Avrupa Bilim ve Teknoloji Dergisi
Sayı 35, S. 548-555, Nisan 2022

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
No. 35, pp. 548-555, April 2022

Research Article

Missing Data Imputation for Solar Radiatıon by Deep Neural Network

Eyyup Ensar Başakın[1*], Mehmet Özger[2]

[1*] İstanbul Technical University, Faculty of Civil, Departmant of Civil Engineering, İstanbul, Turkey, (ORCID: 0000-0002-9045-5302), basakin@itu.edu.tr
[2] İstanbul Technical University, Faculty of Civil, Departmant of Civil Engineering, İstanbul, Turkey, (ORCID: 0000-0001-9812-9918), ozgerme@itu.edu.tr

(First received 9 March 2022 and in final form 30 April 2022)

(**DOI:** 10.31590/ejosat.1085022)

ATIF/REFERENCE: Başakın, E.E. & Özger, M. (2022). Missing Data Imputation for Solar Radiation by Deep Neural Network. *European Journal of Science and Technology*, (35), 548-555.

Abstract

The quality of observations is fundamental issue in natural sciences. Here, the accurate and complete data is required to accomplish satisfactory estimations. There are several factors impairing the quality of measurements, such as a broken or mis-calibrated device and error in reading the measurements. Thus, this study primarily aims the imputation of the missing values in measurement of solar radiation data. Deep Neural Network (DNN) method was used to handle the missing data, and benchmarked with the classical approaches, i.e., Mean Imputation (MI), which one of the most frequently adopted data imputation method in the pertinent literature, the Linear Interpolation (LI) and Spline Interpolation (SI). The overall results highlighted that the DNN method outperformed its counterparts in terms of missing value handling through providing a greater accuracy according to the various performance metrics compared to the classical methods. It is believed that the proposed approach could make valuable contribution to the body of knowledge as well as providing significant overview to the interested researchers by filling the important gap exists in the pertinent literature.

Keywords: Deep learning, Gap filling, Machine learning, Solar radiation.

Eksik Solar Radyasyon Verilerinin Derin Sinir Ağları ile Tamamlanması

Öz

Gözlemlerin kalitesi doğa bilimlerinde önemli bir konudur. Tatmin edici tahminleri gerçekleştirmek için doğru ve eksiksiz veriler gereklidir. Bozuk veya yanlış kalibre edilmiş bir cihaz ve ölçümlerin okunmasındaki hata gibi ölçümlerin kalitesini bozan çeşitli faktörler vardır. Bu çalışmada, güneş radyasyonu verilerinin ölçümünde kayıp değerlerin tamamlanması amaçlanmaktadır. Eksik verileri işlemek için Derin Sinir Ağı (DNN) yöntemi kullanılmış ve ilgili literatürde en sık benimsenen veri atama yöntemlerinden biri olan Ortalama Atama (MI) gibi klasik yaklaşımlarla, Doğrusal İnterpolasyon (LI) ve Spline İnterpolasyon ile kıyaslama yapılmıştır. Genel sonuçlar, DNN yönteminin, klasik yöntemlere kıyasla çeşitli performans ölçütlerine göre daha fazla doğruluk sağlayarak eksik veri tamamlama açısından benzerlerinden daha iyi performans gösterdiğini vurguladı. Önerilen yaklaşımın, ilgili literatürde var olan önemli boşluğu doldurarak ilgili araştırmacılara önemli bir genel bakış sağlamanın yanı sıra bilgi birikimine değerli katkılarda bulunabileceğine inanılmaktadır.

Anahtar Kelimeler: Derin Öğrenme, Eksik Veri Tamamlama, Makina öğrenmesi, Solar Radyasyon.

[*] Corresponding Author: basakin@itu.edu.tr

1. Introduction

Solar radiation values play a key role in the recent instances of hydrological drought. It is known that solar radiation is a critical factor in evaporation (Heck et al., 2020). All meteorological readings that are known to affect drought should be known to determine its extent and take measures. Any analysis or modeling requires such data to be complete and highly accurate. In developing countries, measurement of meteorological data may be subject to some disruptions (Hunziker et al., 2017). These disruptions may be associated with the instruments used, human error, and environmental factors. Measurement of solar radiation data is vulnerable to errors. Instruments should be calibrated and physically cleaned regularly. Otherwise, serious omissions and deviations may occur in the readings.

Some omissions in solar radiation data may be compensated for by classical methods. Such methods principally involve compensation using basic statistical indicators. Mean, mode or median values of a time series are used to compensate for the missing data in the series (Awawdeh et al., 2022; Schneider, 2001). Another classical method is the use of multiple linear regression (MLR) to restore the missing data. Missing data in a dataset may be recovered using independent variables of the same date. Such independent variables may be other meteorological parameters that are known to be associated with solar radiation (Başakın et al., 2021). The major drawback of the MLR method is that it is subject to many prerequisites including normal distribution, stability of variance, significance of coefficients, and normal distribution of errors (Başakın & Ekmekcioğlu, 2021). Another classical method is the interpolation method. Linear or spline interpolation of a time series data can be used to restore the missing data (Stisen & Tumbo, 2015). The stations measuring solar radiation close to the relevant area may also be used for this purpose. Geostatistical methods are used for spatial studies. The most common methods among them are the Krigging and the inverse distance weighting method (Nikroo et al., 2010).

