Handling missing values in greenhouse microclimate dataset using PCA-SARIMAX model

M R Ouamane, A Saboni, O Bennis, F Kratz, H Megherbi, J A Sanchez-Molina

Handling missing values in greenhouse microclimate dataset using PCA-SARIMAX model

M.R. Ouamane[1], A. Saboni[2], O. Bennis[3], F. Kratz[4], H. Megherbi[5], JA. Sanchez-Molina[6]

***Abstract*— The purpose of this paper is to present a novel approach to handle all-sensors losses of the internal greenhouse environmental data due to the power cut throughout the greenhouse. The proposed method is based on the principal components analysis (PCA) and the seasonal autoregressive integrated moving average model with exogenous variables (SARIMAX). The exogenous variables are derived from the external meteorological dataset provided by the weather station of the city where the greenhouse is located. The role of the PCA method is to analyze the correlation between exogenous and the available endogenous variables and then reduce the dimensions of the exogenous dataset. After selecting the best choice of the training set for the SARIMAX model, the obtained results show that the proposed approach represent a promising solution for completing the bulk missing data in internal greenhouse environmental dataset.**

Keywords—SARIMAX model, principal component analysis, greenhouse, missing data, data imputation, time-series.

I. INTRODUCTION

Complete and accurate datasets are essential for statistical, classification, prediction or decision-making tasks in the management of the modern greenhouse environment. Unfortunately, the sensors and the electronic equipment inside and outside the greenhouse operate in very harsh conditions: high solar radiation and humidity, blackouts and floods to name but a few. Consequently, the collected greenhouse environmental datasets usually contain small and wide gaps in certain variables (individual sensors losses) and even in all the variables (all-sensors losses). Hence, the modeling and prediction of greenhouse environment variables using such datasets are complicated and less accurate. It is argued in the data analysis community that the efficient way to handle the missing values is to impute gaps with reliable and accurate data. In the last three decades, numerous research works have been dealt with data imputation in different domains [1-7]. They revealed to be efficient when gaps appear in some variables. The presence of wide gaps in all variables, very frequent in greenhouse application, is a difficult problem and still an open issue. It explains in fact why there are few works devoted for handling bulk missing data in the greenhouse environment. In [6] the authors propose a two-dimensional convolutional neural network called U-Net to impute missing tabular data collected from 27different greenhouses affected by the same climate conditions. The objective of the proposed U-Net architecture is to learn the evolution patterns and interpret the relationship between the five environmental variables (the external and internal temperature, the internal relative humidity, the internal CO2 concentration and the solar radiation). A comparative study has been conducted with linear interpolation (LI), feedforward neural network (FFNN) and long short-term memory (LSTM). The U-Net network and the compared ones were trained using 30% data loss. The obtained results show an acceptable accuracy of the U-Net with Screen size of 50 for the different variables. The LI was unable to impute data of all-sensor losses but had comparable performance to the U-Net for individual sensors losses, while the FFNN and LSTM have failed to be train properly. The encouraging performances of the U-Net architecture in fact depend on the availability of the huge datasets (from 27 greenhouses) which is not always affordable in practice.

In this paper, a method for handling bulk missing data in greenhouse microclimate dataset is developed using the dataset of one greenhouse and the external environment dataset provided by the

[1]Mohamed Ridha OUAMANE is with Department of Electrical Engineering, LESIA Laboratory, University of Biskra, Algeria, and also with PRISME Laboratory, INSA Centre Val De Loire, Bourges, France. (mohamed.ouamane@insa-cvl.fr)

[2]Amine SABONI is with Octo, France.
(saboni.amine@gmail.com)

[3]O. Bennis is with PRISME Laboratory, University of Orleans, 28000 Chartres, FRANCE
(e-mail: ouafae.bennis@univ-orleans.fr).

[4]F. Kratz is with PRISME Laboratory, INSA Centre Val de Loire, Bourges, FRANCE.
(e-mail: frederic.kratz@insa-cvl.fr).

