A case study: Exploratory analysis and proposal for the problem of quality in the meteorological observations of the north of Chile

Un caso de estudio: análisis exploratorio y propuesta para el problema de calidad en las observaciones meteorológicas del norte de Chile

Francisco García Barrera[1] https://orcid.org/0000-0002-2411-9183
David Contreras Aguilar[1*] https://orcid.org/0000-0002-6906-485X
Alonso Inostrosa-Psijas[2] https://orcid.org/0000-0002-2922-0870
Sergio Cerda Lozano[1] https://orcid.org/0009-0008-6684-697X
Xenia Andaur Estica[1] https://orcid.org/0009-0000-1017-3552

Recibido 26 de enero de 2024, aceptado 30 de diciembre de 2024
Received: January 26, 2024 Accepted: December 30, 2024

ABSTRACT

Data quality problems in meteorological variables are a situation that the scientific community constantly faces, mainly because these quality problems materialize as missing data within the time series, which prevents compliance with the established requirements to analyze climate change in a given geographical area. Based on this problem, this article presents an exploratory analysis of the main meteorological variables (Temperature and Precipitation) observed by the meteorological stations distributed in Northern Chile to assess the data quality they present. Data imputation methods are also proposed to address this problem by completing the missing data. In particular, the experiments developed based on the phases of the CRISP-DM methodology are presented in an adapted way considering five different imputation methods of which the residual error closest to zero and the highest positive correction is sought. In the results, CLP, IDC, and RN stand out as the best techniques, which allows us to conclude that these methods can be recommended and proposed as an alternative solution according to the meteorological variable.

Keywords: Time series, meteorological variables, data imputation methods.

RESUMEN

Los problemas de calidad de los datos en variables meteorológicas son una situación que enfrenta constantemente la comunidad científica, principalmente porque estos problemas de calidad se materializan en datos faltantes dentro de las series temporales, lo que impide cumplir con los requisitos establecidos para analizar el cambio climático en un área geográfica determinada. A partir de esta problemática, este artículo presenta un análisis exploratorio de las principales variables meteorológicas (Temperatura y Precipitación) observadas por las estaciones meteorológicas distribuidas en el Norte de Chile para evaluar la calidad de los datos que presentan. Además, se proponen métodos de imputación de datos para abordar este problema completando los datos faltantes. En particular, los experimentos desarrollados con base en las fases de la metodología CRISP-DM se presentan de forma adaptada considerando cinco métodos de imputación diferentes de los cuales se busca el error residual más cercano a cero y la mayor corrección positiva. En los resultados se destacan CLP, IDC y RN como las mejores técnicas,

[1] Universidad Arturo Prat. Facultad de Ingeniería y Arquitectura. Iquique, Chile.
Email: francisco.garcia.barrera@unap.cl; david.contreras@unap.cl; sergio.cerda.lozano@unap.cl; xandaur@unap.cl
[2] Universidad de Valparaíso. Escuela de Ingeniería Informática. Valparaíso, Chile. Email: alonso.inostrosa@uv.cl
* Autor de correspondencia: david.contreras@unap.cl

lo que permite concluir que estos métodos pueden ser recomendados y propuestos como una solución alternativa según la variable meteorológica.

Palabras clave: *Series de tiempo, variables meteorológicas, métodos de imputación de datos.*

INTRODUCTION

The World Meteorological Organization (WMO) proposes the Extreme Climate Indicators (ECI) as a tool to know if a particular geographical area is being affected by climate change. The calculation of the ECI requires using Time Series (TS) data corresponding to temperatures and precipitation - recorded by Meteorological Stations (MS) - which must be homogeneous and have 30 years of information according to the WMO recommendation. The correct calculation of the ECI contributes to having a clear image of the current climate change state, allowing decision-making officials to define the best strategies for avoiding the effects of climate change, which worsen year after year, allowing us to restructure our economies and societies to achieve objectives associated with a better lifestyle.

MS is located in different parts of the world and presents data quality problems. However, in the particular case of the MS located in northern Chile, the situation is worse as it is known that there are also problems with missing data. Consequently, low-quality TS data is recorded, which prevents the ICE from being able to make reliable calculations, as it has to comply with WMO recommendations. Due to the importance of the topic, researchers have developed several methods that can be used to solve these problems and help fill the gaps in the TS.