Recent advancement of machine learning methods has enabled significant improvements in compensating for missing data. One of the key characteristics of machine learning in this regard is that it allows to work flexibly with nonlinear data and make highly accurate estimations. K-nearest neighbor (Ratolojanahary et al., 2019), support vector machine (Gill et al., 2007), decision tree (Hamzah et al., 2021), fuzzy logic (Saplioglu & Kucukerdem, 2018) and artificial neural networks (Coutinho et al., 2018) are the methods used in completing hydrological data. Among them the most frequently used and developed method is ANN. Positive developments in computer technologies have enabled to develop many ANN methods with greater flexibility. Greater amounts of data and increased numbers of variables have given rise to more complex ANNs that yield better results. Today, the most frequently used ANN architectures are deep neural network models. DNN marked a significant progress in ANN.

In this study, DNN was used to restore the missing data of total daily solar radiation. Meteorological variables were used to estimate the missing data. The meteorological variables used were temperature, wind speed, humidity and sunshine hours. Random gaps of different sizes were made in a definite full set. The data at hand were used to train the DNN model, and the gaps were estimated. The study also involved a comparative analysis using classical imputation methods. Then the outcomes were statistically analyzed to pick the most appropriate model.

2. Material and Methods

2.1. Study area and data

This study measures the missing solar radiation values of a station located in the Central Anatolia region of Turkey, where the climate is semi-arid. The station is located at latitude 38.37255, longitude 34.02537 and at an altitude of 980 meters. The studied area is a continental climate zone that is poor in water resources. The yearly average precipitation is 300 mm which is below the Turkey average. The level of ground water has also dropped remarkably in recent years (Demir et al., 2021). The agriculture mode of the area is dry grain farming. 1857 instances of data collected from 2016 to 2021 were used in the study. Descriptive statistical information on all meteorological data is available in Table 1. When table 1 is examined, it is observed that solar radiation values are close to the normal distribution, but other parameters are far from the normal distribution.

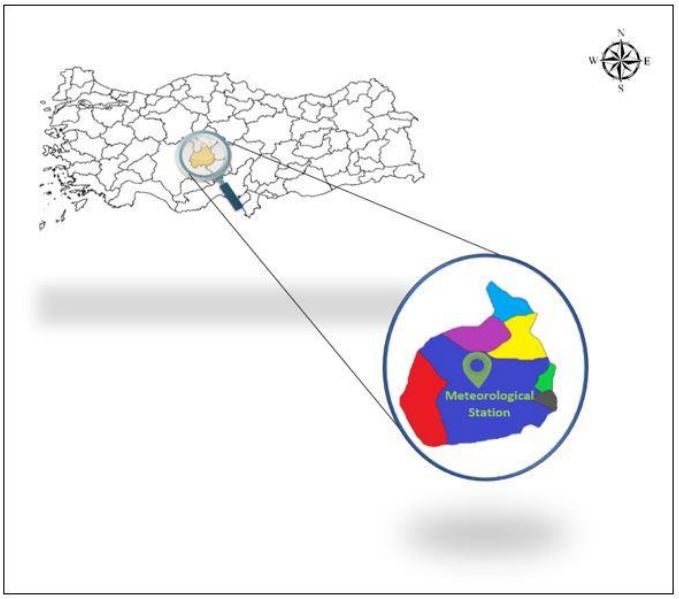


Figure 1. Study area

2.2. Classic imputation methods

2.2.1. Mean Imputation (MI)

Imputation of missing data using mean value can be considered the simplest imputation method. It operates on the principle of filling the gaps using the mean values of the entire data. The most significant disadvantage of this method is that the mean value affected by extreme values. Since variations between maximum and minimum values affect the mean value, undesirable estimations may ocur (Osman et al., 2018)

2.2.2. Linear and Spline Interpolation (LI and SI)

Interpolation methods are usually implemented by generating curve(s) for the existing data points (x_i, f_i). The functions used for this purpose are called interpolation functions (Figure 2). Mostly polynomials of different orders are used as interpolation functions. However, in some cases, more special functions such as logarithmic, exponential and hyperbolic functions, or for periodic data values, trigonometric functions can be used. It would be better to use finite difference-based interpolation methods if the data points are at equal intervals, or linear

interpolation, Lagrange interpolation, etc. if the data points are not at equal intervals.hyperbolic functions, or for periodic data values,

