KYBERNETIKA — VOLUME *59* (2023), NUMBER 6, PAGES 861–879

SEASONAL TIME-SERIES IMPUTATION OF GAP MISSING ALGORITHM (STIGMA)

Eduardo Rangel-Heras, Pavel Zuniga, Alma Y. Alanis, Esteban A. Hernadez-Vargas and Oscar D. Sanchez

This work presents a new approach for the imputation of missing data in weather time-series from a seasonal pattern; the seasonal time-series imputation of gap missing algorithm (STIGMA). The algorithm takes advantage from a seasonal pattern for the imputation of unknown data by averaging available data. We test the algorithm using data measured every 10 minutes over a period of 365 days during the year 2010; the variables include global irradiance, diffuse irradiance, ultraviolet irradiance, and temperature, arranged in a matrix of dimensions $52,560$ rows for data points over time and 4 columns for weather variables. The particularity of this work is that the algorithm is well-suited for the imputation of values when the missing data are presented continuously and in seasonal patterns. The algorithm employs a date-time index to collect available data for the imputation of missing data, repeating the process until all missing values are calculated. The tests are performed by removing 5%, 10%, 15%, 20%, 25%, and 30% of the available data, and the results are compared to autoregressive models. The proposed algorithm has been successfully tested with a maximum of $2,736$ contiguous missing values that account for 19 consecutive days of a single month; this dataset is a portion of all the missing values when the time-series lacks 30% of all data. The metrics to measure the performance of the algorithms are root-mean-square error (RMSE) and the coefficient of determination (R^2). The results indicate that the proposed algorithm outperforms autoregressive models while preserving the seasonal behavior of the time-series. The STIGMA is also tested with non-weather time-series of beer sales and number of air passengers per month, which also have a cyclical pattern, and the results show the precise imputation of data.

Keywords: contiguous missing values, seasonal patterns, time-series

Classification: 62-04, 68Pxx

1. INTRODUCTION

Data play an essential role in different applications such as chemometrics [10], genomics [5], network inference [11], meteorology [26], engineering [24], informatics [17], chemical [6], biochemical [13], pharmaceutical, and industry [12]. From these, meteorological time-series are measured by weather stations to be processed, analyzed, and implemented in the sizing of wind-turbines and photovoltaic systems; or these can also

DOI: 10.14736/kyb-2023-6-0861

be employed to forecast, for example, solar irradiance or wind speed to calculate electric power generated by PV systems or wind-turbines.

This work aims at imputing data in time-series, and to present the literature review we divide the imputation techniques in three groups: a) classic methods like the previous neighbor, linear interpolation, cubic spline, or piecewise cubic Hermite interpolation polynomial; b) time-series methods such as the autoregressive (AR) or autoregressive integrated moving average; and c) nonconventional methods like those based on principal component analysis or other techniques.

Among classic methods, Dan et al. [8] compared the previous neighbor, linear interpolation, cubic spline, piecewise cubic Hermite interpolation polynomial, Akima, and Makima methods for the imputation of pupil diameter missing data. They concluded that the Akima and Makima interpolation methods yield the lowest deviation with a smooth curve fitting that makes them well-suited for slowly varying data. Noor et al. [22] compared the linear interpolation and mean method for missing data in air pollution time-series and showed that linear interpolation results in the smallest RMSE and R^2. Classic methods work well when the missing data is randomly distributed, as in a Missing at Random or Completely Missing at Random form [15][20]; however, when the missing values are contiguous, these techniques tend to fail.

Time-series techniques are also employed to fill gaps, for example, Murad et al. [20] proposed using time-dependent covariates in a Cox model with Multiple Imputations by Chained Equations to complete missing data related to diabetes and some types of cancer. They showed that these methods are only feasible for small groups of contiguous missing data. Regressive, autoregressive, vector autoregressive (VAR), and autoregressive moving average (ARMA) models have also been implemented for the imputation of missing data. Bashir and Wei [3] developed a VAR-based algorithm for handling missing data in multivariate time-series. They tested it on electrocardiogram signals with 10% and 20% Missing Completely at Random data sets, with the disadvantage that time-series must be stationary. Zhang [27] implemented a regressive model for the imputation of clinical data in blood pressure and lactate. The results showed that the relation between missing values and variables is preserved. Liu & Molenaar [16] recovered missing data from electrodermal activity with a VAR model. They first fitted the model to the complete data and the Direct Transfer Functions to examine the directional influence between the child and therapist. Dunsmuir & Robinson [9] developed a method to estimate stationary time-series data in the presence of missing values based on ARMA models; they only tested the models with pollution levels containing little missing data. Anava et al. [2] predicted time-series online with missing data based on AR models. The work focused on studying the time-series prediction problem using AR models in the presence of missing data. Pedreschi et al. [23] evaluated methods to treat missing values in gel-based proteomics data. The methods dealt with missing values during the multivariate analysis with the Nonlinear estimation by Iterative Partial Least Squares algorithm, k-nearest neighbor, and Bayesian Principal Component Analysis (BPCA) before carrying out the multivariate analysis. The authors concluded that there is no absolute truth in terms of which is the most appropriate method to treat missing data, however, from the ones studied, the BPCA showed the best result when applying cross-validation to test the model's performance. Choong et al. [7] proposed a new algorithm

called the Autoregressive Model-based Missing Value Estimation Method, and employed it for the imputation of data of a matrix of gene array values of DNA. The authors used a data set consisting of five different AR processes of order four and tested the method on several microarray time-series. According to the authors, their proposal is well-suited for time-series with many missing values, but it is unclear if they can be contiguous.

In summary, multivariable models (e. g., VARs models) are well-suited when independent variables do not have missing data in the same position as the dependent variable. This is a disadvantage in many cases, such as the one shown in this work, where the dataset shows portions of missing values in all variables at the same time. On the other hand, AR models do not have this problem because they take past values for the imputation of missing data, but their main drawback, as in classic models, comes when the window of missing data is excessively long.

For non-conventional methods, Folch et al. [12] developed a Missing Data Imputation Toolbox for Matlab© based on Principal Component Analysis that exploits the interdependence between variables and works on data matrices with missing values randomly presented in rows and columns. The disadvantage is that it needs additional variables that cannot have missing values in the same periods. Junger & Ponce [15] employed the Expectation-Maximization algorithm for the imputation of multivariate time-series under the assumption of normal distribution. They performed tests with 5%, 10%, 20%, 30%, and 40% Missing at Random data, and the study focused on environmental time-series to develop epidemiological studies of the effect of air pollutants on health. Batista & Monard [4] implemented the k-nearest Neighbor algorithm as an imputation method to treat missing data and compared it to two algorithms based on Decision Trees and the Mean Imputation Method. The models were tested with breast data and resulted in a final model of 10-NN (Nearest Neighbors); the analysis indicates that the k-Nearest Neighbor algorithm outperforms the other techniques. Hui et al. [14] employed Multiple Imputations to fill gaps in the missing data for annual estimations of net ecosystems' carbon exchange, latent heat flux, and sensible heat flux. The algorithm is a Monte Carlo technique in which several simulated values replace the missing ones, but the authors do not indicate the maximum contiguous missing data allowed. These references do not specify the behavior of the models when the window of missing data is very long with contiguous missing values, which is the case presented here.

