VOL. 15, NO. 14, JULY 2020 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

ALGORITHM FOR THE RECOVERY OF MISSING DATA IN THE BOGOTÁ RIVER

Wilson R. López S., César A. Perdomo Ch. and Julián R. Camargo L.
Faculty of Engineering, Universidad Distrital Francisco José de Caldas, Bogotá D.C., Colombia
E-Mail: cperdomo@correo.udistrital.edu.co

ABSTRACT

There are different methods for the estimation of missing data; among the most used is substitution by arithmetic mean and interpolation, which have been used in HFA (Hydrological Frequency Analysis). Interpolation consists of being able to estimate a function f(x) that describes the behavior of already known data and thus estimate an arbitrary x that is within the limits of known values. In this article, a study of different interpolation methods will be carried out to estimate missing parameters related to water quality in the Bogotá River.

Keywords: data mining, HFA, water quality, standardized method, makima.

1. INTRODUCTION

The use of water resources for different industrial and agricultural activities generates the environmental degradation of this resource. Nowadays there are different models that estimate the quality of the water, allowing to take a more detailed control of the state of this resource, however, the taking and handling of the data is one of the biggest limitations of the studies, the samplings are made in very long periods due to the high costs that the companies have to carry out the sampling.

The information obtained from natural phenomena is analyzed with different computational techniques of machine learning and data mining [1] [2]. The extraction of data has been used in applications such as the search for missing parameters and the estimation of parameters [3]. The mathematical models that describe the behavior of physical phenomena use a set of parameters and algorithms [4] [5]. Data extraction can determine the correlation between variables and estimate the missing values of a set of parameters. [6] The use of data mining in the analysis of natural phenomena plays an important role [7], model to investigate earthquakes [8] and to estimate water quality using satellite images [1], [9] and [10]. The estimation of the parameters in drinking water sources has led to several studies to generalize models for many water sources. Some authors create new adaptive predictive models with the advantages of [11].

The missing data in the hydrological data bases are one of the most recurrent problems when studies of the water resources are carried out, for this reason it is fundamental to investigate the different models that allow the estimation of missing data in the hydrological data bases and the possible development of an algorithm that improves the estimation of conventional methods. There are different methods for estimating missing data. The most used are: substitution by arithmetic mean and interpolation, which have been used in HFA (Hydrological Frequency Analysis). Interpolation consists of being able to estimate a function f (x) that describes the behavior of already known data and thus estimate an arbitrary x that is within the limits of known values.

In practice, hydrological studies suffer from missing data (DM) caused, for example, by equipment failures, errors in measurements, budget cuts and natural hazards [12]. In many cases, you do not have many data necessary to perform an adequate planning and management of these data, they are called missing data (MD).

When working with a large amount of data, the data frames that have MD are eliminated, this method does not apply to data related to water quality, since these bases do not have many data. The missing data can be more than 40% of the existing database, this is why this research seeks to propose an algorithm through interpolation that makes an estimate of the missing data that keeps the behavior of the data through weather.

2. METHODOLOGY

The database has 14 water quality variables (BO5, N NO2, SST, P TOTAL, Precipitation, Flow, BOD5 Tributary, N NO2 Tributary, SST Tributary, P Total Tributary, BOD5 Effluent, N NO2 Effluent, SST Effluent, P Total Effluent), each of these variables was taken at 23 stations along the Bogotá River, (Villapinzón, Chocontá, Gachancipá, Suesca, Sesquilé, Tocancipá, Nemocón, Zipaquirá, La Calera, Cajicá, Chía, Cota, Sopó, Bogota, San Antonio, Soacha, Anapoima, Apulo, Viotá, Tena, El Colegio, Tocaima, Girardot), the sampling was carried out from 2008 to 2015 quarterly, for a total of 9016 data.

2.1 Processing of Data

The normalization of the data was made taking into account Table-1 taken from [13].

The database has 10304 rows of data, of which 3478 are missing data (MD) equivalent to 33% of the database. For this study it is assumed that a historical data is for example the data from 2008 to 2015 of the BOD variable of the villa finch station represent a historical data; In total, the database has 322 historical data, because the present study relates a historical data with another to obtain the missing data, eliminates the historical data that does not have any data over time, therefore we have one database of 293 historical data to work.

