7th Intercontinental Geoinformation Days

igd.mersin.edu.tr

Spatial and regression-based missing precipitation data imputation: Western Black Sea region

Seyma Akca*[1], Muhammed Zakir Keskin [2], Ahmad Abu Arra [3] Eyüp Şişman [3]

[1]*Harran University, Geomatic Engineering Department, Sanliurfa, Türkiye*
[2]*Bartın University, Civil Engineering Department, Bartın, Türkiye*
[3]*Yıldız Technical University, Civil Engineering Department, İstanbul, Türkiye*

Keywords
Imputation
IDW
Regression
Black Sea Basin

Abstract

The study of natural phenomena in the environment influences the shaping of human geography. Investigating the occurring physical events is achieved by measuring the magnitudes in nature. These measurements are then structured within certain models, and the resulting outputs are used in engineering applications. However, measurements taken from nature or a system may not provide continuous data due to human and sensor-related errors or inadequacies, resulting in gaps or discontinuities in data acquisition. The success of the method in the missing data completion problem is still an important research topic, as it is influenced by various factors such as the characteristics of the data and the type of missing data. Particularly, the lack of precipitation observation data due to climate change poses serious risks in the planning of water structures. In this study, spatial-based inverse weighted distance (IDW), regression, and statistical methods such as mean and median values are used to fill in and complete missing precipitation data obtained from meteorological stations in the Western Black Sea Region. The results of the study conducted at 10 stations showed that the spatial-based method, IDW, produced more successful results.

1. Introduction

The examination of physical events occurring in nature has an impact on shaping the geography in which humans live. The study of these physical events is achieved through the measurement of quantities in nature. The measurements obtained are structured within specific models and, as a result, the generated outputs are used in engineering applications. However, measurements taken from nature or a system may not be continuous due to human and sensor-related errors or inadequacies, resulting in data gaps or interruptions. When missing data are prevalent in a dataset, they significantly reduce the quality and reliability of that dataset. Effective and accurate planning of water structures and water management require continuous, healthy, and long-term flow data. The success of the missing data completion method depends on various factors, such as the characteristics of the data and the type of missing data, making it an important area of research. Observation data should be complete and extend over many years. In particular, the absence of precipitation observation data related to climate change poses significant risks to the planning of water structures.

Missing data can be categorized into three types: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR). MCAR implies that data is missing randomly, unrelated to any observed or unobserved variables, and the probability of data being missing is independent of all other variables. In contrast, MAR occurs when the missing data is related to observed variables but not to unobserved variables, allowing for flexibility in handling it by including relevant observed variables in the analysis. MNAR is the most challenging type as it suggests that missing data is related to both observed and unobserved variables, requiring complex statistical methods or assumptions. The choice of how to handle missing data depends on the mechanism causing the missingness, significantly impacting the validity of the analysis. There are approaches available for completing

* **Corresponding Author**

*(seymakca@harran.edu.tr) ORCID ID 0000–0002–7888–5078
(mkeskin@bartin.edu.tr) ORCID ID 0009-0005-6724-491X
(ahmad.arra@std.yildiz.edu.tr) ORCID ID0000-0001-8679-1752
(esisman@yildiz.edu.tr) ORCID ID 0000–0003–3696–9967

Cite this study

Akça, Ş., Keskin, M. Z., Arra, A. A., & Şişman, E. (2023). Spatial and regression-based missing precipitation data imputation: Western Black Sea region. Intercontinental Geoinformation Days (IGD), 7, 180-183, Peshawar, Pakistan

missing data, including hot and cold deck, listwise deletion, pairwise deletion, mean imputation, regression imputation, last observation carried forward, stochastic imputation, and multiple imputation (Şeker and Eşmekaya, 2017). When studies in the literature on completing missing data are examined; (Yumus et. al.,2020) For the completion of missing data in HCC survival prediction, they employed methods including (median, mode, mean, decision tree-based regression, and linear regression methods), as well as machine learning-based techniques such as Naive Bayes and decision tree-based classifiers. They achieved the best result with a decision tree regression, with an accuracy of 83%. (Albayrak et. al.,2017) They successfully applied clustering and maximum likelihood methods for completing missing data on a health dataset with a 96.5% success rate. Bakış and Göncü (2015) completed the process of filling in missing data in streamflow measurements in the Zap River basin using correlation-based regression analysis and the Drainage Area Ratio method. (Gümüş and Kavşut, 2013) employed feed-forward back-propagation neural network (FFBPNN), radial based artificial neural network (RBANN), and generalized regression neural network (GRNN) artificial neural network models to predict missing discharge values for the station on the Zamanti River. (Erken and Senyay, 2023) compared the performance of fundamental missing data completion methods such as Listwise Deletion, Last Observation Carried Forward, Mean Imputation, along with machine learning methods like Stochastic Regression, Nearest Neighbor Algorithm, Random Forest Algorithm, and Amelia Algorithm to complete missing data on the Hitters dataset.