Researchers have recently reported algorithms for the imputation of data that employ machine and deep learning techniques. For example, Sun et al. [25] present a review of statistical, machine learning, and deep learning approaches, and discuss the advantages and disadvantages of these methods for the imputation of missing data. The authors test the methods with different amounts of missing data using three kinds of missing mechanisms named the Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing not at Random (MNAR). The deep learning approaches implemented are the Generative Adversarial Imputation Networks (GAIN) and Variational Auto-Encoder (VAE); the authors also test conventional methods like the Multiple Imputation by Chained Equations (MICE) and Miss Forest. The results show that the conventional methods outperform the deep learning methods. Another work by Neves et al. [21] present three novel generative imputation methods based on Generative Adversarial Networks (GAN), the Slim GAIN (SGAIN), the Wasserstein Slim GAIN with

Clipping Penalty (WSGAIN-CP), and Wasserstein Slim GAIN with Gradient Penalty (WSGAIN-GP); from the tests the authors report that the SGAIN outperforms the other techniques. Additionally, Ahn et al. [1] implement six imputation methods named the mean substitution, Last Observation Carried Forward (LOCF), Next Observation Carried Backward (NOCB), Expectation-Maximization (EM), k-Nearest Neighbors (k-NN), and Multiple Imputation by Chained Equation (MICE), where, according to the authors, the best model is the k-NN. It is important to mention that all these methods are not tested in the presence of large sets of continuous missing values, and do not compare directly to the method proposed in this work, since it is focused on time-series with great amounts of continuous missing data.

The literature review shows that many authors in different fields deal with missing data using classic, time-series, or non-conventional methods. Still, they mainly focus on randomly located discontinuous data gaps with algorithms that perform well when the window of contiguous missing points is small. To the authors' best knowledge, there is no specific technique to treat cyclic and seasonal time-series with many contiguous missing values. Therefore, this work aims at solving the problem of treating vast quantities of contiguous missing data in weather time-series under the assumption that sensors can stop working for long periods, for example, when systems are under maintenance or simply fail.

The data used here comes from a weather station that has been recording information from the year 2010 up to the present day in Temixco, Morelos, México. We use the time-series of the year 2010 because it has no missing values, as opposed to other years; however, there are data windows of atypical values, as will be shown latter. The data of the year 2010 results in a matrix of $52,560 \times 4$ (four vectors, each one with a length of $52,560$), with data-points measured every 10 minutes (144 values per day); the tests consider a window with a minimum of 144 contiguous missing values (a full day). The variables include global irradiance in $\mathrm{W/m^2}$, diffuse irradiance in $\mathrm{W/m^2}$, ultraviolet irradiance in $\mathrm{W/m^2}$, and temperature in C. We found that some years, other than 2010, have full days of missing data, representing large windows of contiguous missing values. Moreover, some days with missing values are holidays, from which we speculate that these may be maintenance days.

The proposal is based on techniques such as persistence, moving averages, and the random behavior inherent to the weather time-series with cyclic and seasonal patterns. First, the algorithm takes advantage of a persistence model in which data from variables such as global irradiance or temperature have similar values in a particular time frame, in this case, a month. The algorithm then finds all non-missing values of the month and at the time of each missing point. When the total number of data used for the imputation of the missing values is greater than ten, the algorithm randomly takes a group of values to represent the random behavior inherent to the time-series and computes its average (moving average technique). Finally, the average is used for the imputation of a missing value, and the procedure is repeated until all missing values are computed.

The work is presented in three sections. Section 1 gives an introduction and related work; Section 2 explains the development of the proposed algorithm; and Section 3 exposes the tests to evaluate its performance.

2. DEVELOPMENT OF THE STIGMA

The proposed algorithm operates over a matrix that contains data of weather variables and their corresponding dates. All missing values must be identified with a date index that include the month, day, and time of each missing value, tied to their row position within the data matrix. The date index helps to locate all non-missing values over a certain period to be used to calculate an unknown point. A set of these values is randomly selected to calculate its average and fill a clear space. The steps of the algorithm are given by:

1. Build a matrix of weather data and identify the date and time of its entries;
2. Identify all missing values by their date and time;
3. Relying on a persistence model, find all the non-missing data that match the time of the missing values for each month, and randomly select n points to compute its average value;
4. Use the average value to fill the missing data according to their time of occurrence;
5. Repeat Steps 2 to 4 for every point of every variable until all missing data are computed.

As an example, Table 1 shows a window of five contiguous missing values corresponding to January first from 09:50 to 10:30 hours. The algorithm identifies the row with the missing values in the matrix and finds available data at the same time in every day of the month. Then it randomly selects data points and calculates the average value, as shown in Table 2, to finally use it to fill the missing data in the rows of Table 1. The process is repeated until all missing data are calculated.

	Unknown Dates	Global irradiance (W/m^2)	Diffuse irradiance (W/m^2)	UV irradiance (W/m^2)	Temperature (°C)
Window of missing data	01/01/2010 09:50:00	NaN→305.02	NaN→114.44	NaN→14.42	NaN→17.96
	01/01/2010 10:00:00	NaN	NaN	NaN	NaN
	01/01/2010 10:10:00	NaN	NaN	NaN	NaN
	01/01/2010 10:20:00	NaN	NaN	NaN	NaN
	01/01/2010 10:30:00	NaN	NaN	NaN	NaN

Tab. 1. Unknown data.

	Known Dates	Global irradiance (W/m^2)	Diffuse irradiance (W/m^2)	UV irradiance (W/m^2)	Temperature (°C)
Random dates	24/01/2010 09:50:00	532.2	64.8	21.9	20.2
	15/01/2010 09:50:00	189.3	168.1	10.8	18.5
	08/01/2010 09:50:00	110.8	102.9	7.5	15.2
	25/01/2010 09:50:00	511.0	76.6	20.9	20.9
	31/01/2010 09:50:00	181.8	159.8	11.0	15.0
	Average	305.02	114.44	14.42	17.96

Tab. 2. Known data

3. TESTS AND RESULTS

3.1. Performance measurements

To measure the performance of the STIGMA we employ the root-mean-square error (RMSE) and the coefficient of determination (R^2), respectively given by

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} \tag{1}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} \tag{2}$$

where Y_i is the actual data, $\hat{Y}_i$ is the calculated data, $\overline{Y}$ is the average of the actual data, and n is the number of samples.