Table-1. Standardized methods applied to environmental quality (taken from [13]).

INTERVAL	STANDARDIZED METHOD		CONDITIONS
$7\ \frac{mg}{L} \leq DBO_5 \leq\ 135\ \frac{mg}{L}$	$CA_i = \frac{\text{Máximo } DBO_i - DBO_i}{\text{Máxima } DBO_i - \text{Mínimo } DBO_i}$	$CA_i = 1.0157 - 0.0068 * DBO$ $r^2 = 0.9982$	Si $DBO_5 \geq 135\frac{mg}{L}, CA_i = 0.1$
$10\ \frac{mg}{L} \leq N_{NO2} \leq\ 55\ \frac{mg}{L}$	$CA_i = \frac{\text{Máximo } N_i - N_i}{\text{Máxima } N_i - \text{Mínimo } N_i}$	$CA_i = 1.2 - 0.02 * N$ $r^2 = 1.0000$	Si $N_{NO2} \geq 55\frac{mg}{L}, CA_i = 0.1$
$50\ \frac{mg}{L} \leq SST \leq\ 500\ \frac{mg}{L}$	$CA_i = \frac{\text{Máximo } SST - SST}{\text{Máximo } SST - \text{Mínimo } SST}$	$CA_i = 1.1 - 0.002 * SST$ $r^2 = 1.0000$	Si $SST \geq 500\frac{mg}{L}, CA_i = 0.1$
$0.5\ \frac{mg}{L} \leq P_T \leq\ 17\ \frac{mg}{L}$	$CA_i = \frac{\text{Máximo } P_i - P_i}{\text{Máxima } P_i - \text{Mínimo } P_i}$	$CA_i = 0.9927 - 0.0537 * P_{TOTAL}$ $r^2 = 0.9959$	Si $P_T \geq 17\frac{mg}{L}, CA_i = 0.1$
$5\ mm \leq P \leq\ 1000\ mm$	$CA_i = \frac{P - \text{Mínimo } P}{\text{Máximo } P - \text{Mínimo } P}$		Si $P \leq 5\ mm, CA_i = 0.1$

In Figure-1, a graph of all the normalized variables is presented where the blank spaces represent the missing values. It is observed that the measurements between stations vary, however over the years the same behavior is established, this allows that different methods can be proposed to estimate the missing data.