In this study, missing data in July precipitation records obtained from the stations located in the Western Black Sea basin, including Bartın (17020), Amasra (17602), Ulus (17615), Akçakoca (17015), Çerkeş (17646), Düzce (17072), Cide (17604), Devrekani (17618), İnebolu (17024), KaradenizEreğli (17611), were aimed to be filled using a spatial interpolation method, Inverse Weighted Distance (IDW), unlike statistical methods (mean, mode, median), regression, and other studies. The relevant results are presented in the continuation of the study.

2. Method

2.1. Study Area

The Western Black Sea basin, one of Türkiye's regions, receives a substantial amount of rainfall, as depicted in Figure 1. Encompassing an expanse of 28,855 km^2, this region extends from east to west and is known for its significant precipitation. The information related to the stations used within the study area is provided in Table 1.

Between 2019 and 2022, randomly selected data points from the precipitation data were omitted, and predictions were made using statistical methods (mean and median), regression, and IDW. These predictions were then verified against actual precipitation data using validation metrics to assess their accuracy.

Table 1. Station details

Station Code	Station Name	Lat. (N)	Lon. (E)	Elev. (m)
17020	Bartın	41.62	32.36	33
17602	Amasra	41.75	32.38	73
17615	Ulus	41.58	32.64	162
17015	Akçakoca	41.09	31.14	10
17646	Çerkeş	40.82	32.88	1126
17072	Düzce	40.84	31.15	146
17604	Cide	41.88	32.95	36
17618	Devrekani	41.60	33.83	1050
17024	İnebolu	41.98	33.76	64
17611	KaradenizEregli	41,26	31,43	19

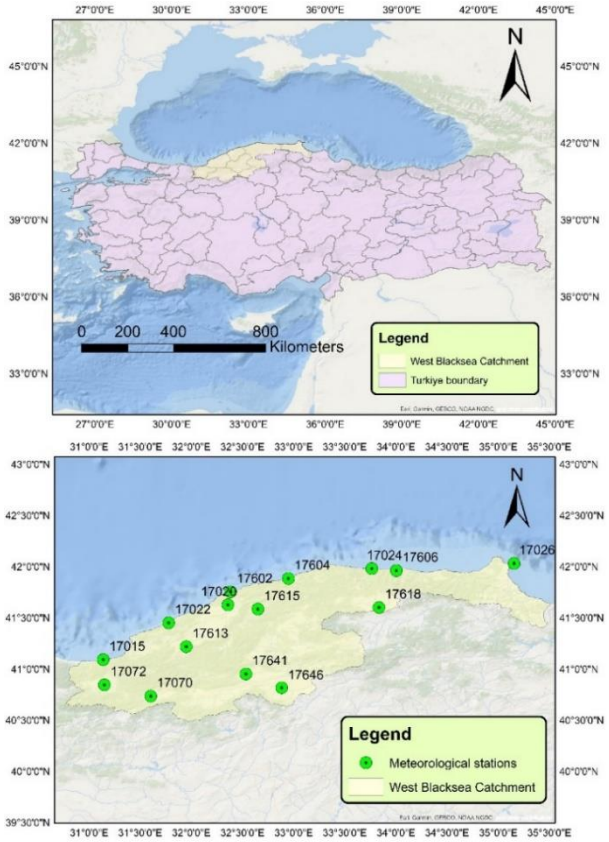

Figure1. The Western Black Sea basin (Türkiye

2.2. Statistical Data Imputation (Mean and Median)

The mean is obtained by summing all the values in the entire series and dividing by the number of data points. The mean of a series is calculated as shown in Equation 1.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad (1)$$

The median (middle) is the value that divides a series or sample data into two equal halves when sorted from lowest to highest.