3.2. Testing the algorithm with weather time-series

We propose six tests to validate the performance of the algorithm by removing 5%, 10%, 15%, 20%, 25%, and 30% of the data and comparing it with an autoregressive model. For this time-series with contiguous missing values we choose an autoregressive model since it takes past values to compute new ones that can also be used to calculate a new set of missing values. However, this kind of model fails when the window of contiguous missing data is large, as will be shown by the results. The autoregressive model of order p used in this work is defined as

$$X_t = \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t \tag{3}$$

where φ_i are the parameters of the model, ε_t is white noise, and X_t is the time series [18].

Table 3 shows the dates and number of missing points in the headers "Days" and "Missing Data", respectively. A close examination of Table 3 shows that the biggest window of consecutive missing values is 2,736 (February 10-28) and takes place when 30% of the data is missing; on the other hand, the smallest window is of 144 values, which corresponds to an entire day. This large set of 2,736 contiguous missing values is selected in this work because it presents a challenge to classical and autoregresive algorithms. Under the assumption that at least some of the missing data takes place on holidays, we remove very little or no data in June and July because we managed to accommodate all missing data in months were there are in fact holidays. As mentioned before, we speculate that maintenance days may be placed on holidays and that this can result in missing values. However, from the point of view of the proposed algorithm, there is no distinction among missing values that result from maintenance days or any other reason, therefore, all of them are treated in the same way.

We compared some of the statistical data distribution parameters like mean, standard deviation, asymmetric coefficient (skewness), and kurtosis to determine how close the calculated data is to the actual data in terms of statistical distribution. First we present

5% of missing data			**10% of missing data**		**15% of missing data**	
Month	Days	Missing Data	Days	Missing Data	Days	Missing Data
Jan	1,6	288	1-6	864	1-14	2016
Feb	24	144	24-28	720	24-28	720
Mar	21	144	21	144	15-25	1584
Apr	30	144	27-30	576	27-30	576
May	1,5,15	342	1, 5-7	720	1,5-8	720
Jun	--	--	--	--	--	--
Jul	--	--	--	--	--	--
Aug	15	144	5,15	288	15	144
Sep	1,16	288	1,16-18	576	1,16-18	576
Oct	12	144	8-9,12	432	8-9,12	432
Nov	1,2,20	342	1-2,20	432	1-2,20	432
Dec	12,24-26	576	12,24-26,31	720	12,24-26,31	720
20% of missing data			**25% of missing data**		**30% of missing data**	
Month	Days	Missing Data	Days	Missing Data	Days	Missing Data
Jan	1-14	2016	1-14	2016	1-14	2016
Feb	16-28	1872	16-28	1872	10-28	2736
Mar	15-25	1584	15-25	1584	15-25	1584
Apr	27-30	576	14-30	2448	14-30	2448
May	1,5-8	720	1,5-8	720	1,5-8	720
Jun	--	--	--	--	4-16	1872
Jul	--	--	--	--	--	--
Aug	15-21	1008	15-21	1008	15-21	1008
Sep	1,16-18	576	1,16-18	576	1,16-18	576
Oct	8,9,12	432	8,9,12	432	8,9,12	432
Nov	1,2,20	432	1,14-20	1152	1,14-20	1152
Dec	8-12,24-26,31	1296	8-12,24-26,31	1296	8-12,24-26,31	1296

Tab. 3. Dates and quantity of missing data.

the descriptive statistics in Table 4 to Table 9 to compare the values of the calculated and actual data for missing value sets of 5%, 10%, 15%, 20%, 25%, and 30%, respectively. The results for the test with 10% of missing data are slightly different from the rest, as will be explained. In Table 4 to Table 9, GI stands for global irradiance, DIF for diffuse irradiance, UV for ultraviolet irradiance, and T for temperature.

Variable	RMSE	Average		Standard deviation		Skewness		Kurtosis	
		Imputed	Actual	Imputed	Actual	Imputed	Actual	Imputed	Actual
GI	79.42	222.75	231.59	295.32	316.77	0.92	1.04	2.29	2.62
DIF	35.58	54.27	58.13	64.63	74.67	0.63	1.05	1.80	3.04
UV	3.68	10.86	11.19	14.93	15.89	1.05	1.20	2.60	3.08
T	1.83	21.51	22.06	4.64	4.69	0.34	0.31	2.67	2.58

Tab. 4. Removing 5% of the data. Descriptive statistics of the calculated and actual data.

Variable	RMSE	Average		Standard deviation		Skewness		Kurtosis	
		Imputed	Actual	Imputed	Actual	Imputed	Actual	Imputed	Actual
GI	86.36	224.54	233.80	296.50	317.56	0.93	1.03	2.57	2.32
DIF	37.60	54.83	58.60	65.09	75.42	0.63	1.09	3.14	1.83
UV	3.81	11.08	11.37	15.13	15.92	1.03	1.15	2.90	2.57
T	1.67	21.90	22.32	4.72	4.88	0.34	0.42	2.64	2.57

Tab. 5. Removing 10% of the data. Descriptive statistics of the calculated and actual data.

Variable	RMSE	Average		Standard deviation		Skewness		Kurtosis	
		Imputed	Actual	Imputed	Actual	Imputed	Actual	Imputed	Actual
GI	101.96	231.46	225.72	304.45	313.70	0.90	1.10	2.26	2.74
DIF	38.52	51.31	55.96	61.06	73.55	0.66	1.18	1.92	3.42
UV	4.45	11.46	11.17	15.63	15.91	1.02	1.20	2.54	3.06
T	2.10	22.05	21.85	4.62	4.98	0.28	0.36	2.36	2.50

Tab. 6. Removing 15% of the data. Descriptive statistics of the calculated and actual data.

Variable	RMSE	Average		Standard deviation		Skewness		Kurtosis	
		Imputed	Actual	Imputed	Actual	Imputed	Actual	Imputed	Actual
GI	108.20	227.44	222.29	298.57	311.80	0.90	1.12	2.25	2.81
DIF	38.59	51.18	55.41	61.07	73.68	0.67	1.23	1.94	3.57
UV	4.80	11.31	11.08	15.39	15.87	1.01	1.22	2.52	3.11
T	2.18	21.72	21.33	4.42	4.90	0.31	0.33	2.47	2.63

Tab. 7. Removing 20% of the data. Descriptive statistics of the calculated and actual data.

Variable	RMSE	Average		Standard deviation		Skewness		Kurtosis	
		Imputed	Actual	Imputed	Actual	Imputed	Actual	Imputed	Actual
GI	102.93	233.92	231.33	306.06	318.97	0.90	1.07	2.27	2.66
DIF	38.94	54.03	54.96	64.25	72.13	0.67	1.22	1.95	3.63
UV	4.59	11.62	11.53	15.76	16.28	1.02	1.17	2.54	2.97
T	2.08	22.23	21.89	4.60	5.00	0.24	0.26	2.38	2.50

Tab. 8. Removing 25% of the data. Descriptive statistics of the calculated and actual data.