This work first performs an exploratory analysis of the recorded data of meteorological variables in the MS located in northern Chile. Then, different methods for the imputation of data in TS are evaluated using data analysis techniques to identify a method that can improve the quality of the TS regarding its completeness and representativeness in the behavior of the meteorological variables, evaluated and validated by a correlation coefficient between the data created by the imputation methods and the actual TS data with a minimum residual error.

In the following sections of this article, we provide a synthesis of related works, followed by a description of the materials and methods - detailing the geographical area and study data for the work - together with the methodology and data imputation methods used. Additionally, the configuration of the experiments, the exploratory analysis, the data quality analysis, and the analysis of results are also presented. Finally, deliver the study's conclusions and possible lines for future work.

RELATED WORK

The need to address the problem of missing data that may be present in data sources drives the scientific community's constant search, study, evaluation, and comparison of data imputation methods; this is reflected in the works of Aljuaid and Sasi [1], Khan and Hoque [2] and Jadhav, D. Pramod, and K. Ramanathan [3], where the authors evaluate different methods on test data sets.

Considering that most meteorological data come from MS and measurement equipment, according to Papale [4], the most common problems that can cause missing data are power outages, damage to measurement instruments, incorrect calibrations, lack of maintenance, and human actions such as theft and vandalism. Additionally, missing data can occur in the data acquisition phase due to poor filtering of data quality.

Generally, researchers evaluate methods on a test data set or over actual data of some particular context. The latter is because in medical, biological, and climate sciences, among others, the missing data problem is latent Osman, Abu-Mahfouz, and Page [5]. Examples of this are the works conducted by [5]-[8] of techniques applied to data from different fields of knowledge, especially in hydrometeorological sciences [5], [9]-[12].

Studies have also been conducted in Ecuador [13] and Chile [14] to evaluate data imputation methods. In particular, the Carrera's [13] research work addresses the problem of collecting and publishing meteorological data in Ecuador by the National

Institute of Meteorology and Hydrology. This work calculates the missing data using the linear regression and normal ratio method. The results obtained from the data imputation were validated using the runs test and double mass curves. Specifically, the results show that the best methodology for data imputation in TS for Ecuador's Andean and coastal regions is the simple linear regression method.

Furthermore, Pizarro [14]compares the linear correlation methods, linear regression, distance ratios, neighborhood averages, average ratios, correlation of neighboring MS, and multiple regressions. The methods are analyzed through the coefficient of determination (R^2), standard error of estimation (SEE), Bland and Altman concordance test, and analysis of variance, with which it is determined which method presents the best fit for the region. It was concluded that the method that obtained the best results was the multiple regression method. Concerning the simple linear regression method, this is one of the most used in Chile due to its easy application; however, despite being the most recommended in the literature, it did not obtain the best results in practice for the MS used in this study. In addition, it is recommended that this type of study be extended to arid and semi-arid zones.

Other works seek to reconstruct missing data in monitoring records with techniques such as K-Nearest Neighbor, multiple imputations using chained equations, linear interpolation, spline interpolation, and iterative singular value decomposition [15]. Missing data are also present in high-resolution meteorological records, so techniques such as ARIMA and multiple regression modeling are proposed for this problem [16]. Outliers are often a problem in time series analysis, so using meteorological data from Colombia, the authors propose a methodology that removes outliers and imputes the missing data generated by the process by applying the discrete Fourier transform [17]. In India, a case study was developed with 33 years of meteorological data from Kozhikode City, in which univariate and multivariate models were applied to address the missing data problem [18].

Finally, it is necessary to indicate that the present research differs from previous works because of the many methods considered, which differ from those applied in other studies. However, on the other hand, it is a work of comparative analysis of the imputation methods, evaluating different meteorological variables such as mean, minimum, maximum temperature, and precipitation of northern Chile, whose geomorphology makes these variables have different behaviors according to the location from where they are measured and recorded, in addition to the desert characteristics of extreme aridity.

MATERIALS AND METHODS

Study area and time series of meteorological variables
The study area for the present work corresponds to the Norte Grande of Chile (Arica and Parinacota Region, Tarapacá Region, and Antofagasta Region), in which five of the noteworthy primary forms of Chilean relief can be distinguished: Coastal Plain, Coastal Range, Intermediate Depression, Altiplano and Andes Mountains.

The Norte Grande of Chile is generally characterized by an extremely arid desert climate with the following specific types of climate: Cloudy coastal desert climate, inland desert climate, high-altitude marginal desert climate, and high-altitude steppe climate.