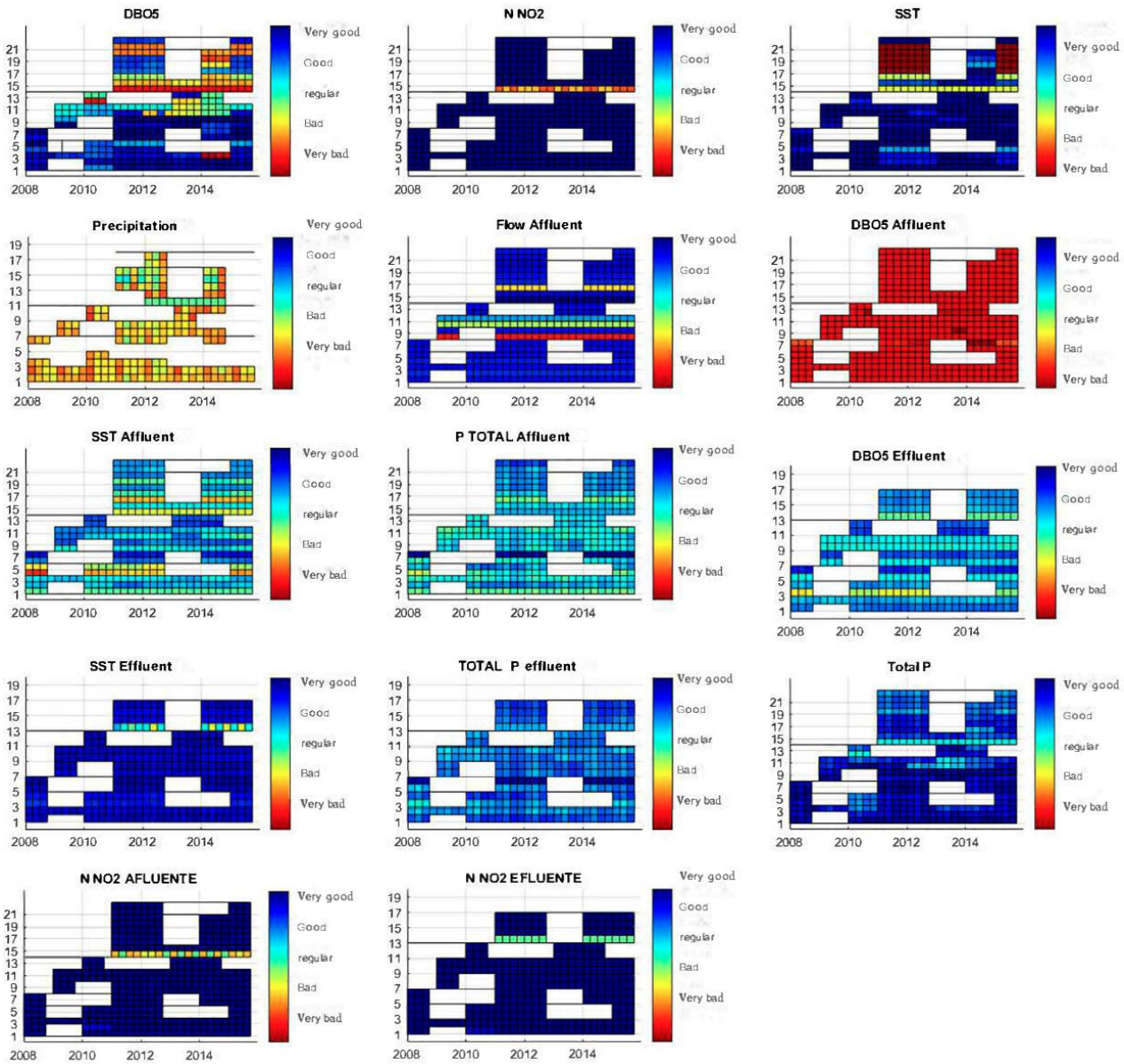

Figure-1. Standardized variables with missing data.

The average consists of obtaining the arithmetic mean of the existing data and filling in the missing data with the obtained value. The interpolation corresponds to mathematical models that estimate the missing data taking into account the existing data, for this case study will be used eight types of interpolation.

To evaluate the accuracy of the methods used in data recovery, performance is evaluated through a Jackknife re-sampling procedure. This consists of taking complete data series and eliminating some data to consider them as missing. The criteria used to evaluate the yields are the relative mean square error [14] and the mean relative bias or true error [15].

We start by making a comparison between the interpolation method and the media method, for this we take a parameter and a sampling station, in this case we will work with the Chocontá station and the BOD5 parameter, because it does not present values constants over time and presents missing data in the year 2009, in Figure-2 the real data represented by points are presented, in red the result of the estimation by means of interpolation and in blue calculating the average.

Given that in hydrological studies one of the most important aspects is to observe the variables over time, it can be observed in *Figure-2* that the most suitable method for data recovery is interpolation by trimesters since this is the most important resembles the morphology of original grace. The interpolation method is the one that most closely approximates the morphology of the parameter with missing data, taking into account this result; the different interpolation models are studied to estimate the missing data of the case study.