Variable	RMSE	Average		Standard deviation		Skewness		Kurtosis	
		Imputed	Actual	Imputed	Actual	Imputed	Actual	Imputed	Actual
GI	105.83	230.94	236.97	301.70	323.65	0.91	1.03	2.31	2.54
DIF	40.63	55.64	56.81	67.01	74.03	0.76	1.19	2.16	3.49
UV	4.81	11.49	11.85	15.51	16.57	1.02	1.13	2.57	2.83
T	2.21	22.34	22.27	4.64	5.04	0.21	0.24	2.32	2.46

Tab. 9. Removing 30% of the data. Descriptive statistics of the calculated and actual data.

As can be seen, the calculated and actual data present similar average values as well as standard deviation, which confirms the accuracy of the algorithm; this means that the calculated and actual data share a very similar central value and that the dispersion of the data is also similar. Additionally, the proposed algorithm results in a lower standard deviation with respect to the actual data, indicating less variation due to how the information is processed. Other important metrics are skewness and kurtosis, where the calculated and actual data also have similar values; this means that the calculated data preserves the distribution of the original time series. In the case of skewness, the values are always positive, which denotes that the data is concentrated to the left of

the distribution with the tail on the right. For the case of kurtosis, the values are, in general, close to three; a value of three indicates data concentrated in the center of the distribution (normal distribution); a value lower than three means that the distribution of the data is wide; and a value higher than three results in the opposite, i. e., a very narrow data distribution [19].

For the test cases in Table 4, and Table 6 to Table 9, we observe the same trend in the variation of skewness and kurtosis for the diffuse irradiance. This shows that the proposed algorithm tends to reduce the value of these metrics, which means a more symmetric distribution with fewer data in its center; kurtosis indicates the quantity of data in the Gaussian bell whereas skewness accounts for its symmetry. However, in Table 5, we observe the opposite for kurtosis, i. e., the proposed algorithm concentrates the data distribution. In general, these results indicate that the statistical properties of the calculated data are maintained, which is a desirable feature when performing studies with the resulting time series, for example, forecasting.

We now compare the proposed algorithm with the autoregressive model in equation 3 using the RMSE in equation 1 for each variable, as shown in Figure 1. In general, the RMSE for the proposed algorithm increases with the number of missing data, except for the global irradiance and temperature, where it decreases when going from 20% to 25% of missing data (Figure 1a) and d)); this means that the calculated and actual data are very similar, which is desirable. According to the RMSE, the proposed algorithm outperforms the autoregressive model by resulting in a lower value in every test, which means that calculated and actual data have very similar values, also a desirable result.

The R^2 performance index in equation 2 is shown in Figure 2 for the case with 30% of missing data since it has the highest RMSE according to Figure 1 (worst case); an R^2 value of 100% denotes a perfect fit between the calculated and original data. The highest values of R^2 correspond to the ultraviolet irradiance with 91.6% and 83.0%; followed by the global irradiance with 89.3% and 78.0%; next is the temperature with 80.9% and 75.0%; and last the diffuse irradiance with 69.9% and 51.7%. For all variables, the first result for R^2 corresponds to the proposed algorithm and the second to the autoregressive model. As with the descriptive statistics in Table 4 to Table 9, the diffuse irradiance shows a decreased performance when compared to the other variables, and this will be studied later in more detail. Still, the R^2 value is higher for all cases where the STIGMA is employed for the imputation of the missing data.

In Figure 2 we also show a linear regression of the data, which gives information on how the calculated data fits the actual data. The continuous line represents the linear regression model ($\hat{\mathrm{Y}}$) of the calculated data versus the actual data, and the discontinuous line (Y) stands for the actual data versus the actual data (hypothetical case of a perfect fit). The plots in Figure 2a), c), e), and g) show the results of the proposed algorithm and the plots in Figure 2b), d), f), and h) show the ones of the autoregressive model. The linear regression of the calculated data ($\hat{\mathrm{Y}}$) is obtained by plotting the calculated and actual data on the y and x axes, respectively, whereas the perfect fit (Y) is the real data plotted on the y and x axes; the closer these two lines are, the better the calculated data fits the actual data. It is evident from all plots that the proposed algorithm performs better than the autoregressive model, as these two lines are closer.

The plots of the calculated and actual data for all variables in the case of 30%

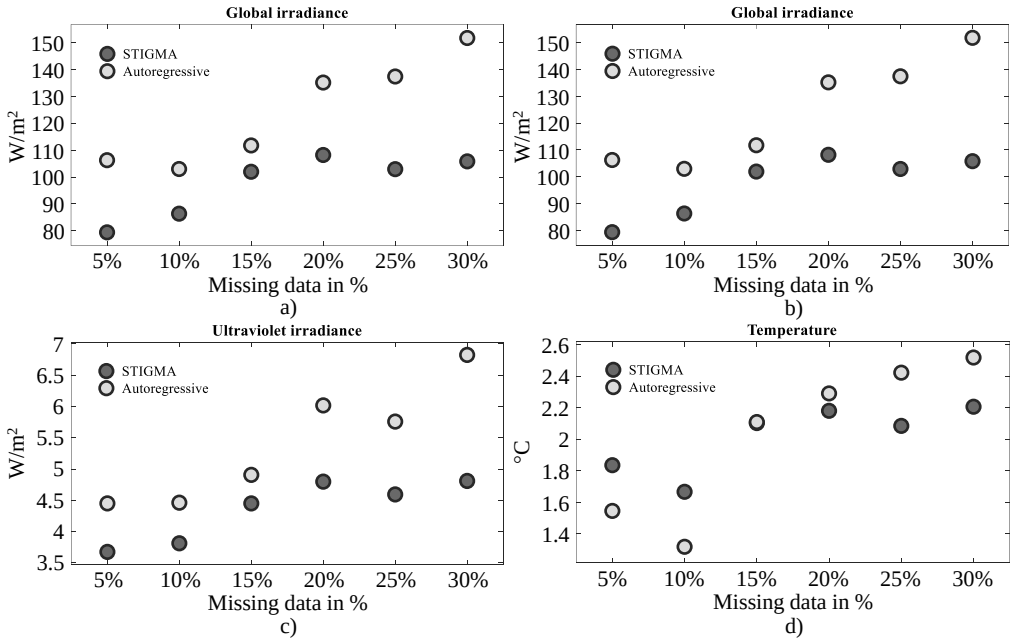

Fig. 1. RMSE for the calculated data; using the STIGMA algorithm and the Autoregressive model, a) global irradiance, b) diffuse irradiance, c) ultraviolet irradiance, d) temperature.

of missing data are depicted in Figure 3, where the calculated data results from the proposed algorithm. As shown in the plots, the STIGMA maintains the trend of the corresponding time-series even for large gaps of missing data for the global irradiance, ultraviolet irradiance, and temperature. However, for the case of the diffuse irradiance the results do not perform as expected and will be analyzed in detail. For comparison, Figure 4 depicts the actual and calculated data using the autoregressive model for all variables, where a visual inspection shows that the calculated data fails to maintain the trend of the known data; we see that the shape of the calculated and known data are not similar. The results for the diffuse irradiance data calculated using the proposed algorithm are now given using Figure 5, where the "best" and "worst" three days are selected based on their individual RMSE. The "best" three days, ordered from lowest to highest by their individual RMSE, are December 31, December 08, and October 08 (Figure 5 a)), and have a combined RMSE of 7 W/m^2 and an R^2 of 98.6% for the complete set of 432 data points (144 for each day).