Under this context and the problem introduced at the beginning of the article, it becomes relevant to analyze the TS of the main meteorological variables that characterize the climate within climate change and the ECI. For this reason, the present investigation considers the data of the MS that record precipitation (PP), minimum temperature (MINT), mean temperature (MEDT), and maximum temperature (MAXT).

A TS corresponds to a collection of observations of a variable collected sequentially over time. These observations are usually collected in equispaced time instants when they must be captured continuously, or the accumulated values are recorded over defined time intervals for observation [19].

Methodology and imputation methods
Regarding the methodology used to carry out the research work, the data mining methodology called CRISP-DM is used, and its phases are Understanding the problem, Understanding the data, Data Preparation, Modeling, Evaluation, and Deployment.

The modeling phase is where the data imputation experiment is performed, which in its most basic definition indicates using the information contained in the sample to assign a value to those variables that have records with a missing value, either because there is a lack of information or because it is detected that some of the values collected do not correspond to the expected behavior. Five deterministic imputation methods for filling TS in meteorological variables stand out among the principal methods. They are described in the following sections.

Weighted linear combination (WLC)

This method consists of substituting the missing data from the statistically close TS data, known as neighboring MS so that each incomplete data is obtained by the weighted linear combination of the data from the TS to be used for the completion. These data weight in the WLC proportional to the Pearson correlation coefficient with the incomplete TS as long as it is higher than an acceptable critical value.

On the one hand, the value of the correlation coefficient to accept an MS as a neighbor depends on the type of variable to be completed and the climatic affinity and scope of the spatial correlation between MS. For example, the value usually chosen for the PP variable (a variable that presents a low correlation between neighboring MS) is between: $0{,}7 < r < 0{,}8$ [20], On the other hand, for variables such as temperature, insolation or atmospheric pressure, values can be chosen should be de $r > 0{,}8$ [21]. This method corresponds to equation 1, where, for a given month t, the incomplete data $x(t)$ is defined as:

$$x(t)=\frac{r_1\cdot x_1(t)+r_2\cdot x_2(t)+r_3\cdot x_3(t)+\ldots r_n\cdot x_n(t)}{r_1+r_2+r_s+\ldots+r_n} \quad (1)$$

Where r_i is the Pearson's correlation coefficient between the TS i and the incomplete TS, $x_i(t)$ is the value at the time t of the TS. The number of TS used for completion is arbitrary in principle. The recommended number is around five and never less than two [22].

Distance ratios (DR)

This method is based on the relationship of a meteorological variable with the distance in kilometers between the MS to be filled with two aligned confidence MS. These must have a linear spatial arrangement, so it is only reliable for flat areas, not suitable for mountainous areas [23]; thus, and using equation (2), the missing data can be estimated [24].

$$PX=PA+a\times\frac{(PB-PA)}{(a+b)} \quad (2)$$

Where PX represents the position of an MS with lack of information; *PA y PBPA* is the presence of an MS with complete information; *a y ba* represents the distance on a plane from the MS X. The MS with missing data must lie between two MS that present complete statistics.

Ratios of average (RA)

This method estimates the missing data as the average recorded in n different reliable MS. The missing data is determined with equation (3) [24].

$$PX=\frac{\overline{PX}}{N}\times\left[\frac{PA}{\overline{\overline{PA}}}+\frac{PB}{\overline{\overline{PB}}}+\ldots+\frac{PN}{\overline{\overline{PN}}}\right] \quad (3)$$

Where $\overline{PX},\overline{PA},\overline{PB},\ldots,\overline{PN}$ are the expected annual average values recorded for such MS; *PA, PB, …, PN* correspond to the values recorded in each of the N MS during the period where data is missing in the X MS. What is achieved by applying this method is to estimate the missing MS based on the relationships between the PP of a period under study and the average PP [14].