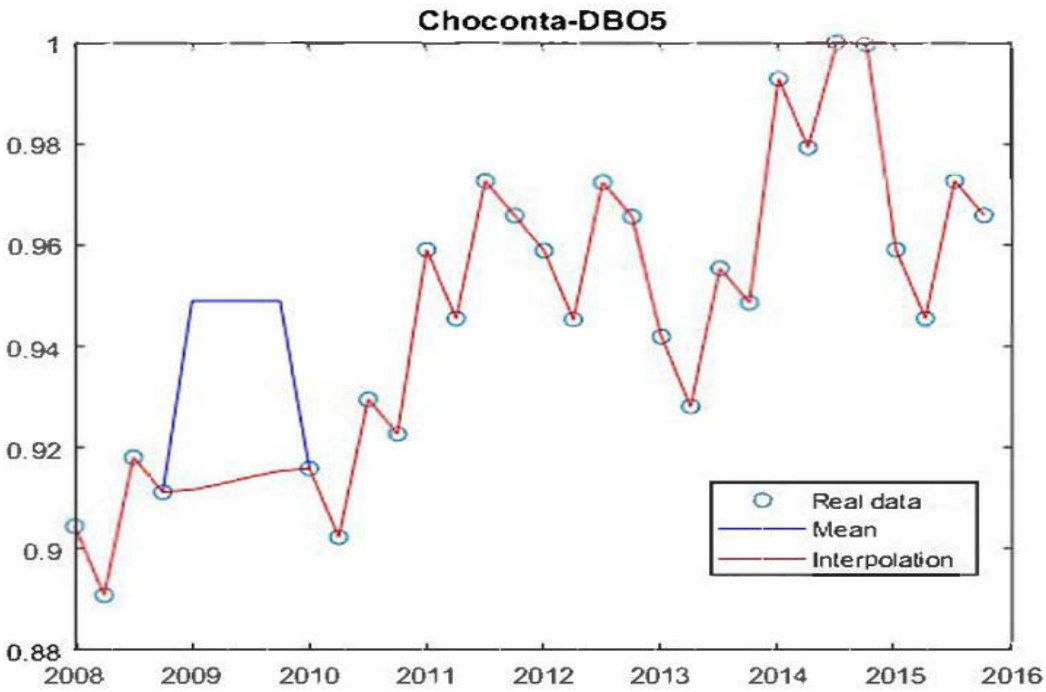

Figure-2. Estimation of missing data by means and interpolation.

Three historical data are taken that do not present missing data, the 2009 data are eliminated and three methods are applied to make the estimates of said data. The procedures to be followed are the following: calculate the arithmetic average, use the interpolation applied to the whole time series, and the same type of interpolation, but applied independently to each quarter. The results are shown in Figure-3.

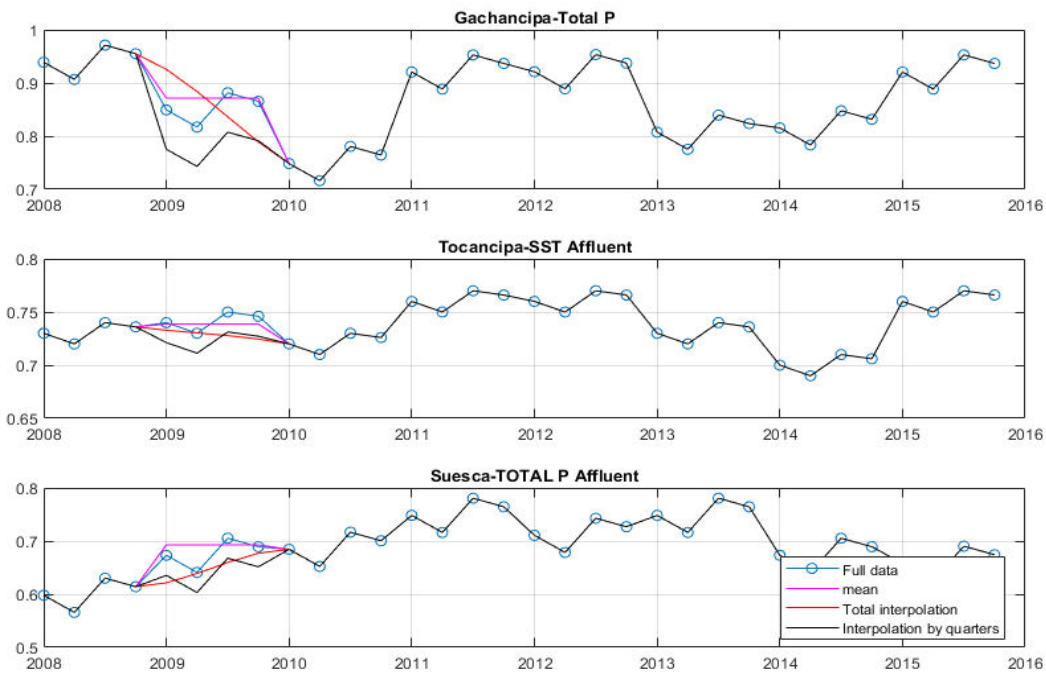

Figure-3. Comparison of data filling by mean, interpolation and interpolation by quarters.