Table 1. Descriptive Statistics

Parameters	Unit	Maximum	Minimum	Mean	Median	St. Dev.	Skewness
Maximum Temperature	℃	38.7	-5.5	20.28	21.2	9.83	-0.32
Minimum Temperature	℃	24.8	-16.5	8.09	8.3	7.76	-0.23
Average Temperature	℃	31	-11.1	13.91	14.15	8.91	-0.21
Maximum Relatiive Humidity	%	96	24	71.12	73	15.95	-0.41
Minimum Relatiive Humidity	%	91	5	29.91	25	16.44	0.98
Average Relatiive Humidity	%	94.9	14.1	50.6	48.7	17.42	0.26
Sunshine Duration	Hr	14	0	7.31	7.8	4.1	-0.31
Wind Sped	m/s	4.9	0.4	1.75	1.6	0.69	1.04
Total Solar Radiation	W/m^2/day	532578	6000	282429	285600	144874	-0.07

trigonometric functions can be used. It would be better to use finite difference-based interpolation methods if the data points are at equal intervals, or linear interpolation, Lagrange interpolation, etc. if the data points are not at equal intervals.

It is possible to modify the attained equation to make them suitable for direct interpolation rather than solving a linear equation set to interpolate by running a polynomial through a group of points.

$$f(x) = a_0 + a_1 x \tag{1}$$

Let the points (x_0, f_0) and (x_1, f) are the cartesian coordinates. Since the interpolation based on these two consecutive points satisfies the Eq. (1), Eq. (2) and Eq. (3) can be obtained as follows:

$$f_0 = a_0 + a_1 x_0 \tag{2}$$

$$f_1 = a_0 + a_1 x_1 \tag{3}$$

Here, the coefficients, i.e., a_0 and a_1, can be expressed as follows:

$$a_0 = \frac{f_0 x_1 - f_1 x_0}{x_1 - x_0}; \qquad a_1 = \frac{f_1 - f_0}{x_1 - x_0};$$

Thus, one can identifies the equation for a linear system as follows:

$$f(x) = \frac{f_0 x_1 - f_1 x_0}{x_1 - x_0} + \frac{f_1 - f_0}{x_1 - x_0} x \tag{4}$$

The final equation can be obtained through Eq. (5)

$$f(x) = L_0 f_0 + L_1 f_1 \tag{5}$$

$$L_0 = \frac{x - x_1}{x_0 - x_1}; \quad L_1 = \frac{x - x_0}{x_1 - x_0}$$

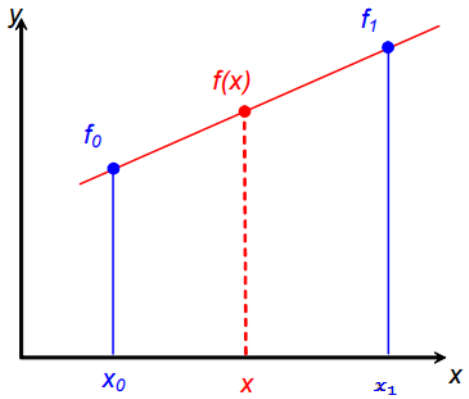


Figure 2. Graphical representation of Linear Interpolation

2.3. Deep Neural Network (DNN)

Deep learning is a field of machine learning that stands on the intersection of neural networks, artificial intelligence,

graphic modeling, optimization, pattern recognition and signal processing. Deep learning networks represent a revolutionary development in neural networks, and used to make more powerful estimations (Lecun et al., 2015). Deep learning is about supervised or unsupervised learning from data using multi-layer machine learning models (Figure 3). Layers in these models are made up of multiple stages of non-linear data transformations where properties of the data are represented in increasing and more abstract layers.

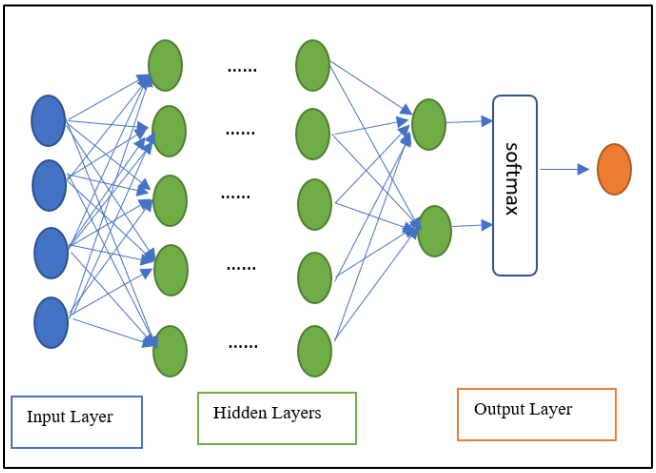


Figure 3. Architecture of DNN

2.4 Performance Criteria

In this study, the root mean square error (RMSE),mean absolute percentage error (MAPE) and Nash-Sutcliffe efficiency (NSE) (Nash & Sutcliffe, 1970) coefficient was used to evaluate the accuracy of the proposed models. The MSE value ranges from 0 to ∞ and the best value is 0. Zero indicates that the prediction process is performed without error. The NSE is a ratio of the mean square errors and the variance of the observed values. NSE is calculated by subtracting this ratio from 1. The resulting coefficient ranges from 1 to -∞, while 1 represents the highest accuracy. The equations of RMSE and NSE are given as follows in Equation 6 and Equation 7, respectively:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(R_{pi} - R_{oi})^2} \quad (6)$$

$$NSE = 1 - \left[\frac{\sum_{i=1}^{n}(R_{pi}-R_{oi})^2}{\sum_{i=1}^{n}(R_{oi}-\bar{R}_o)^2}\right] \quad (7)$$