[5]H. Megherbi is with LESIA Laboratory, University of Biskra, Algeria.
(h.megherbi@univ-biskra.dz)

[6]JA. Sanchez-Molina is with UAL-ARM-TEP197, Laboratory, University of Almeria.
(jorgesanchez@ual.es)

meteorological station of the city where the greenhouse is located (in our case Almeria, Spain). To tackle the daily seasonality of 24 hours in the greenhouse environmental variables, the proposed method is based on seasonal autoregressive integrated mobile average with exogenous variable (SARIMAX) modeling technique, in addition to the principal component analysis (PCA) for data preprocessing and dimensionality reduction of the exogenous variables.

The remainder of this paper is organized as follows. In section II, the available greenhouse environmental dataset is described in addition to the details of the proposed data restoration approach including the data analysis based on PCA, SARIMAX model and the model parameters selection method. The obtained results and their discussions are presented in section III. The main conclusions drawn from the study are illustrated in the last section.

II. MATERIALS AND METHODS

A. *AVAILABLE DATASET*

The investigation of the developed method was performed by an experimental dataset provided by the University of Almeria in Spain within a framework of cooperation with PRISME laboratory, France. The dataset is collected from a greenhouse located at the Cajamar Foundation (El Ejido, Almería, South-East Spain) covered by PE film of 200 µm thickness. This traditional "parral-type" greenhouse has a surface of 877 m^2 (37.80 m × 23.20 m). The greenhouse is equipped with a hinged roof window with a maximum opening angle of 45°, and a lateral window with a length of 37 m, and an opening from 0◦ to 45◦. The greenhouse is instrumented with several sensors to gain environmental data. In addition, a meteorological station is installed at a height of 6 m for measuring different meteorological variables such as outside temperature, relative humidity, global radiation, photosynthetic radiation and wind speed, and direction. The database is collected in 2007 from February 15th to June 15th, with time step of 1 min. A meteorological dataset with nine variables of the city of Almeria covering all the periods of the available inside greenhouse dataset, with 1 hour time step, is introduced in this study and their variables are used as exogenous factors.

B. *IMPUTATION METHODOLOGY*

The proposed method is based on SARIMAX model, which is an extension of Seasonal Autoregressive Integrated Mobile Average (SARIMA) model, upgraded with the ability to integrate exogenous variables. In the present work, the used exogenous variables represent the external environment of the greenhouse, in order to make the model more accurate and to increase the forecasting performance, especially for bulk data missing.

Firstly, using the PCA, the relationship between datasets variables (endogenous and exogenous) is analyzed; then, the exogenous dataset is reduced to remove redundancy from the data. Once this pre-processing operation is carried out, the compressed data is used as exogenous factors for the SARIMAX model. Using SARIMAX model suppose that the endogenous time series is stationary. Therefore, the unit root test is performed to verify the stationarity requirement of the input time series. To identify the SARIMAX model parameters, the corrected Akaike-information-criterion is used.

C. *PRINCIPAL COMPONENT ANALYSIS (PCA)*

The PCA method is one of the popular dimension reducing techniques. It removes redundancy and complexity from correlated data and leads to a smaller number of dimensions. A data distribution containing the descriptive variables can be transformed into another one with the necessary characteristic variables known as principal components describing most of the information of the original dataset. Indeed, more variables are correlated less number of principal components are needed for representing the data.

In the present study, the PCA is used to select pertinent variables from the set of correlated data provided by the meteorological station of ALMERIA city. The dataset of correlated variables is transformed into an uncorrelated component.

The original dataset is m × n matrix noted by X, where m is the number of variables (rows) and n number of samples (columns). It is normalized as follows:

$$X_{centred}(i) = X_i - \acute{X}_i \quad (1)$$

X_i : The ith variable of the dataset.

$\acute{X}_i$: Mean of the i[th] variable of the dataset.