To have an effectiveness of the different methods used in the present study, we will work with the historical data that are complete in this case 79 data. The methodology to check the error of the methods is the following:

- It is removed from a section of the data to be analyzed.
- The methods of recovering missing data are applied.
- The true error (abs (true value-estimated value)) of each historical data is calculated.

- The error obtained in each of the data is averaged.
- A comparison is made confirming which method obtained the smallest error.
- The best methods to recover dependent data and the location of missing data are selected.

For this study, different interpolation methods were applied every quarter. Table-2lists the types of interpolation used.

Table-2. Types of interpolations.

Method	Description
linear	Linear interpolation. The interpolated value at a query point is based on linear interpolation of the values at neighboring grid points in each respective dimension.
nearest	Nearest neighbor interpolation. The interpolated value at a query point is the value at the nearest sample grid point.
next	Next neighbor interpolation. The interpolated value at a query point is the value at the next sample grid point.
previous	Previous neighbor interpolation. The interpolated value at a query point is the value at the previous sample grid point.
pchip	Shape preserving piecewise cubic interpolation. The interpolated value at a query point is based on a shape preserving piecewise cubic interpolation of the values at neighboring grid points.
v5cubic	Cubic convolution used in MATLAB® 5.
makima	The interpolated value at a query point is based on a piecewise function of polynomials with degree at most three.
spline	Spline interpolation using not-a-knot end conditions. The interpolated value at a query point is based on a cubic interpolation of the values at neighboring grid points in each respective dimension.

3. RESULTS

For all the case studies, the 79 historical data were taken and the data that was in the same position for each historical data was eliminated, the interpolation was applied and the true error of each historical data was extracted, finally for each case studied it was they averaged the true errors and took the one that had the least error.

3.1 Data Recovery Frame Start

In this case, the interpolations makima, Pchip and Spline are evaluated, which allow interpolation when there is no data at the beginning of the frame. For these point 15 experiments were performed in the first, the first data was eliminated and the true average error of the experiment was calculated, for the second the first two values were eliminated in this way until the first 15 values of the plot in Table-3 were eliminated the results of these experiments are shown:

Table-3. Average true error for data lost from positions 1 to 15.

Missing data position	**1**	**1 to 2**	**1 to 3**	**1 to 4**	**1 to 5**	**1 to 6**	**1 to7**	**1 to 8**
Type of interpolation								
makima	0,0951	0,0966	0,0972	0,0987	0,0981	0,0991	0,0985	0,0969
pchip	0,139	0,1403	0,1411	0,1426	0,1217	0,115	0,1079	0,1013
spline	0,306	0,3117	0,3132	0,3146	0,3712	0,3883	0,3958	0,3972
Better interpolation	makima	makima	makima	makima	makima	makima	makima	makima
Missing data position	1 to 9	1 to 10	1 to 11	1 to 12	1 to 13	1 to 14	1 to 15	
Type of interpolation								
makima	0,1199	0,1291	0,1344	0,1388	0,2514	0,3083	0,3553	
pchip	0,122	0,1297	0,1378	0,1414	0,2509	0,306	0,3596	
spline	0,6942	0,8251	0,9344	1,0235	1,2161	1,3119	1,396	
Better interpolation	makima	makima	makima	makima	pchip	Pchip	makima	

3.2 End of Frame Data Recovery

For this case, the same type of interpolations applies as the previous case makima, pchip and spline. For this experiment section, the data in positions 19 to 32 will be eliminated and the average true error is calculated, the same procedure is performed with positions 20 to 32 until having only position 32 as missing data the results of this experimentation they are presented in Table-4:

Table-4. Average true error for lost data from positions 19 to 32.