On the other hand, Figure 5 b) shows the "worst" three days, again ordered from lowest to highest by their individual RMSE, that are June 16, January 04, and February 24, and have a combined RMSE of 72.5W/m^2 and an R^2 of -9.9%; a negative value of R^2 means that the error between the calculated and actual data is much larger than the error between the actual data and its average, which is not desirable. As mentioned before, we notice that the "worst" three days show atypical data points that do not

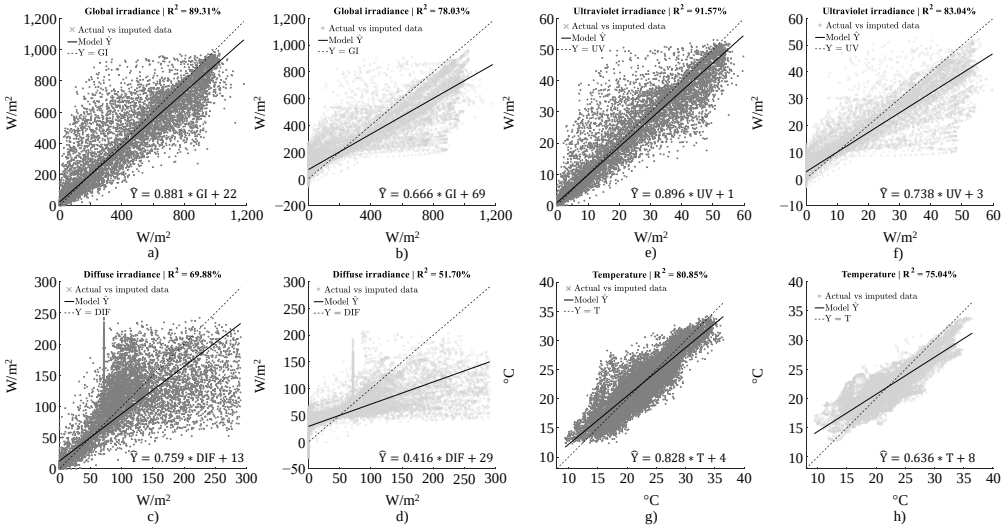

Fig. 2. Linear regression to measure the data fitness for a 30% of the missing data using the STIGMA algorithm for the a) global irradiance, c) diffuse irradiance, e) ultraviolet irradiance, and g) temperature; and using the autoregressive technique for the b) global irradiance, d) diffuse irradiance, f) ultraviolet irradiance, and h) temperature.

follow the trend of the signal. These atypical points are clearer on June 16 were we see a data window of constant values (see the data marked in blue in Figure 5 b)); the authors do not know why the data behaves in this manner.

As an example, let us consider the data from February, where only 9 days are known and the rest have to be calculated (19 days). Figure 5 c) shows the available data used for the imputation of unknown values, from February 01 to February 09, where we see that data is not uniform among days, and some of them present very high or low values that fall outside what would be considered normal, i. e. a smooth sinusoidal like behaviour as in February 06 and February 07; a good practice could be to delete the atypical values (February 01 to February 05 and February 08 to February 09) and calculate new data using days with typical behaviour. Also, Figure 5 d) depicts the calculated data, from February 10 to February 28, where we can see the same kind of atypical high and low values compared to the other time-series (global irradiance, ultraviolet irradiance, and temperature). Since the proposed algorithm randomly selects data points for the imputation of missing values, if the available data is not accurate it will result in inaccurate calculated values.

Now, as with the results obtained using the proposed algorithm, Figure 6 shows the "best" and "worst" three days from Figure 4 for the diffuse irradiance data calculated with the autoregressive model; these days are selected based on their individual RMSE and ordered from lowest to highest. The "best" three days have a combined RMSE of

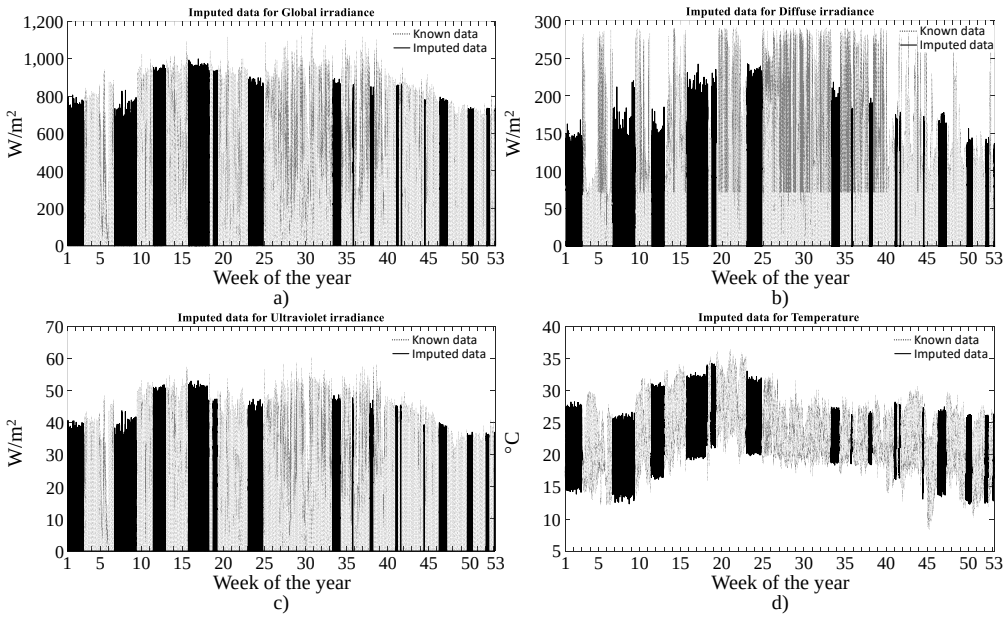

Fig. 3. Time-series with 70% of known data and 30% of the values calculated using the STIGMA for the a) global irradiance, b) diffuse irradiance, c) ultraviolet irradiance, and d) temperature.

9.6 W/m^2 and an R^2 of 94.8% for the complete set of 432 data points (144 for each day), and correspond to November 20, May 01, and March 25 (Figure 6 a)). On the other hand, the "worst" three days are March 16, June 05, and June 11 with a combined RMSE of 91.4 W/m^2 and an R^2 of -447.4% (Figure 6 b)). Comparing these results with the ones calculated with the proposed algorithm (Figure 5), we can observe that the proposal performs better, showing lower RMSE and less negative R^2 values, which, as mentioned before, means that the error between the calculated and actual data is very large.