The resultant normalized data is used to calculate the covariance matrix S defined by the following expression:

$$S=\frac{1}{1-n}X_{centred}^{T}X_{centred}\in R^{m\times m} \quad (2)$$

Finally, the eigenvalues and eigenvectors of the correlation matrix are calculated to identify the PCA model parameters. The variance part of each principal component is related to the eigenvalues as fellows:

$$W_{PC}(i)=\frac{\lambda_i}{\sum_{j=1}^{m}\lambda_j} \quad (3)$$

The selection of the representative principal component is based on the sum of the preserved variance for each variable. The dispersion of data is sufficiently preserved when $\sum_{i=0}^{m} W_{PC}(i)>0.8$ where $i\epsilon(1,2,\ldots,m)$.

Hence, the transformation of correlated variables x_i into new uncorrelated variables z_i is performed as follows:

$$Z=U^{T}X_{centred} \quad (4)$$

These uncorrelated variables are called the principal components of $X_{centred}$, the i[th] component is obtained as follow:

$$z_i=u_i^{T}X_{centred} \quad (5)$$

With u_i the i[th] eigenvector of the covariance matrix S.

The vectors of maximum variation can be represented in a visualization space called “factorial space” or “correlation circle”. This graphical representation shows how the variables are correlated.

The coefficient of correlation between two variables is given by the cosinus of the angle between these two variables:

$$\rho_{i,j}=\cos\left(Angle(x_i,x_j)\right) \quad (6)$$

The purpose of this analysis is to define a simplified model with only pertinent variables. The reduced data is used as exogenous factors. The new reduced system is also used for clustering data, to verify the relationship between exogenous and endogenous factors.

D. Box and Jenkins methodology

Box-Jenkins methodology aims to develop a mathematical model that describes the behavior of observed past values of time series to be used in forecasting future values. The used approach describes the effect of trend and seasonal components in time series, such approach is quite different from that used in regression or exponential smoothing [7]. Box and Jenkins methodology is described in a three-step process as follows: first check the stationary of the time series, by verifying if the statistical properties of the series (variance and level) are constant. The second step is the identification of the model, making an appropriate choice of model parameters. The last step consists of the fitting of a selected model and diagnostics of results.

E. SARIMAX model

SARIMAX is a mathematical model used for time series forecasting, that account for time series dependence across season, with the ability to integrate exogenous variables. The model consists of two main parts autoregressive and mobile average, with a differentiating operator[11].

In the autoregressive part, the current data is estimated based on a linear combination of past data. This part is denoted as AR (p), with p the order of the model. The formula of the model is expressed as follows:

$$X_t=\varphi_1X_{t-1}+\varphi_2X_{t-2}+\ldots+\varphi_pX_{t-p}+\varepsilon_t \quad (7)$$

$\varphi_1,\varphi_2,\ldots,\varphi_p$: The parameters of the autoregressive part of the SARIMAX model.

ε_t : White noise $\varepsilon_t\ N(0,\sigma_\varepsilon^2)$

The moving average model uses past forecast errors $t-¿,\ldots,\varepsilon_{t-p}$ to predict the current data X_t . Each $\varepsilon_{t-1},\varepsilon_¿$ value of X_t is considered as a weighted moving average of the past errors. This model is referred, as MA (q) with q is the order of the model. The model is given by:

$$X_t=\varepsilon_t-\theta_1\varepsilon_{t-1}-\theta_2\varepsilon_{t-2}-\ldots-\theta_q\varepsilon_{t-q} \quad (8)$$

$\theta_1, \theta_2, \ldots, \theta_p$: The parameters of the mobile average part of the SARIMAX model.