Missing data position	**19 to 32**	**20 to 32**	**21 to 32**	**22 to 32**	**23 to 32**	**24 to 32**	**25 to 32**
Type of interpolation							
makima	0,0951	0,0966	0,0972	0,0987	0,0981	0,0991	0,0985
pchip	0,139	0,1403	0,1411	0,1426	0,1217	0,115	0,1079
spline	0,306	0,3117	0,3132	0,3146	0,3712	0,3883	0,3958
Better interpolation	makima	makima	makima	makima	makima	makima	makima
Missing data position	26 to 32	27 to 32	28 to 32	29 to 32	30 to 32	31 to 32	32
Type of interpolation							
makima	0,0969	0,1199	0,1291	0,1344	0,1388	0,2514	0,3083
pchip	0,1013	0,122	0,1297	0,1378	0,1414	0,2509	0,306
spline	0,3972	0,6942	0,8251	0,9344	1,0235	1,2161	1,3119
Better interpolation	makima	makima	makima	makima	makima	makima	makima

3.3 Data Recovery Intermediate Frame

In this case, the data is eliminated from position 5 to position 20 following the same pattern as the two cases presented in the previous paragraphs, in Table-5, the result of this training set is shown:

Table-5. True average error for lost data from positions 5 to 20.

Missing data position	**5 to 5**	**5 to 6**	**5 to 7**	**5 to 8**	**5 to 9**	**5 to 10**	**5 to 11**	**5 to 12**
Type of interpolation								
pchip	0,078	0,067	0,062	0,082	0,076	0,07	0,066	0,164
linear	0,038	0,035	0,035	0,035	0,035	0,036	0,036	0,036
nearest	0,04	0,036	0,035	0,035	0,039	0,042	0,044	0,044
previous	0,049	0,046	0,047	0,047	0,048	0,048	0,049	0,049
v5cubic	0,072	0,07	0,069	0,069	0,066	0,059	0,058	0,058
makima	0,071	0,062	0,058	0,081	0,074	0,069	0,065	0,156
Better interpolation	nearest	nearest	next	next	previous	previous	previous	previous
Missing data position	5 to 13	5 to 14	5 to 15	5 to 16	5 to 17	5 to 18	5 to 19	5 to 20
Type of interpolation								
pchip	0,153	0,142	0,133	0,252	0,239	0,226	0,214	0,078
linear	0,035	0,033	0,032	0,032	0,034	0,034	0,035	0,038
nearest	0,04	0,035	0,032	0,032	0,036	0,039	0,041	0,04
previous	0,049	0,049	0,05	0,05	0,05	0,049	0,05	0,049
v5cubic	0,087	0,109	0,124	0,124	0,126	0,126	0,126	0,072
makima	0,146	0,137	0,129	0,101	0,097	0,093	0,09	0,071
Better interpolation	previous	previous	previous	previous	previous	previous	previous	previous

3.4 Comparison of Experiments

In this section, we will compare the result of the three experiments and in this way determine if it is better to apply the makima method when there are more than 4 lost data at the beginning or end of the plot. Table-6 compares the makima method for missing data at the beginning and the previous interpolation for intermediate values:

Table-6. Comparison between makima vs previous interpolation for missing data at the beginning of the frame.

Lost data position	1 to 6	1 to 7	1 to 8	1 to 9	1 to 10	1 to 11	1 to 12	1 to 13	1 to 14	1 to 15
Makima start data	0,0991	0,0985	0,0969	0,1199	0,1291	0,1344	0,1388	0,2514	0,3083	0,3553
Lost data position	5 to 6	5 to 7	5 to 8	5 to 9	5 to 10	5 to 11	5 to 12	5 to 13	5 to 14	5 to 15
previous	0,049	0,046	0,047	0,047	0,048	0,048	0,049	0,049	0,049	0,049
Makima-previus difference	0,05	0,053	0,05	0,073	0,081	0,086	0,09	0,202	0,259	0,306

The Table-7 shows a comparison between the makima interpolation and previous ones for the estimation of missing data at the end of the data frame.

Table-7. Final makina comparison and previous interpolation for data recovery at the end of the frame.

Lost data position	19-32	20-32
Makima final data	0,0951	0,0966
Lost data position	5 a 19	5 a 20
Previous	0,05	0,049
Makima-previus difference	0,045	0,048

3.5 Algorithm Proposal to Complete Data

An algorithm was developed in Matlab that allows the estimation of data with better results than those obtained with the interpolations previously analyzed individually.