Normal ratio (NR)

This method consists of calculating the incompleteness, $x(t)$, of a TS from the data of the TS of neighboring and contemporary MS, thus presenting a high degree of correlation with the TS to be completed employing the expression presented in equation (4) [21].

$$x(t)=\frac{1}{n}\left[\frac{\overline{x}}{\overline{x}_1}x_1(t)+\frac{\overline{x}}{\overline{x}_2}x_2(t)+\ldots+\frac{\overline{x}}{\overline{x}_n}x_n(t)\right] \quad (4)$$

Where $\overline{x},\overline{x}_1,\overline{x}_2,\ldots,\overline{x}_n$ are the means of the variables in question for the incomplete and the neighboring TS, respectively, while $x_1(t)$, $x_2(t)$,…, $x_n(t)$ are the data corresponding to the neighboring TS. This method uses the variability recorded in other MS and with the proportional ratio between them. It is recommended that at least three auxiliary MS be used. Using three MS softens the influence that an error in one of them could have.

Inverse squared distance (ISD)
This method is similar to RA but with the addition of distance weighting. The information for applying of this method does not go beyond having the exact location of the measurement points. In addition, its estimation will depend significantly on the separation of the MS in the respective area. Therefore, the missing data is determined with equation (5) [25].

$$P_x = \frac{\sum_{i=1}^{n} P_i \left(\frac{1}{D_i^{\,2}} \right)}{\sum_{i=1}^{n} \left(\frac{1}{D_i^{\,2}} \right)} \tag{5}$$

Where P_i represents the data recorded in the MS i, n is the total number of MS, and D_i is the distance between the site to be estimated and the MS.

RESULTS

This section discusses the results of the analysis of the variables evaluated. First, the experimental setup is explained, then the experimental activity is described, and finally, the results obtained from the experiments are analyzed.

Experimental setup - data
The database used is composed of 2,665,137 records. These belong to 638 different MS in northern Chile. Of the identified MS, 586 were filtered beforehand, leaving them as the basis for the subsequent phases of processing and selection of MS in the experiment. These MS belong to 14 organizations, as shown in Table 1, both governmental and private, distributed in the study area; the information regarding these MS dates from 1935 to 2014. However, these records are not homogeneous in that date range. For example, MS M_230_AA, located in the Antofagasta region, whose owner is NOAA, stands out as it has the oldest data. Likewise, MS M_002_W, located in the Tarapacá region in the commune of Pozo Almonte (owned by the company SQM), presents data in more recent time ranges. The MS's average amount of information in the database is 16.3 years.

As mentioned above, the meteorological variables PP, MINT, MEDT, and MAXT were analyzed for this research. In addition, tests were also carried out to find the best range of dates where there was a considerable amount of data to perform the necessary experiments, where for each meteorological variable analyzed, the best range of dates in terms of the number of records was found to be as shown in Table 2.

For the implementation of the methods of this work, a Microsoft SQL Server database was used to consolidate the information from the different MS, whose records were examined and processed through the Knime Analytics Platform software. Additionally, an automated process was developed in the latter where the records were filtered, and those belonging to the MS to be used in the experiments were selected. On the other hand, the

Table 1. MS owners used for the research.

Acronym Owner	Owner Name	Number of stations
CMCC	Cerro Colorado Mining Company.	2
CMDIC	Doña Inés de Collahuasi Mining Company.	6
CMDIC/DGA	Compañía Minera Doña Inés de Collahuasi/Dirección General de Aguas (General Water Directorate).	5
DGA	General Directorate of Water.	117
DMC	Meteorological Directorate of Chile.	4
NOAA	National Oceanic and Atmospheric Administration National Office of Oceanic and Atmospheric Administration.	294
SENAMHI	National Service of Meteorology and Hydrology of Peru.	151
SQM	Chemical and Mining Society of Chile.	2
UNAP	Arturo Prat University	5

Table 2. Date ranges of meteorological variables where experiments were carried out.

Meteorological variable	Date range of study
Precipitation	1982 - 2008
Minimum Temperature	1996 - 2012
Mean Temperature	1985 - 2011
Maximum Temperature	1993 - 2009

TS imputation methods were implemented in the Python programming language for later use in the automated process mentioned above.

Experimental activity

In the experiments, the daily data of the meteorological variables PP, MINT, MEDT, and MAXT were used to perform 15 iterations in the selected TS, where in each iteration, an amount of empty data is generated to be filled in by the imputation methods, and thus compared with the real data. The following is an excerpt of how the experiments performed for MAXT were carried out to exemplify this.

When the first iteration is executed, a temporary void is generated by eliminating one record; in the second iteration, two records are eliminated, and so on. As the iterations progress, an amount of empty data equivalent to the iteration is generated. Figure 1a shows the TS with five missing data corresponding to the fifth iteration.

