m and 40m (wind velocity, standard deviation of the wind velocity and maximum wind velocity), several regression models are performed. With the idea to understand the behavior of the data, univariable and multivariable, and also, linear (Linear Regression) and non linear regressors (Support Vector Machines for regression and MultiLayer Perceptron) were tested.

Mathematically, a regression problem can be defined as follows. Given N input vectors x_1, x_2, \dots, x_N in a d -dimensional space $x_i \in \mathbb{R}^d$, and N output values y_1, y_2, \dots, y_N where $y_i \in \mathbb{R}$, it is wanted to find the function f that minimizes the difference between $f(x_i)$ and y_i .

Simple linear regression [7] predicts a quantitative response on the basis of a single predictor variable x_i . It means, for this model $x_i \in \mathbb{R}$. It assumes there is approximately a linear

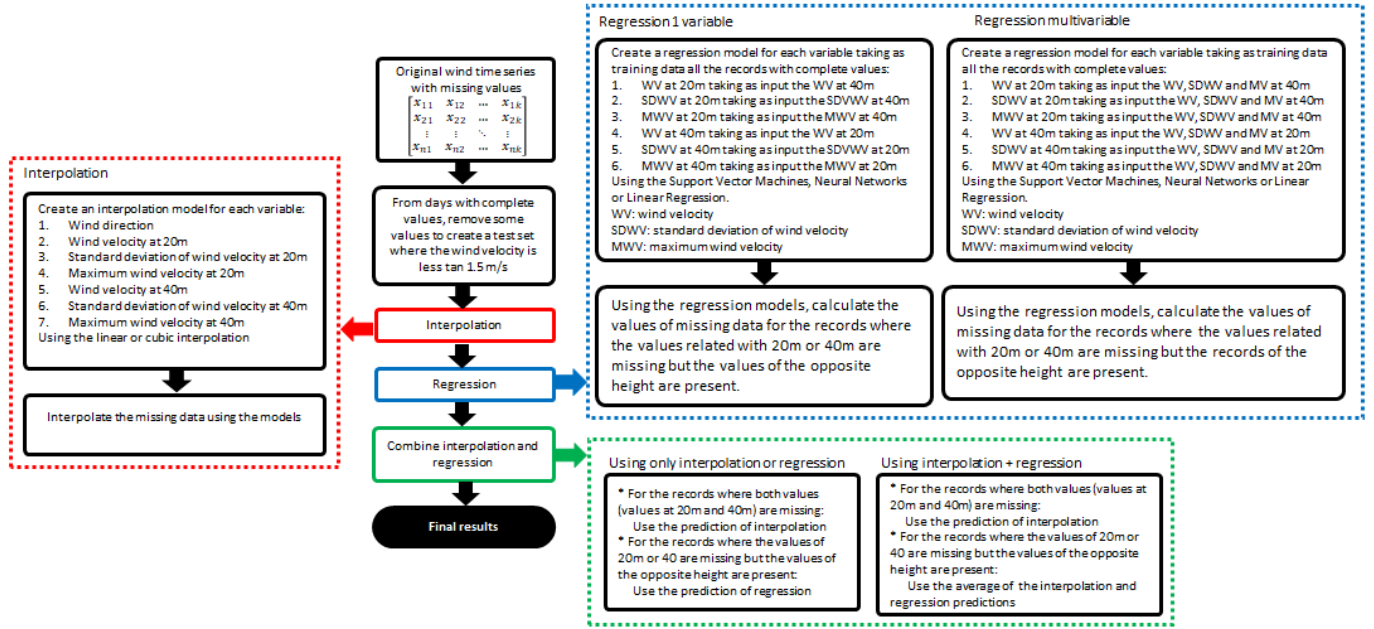


Fig. 4. The overall framework of the proposed technique.