To evaluate the effectiveness of the algorithm, the records that do not present missing data are used, the last five data of the plot are eliminated for all the study cases, the first data of the plot are also eliminated and some intermediate data the results are presented in Table-8.

Table-8. Comparison between conventional interpolation methods and proposed recovery algorithm**.**

Missing data position	**1 a 1**	**1 a 2**	**1 a 3**	**1 a 4**	**1 a 5**	**1 a 6**	**1 a 7**	**1 a 8**
Type of interpolation								
pchip	0,094	0,100	0,104	0,107	0,104	0,100	0,097	0,096
makima	0,075	0,078	0,080	0,081	0,082	0,083	0,083	0,083
spline	0,281	0,285	0,285	0,280	0,292	0,310	0,316	0,310
algorithm	0,064	0,065	0,066	0,066	0,063	0,063	0,061	0,060
Better interpolation	alg	alg	alg	alg	alg	alg	alg	alg
Missing data position	1 a 9	1 a 10	1 a 11	1 a 12	1 a 13	1 a 14	1 a 15	
Type of interpolation								
pchip	0,102	0,113	0,118	0,121	0,172	0,216	0,256	
makima	0,091	0,098	0,104	0,104	0,112	0,120	0,126	
spline	0,375	0,440	0,486	0,439	0,395	0,344	0,305	
algorithm	0,061	0,061	0,062	0,061	0,056	0,053	0,049	
Better interpolation	alg	alg	alg	alg	alg	alg	alg	

4. DISCUSSIONS

The results obtained in Tables3 to 7 show that depending on the location of the missing data in the plot there are different types of interpolations that give better results than others. Among the results we can highlight the following:

- Table-3 shows that the makima type interpolation is the best to rescue the first data of a temporary plot.
- For the set of experiments in Table-4 it is observed that the best interpolation for the data recovery at the end of the plot is makima.
- In the cases of Table-4 and Table-5, it was obtained that the makima type interpolation is the best for data recovery, but in the last case intermediate frame the previous one is that it is closest to the real data, to be able to apply a previous type interpolation is necessary that there be at least 4 data the beginning of the plot and 4 at the end of the plot.
- For the case of study of Table-5 it is observed that the best interpolation is the previous one for most of the cases.
- Table-6 shows that the estimation of data for positions higher than the one removed by the previus method is more effective than the makima.
- Table-7 shows that for this case it is better to use the previous interpolation for the recovery of data near the end of the frame.
- Finally, in Table-8, for all the case studies, the developed algorithm presents better results than the other interpolation methods.

The Figure-4shows a graphical comparison between the makima type interpolation and the developed algorithm.

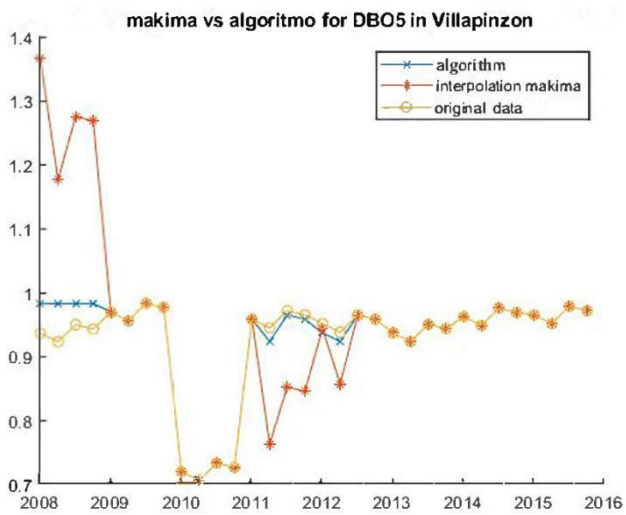

Figure-4. Comparison between the makima type interpolation and the algorithm developed for BOD5 at the Villa Pinzon station.

5. CONCLUSIONS

The estimation of missing data with the interpolation method presents a better result when performing the importation on a quarterly basis and not in the complete set of the time series.

The developed algorithm allows to solve errors that present the different types of interpolation as the estimation of data greater than one.