Subsequently, the missing data generated using the imputation methods are filled in. For example, in Figure 1b, the results generated with the WLC method are displayed versus the original TS values for the same iteration in Figure 1a.

Once the data imputation has been performed, Pearson, Spearman, and Kendall correlation coefficients are calculated between the data generated by the imputation method and the real TS data to validate that the data created by the imputation methods maintain the same behavior as the original TS.

In addition, the Anderson-Darling, Lilliefors, and Kolmogorov-Smirnov normality tests with 95% confidence are also performed to analyze the corresponding correlation coefficient according to the results of the normality tests. If the TS has a normal distribution, the correlation coefficient to be analyzed is Pearson's, whereas if the data distribution is non-normal, the correlation coefficients to be analyzed are Spearman and Kendall.

Regardless, the residual errors are calculated for each iteration and their descriptive statistics (Table 3),

Table 3. Descriptive statistics of residual errors between WLC-generated data and actual MAXT data.

Descriptive Statistics	Value
Mean	–1.74259
Standard Deviation	0.737686
Minimum	–2.95822
Q1	–2.01653
Median	–1.51377
Q3	–1.51283
Maximum	–0.71159
IQR	0.503694
Kurtosis	2.245484
Asymmetry Coefficient	–0.34723

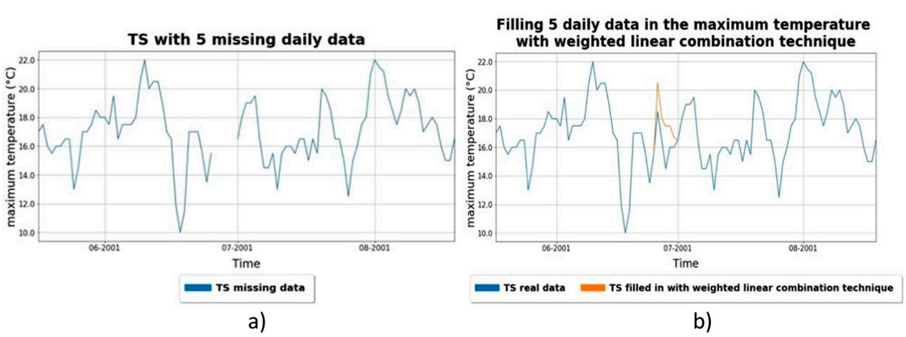


Figure 1. Weighted linear combination for MAXT.

allowing us to observe how far apart the values generated by the imputation method and the real TS data are.

The residual error analysis calculates the average residual errors in the different experiments performed. As the amount of eliminated data increases, so does the amount of residual errors calculated in each iteration. Thus, the mean of these is calculated to find a representative residual error for each experiment. Once each experiment has its respective residual error mean, descriptive statistics are elaborated, considering the residual error means of all iterations to analyze the behavior of the imputation method.

The criteria for evaluating and validating the imputation methods include the residual error, the positive and significant correlation coefficients obtained between the data created by the imputation methods, and the actual data from the ST. The best results are represented by the imputation methods with the residual error closest to zero and the highest positive and significant correlation coefficient considering a value equal to or greater than 0.7. The entire process described above, which generates the experiments' results, evaluates and validates them, is repeated in each iteration and for each imputation method.

As the methods use neighboring EMs in the imputation of data, they must be selected under criteria that allow them to be validated as reliable sources of information. For this, an analysis of correlation and geographic locations between EM is carried out, accepting those that present a correlation coefficient equal to or greater than 0.7 and have coherence in the geographical characteristics where they are located. Given the above, the group of EMs selected for each meteorological variable is different since it depends on the correlation analysis, geographic location, and if they have records of the variable of interest, creating four groups of EM where each one is used to a specific variable.

Exploratory analysis

In the first part of the exploratory analysis, the quality of the database records, in general, is evaluated to determine the amount of missing data that the MS presents. In Table 4, an excerpt of the analysis is shown.

Also, the exploratory statistical analysis of the records for each meteorological variable is considered within the same process. An example of this is summarized in Table 5, Table 6, Table 7, and Table 8, corresponding to the meteorological variables PP, MINT, MEDT, and MAXT, respectively. These tables show the missing values in the different variables by presenting values of -99.9 in the descriptive statistics.

Table 4. Extract of quality analysis of records from different MS.