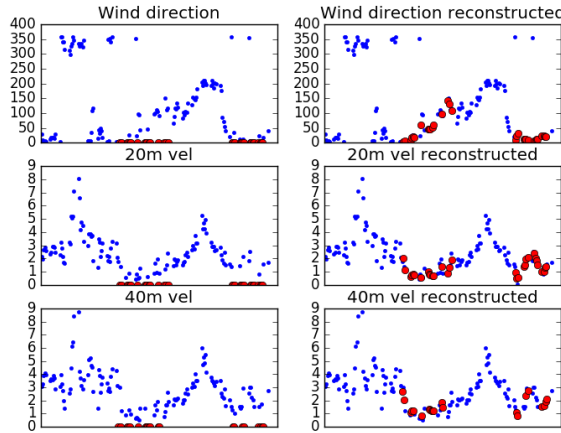


Fig. 5. Reconstruction of missing data in the 166th day

relationship between the values x_i and y_i . Mathematically, it can be written as $f(x_i) = \beta_0 + \beta_1 x_i$. However, in practice it is often needed to have more than one predictor ($x_i \in \mathbb{R}^d$). Multiple linear regression model [7] takes the form: $f(x_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_d x_{id}$, where x_{ji} represents the j th predictor and β_j quantifies the association between that variable and the response. The parameters of linear regression β are estimated using least squares.

Support Vector Machines (SVM) has been used successfully to solve supervised learning problems because given a training set it can learn quickly a linear model. Smola and Scholkopf [14] describe how SVM can be used specifically to regression problems. The model proposed by SVM defines

$f(x_i) = \langle w, x_i \rangle + b$, where $\langle \cdot, \cdot \rangle$ represents the dot product, and the vector w and the scalar value b are the parameters. To find the values of w and b , SVM solves an optimization problem that constructs a margin that wraps the linear model. SVM tries to minimize the distance between the margin and the value $f(x_i)$ for those input-output training data (x_i, y_i) where $|f(x_i) - y_i| \geq \epsilon$. It means all the pairs of input-output training where $|f(x_i) - y_i| < \epsilon$ do not contribute to the optimization function. The optimization problem is defined in equation 3.

$$\begin{aligned} \min_{w, b, \zeta, \zeta^*} \quad & \frac{1}{2} \langle w, w \rangle + C \sum_{i=1}^n (\zeta_i + \zeta_i^*) \\ \text{subject to} \quad & y_i - f(x_i) \leq \epsilon + \zeta_i \\ & f(x_i) - y_i \leq \epsilon + \zeta_i^* \\ & \zeta_i, \zeta_i^* \geq 0, i = 1, 2, \dots, n \end{aligned} \quad (3)$$

The optimization problem is divided in two terms: the first one $\frac{1}{2} \langle w, w \rangle$ prevents the overfitting and the second $\sum_{i=1}^n (\zeta_i + \zeta_i^*)$ minimizes the distances between the margin and $f(x_i)$. The value of C gives the relevance to each term (In this case, $C=1.0$). One of the main characteristics of SVM is that it allows to add kernel functions to learn non linear patterns. In this case, an exponential kernel is used $\exp(-\gamma \|x - x'\|^2)$, where $\gamma = 1/n_{\text{features}}$.

The term "neural network" has its origins in attempts to find mathematical representations of information processing in biological systems [5]. The Neural Networks (NN) have been used very broadly to solve a wide range of different problems with distinct types of mathematical models and structures. The perceptron is the basic processing element [1]. Association with each input x_{ji} is a connection weight or synaptic weight w_j , and the output in the simplest case is a weighted sum of

the inputs. The NN adds an activation function to change the behavior of the linear combination:

$$f(x_i) = f\left(\sum_{j=1}^d w_j x_{ji} + w_0\right) \quad (4)$$

w_0 is the intercept value to make the model more general. For this work, the activation function is the rectified linear unit function $f(x) = \max(0, x)$. A MultiLayer Perceptron (MLP) is a network of several perceptrons structured in layers. The first layer receives the original inputs. The hidden (middle) layers receive as input the outputs of the first layer and provide their answers to the next layer. The last layer generates the final output. For this work, only a hidden layer of 100 neurons is used. There are different algorithms to learn or calculate the weights, in this case, the stochastic gradient-based optimizer is used.

All the regressors were performed using the python library scikit-learn [13]. The regression models were trained using all the complete records in the dataset.

The 1 variable linear regression models are trained as follows:

- Wind velocity at 40m, taking as input: wind velocity at 20m.
- Wind velocity at 20m, taking as input: wind velocity at 40m.
- Standard deviation of the wind velocity at 40m, taking as input: standard deviation of the wind velocity at 20m.
- Standard deviation of the wind velocity at 20m, taking as input: standard deviation of the wind velocity at 40m.
- Maximum wind velocity at 40m, taking as input: maximum wind velocity at 20m.
- Maximum wind velocity at 20m, taking as input: maximum wind velocity at 40m.