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{R_{oi}-R_{pi}}{R_{oi}}\right| \quad (8)$$

where, n represents number of observation and prediction, R_{pi} represents the ith solar radiation predicted value, R_{oi} denotes ith solar radiation observed value, $\bar{R}_o$ denotes the average of observed solar radiation.

Table 2. Model Performance

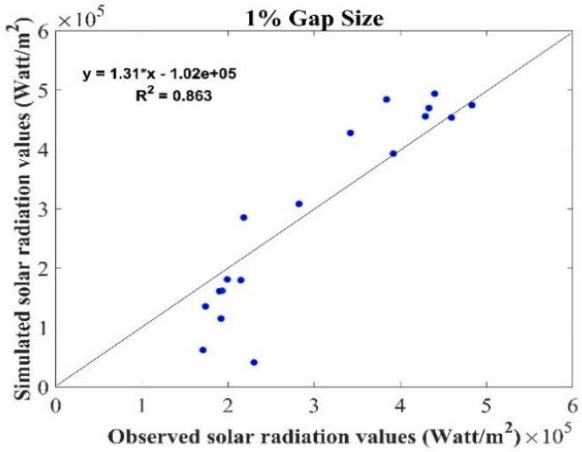
Model	Gap Size					Gap Size					Gap Size				
	1%	5%	10%	20%	30%	1%	5%	10%	20%	30%	1%	5%	10%	20%	30%
	RMSE					NSE					MAPE				
DNN	69014	**24431**	60789	**50048**	69624	0.821	**0.969**	0.819	**0.877**	0.775	21	**19**	20	**19**	23
LI	52022	63405	**60484**	65583	**66134**	0.898	0.796	**0.821**	0.790	**0.797**	20	21	**20**	23	**22**
SI	**50232**	75797	72005	77660	82461	**0.905**	0.705	0.746	0.709	0.684	**20**	28	28	31	35
MI	159313	139794	142651	143862	146792	0.050	-7.1E-05	0.005	0.002	7.61E-05	97	59	87	90	100

3. Results and Discussion

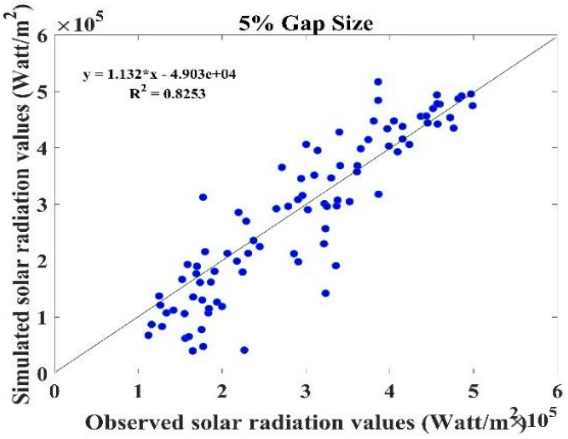
Three classic methods and a modern method were used to impute missing data in the study. A verified dataset was selected for using the methods. In the dataset, inputs were meteorological data, and outputs were daily total solar radiation data. Random gaps were created in the solar radiation dataset. Gaps were categorized in five different groups: 1%, 5%, 10%, 20%, and 30%. A fixed distinction was not made between the training and test datasets. Only for the DNN method, a distinction of 80-20% was made in the verification phase, then a test was run for the parts where there were missing data. Only solar radiation time series data were used for the imputation done by classical methods. One of the classic methods, mean imputation was calculated by taking the arithmetic average of the solar radiation time series values. This value was assigned to all missing data to complete the simulation. The mean imputation method did not yield good results due to the nonlinear nature of the radiation data. When the root mean square error (RMSE) and Nash- Sutcliffe (NSE) values based on the magnitudes of the missing data, the mean imputation method performed poorly in all groups.

Linear and spline interpolation methods were selected for the completion using interpolation. They performed better than the DNN model in the estimations for the 1%, 10% and 30% gap rates. The SI method performed the best in the 1% gap rate with simulation RMSE of 50232. In the 10%-gap group, LI performed the best simulation with RMSE of 60585. Lastly, LI performed the best in the 10%-gap group with 66134 RMSE. It is thought that the main reason the aforementioned models performed better than the DNN model is the distribution of the gaps.