It is important to note that the autoregressive model takes past values for the imputation of new ones (Equation (3)), but when the window of missing data is large, it is clear that it will eventually use calculated values as data for new ones, which diminishes accuracy. On the other hand, the proposed algorithm finds all similar known data corresponding to a certain period and takes n values to calculate the average employed for the imputation of the missing data; the process is repeated until all missing values are completed.

3.3. Other applications of the STIGMA

To further confirm the results, the STIGMA is employed for the imputation of the beer sales and air-passenger time-series shown in Figure 7 and Figure 8, respectively; these time series also present cyclic patterns. There are no missing values in any of the two time-series. However, we randomly remove 30% of the data points in both cases to test

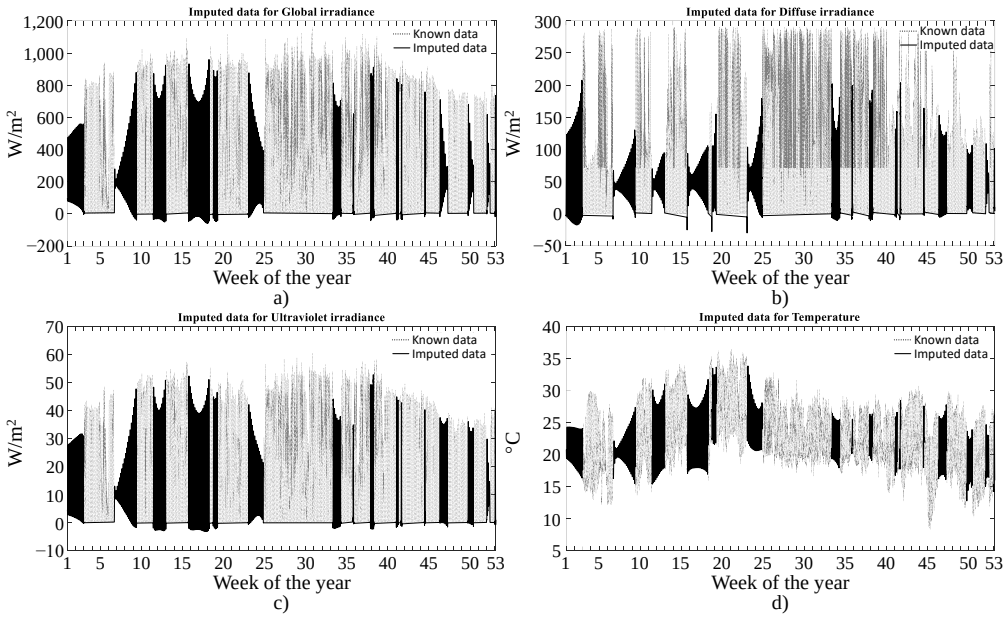

Fig. 4. Time-series with 70% of known data and 30% of the values calculated using the autoregressive model for the a) global irradiance, b) diffuse irradiance, c) ultraviolet irradiance, and d) temperature.

the proposed algorithm.

The beer sales time-series has 154 measurements every three months from January 01, 1956, to April 01, 1994. In Figure 7 a) we show the complete time-series along with the 30% randomly removed data points that are going to be calculated.

Since the proposed algorithm works from a seasonal pattern, we divide the complete time-series into a set of sections marked by discontinuous lines, as shown in Figure 7 a). The criterion used was to group data points that present a somewhat periodic behavior and run the proposed algorithm on each one of them. First, we identify the window that presents a positive trend corresponding to January 01, 1956, to January 01, 1973 (first 72 data points) to determine the number of sections. To reduce the effect of this trend, we divide the window in four sections with 18 points in each one. The rest of the time-series (last 82 points) is fairly stationary and is treated as one section.

Figure 7 b) shows the actual versus the calculated data using the proposed algorithm, where a visual inspection indicates good agreement between them. This is confirmed by an RMSE of 13.2×10^6 liters and an R^2 of 98.2%, which indicates that the error between the calculated and actual data is small.

On the other hand, the time-series for air passengers has 144 data points, measured every month from January 01, 1949, to December 01, 1960. Similar to the time-series of beer sales, Figure 8 a) shows the complete time-series along with the 30% randomly removed data points that are to be calculated; the proposed algorithm also runs over sections, in this case, three.

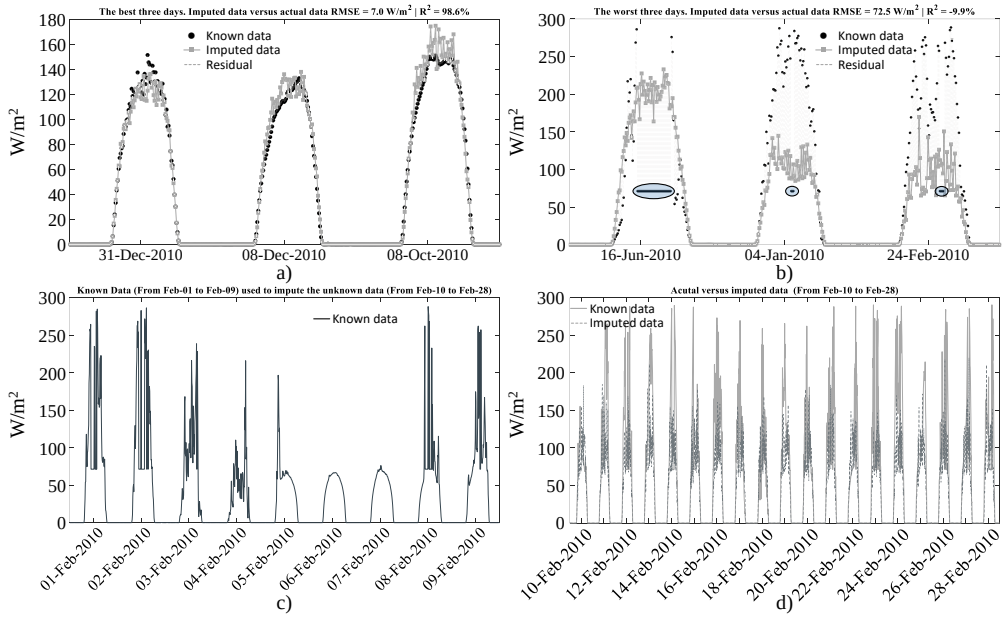

Fig. 5. Calculated data with STIGMA for Diffuse irradiance: a) The best three days according to the performance tests, b) The worst three days according to the performance tests, c) Known data, d) Unknown data.