The multivariable regression models are trained as follows:

- Wind velocity at 40m, taking as input: wind velocity, standard deviation and maximum wind velocity at 20m.
- Wind velocity at 20m, taking as input: wind velocity, standard deviation and maximum wind velocity at 40m.
- Standard deviation of the wind velocity at 40m, taking as input: wind velocity, standard deviation and maximum wind velocity at 20m.
- Standard deviation of the wind velocity at 20m, taking as input: wind velocity, standard deviation and maximum wind velocity at 40m.
- Maximum wind velocity at 40m, taking as input: wind velocity, standard deviation and maximum wind velocity at 20m.
- Maximum wind velocity at 20m, taking as input: wind velocity, standard deviation and maximum wind velocity at 40m.

E. Combination interpolation and regression

Two ways of combining the interpolation and regression outputs are proposed:

- Interpolation or regression. For the records where both values (values at 20m and 40m) are missing: use the prediction of interpolation. For the records where the values of 20m or 40m are missing, but the values of the opposite height are present: use the prediction of regression.
- Interpolation + regression. For the records where both values (values at 20m and 40m) are missing: use the prediction of interpolation. For the records where the values of 20m or 40m are missing, but the values of the opposite height are present: use the average of the interpolation and regression predictions.

For the wind direction values reconstruction, only the interpolation techniques are applied because the data only have one time series for those values.

IV. EXPERIMENTS AND RESULTS

Several combinations of interpolation and regressions models were performed. To measure the effectiveness of each approach, we use the mean absolute error (MAE) [9] between the real values y_i and the predicted ones \hat{y}_i using the records whose values were intentionally removed. MAE is defined as $\frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i|$. In general, smaller values indicate better performance.

Table II shows the results of the techniques combinations. To simplify the reading of results the next abbreviations are used. CI: cubic interpolation, LI: linear interpolation, LR: linear regression, SVM: support vector machines regression and MLP: MultiLayer Perceptron regression. The word "or" indicates the combination technique "interpolation or regression", and the "+" the technique "interpolation + regression". The best results are in bold.

First, it can be observed, the linear interpolation by itself produces better results than cubic interpolation to calculate the missing values in all the variables. In the multivariable models, the linear regression models generate better results than the non linear ones. The combination of linear interpolation + multivariable linear regression generates the best results in general, but it is too close to the results of the same approach but univariable.

The error values obtained were considerably small taking into account the maximum values and the meaning of the variables.

V. PREDICTION OF MISSING DATA

Once the efficiency of our methodology has been proved, the next step consists of calculating the values of all the missing data. In base of our results we propose:

- 1) Create seven linear interpolation models, one for each variable.
- 2) Create six multivariable linear regression models to predict the variables: wind velocity, standard deviation of the wind velocity and maximum wind velocity at 20 and 40m, taking as input: wind velocity, standard deviation and maximum wind velocity in the opposite height.

TABLE II
AVERAGE ERROR

Descripción	Max value	1 variable models				Multivariable models					
		CI	LI	LI or LR	LI + LR	LI or SVM	LI or MLP	LI or LR	LI + SVM	LI + MLP	LI + LR
Wind direction (grades)	360	20.01	17.97	-	-	-	-	-	-	-	-
20m Vel. (m/s)	16.12	0.59	0.38	0.27	0.29	0.28	0.58	0.27	0.31	0.39	0.27
20m Std. dev. vel. (m/s)	2.69	0.19	0.12	0.09	0.09	0.08	0.10	0.08	0.10	0.10	0.10
20m Vel max (m/s)	22.53	0.64	0.42	0.32	0.32	0.38	0.37	0.32	0.35	0.35	0.30
40m Vel (m/s)	18.60	0.66	0.34	0.34	0.30	0.35	0.49	0.32	0.25	0.48	0.28
40m Std. dev. vel. (m/s)	2.75	0.23	0.09	0.12	0.09	0.12	0.13	0.11	0.10	0.11	0.09
40m Vel max (m/s)	24.82	1.08	0.39	0.41	0.33	0.45	0.39	0.40	0.34	0.32	0.33

3) For each missing value.