ID MS	Amount of data	Number of years	Missing data
M_001_A_XV	18479	52	636
M_001_AH	6939	31	4384
M_001_AJ	7549	20	31
M_002_B_II	12053	32	0
M_003_A_II	14723	44	1444
M_004_A_II	6433	18	1
M_006_A_XV	7426	21	91
M_009_A_XV	4153	12	169
M_010_A_XV	14319	43	1660
M_011_A_II	15878	44	436
M_011_B	11964	32	0
M_015_A	12905	38	975
M_016_A	18583	52	440

The Anderson-Darling, Lilliefors, and Kolmogorov-Smirnov normality tests were applied with 95% confidence in different TS located in different areas of northern Chile, which means that its significance (a) is 0.05 to determine the type of data distribution of meteorological variables. All p-values obtained in the normality tests are less than a, suggesting that the data do not follow a normal distribution. Given this, it can be inferred that all MS data have this type of distribution regardless of whether they are located in different regions of northern Chile or different geographic areas of Chile. We used the MS detailed in Figure 2 and Figure 3.

Data quality analysis

The following is an analysis of some of the most representative MS for each meteorological variable analyzed to check the quality and consistency of the data.

Meteorological variable precipitation

For the meteorological variable PP, we used the range of dates between 1982 and 2008, where two

Table 5. Extract of exploratory analysis on records of different MS for the meteorological variable PP.

ID_MS	MEAN	DESV_STANDARD	MIN	Q1	MEDIAN	Q3	MAX	IQR
M_001_A_XV	0.0364	0.91990744	–99.9	0	0	0	29.5	0
M_001_AH	0.2719	1.62428492	0	0	0	0	33.2	0
M_001_AJ	–1.641	12.9192102	–99.9	0	0	0	27	0
M_002_B_II	–7.627	26.5428246	–99.9	0	0	0	14.1	0
M_003_A_II	–2.225	14.7444882	–99.9	0	0	0	3	0
M_004_A_II	–4.905	21.5928208	–99.9	0	0	0	5	0
M_006_A_XV	0.2013	6.92367604	–99.9	0	0	0	37.2	0
M_009_A_XV	–0.357	8.83113347	–99.9	0	0	0	24.4	0
M_010_A_XV	–0.046	11.50167	–99.9	0	0	0	92	0
M_011_A_II	–0.077	6.74795248	–99.9	0	0	0	52	0
M_011_B	0.0025	0.10195542	0	0	0	0	9.1	0
M_015_A	0.3529	1.90958466	0	0	0	0	41.1	0
M_016_A	0.2795	3.80104363	–99.9	0	0	0	42	0

Table 6. Extract of exploratory analysis on records from different MS for the meteorological variable MINT.

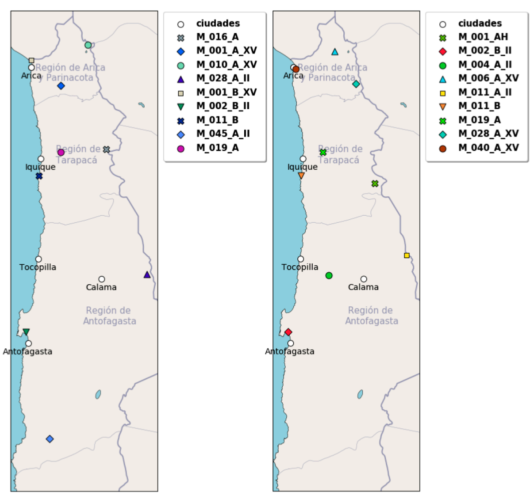
ID_MS	MEAN	DESV_STANDARD	MIN	Q1	MEDIAN	Q3	MAX	IQR
M_018_A	–79.19	42.05	–99.9	–99.90	–99.90	–99.90	14.00	0.00
M_018_A_II	–99.41	7.23	–99.9	–99.90	–99.90	–99.90	6.50	0.00
M_018_A_XV	–99.90	0.00	–99.9	–99.90	–99.90	–99.90	–99.90	0.00
M_018_AA	–36.29	52.35	–99.9	–99.90	4.00	7.00	17.00	106.90
M_019_A	–7.22	33.46	–99.9	1.00	3.00	6.80	22.00	5.80
M_019_A_II	–99.90	0.00	–99.9	–99.90	–99.90	–99.90	–99.90	0.00
M_019_A_XV	–99.90	0.00	–99.9	–99.90	–99.90	–99.90	–99.90	0.00
M_019_AA	–99.90	0.00	–99.9	–99.90	–99.90	–99.90	–99.90	0.00
M_020_A	–99.90	0.00	–99.9	–99.90	–99.90	–99.90	–99.90	0.00
M_020_A_II	–99.84	2.53	–99.9	–99.90	–99.90	–99.90	8.10	0.00
M_020_A_XV	–15.17	31.96	–99.9	–7.00	–4.00	–1.00	16.20	6.00
M_020_AA	–99.90	NULL	–99.9	–99.90	–99.90	–99.90	–99.90	0.00
M_021_A	–99.90	0.00	–99.9	–99.90	–99.90	–99.90	–99.90	0.00