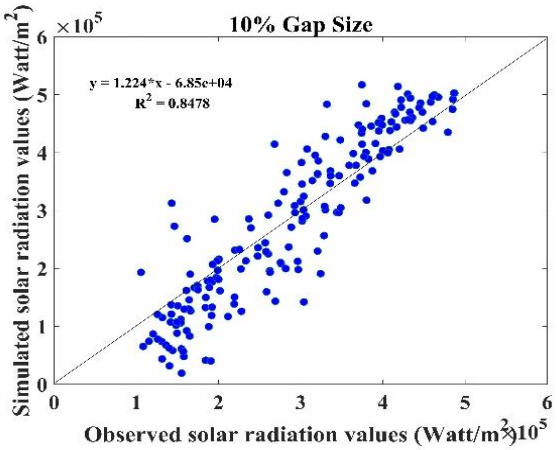
Five different models for five different gap groups using the DNN model. Three hidden layers, and 100, 50, and 30 hidden neurons were used in each of the hidden layers. The "ReLu" activation function was used as the activation function. A grid search algorithm was used to optimize the hyperparameters. DNN was the third best model in the 1% gap group after LI and SI with RMSE = 69624. For the 5% gap group, the most successful method was DNN with RMSE =50048. It ranked second by a narrow margin in the 10% gap group, and outperformed the other estimation models by a considerable margin in the 20% gap group. In the last group, DNN made the second-best estimations again by a small margin for a gap rate of 30% (Table 2).



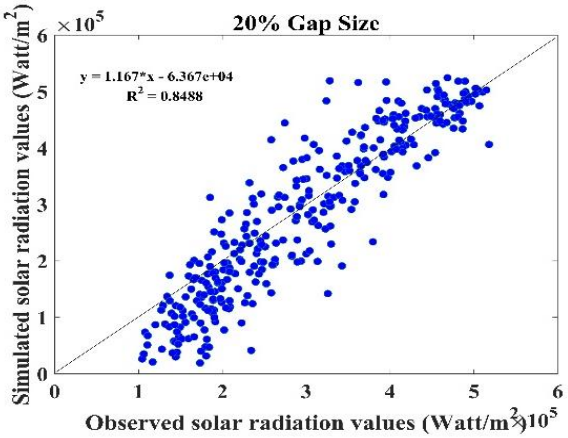
a)



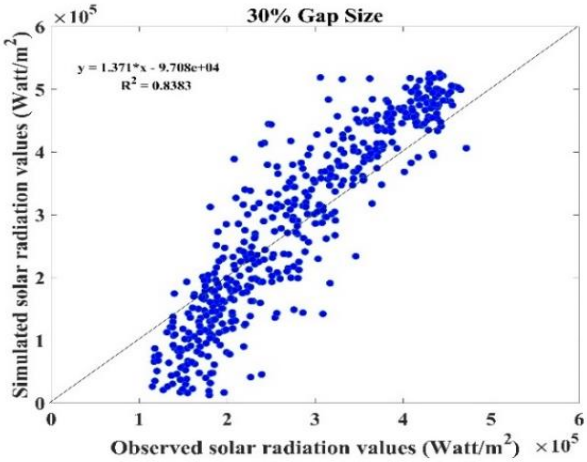
b)



c)



d)

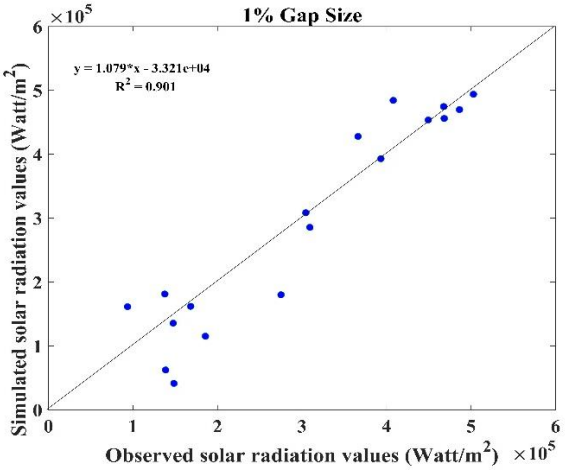


e)

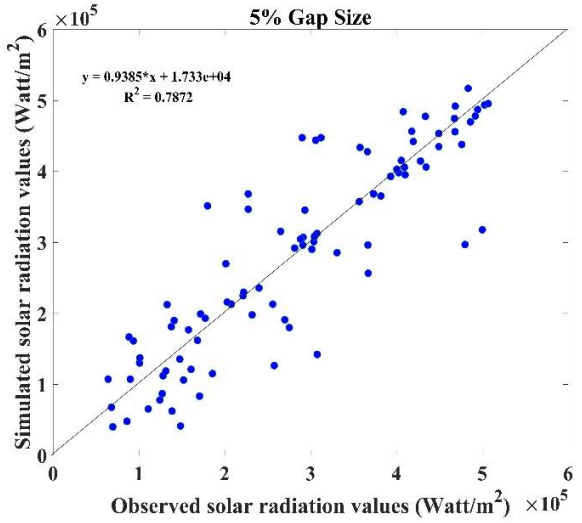
Figure 4. Scatter plot of DNN model different gap size a)1%, b)5%, c)10%, d)20%, e)30%