The combined model SARIMAX is described mathematically as follows:

$$\varphi_p(B)\Phi_P(B^s)\nabla^d \nabla_s^D X_t = \theta_q(B)\Theta_Q(B^s)\varepsilon_t$$

Where, $\varphi_p(B)$ is the regular AR model of order p. $\Phi_P(B^s)$ is the seasonal AR model of order P and s is the time span of the repeating seasonal pattern. $\theta_q(B)$ is the regular MA model of order q. $\Theta_Q(B^s)$ is the seasonal MA model of order Q. B is the backward shift operator. The model is designed by $SARIMAX(p,d,q)\times(P,D,Q)_S$, where (p,d,q) are the non-seasonal orders and $(P,D,Q)_S$ are the seasonal orders. The differentiating operator ∇^d is introduced in the model to remove the non-seasonal non-stationarity, whereas the seasonal differentiating operator ∇_s^D is used to eliminate the seasonal non-stationarity.

Before undertaking the model identification, in this study, a test of unit root called “Augmented Dickey-Fuller test” is used for checking the stationarity of the time series. The test suggests an alternative equation by subtracting Y_{t-1} from both sides of the equation (10) which represent an autoregressive model with the p order equal to 1.

$$Y_t = \rho Y_{t-1} + \epsilon \quad (10)$$

$$\Delta Y_t = (\rho - 1) Y_{t-1} + \epsilon \quad (11)$$

This test is performed under two hypotheses: the null hypothesis (H_0: $(\rho-1)=0$) which suppose that the data is non-stationary and needs to be differentiated to make it stationary; and the alternative hypothesis (H_1 : $(\rho-1)<0$) suppose that the data is stationary and there is no need for its differentiation [8]. The next step is to estimate the SARIMAX model orders, both seasonal and non-seasonal. A traditional Box-Jenkins framework for identifying a suitable model is based on autocorrelation and partial-autocorrelation. This method is not always informative, especially in small data size, the interpretation of the correlogram is more complicated when the data is differentiated, hence inappropriate models are often fitted [9]. A more efficient method for the selection of the best model (structure and parameters) is used in this paper. It is an automatic procedure based on the corrected AIC (Akaike information criteria). This procedure is an implemented Python function.

F. *Test and Diagnostic Procedure*

The diagnostic of the used model is crucial to evaluate the performance of prediction methods. For this purpose, the dataset is split into training and testing sets, where the testing set succeeding the training set. The measurement of prediction accuracy is defined as the root-mean-square error (RMSE), expressed as follow:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n} (\hat{y}_t - y_t)^2}{n}} \quad (12)$$

Where $\hat{y}_t$ is the predicted value, y_t is the observed value.

The tests used are presented as follow:

- Testing the model adequacy based on AIC criterion. The best model corresponds to the lower value [8].
- Analyzing the residuals to check the adequacy of the model, by representing the correlogram and the quantile-quantile plot.
- Analyzing the robustness of the model by changing the test set.
- Testing of the accuracy of the model.

III. Results and Discussion

The variables concerned by imputation are the inside temperature and solar irradiation of the greenhouse. At first, the dataset for training and analyzing has been formed with fifteen days from 1st March to 15th March. As the phenomena in question have slow dynamics, the time step is ch[illegible] the co[illegible]ples is

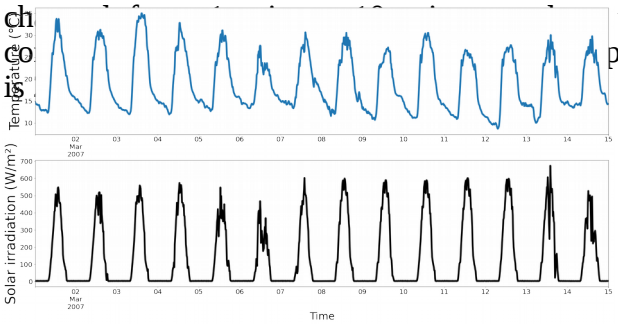


Fig. 1. Temperature and Solar irradiation for training stage.