Table 7. Extract of exploratory analysis on records from different MS for the meteorological variable MEDT.

ID_MS	MEAN	DESV_STANDARD	MIN	Q1	MEDIAN	Q3	MAX	IQR
M_042_AA	26.09	5.61	–99.90	24.20	26.88	28.68	38.00	4.48
M_043_A_II	–99.90	0.00	–99.90	–99.90	–99.90	–99.90	–99.90	0.00
M_043_AA	20.82	22.87	–99.90	23.00	25.50	27.33	38.00	4.33
M_044_A_II	2.47	24.90	–99.90	3.90	8.20	11.90	19.90	8.00
M_045_A_II	–2.36	40.78	–99.90	10.80	14.40	16.40	25.00	5.60
M_045_AA	24.88	3.96	6.03	23.70	25.75	27.45	32.63	3.75
M_046_A_II	–99.90	0.00	–99.90	–99.90	–99.90	–99.90	–99.90	0.00
M_046_AA	24.06	6.88	–99.90	22.48	25.00	27.00	34.50	4.53
M_047_A_II	–99.90	0.00	–99.90	–99.90	–99.90	–99.90	–99.90	0.00
M_047_AA	24.64	3.97	8.90	22.73	25.45	27.18	35.75	4.45
M_048_A_II	–99.90	0.00	–99.90	–99.90	–99.90	–99.90	–99.90	0.00
M_048_AA	19.00	0.00	19.00	19.00	19.00	19.00	19.00	0.00
M_049_AA	24.90	3.57	6.18	23.68	25.64	27.09	32.32	3.41

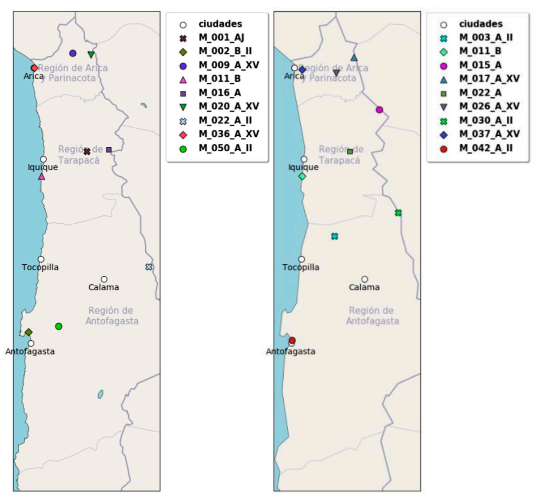
Table 8. Extract of exploratory analysis on records from different MS for the meteorological variable MAXT.

ID_MS	MEAN	DESV_STANDARD	MIN	Q1	MEDIAN	Q3	MAX	IQR
M_061_AA	–97.99	14.750	–99.9	–99.9	–99.9	–99.9	18	0.00
M_062_AA	0.62	37.856	–99.9	12.5	14.5	16.1	37.7	3.60
M_063_AA	–4.67	43.293	–99.9	12.2	14.2	15.7	37.5	3.50
M_064_AA	–99.90	0.000	–99.9	–99.9	–99.9	–99.9	–99.9	0.00
M_065_AA	–99.90	0.000	–99.9	–99.9	–99.9	–99.9	–99.9	0.00
M_066_AA	–99.90	0.000	–99.9	–99.9	–99.9	–99.9	–99.9	0.00
M_067_AA	15.54	41.616	–99.9	27	31	32.8	36.3	5.80
M_068_AA	–4.65	57.775	–99.9	–99.9	29.6	32	38.8	131.90
M_069_AA	–85.86	40.278	–99.9	–99.9	–99.9	–99.9	37.8	0.00
M_070_AA	3.64	52.719	–99.9	23	29.8	32.2	38.6	9.20
M_071_AA	–10.02	60.928	–99.9	–99.9	29.8	33	36	132.90
M_072_AA	–0.77	56.539	–99.9	20.7	30.8	33	37.8	12.30
M_073_AA	7.67	49.664	–99.9	24.4	30.5	32.6	38.4	8.20



a) b)