Figure 4, 5, and 6 show the scatter plots of the estimated and observed values. Figure 4 indicates that almost all DNN models scatter closely to the ideal estimation line. This implies that the model is not subject to any overfitting and the models do not have errors entailing and affecting each other. Only the number of hidden layers and hidden neurons may be considered

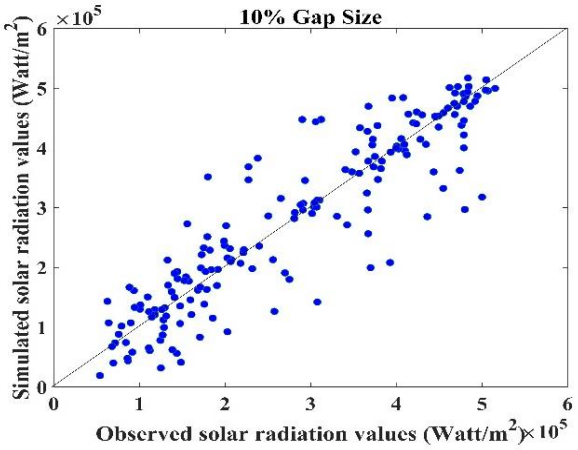
inadequate for the model devised for the gap rate of 30%. The model performed poorly in estimating small values. An examination of the figure 5 reveals distributions far from the ideal estimation line for any gap rate. Even if estimation errors show a normal distribution, it is fair to say that model variances have some excess. An examination of the observed and simulated values of the SI method reveals underestimation for all gap groups (Figure 6). The estimated values were mostly below the observed values. This implies that it is likely for a systematic series of errors to arise from the outcomes derived from this method.



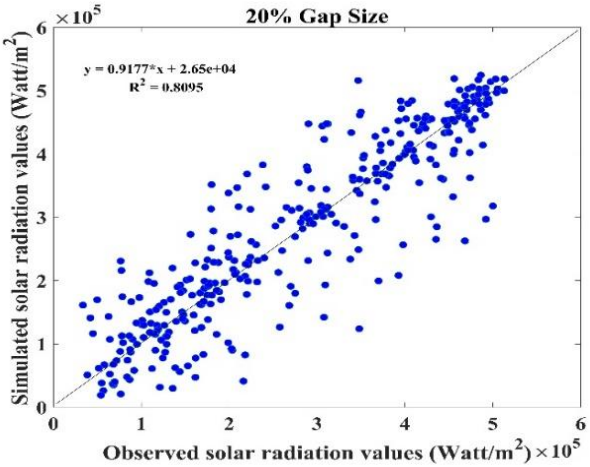
a)



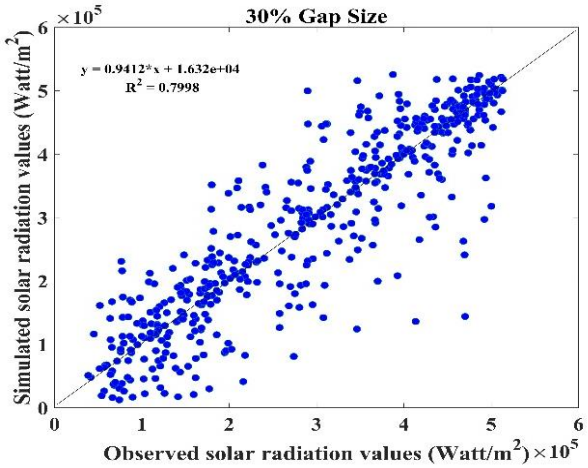
b)



c)



d)

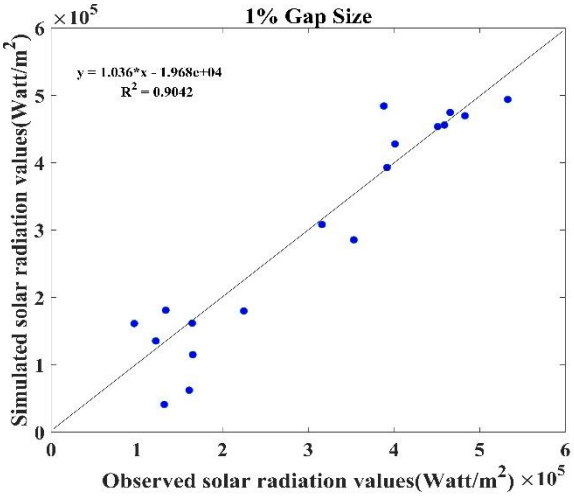


e)