G. *Selection of patients variables*

Analyzing data correlation with PCA allows us to select the pertinent variables to use as exogenous factors. Before analyzing data with PCA, the number of necessary principal components to represent data must be determined. This choice is essentially depending on the cumulative explained variance. Fig. 2 shows the preserved percentage of each component on the y-axis and the number of components on the x-axis.

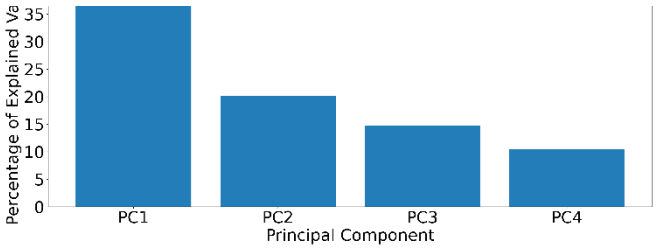


Fig. 2. Scree plot for the four components.

As illustrated in Fig 2, the preserved variance percentage for the first component PC1 is about 39% and the second component PC2 represents 16%. However, the four first components allow us to explain a cumulative percentage of 80%. Hence, the number of necessary components to represent the weather station dataset into a new space of visualization is four components.

The representation of the eigenvectors of the variables on the plane of PC1-PC2 is showed in Fig 3. It represents in fact the direction of the maximal variability for each variable in the two-principal component plane. In this plane, the correlation between variables is proportional with the angle between eigenvectors. Specifically, a small angle value represents a high correlation.

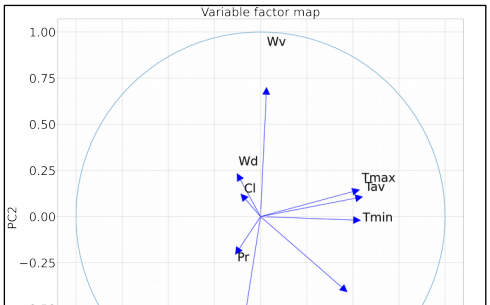
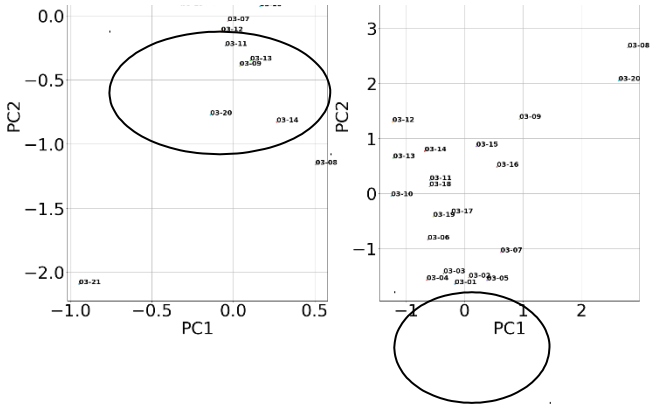


Fig. 3. Correlation circle with PC1-PC2

The representation of the direction of maximal variability for each variables in 2D plane, shows four cluster of correlated variables: {T_{max} (max atmospheric temperature), T_{min} (min atmospheric temperature) , T_{av} (average atmospheric temperature)} , {H_{ex} (external humidity), P_r (atmospheric pressure)} , {F_l (feels like temperature} and { C_l (density of clouds) , W_d (wind direction) , W_v (wind velocity)}. Correlated variables provide the same information to the system with redundancy, for the reason a selection of variables from each cluster is carried out. From the first cluster the T_{av} is selected, which represents the average atmospheric temperature, in the second cluster, we chosed the most preserved variable H_{ex} and from the last cluster the most correlated variable (W_v) with PC2 is selected.

G. *Relationship between exogenous and endogenous datasets*



(a) (b)

Fig. 4. Plot of PCA scores on the plane PC1-PC2

Concerning the internal data (a), the plot of scores obtained by PCA shows a cluster of days with a similar weather, and some days with particular weather pattern. In the other side, the data of external environment (b) confirm that these days represent a particular weather. This

coherence verified by the PCA scores, indicate that there is a strong relationship between external and internal database.