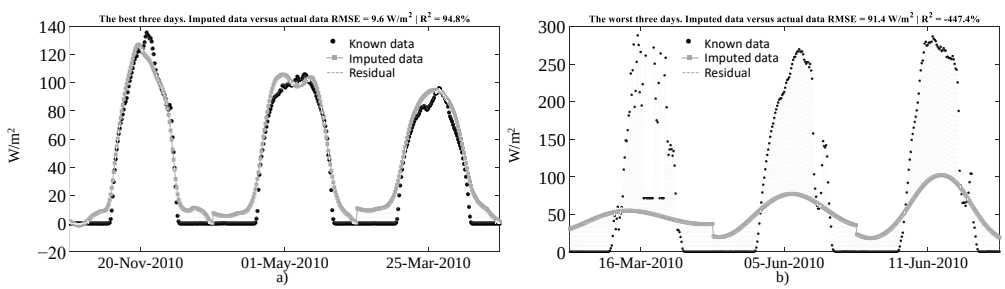

Fig. 6. Calculated data with the autoregressive model for Diffuse irradiance: a) The best three days according to the performance test, b) The worst three days according to the performance test.

Figure 8 b) shows the actual versus the calculated data using the proposed algorithm, where it is clear to see the good agreement between them; again, this is confirmed by an RMSE of 283.7 passengers per month and an R^2 of 96%, which also indicates that the error between the calculated and actual data is small.

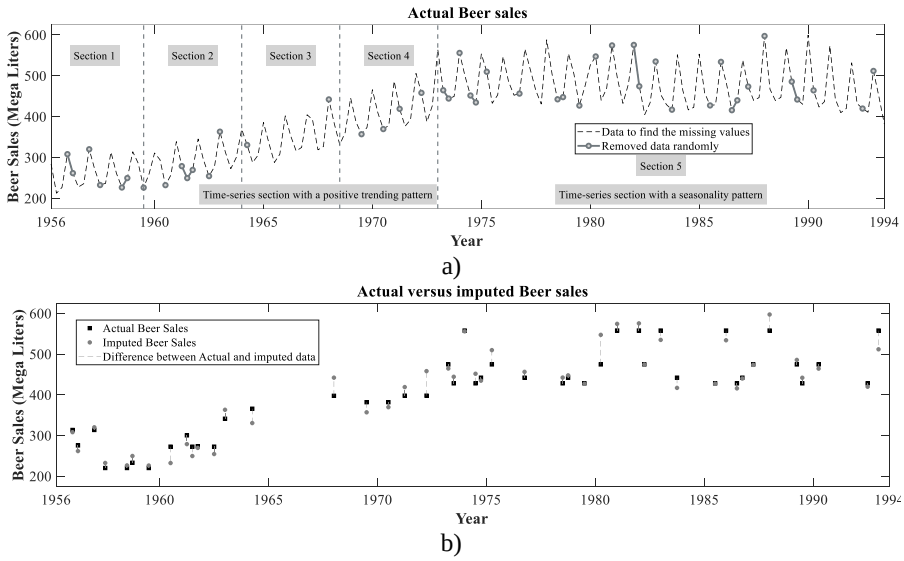

Fig. 7. a) Known data along with values calculated using the STIGMA, and b) Known vs. calculated data error for beer sales.

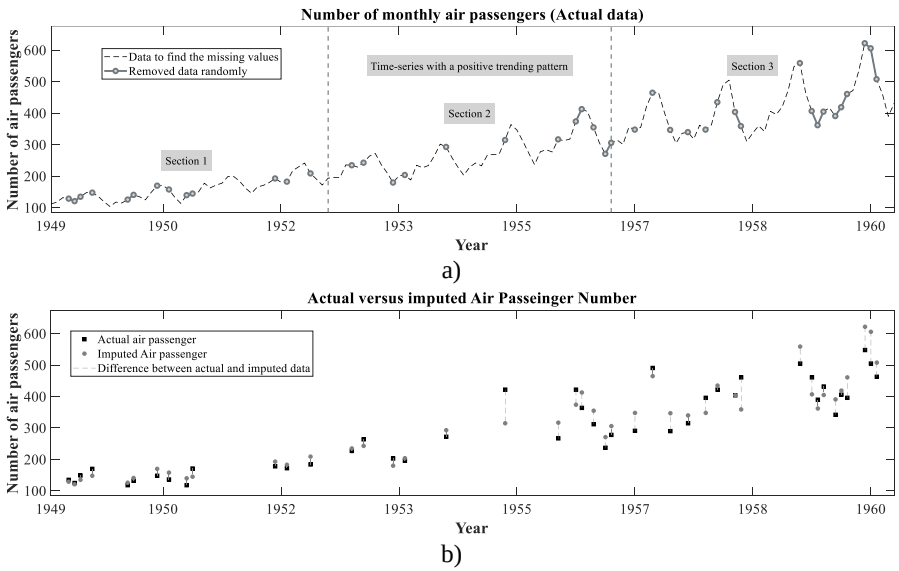

Fig. 8. a) Known data along with values calculated using the STIGMA, and b) Known vs. calculated data error for the monthly number of air passengers.

4. CONCLUSIONS

The main problem researchers and engineers face using time-series from weather stations is filling gaps when data is missing. When these gaps are small, several classic methods can be employed; however, they fail when the window length of missing data is large because they need available data close to the missing values. On the other hand, autoregressive models are well-suited in these cases because they take past data to determine the missing values and use these to compute new ones. However, these methods also fail when the window of missing data is very large and the points are consecutive.

We developed a new approach for the imputation of missing data from seasonal and cyclic time-series that, according to the results, is straightforward to implement and outperforms an autoregressive model employed for testing. The proposed model is based on the persistence and moving average models, and it randomly finds available data using the date and time of the missing values, thus better representing the behavior of the time-series. The algorithm was developed for the imputation of large windows of contiguous missing data, and, for example, in the tests presented in this work, we successfully calculated a maximum window of 2,736 data points.

The results show that the statistical properties of the calculated and available data are very similar even for large windows of missing values, thus confirming the accuracy of the proposed algorithm. The tests in this work also show that the proposed algorithm is adequate for the imputation of data on time-series with seasonal and cyclic patterns, achieving lower RMSE and higher R^2 values when compared to an autoregressive model employed for testing. Moreover, the algorithm is also tested with time-series of beer sales and the number of air passengers, also showing good performance. The results suggest that the proposed algorithm can be used to complete missing data in weather stations to develop, for example, forecasting models based on statistical and machine learning techniques for the sizing of PV systems or wind parks.

ACKNOWLEDGEMENT

This research was funded by the Postdoctoral Scholarship Program 2022 under grant CVU-410775 and by the Project PCC-2022-319619, both awarded by CONAHCYT, México.

The authors also thank Universidad de Guadalajara for supporting the research reported in this paper.

(Received September 8, 2023)

REFERENCES

[1] H. Ahn, K. Sun, and K. P. Kim: Comparison of missing data imputation methods in time series forecasting. Computers Materials Continua *70* (2022), 767–779. DOI:10.32604/cmc.2022.019369

[2] O. Anava, E. Hazan, and A. Zeevi: International Conference on Machine Learning. Proc. Machine Learning Research, Lille 2015.