The methods of data recovery through interpolation allow to observe the morphology of the different time series as opposed to the estimation of data by the arithmetic mean method

The development of an algorithm for the estimation of missing data was achieved, with better results than those obtained with conventional methods such as the case of the arithmetic mean or the different interpolation methods.

ACKNOWLEDGMENTS

The authors would like to thank to the Universidad Distrital Francisco José de Caldas and the LASER research group that supported the development and testing of the project.

REFERENCES

[1] C. Doña, N. Chang, V. Caselles, J.M. Sánchez, A. Camacho, J. Delegido and B. W. Vannah.15 March 2015. Integrated satellite data fusion and mining for monitoring lake water quality status of the Albufera de Valencia in Spain. Journal of Environmental Management. 151: 416-426.

[2] M. K. Sohrabi and S. Akbari. July 2016. A comprehensive study on the effects of using data mining techniques to predict tie strength. Computers in Human Behavior, 60: 534-541,http://dx.doi.org/10.1016/j.chb.2016.02.092.

[3] G. Ssali and T. Marwala. 2008. Computational intelligence and decision trees for missing data estimation. IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)

[4] S. Chapra. 1987. Surface Water Quality Modelling", McGraw Hill, Brown, L.C. and Barnwell, T.O., The Enhanced Stream Water Quality Models QUAL2E and QUAL2E-UNCAS, EPA/600/3-87- 007, U.S. Environmental Protection Agency, Athens, p. 189.

[5] S. Chapra, G. Pelletier and H. Tao. 2008. QUAL 2K: A Modeling Framework for Simulating River and Stream Water Quality. Documentation.

[6] A. Toor Umair, A. Zubair and K. Dong-Jin. July 2014. Estimation of N2O emission during wastewater nitrification with activated sludge: Effect of ammonium and nitrite concentration by regression analysis. Journal of Industrial and Engineering Chemistry, 20(4): 2574-2579, http://dx.doi.org/10.1016/j.jiec.2013.10.042.

[7] A. Erturk, M. Gurel, A. Ekdal, C. Tavsan, A. Ugurluoglu, D. Zafer Seker, A. Tanik and I. Ozturk. July 2010. Water quality assessment and meta model development in Melen watershed – Turkey. Journal of Environmental Management. 91(7): 1526-1545.

[8] S. L. Nimmagadda and H. Dreher. 2007. Ontology based data warehouse modeling and mining of earthquake data: prediction analysis along Eurasian-Australian continental plates. 2007 5th IEEE International Conference on Industrial Informatics, Vienna. pp. 597-602.

[9] M. Bonansea, M. C. Rodriguez, L. Pinotti and S. Ferrero. March 2015. Using multi-temporal Landsat imagery and linear mixed models for assessing water quality parameters in Río Tercero reservoir (Argentina)", Remote Sensing of Environment. 158: 28-41.

[10] E. T. Harvey, S. Kratzer and P. Philipson. March 2015. Satellite-based water quality monitoring for improved spatial and temporal retrieval of chlorophyll-a in coastal waters. Remote Sensing of Environment. 158: 417-430.

[11] X. Wang, J. Zhang and V. Babovic. July2106. Improving real-time forecasting of water quality indicators with combination of process-based models and data assimilation technique. Ecological Indicators. 66: 428-439,http://dx.doi.org/10.1016/j.ecolind.2016.02.016.

[12] A. M. Kalteh and P. Hjorth. 2009. Imputation of missing values in a precipitation- runoff process database. Hydrology Research. 40(4): 420-432.

[13] J. P. Rodríguez. Desarrollo de un modelo de planificación ambiental para la calidad de los recursos hídricos superficiales considerando su variabilidad climática estacional mediante implementación computacional. Universidad Distrital Francisco José de Caldas, Doctorado en Ingeniería

[14] F. Chebana and T. B. M. J. Ouarda. Depth and homogeneity in regional flood frequency analysis. Water Resources Research. 44(11).

[15] C. Beaulieu, S. Gharbi, T. Ouarda, C. Charron and M. A. B. Aissia. 2012. Improved model of deep-draft ship squat in shallow waterways using stepwise regression trees. Journal of Waterway, Port, Coastal, and Ocean Engineering. 138(2): 115-121.