H. Stationarity Check

As referenced above, the used test of stationarity is Dickey Fuller test. The obtained T-statistical value is showed in TABLE Ⅰ.

TABLE Ⅰ.

T-statistical for Dickey Fuller test

Time series	**Temperature**	**Solar irradiation**
T-statistical	-8,78	-7,17

The T-statistical values for each time series is less than the critical values for no trend case at 1% (-3.43) based on MacKinnon [10]. Hence, the two time series for temperature and solar radiation are stationary.

Table Ⅱ represent the summary evaluation of four candidates models, based on Corrected Akaike information. The best-fitted model corresponds to the minimum value of AIC.

TABLE Ⅱ.

AIC values for fitted model

SARIMAX model	AIC	
	Temperature	Solar irradiation
(0,0,0)(2,1,0)	2518,57	6264,71
(0,0,0)(3,1,0)	2053,04	5160,97
(0,0,0)(3,1,1)	2005,81	**5067,79**
(0,0,0)(3,1,2)	**1981,09**	5076,40
(2,1,0)(3,1,2)	15740,64	16139,58

Based on the values of AIC criterion, the best fitted model is which corresponds to the minimal AIC values, Hence the used model is SARIMAX (0,0,0)×(3,1,2) for temperature and SARIMAX (0,0,0)×(3,1,1) for solar radiation. The estimated parameters of the model by Maximum Likelihood optimization with Python, is given in TABLE Ⅲ.

TABLE Ⅲ.

Optimized parameters for SARIMA model

Parameter	θ_1	θ_2	θ_3	Θ_1	Θ_2
Value for temperature model	0.4435	-0.2042	-0.0064	-1.66	0.99
Value for solar irradiation model	0.1	0.0046	-0.1	-1	

Three type of model are used to improve the impact of seasonal component and exogenous factors. The results are illustrated as follows:

- The real curve of internal temperature and solar irradiation.
- The estimated curves by each model (ARIMA, SARIMA and SARIMAX).

A. Impact of Seasonal Component

The prediction results of the ARIMA model are given in Fig. 5. It is clear that this model is unable to predict the seasonal component. This means that the ARIMA model is unsuitable to impute the bulk

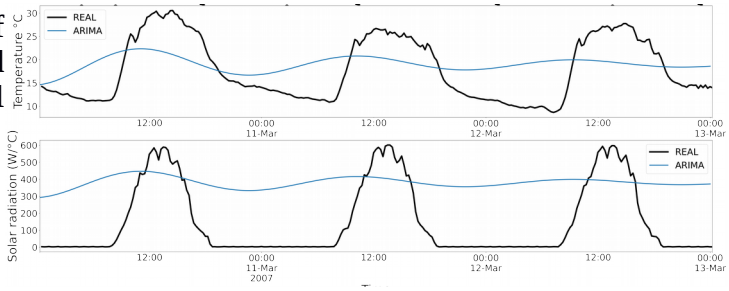


Fig. 5. Illustration of the results of the ARIMA model.

I. Impact of Exogenous Factors

For the investigation of the impact of exogenous variables, the performance of the forecasting using SARIMA and SARIMAX models are compared for two different time periods. The first period represents two days with particular weather pattern, whereas the second period represents three days with normal weather pattern. The selection of this periods is done by the PCA method and particularly using the score plot of principal component. The obtained results for the first and second time periods are illustrated in Fig. 6 and Fig.7, respectively.