2.3. Lineer Regression

The direction and degree of the relationship between two numerical variables can be determined through correlation analysis. In the scope of this study, Pearson correlation coefficients were calculated for the annual total rainfall series of the meteorological stations mentioned in the study. The identification of the independent variable for completing missing observations with Regression Analysis was done based

on the Pearson correlation coefficient. For the meteorological station with missing observations (Y), the primary station (x) with the highest observed correlation at the same time is determined, and the missing data in the annual total rainfall series can be calculated using the traditional linear regression model provided in Eq. 3. In the research scope, a linear regression model is established between stations with missing data (Y) and meteorological stations without missing data (x). In the traditional regression model, monthly rainfall values observed simultaneously at the stations are used.

$$Y = \beta_0 + \beta_1 x + \xi \quad (3)$$

Where, Y, x, β_1 and ξ epresent the dependent and independent variables (simultaneously observed annual rainfall heights at the stations), regression coefficients, and the error term, respectively.

2.4. Inverse Distance Weighted Enterpolation

Inverse Distance Weighting (IDW) is a common interpolation technique that estimates values for unsampled points by considering nearby sampled points at varying distances. It calculates cell values based on proximity, where closer points hold more influence while points farther away have diminishing impact. IDW assesses data characteristics like distribution, trend, anisotropy, and clustering, making localized comparisons. This deterministic method is popular for weighted moving average interpolation.

The IDW estimator is represented as Equation 4,

$$z_p = \frac{\sum_{i=1}^{n}(\frac{z_i}{d_i^p})}{\sum_{i=1}^{n}(\frac{1}{d_i^p})} \quad (4)$$

The $z_p$location for predictions depends on neighboring measurements (i = 1, 2, ..., n) with p as the assigned range for each observation at location d. Increasing the exponent reduces the weight of observations farther from the prediction location, making predictions more closely resemble nearby observations (Taylan and Damçayırı, 2016). IDW interpolation is based on the principle that nearby points have a greater influence, performing interpolation from the point of interest by decreasing the weight as it moves away, relying on a weighted average of sample points (İlker et al., 2019).

2.5. Evaluation Metrics

In the validation of the models used, the following criteria have been utilized: proportionality bias (Eq. 5), percentage prediction error (PE%) (Eq. 6), root mean square error (RMSE) (Eq. 7), and mean absolute difference (MAD) (Eq. 8), as described by Ryan and Cryer (2005).

$$\boldsymbol{Bias} = \frac{\sum_{i=1}^{n}(\frac{\hat{x}_i - x_i}{x_i})}{n} \quad (4), \qquad \boldsymbol{PE}_{\%} = \frac{\sum_{i=1}^{n}\left|\frac{\hat{x}_i - x_i}{x_i}\right| \times 100}{n} \quad (5)$$

$$\boldsymbol{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\frac{\hat{x}_i - x_i}{x_i})^2}{n}} \quad (6), \qquad \boldsymbol{MAD} = \frac{\sum_{i=1}^{n}|x_i - \hat{x}_i|}{n} \quad (7)$$

The rainfall values obtained from the stations for the month of July between 2019 and 2022 are represented by (x_i) while $\hat{x}_i$ represents the estimated rainfall values. In this context, n represents the number of missing data.

3. Results

The results of the IDW interpolation used to predict missing data between 2019 and 2022 are presented in Figure 2. The results of the accuracy assessment metrics for missing data predicted using statistical, regression, and IDW methods are presented in Table 2.