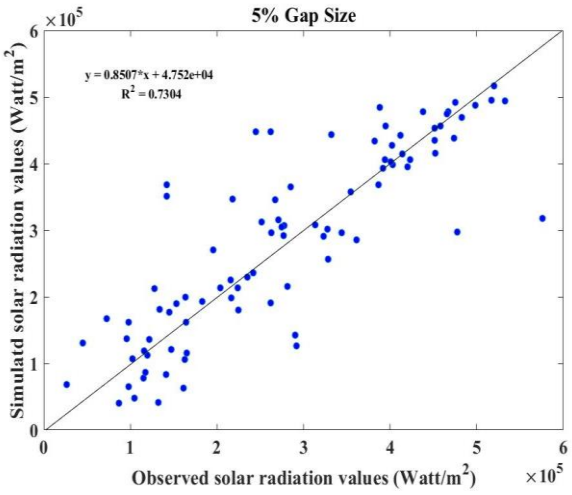
Figure 5. Scatter plot of Linear Interpolation model different gap size a)1%, b)5%, c)10%, d)20%, e)30%

4. Conclusions

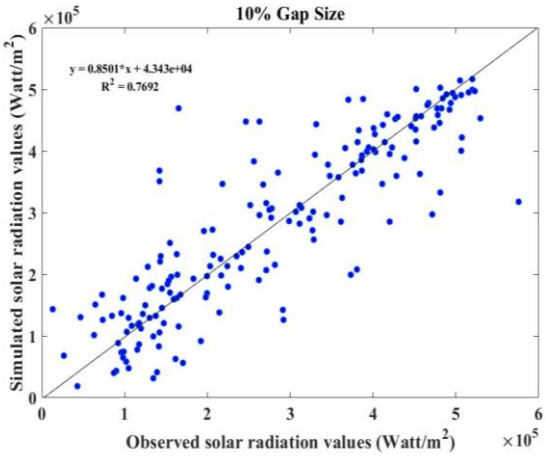
Data quality is essential in natural sciences. Data engineering favors quality over quantity of data. This study was about imputation of missing solar radiation data using several methods. Three of the methods were classical methods which involved estimation of missing data using solar radiation values only. In addition to the classical methods, the DNN method that has become popular in recent years was used. Inputs of the DNN method were the meteorological parameters that were known to affect the solar radiation data. Using meteorological data obtained from the same station on the same date, the solar radiation values were estimated to a statistically successful extent. The DNN method performed well in four models out of five generated for the gap rates, and was outperformed significantly by the classical methods in only one of the models. It was concluded from the study that distribution of missing data may affect the estimation outcomes. Considering the structure of the randomly generated gaps, classical methods may also yield good outcomes. In future studies, data sets with different gap distributions will be generated to investigate which model can yield better outcomes for each distribution type. The study highlighted that the DNN algorithm, which broke new grounds in natural sciences as well as computer science, is useful for imputation of missing data and superior to the classical methods in some respects.



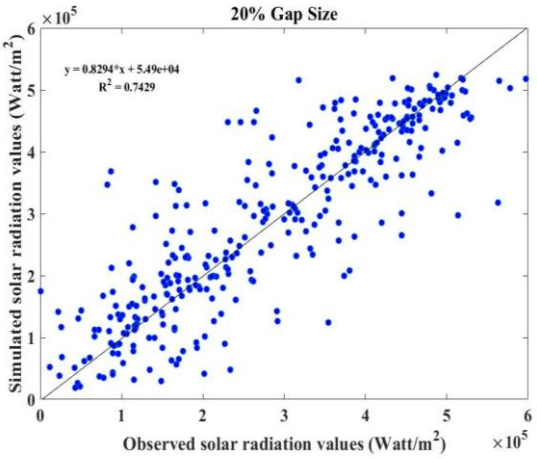
a)



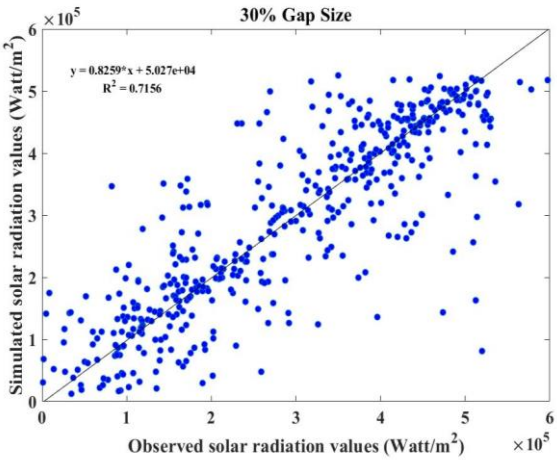
b)



c)



d)



e)

Figure 6. Scatter plot of Spline Interpolation model different gap size a)1%, b)5%, c)10%, d)20%, e)30%

5. Acknowledge

We would like to thank Meteorological General Institution for providing meteorological data. The authors would like to acknowledge that this paper is submitted in partial fulfillment of the requirements for PhD degree of Eyyup Ensar Başakın at Istanbul Technical University.