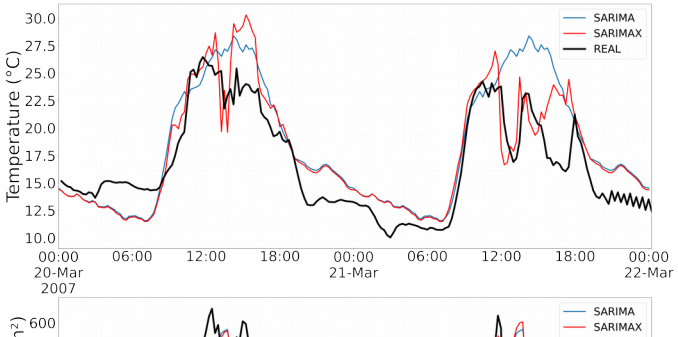


Fig. 6. Illustration of the results of the SARIMA and SARIMAX model for days with particular weather.

The evaluation in term of RMSE of the SARIMA and SARIMAX models in both time periods for the temperature and solar radiation series is given in Table IV.

From Fig. 6 and Table IV, when the days present particular weather pattern, it is obvious that the SARIMAX model performs better than SARIMA model. However, when the days presents normal weather pattern, the SARIMA model outperforms the SARIMAX model, see Fig. 7 and Table IV.

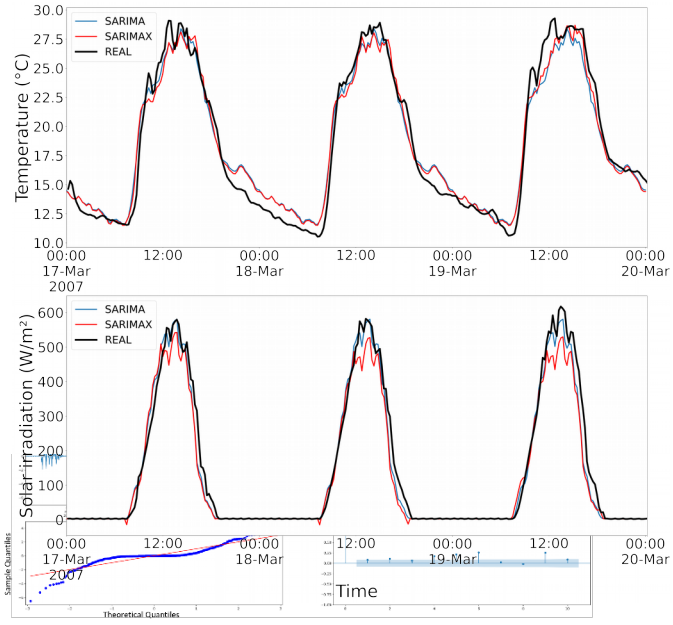


Fig. 7. Illustration of the results of the SARIMA and SARIMAX model for normal days.

TABLE Ⅳ.

RMSE values for each model

Forecast model	RMSE			
	Days with particular weather		Days with regular weather	
	Temp	RG	Temp	RG
SARIMA	3.3	121.8	**1.1**	**20.2**
SARIMAX	**2.5**	**116.7**	1.2	33.8

In the days with particular weather pattern, the exogenous factors provide additional information to the imputation mechanism. Indeed, a significant climate change has an important impact on the internal environment of the greenhouse. Therefore, the use of exogenous variables is necessary in days with particular weather pattern which constitutes usually the conditions of the occurrence of the bulk loss of data. However, in the days with normal weather pattern there is neglected impact on the internal protected environment of the greenhouse. So, in this case, the prediction of the greenhouse internal environment variables using exogenous factors can limit the performance of the model.Residual analysis

The residuals of the model are tested to improve the accuracy of the fitted model. In the Fig 8 the Q-Q plot is presented, to check whether the residuals are a Gaussian nois

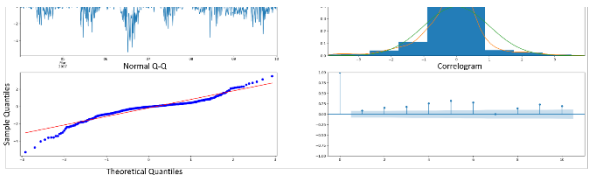


Fig. 8. Diagnostics plot for fitted models

According to the diagnostics plot the residual of the both models are not correlated. However,

the Q-Q plot shows that the residuals have the nature of Gaussian noise.