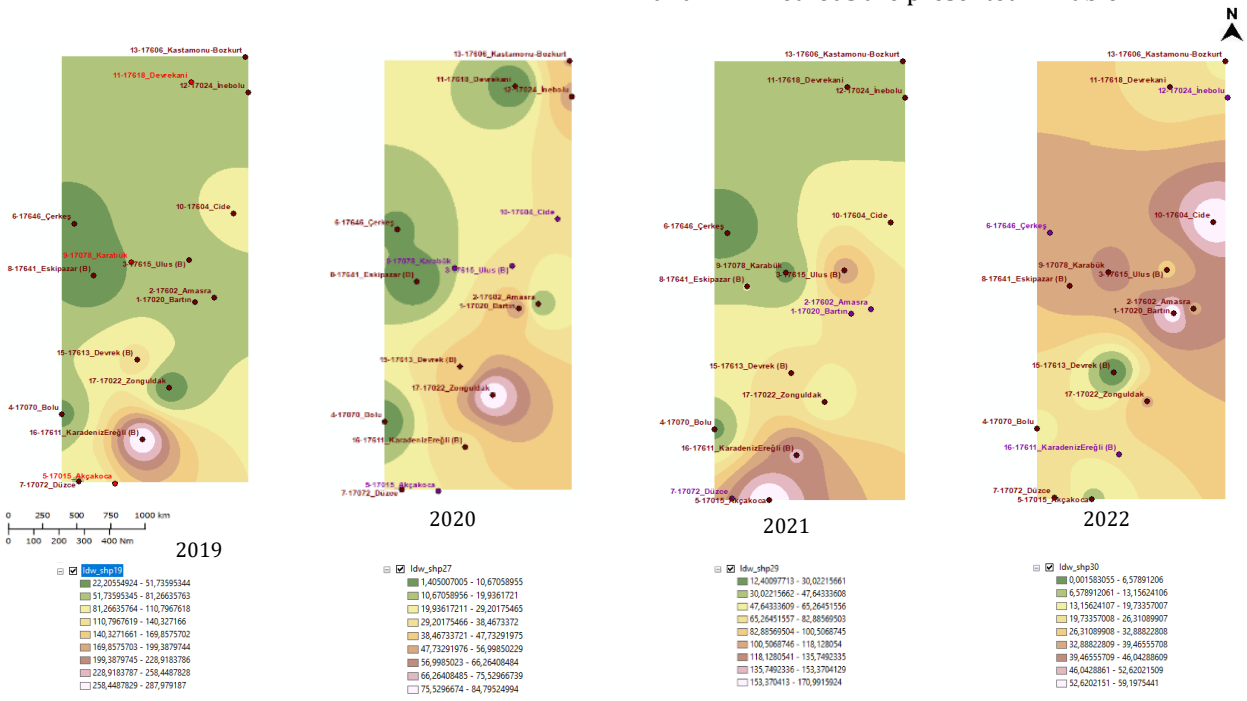

Figure 2. IDW Interpolation results

Tablo 2. Evaluation Metrics Results

Methods	Metrics.	Bartın	Amasra	Ulus	Akça koca	Çerkeş	Düzce	Cide	Devrekani	İnebolu	KaradenizE
Average	Bias	-0.40	-0.26	0.84	-0.36	-0.70	-0.73	**0.76**	-0.36	4.40	5.38
	PE%	39.81	25.91	**84.03**	36.34	70.30	73.39	131.16	36.22	439.80	537.80
	RMSE	0.40	0.26	**0.84**	0.42	0.70	0.73	3.13	0.36	4.40	5.38
	MAD	37.9	19.9	**12.1**	57.05	61	115.2	38.8	28.8	28.8	62.5
Median	Bias	-0.58	-0.50	2.33	-0.46	-0.75	-0.82	2.55	-0.57	2.94	4.92
	PE%	57.77	50	233.33	45.78	75.50	81.84	255.26	56.69	293.88	492.13
	RMSE	0.58	0.50	2.33	0.49	0.75	0.82	2.55	0.57	2.94	4.92
	MAD	55	38.4	33.6	57.05	61	117.2	38.8	28.8	28.8	62.5
Regression	Bias	-0.53	0.35	3.06	**-0.35**	-0.62	-0.78	0.80	**-0.17**	4.38	3.76
	PE%	53.15	34.77	305.56	**35.06**	61.63	77.72	**80.26**	**17.13**	437.76	375.59
	RMSE	0.53	0.35	3.06	**0.35**	0.62	0.78	**0.80**	**0.17**	4.37	3.76
	MAD	50.6	26.7	44	36.65	49.8	111.3	**12.2**	**8.7**	42.9	47.7
IDW	Bias	**-0.26**	**0.01**	**1.28**	-0.44	**-0.56**	**-0.14**	0.99	0.35	**1.39**	**0.74**
	PE%	**20.62**	**0.57**	127.85	43.55	**56.05**	**12.00**	99.54	34.76	**138.88**	**74.25**
	RMSE	**0.21**	**0.01**	1.28	0.62	**0.56**	**0.12**	0.99	0.35	**1.39**	**0.74**
	MAD	**19.63**	**0.44**	18.41	**36**	**45.29**	**17.19**	15.13	17.66	**13.61**	**9.43**

4. Discussion

When examining the results presented in Table 2, it can be observed that the best Bias value was obtained using the IDW method for the Bartın, Amasra, Ulus, Çerkeş, Düzce, Cide, İnebolu, and Karadeniz Ereğli station. The best Bias value for the Akçakoca was achieved with the regression method, and for Devrekani, it was also the regression method.