References

Awawdeh, S., Faris, H., & Hiary, H. (2022). EvoImputer: An evolutionary approach for Missing Data Imputation and feature selection in the context of supervised learning. Knowledge-Based Systems, 236, 107734. https://doi.org/10.1016/j.knosys.2021.107734

Başakın, E. E., & Ekmekcioğlu, Ö. (2021). Letter to the Editor "Estimation of global solar radiation data based on satellite-derived atmospheric parameters over the urban area of Mashhad, Iran." Environmental Science and Pollution Research, 28(15), 19530–19532. https://doi.org/10.1007/s11356-021-13201-4

Başakın, E. E., Ekmekcioğlu, Ö., Özger, M., Altınbaş, N., & Şaylan, L. (2021). Estimation of measured evapotranspiration using data-driven methods with limited meteorological variables. Italian Journal of Agrometeorology, 2021(1), 63–80. https://doi.org/10.36253/ijam-1055

Coutinho, E. R., da Silva, R. M., Madeira, J. G. F., Coutinho, P. R. de O. dos S., Boloy, R. A. M., & Delgado, A. R. S. (2018). Application of artificial neural networks (ANNs) in the gap filling of meteorological time series. Revista Brasileira de Meteorologia, 33(2), 317–328. https://doi.org/10.1590/0102-7786332013

Demir, V., Uray, E., Orhan, O., Yavariabdi, A., & Kusetogullari, H. (2021). Trend Analysis of Ground-Water Levels and The Effect of Effective Soil Stress Change: The Case Study of Konya Closed Basin. European Journal of Science and Technology, 24, 515–522. https://doi.org/10.31590/ejosat.916026

Gill, M. K., Asefa, T., Kaheil, Y., & McKee, M. (2007). Effect of missing data on performance of learning algorithms for hydrologic predictions: Implications to an imputation technique. Water Resources Research, 43(7), 1–12. https://doi.org/10.1029/2006WR005298

Hamzah, F. B., Hamzah, F. M., Razali, S. F. M., & Samad, H. (2021). A comparison of multiple imputation methods for recovering missing data in hydrological studies. Civil Engineering Journal (Iran), 7(9), 1608–1619. https://doi.org/10.28991/cej-2021-03091747

Heck, K., Coltman, E., Schneider, J., & Helmig, R. (2020). Influence of Radiation on Evaporation Rates: A Numerical Analysis. Water Resources Research, 56(10). https://doi.org/10.1029/2020WR027332

Hunziker, S., Gubler, S., Calle, J., Moreno, I., Andrade, M., Velarde, F., Ticona, L., Carrasco, G., Castellón, Y., Oria, C., Croci-Maspoli, M., Konzelmann, T., Rohrer, M., & Brönnimann, S. (2017). Identifying, attributing, and overcoming common data quality issues of manned station observations. International Journal of Climatology, 37(11), 4131–4145. https://doi.org/10.1002/joc.5037

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444. https://doi.org/10.1038/nature14539

Nash, E., & Sutcliffe, V. (1970). River flow forecasting Through conceptual models PART I- A Discussion of principles. Journal of Hydrology, 10, 282–290.

Nikroo, L., Kompani-Zare, M., Sepaskhah, A. R., & Fallah Shamsi, S. R. (2010). Groundwater depth and elevation interpolation by kriging methods in Mohr Basin of Fars province in Iran. Environmental Monitoring and Assessment, 166(1–4), 387–407. https://doi.org/10.1007/s10661-009-1010-x

Osman, M. S., Abu-Mahfouz, A. M., & Page, P. R. (2018). A survey on data imputation techniques: Water distribution system as a use case. IEEE Access, 6, 63279-63291. https://doi.org/10.1109/access.2018.2877269

Ratolojanahary, R., Houé Ngouna, R., Medjaher, K., Junca-Bourié, J., Dauriac, F., & Sebilo, M. (2019). Model selection to improve multiple imputation for handling high rate missingness in a water quality dataset. Expert Systems with Applications, 131, 299–307. https://doi.org/10.1016/j.eswa.2019.04.049

Saplioglu, K., & Kucukerdem, T. S. (2018). Estimation of missing streamflow data using anfis models and determination of the number of datasets for anfis: The case of yeŞİlirmak river. Applied Ecology and Environmental Research, 16(3), 3583–3594. https://doi.org/10.15666/aeer/1603_35833594

Schneider, T. (2001). Analysis of incomplete climate data: Estimation of Mean Values and covariance matrices and imputation of Missing values. Journal of Climate, 14(5), 853–871. https://doi.org/10.1175/1520-0442(2001)014<0853:AOICDE>2.0.CO;2

Stisen, S., & Tumbo, M. (2015). Interpolation des données pluviométriques journalières pour la modélisation hydrologique dans des régions à données clairsemées en utilisant des informations issues de données satellitaires. Hydrological Sciences Journal, 60(11), 1911–1926. https://doi.org/10.1080/02626667.2014.992789