IV. CONCLUSION

This paper investigates a PCA-SARIMAX method for handling bulk loss data in greenhouse internal environment dataset. The proposed approach is evaluated using an experimental greenhouse dataset. The exogenous dataset of this greenhouse is analyzed using PCA, to simplify the model input variables while keeping the relevant ones. The PCA method is also used for clustering the days according to the weather pattern. The selected relevant variables are reduced into four principal components to be used as exogenous input of the SARIMAX model. The evaluation of the SARIMAX model results is carried out on different dataset with different weather patterns. A comparative study is performed with these different datasets and revealed that the use of exogenous factors is effective only in days representing a particular weather pattern.

In future research work, we suggest the combination of SARIMA and SARIMAX models in addition to PCA method for more accuracy in handling the missing data in different weather conditions. We also suggest the use of more advanced parameters estimation methods and models based on artificial intelligence techniques. This imputed data will be used to develop a virtual sensor for evapotranspiration estimation, used in the regulation of the greenhouse crops irrigation.

REFERENCES

[1] G. Molenberghs, M. Fitzmaurice, A. Kenward, A. Tsiatis and G. Verbeke. Handbook of Missing Data Methodology. Boca Raton : CRC Press, Taylor & Francis Group, 2015.

[2] P.S. Raja and K. Thangavel, “Missing Value Imputation using Unsupervised Machine Learning Techniques”. *Soft Computing*, vol. 24, pp. 4361–4392, 2020.

[3] J. Ke, S. Zhang, H. Yang and X. Chen, “PCA-Based Missing Information Imputation for Real-Time Crash Likelihood Prediction under Imbalanced Data”, *Transportmetrica A: Transport Science*, vol. 15, no. 2, pp. 872-895, 2019.

[4] R. Barrela, C. Amado, D. Loureiro and A. Mamade; “Data Reconstruction of Flow Time Series in Water Distribution Systems – A New Method that Accommodates Multiple Seasonality”. *Journal of Hydroinformatics*, vol. 19, no.2, pp. 238-250, 2017.

[5] H. Adanacioglu, M. Yercan, “An Analysis of Tomato Prices at Wholesale Level in Turkey: an Application of SARIMA Model”, *Custos egronegócio*, vol. 8, no.4, pp. 52–75, 2012.

[6] Moon, T., Lee, J.W. and Son, J.E., “Accurate Imputation of Greenhouse Environment Data for Data Integrity Utilizing Two-Dimensional Convolutional Neural Networks”, *Sensors*, 2021.

[7] A. Kocian, G. Carmassi, F. Cela, L. Incrocci, P. Milazzo and S. Chessa “Bayesian Sigmoid-Type Time Series Forecasting with Missing Data for Greenhouse Crops”, *Sensors*, 20, 3246, 2020

[8] S. Ali, W. Dinata, M. Azka, M. Faisal, Suhartono, R. Yendra, and M. D. H. Gamal , "Short-Term Load Forecasting Double Seasonal ARIMA Methods: An Evaluation Based on Mahakam-East Kalimantan Data", *AIP Conference Proceedings*, 2020.

[9] A. C. Harvey and P. H. J. Todd, “Forecasting Economic Time Series With Structural and Box-Jenkins Models: A Case Study”, *Journal of Business & Economic Statistics*, vol.1, no. 4, pp.299-307, 1983.

[10] J. G. MacKinnon, "Critical Values For Cointegration Tests," *Working Paper 1227*, Economics Department, Queen's University, 2010.

[11] S. I. Vagropoulos, G. I. Chouliaras, E. G. Kardakos, C. K. Simoglou and A. G. Bakirtzis, "Comparison of SARIMAX, SARIMA, modified SARIMA and ANN-based models for short-term PV generation forecasting," *2016 IEEE International Energy Conference (ENERGYCON)*, 2016.