[3] F. Bashir and H. L. Wei: Handling missing data in multivariate time series using a vector autoregressive model-imputation (VAR-IM) algorithm. Neurocomputing *276* (2018), 23–30. DOI10.1016/j.neucom.2017.03.097

[4] G. E. A. P. A. Batista and M. C. Monard: An analysis of four missing data treatment methods for supervised learning. Appl. Artific. Intell. *17* (2003), 519–533. DOI:10.1080/713827181

[5] L. P. Bras and J. C. Menezes: Dealing with gene expression missing data. IEE Proceedings - Systems Biology, *153* (2006), 105–119. DOI: 10.1049/ip-syb:20050056

[6] S. Brown, R. Tauler, and B. Walczak: Comprehensive Chemometrics: Chemical and Biochemical Data Analysis. (Second edition.) Elsevier, Smsterdam 2020.

[7] M. K. Choong, M. Charbit, and H. Yan: Autoregressive-model-based missing value estimation for DNA microarray time series data. IEEE Trans. Inform. Technol. Biomedicine *13* (2009), 131–137. DOI:10.1109/TITB.2008.2007421

[8] E. L. Dan, M. Dînşoreanu, and R. C. Mureşan: 2020 IEEE International Conference on Automation, Quality and Testing, Robotics (AQTR). IEEE, London 2020.

[9] W. Dunsmuir and P. M. Robinson: Estimation of time series models in the presence of missing data. J. Amer. Statist. Assoc. *76* (1981), 560–568. DOI:10.1080/01621459.1981.10477687

[10] A. Folch-Fortuny, F. Arteaga, and A. Ferrer: Enabling network inference methods to handle missing data and outliers. BMC Bioinformatics *16* (2015), 1–12. DOI:10.1186/s12859-015-0717-7

[11] A. Folch-Fortuny, F. Arteaga, and A. Ferrer: PCA model building with missing data: New proposals and a comparative study. Chemometr. Intell. Labor. Systems *146* (2015), 77–88. DOI:10.1016/j.chemolab.2015.05.006

[12] A. Folch-Fortuny, F. Arteaga, and A. Ferrer: Missing data imputation toolbox for MATLAB. Chemometr. Intell. Labor. Systems *154* (2016), 93–100. DOI:10.1016/j.chemolab.2016.03.019

[13] J. M. González-Martíneza, O. E. de Noord, and A. Ferrer: Multisynchro: a novel approach for batch synchronization in scenarios of multiple asynchronisms. J. Chemometr. *28* (2014), 462–475. DOI:10.1002/cem.2620

[14] D. Hui, S. Wan, B Su, G. Katul, R. Monson, and Y. Luo: Gap-filling missing data in eddy covariance measurements using multiple imputation (MI) for annual estimations. Agricultur. Forest Meteorology *121* (2004), 93–111. DOI:10.1016/S0168-1923(03)00158-8

[15] W. L. Junger and A. Ponce de Leon: Imputation of missing data in time series for air pollutants. Atmosph. Environment *102* (2015), 96–104. DOI:10.1016/j.atmosenv.2014.11.049

[16] S. Liu and P. C. M. Molenaar: iVAR: A program for imputing missing data in multivariate time series using vector autoregressive models. Behavior Res. Methods *46* (2014), 1138–1148. DOI:10.3758/s13428-014-0444-4

[17] R. Magán-Carrión, F. Pulido-Pulido, J. Camacho and P. García-Teodoro: Tampered data recovery in WSNs through dynamic PCA and variable routing strategies. J. Commun. *8* (2013), 738–750. DOI:10.12720/jcm.8.11.738-750

[18] S. Makridakis, S. C. Wheelwright, and R. J. Hyndman: Forecasting: Methods and Applications. (Third edition.) Wiley, India 2008.

[19] D. C. Montgomery: Statistical Quality Control. (Sixth edition.) Wiley, New York 2005.

[20] H. Murad, R. Dankner, A. Berlin, L. Olmer, and L. S. Freedman: Imputing missing time-dependent covariate values for the discrete time Cox model. Statist. Methods Medical Res. *29* (2020), 2074–2086. DOI:10.1177/0962280219881168

[21] D. T. Neves, J. Alves, M. G. Naik, A. J. Proenca, and F. Prasser: From missing data imputation to data generation. J. Comput. Sci. *61* (2022), 101640. DOI:10.1016/j.jocs.2022.101640

[22] N. M. Noor, M. M. Al Bakri-Abdullah, A. Shukri Yahaya, and N. A. Ramli: Comparison of Linear Interpolation Method and Mean Method to Replace the Missing Values in Environmental Data Set. Trans Tech Publications, Switzerland 2014.

[23] R. Pedreschi, M. L. A. T. M. Hertog, S. C. Carpentier, J. Lammertyn, J. Robben, J. P. Noben, B. Panis, R. Swennen, and B. M. Nicola: Treatment of missing values for multivariate statistical analysis of gel-based proteomics data. Proteomics *29* (2008), 1371–1383. DOI:10.1007/978-1-4020-6754-9_11728

[24] J. Quevedo, V. Puig, G. Cembrano, J. Aguilar, C. Isaza, D. Saporta, G. Benito, M. Hedo, and A. Molina: Estimating missing and false data in flow meters of a water distribution network. IFAC Proc. Vol. *39* (2006), 1181–1186. DOI:10.3182/20060829-4-CN-2909.00197

[25] Y. Sun, J. Li, Y. Xu, T. Zhang, and X. Wang: Deep learning versus conventional methods for missing data imputation: A review and comparative study. Expert Systems Appl. *227* (2023), 120–201. DOI:10.1016/j.eswa.2023.120201

[26] M. Zarzo and P. Martí: Modeling the variability of solar radiation data among weather stations by means of principal components analysis. Appl. Energy *88* (2011), 2775–2784. DOI:10.1016/j.apenergy.2011.01.070

[27] Z. Zhang: Missing data imputation: focusing on single imputation. AME Publ. *4* (2016), 1–8. DOI:10.21037/amj.2016.12.02

Eduardo Rangel-Heras, Universidad de Guadalajara, Blvd. Gral. Marcelino García Barragán 1421, Olímpica, C.P. 44430 Guadalajara. México.
e-mail: eduardo.rangel@academicos.udg.mx

Pavel Zuniga, Universidad de Guadalajara, Blvd. Gral. Marcelino García Barragán 1421, Olímpica, C.P. 44430 Guadalajara. México.
e-mail: pavel.zuniga@academicos.udg.mx

Alma Y. Alanis, Universidad de Guadalajara, Blvd. Gral. Marcelino García Barragán 1421, Olímpica, C.P. 44430 Guadalajara. México.
e-mail: alma.alanis@academicos.udg.mx

Esteban A. Hernandez-Vargas, Department of Mathematics and Statistical Science, University of Idaho, Moscow, Idaho. U. S. A.
e-mail: esteban@uidaho.edu

Oscar D. Sanchez, Universidad de Guadalajara, Blvd. Gral. Marcelino García Barragán 1421, Olímpica, C.P. 44430 Guadalajara. México.
e-mail: didier.sanchez@academicos.udg.mx