The lowest PE% value was obtained with the IDW method for the Bartın, Amasra, Çerkeş, Düzce, İnebolu, and Karadeniz Ereğli station. For Ulus station, the lowest PE% value was achieved using the average method, and for Akçakoca, Devrekani, and Cide station, it was the regression method.

In terms of RMSE, the IDW method produced the lowest value for the Bartın, Amasra, Çerkeş Düzce, İnebolu, and Karadeniz Ereğli station. The lowest RMSE value for the Ulus station was obtained using the average method, while the Akçakoca and Cide station achieved the lowest RMSE with the regression method, and the Devrekani station also had the lowest RMSE with the regression method.

IDW had the lowest MAD values for several stations, including Bartın, Amasra, Çerkeş, Düzce, İnebolu, and Karadeniz Ereğli, while the Ulus station achieved the lowest MAD using the average method. Akçakoca and Devrekani stations had their lowest MAD values with the IDW method, and Cide station with the regression method. Although the Ulus station performed well with the average method, the IDW method ranked second. Stations like Cide and İnebolu saw better results with the regression model, where IDW was the second-best method. Differences in station results can be attributed to factors such as elevation, rainfall distribution, and inter-station correlations. In general, considering all the results, it has been observed that the IDW method generally produces better results compared to other methods in completing missing data.

5. Conclusion

In this study, for a dataset created by randomly removing rainfall data from 10 different stations in the Western Black Sea Basin between 2019 and 2022, missing data were estimated using statistical methods (mean and median), regression, and IDW techniques. When measuring the accuracy of the models created according to four different evaluation metrics, IDW produced the best results in 6 out of 10 stations. For the other stations, the second-best performing method is again IDW.

References

Albayrak, M., Turhan, K., & Kurt, B. (2017, October). A missing data imputation approach using clustering and maximum likelihood estimation. In 2017 Medical Technologies National Congress (TIPTEKNO), 1-4

Bakış, R., & Göncü, S. (2015). Akarsu debi ölçümlerinde eksik verilerin tamamlanması: Zap Suyu Havzası örneği.

Erken, Ş., & Şenyay, L. (2023). Makine Öğrenmesi İle Eksik Veri Tamamlama Yöntemlerinin Sınıflandırma Performansına Etkileri. Kayseri Üniversitesi Sosyal Bilimler Dergisi, 5(1), 51-71.

Gümüş, V., & Kavşut, M. E. (2013). Zamanti Nehri-Ergenusağı Đstasyonu Eksik Aylık Akım Verilerinin Tahmini. Gazi University Journal of Science Part C: Design and Technology, 1(2), 81-91

İlker, A., Terzi, Ö., & Şener, E. (2019). Yağışın Alansal Dağılımının Haritalandırılmasında Enterpolasyon Yöntemlerinin Karşılaştırılması: Akdeniz Bölgesi Örneği. Teknik Dergi, 30(3), 9213-9219.

Ryan, B. F. & Cryer, J. (2005). Minitab Handbook. Fifth Edition,Regression and Correlation, 313-349, Belmont,California, p.505.

Şeker, Ş. E., & Eşmekaya, E. (2017). Eksik verilerin tamamlanması (imputation). YBS Ansiklopedi, 4(3), 10-17.

Taylan, E. D., & Damçayiri, D. (2016). Isparta bölgesi yağış değerlerinin IDW ve Kriging enterpolasyon yöntemleri ile tahmini. Teknik Dergi, 27(3), 7551-7559.

Yumuş, M., Apaydın, M., Değirmenci, A., & Karal, Ö. (2020, October). Missing data imputation using machine learning based methods to improve HCC survival prediction. 28th Signal Processing and Communications Applications Conference (SIU), 1-4.