- If the value corresponds to wind direction: use the prediction of interpolation. This is because it is not possible to generate a regression model for this variable.
- Otherwise. For the records where both values (at 20m and 40m) are missing: use the prediction of interpolation. For the records where the values of 20m or 40m are missing, but the values of the opposite height are present: use the average of the interpolation and regression predictions.

VI. CONCLUSION

The use of the combination of linear interpolation and linear regression techniques was successfully applied to solve the problem of missing data from the database proportioned by the Instituto Nacional de Electricidad y Energías Limpias (INEEL) provided by a particular wind farm located in the Isthmus of Tehuantepec in Oaxaca, Mexico.

The linear interpolation method achieves acceptable and interesting results to predict missing values. However, the output of the sensors positioned around of 20 and 40 meters are correlated time series, and it is helpful to use the values of the one sensor to predict the values of the second sensor when one of them has problems. The combination of linear regression with linear interpolation generates the best performance. The average absolute errors of the predicted values of the wind speed are 0.27 and 0.28 m/s around of 20 and 40 meters respectively, providing an average error of the wind direction around of 17.97 degrees.

The prediction of the missing values is useful to understand the wind behavior in the Isthmus of Tehuantepec, which allows reducing the dangerous events in several sections of a wind turbine. The wind database of a wind farm is used to statistically determine the behavior of wind for several years providing information related to the mean wind speed and dominant wind direction. This way, it is possible to know months even days, where the wind is more dangerous and to reduce damages of the overall wind turbines. The future work is related with the application of Genetic Programming with the intention of reducing the average errors of the predicted values of the wind speed and direction of wind farms located into the Isthmus of Tehuantepec, Oaxaca, Mexico.

ACKNOWLEDGMENT

The authors want to thank the Instituto Nacional de Electricidad y Energías Limpias (INEEL) from Cuernavaca, Morelos for providing the data used in this research and the anonymous reviewers for their valuable comments that helped improving the quality of this manuscript. The first author also thanks "Red temática en Inteligencia Computacional Aplicada" (Red-ICA) of CONACyT for her traveling funding.

REFERENCES

- [1] Ethem Alpaydin. *Introduction to machine learning*. MIT press, 2009.
- [2] Aida Alvera-Azcárate, Alexander Barth, J-M Beckers, and Robert H Weisberg. Multivariate reconstruction of missing data in sea surface temperature, chlorophyll, and wind satellite fields. *Journal of Geophysical Research: Oceans*, 112(C3), 2007.
- [3] Alfredo Arcos Jiménez, Carlos Quiterio Gómez Muñoz, and Fausto Pedro García Márquez. Machine learning for wind turbine blades maintenance management. *Energies*, 11(1):13, 2017.
- [4] Roslan Armina, Azlan Mohd Zain, Nor Azizah Ali, and Roselina Sallehuddin. A review on missing value estimation using imputation algorithm. In *Journal of Physics: Conference Series*, volume 892, page 012004. IOP Publishing, 2017.
- [5] C.M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.
- [6] Chandan Gautam and Vadlamani Ravi. Data imputation via evolutionary computation, clustering and a neural network. *Neurocomputing*, 156:134–142, 2015.
- [7] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.
- [8] Eric Jones, Travis Oliphant, and Pearu Peterson. {SciPy}: open source scientific tools for {Python}. 2014.
- [9] Tianhong Liu, Haikun Wei, and Kanjian Zhang. Wind power prediction with missing data using gaussian process regression and multiple imputation. *Applied Soft Computing*, 2018.
- [10] Xiao Liu, Xu Lai, and Jin Zou. A new mcp method of wind speed temporal interpolation and extrapolation considering wind speed mixed uncertainty. *Energies*, 10(8):1231, 2017.
- [11] Antonino Marvuglia and Antonio Messineo. Monitoring of wind farms power curves using machine learning techniques. *Applied Energy*, 98:574–583, 2012.
- [12] Mohamed Noor Norazian, Yahaya Ahmad Shukri, Ramli Nor Azam, and Abdullah Mohd Mustafa Al Bakri. Estimation of missing values in air pollution data using single imputation techniques. *ScienceAsia*, 34(3):341–345, 2008.
- [13] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.
- [14] Alex J Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222, 2004.
- [15] Zhiling Yang, Yongqian Liu, and Chengrong Li. Interpolation of missing wind data based on anfis. *Renewable Energy*, 36(3):993–998, 2011.