Figure 2. Map of MS used MINT (a) and MEDT (b) for normality tests.



a) b)

Figure 3. Map of MS used MAXT a) and PP b) for normality tests.

of the best MS in terms of number of records are the following:

- M_001_B_XV: This MS is located in the region of Arica and Parinacota, in the commune of Arica. Its owner is the DMC, and for the selected date range, it has 9862 records (Figure 4a).
- M_011_B: This MS is located in the Tarapacá region, in the commune of Iquique. Its owner is the DMC, and for the selected date range, it has 9862 records (Figure 4b).

Furthermore, among the MS with a large number of missing records for the same meteorological variable are the following:

- M_001_AH: This MS is located in the Tarapacá region, in the commune of Pica. Its owner is CMDIC/DGA, and for the selected date range, it has 5205 records (Figure 4c).
- M_030_A_II: This MS is located in the Antofagasta region, in the Ollagüe commune. Its owner is the DGA, and for the selected date range, it has 5268 records (Figure 4d).

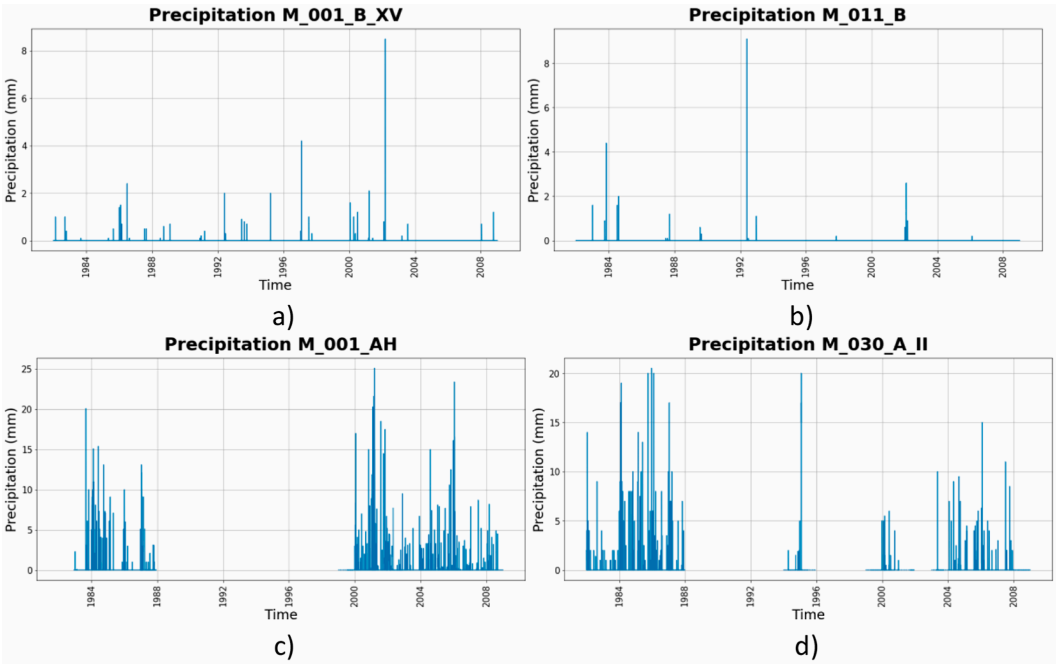


Figure 4. Quality of information in meteorological variable of PP.

Weather variable minimum temperature

For the meteorological variable MINT, the range of dates between 1996 - 2012 was used, where two of the best MS in terms of number of records are the following:

- M_002_B_II: This MS is located in the Antofagasta region, in the Antofagasta commune. Its owner is the DMC, and for the selected date range, it has 7302 records (Figure 5a).
- M_001_B_II: This MS is located in the Antofagasta region. Its owner is the DMC, and for the selected date range, it has 7283 records (Figure 5b).

Among the MS with a large number of missing records and for the same meteorological are the following:

- M_028_A_II: This MS is located in the Antofagasta region, in the Calama municipality. Its owner is the DGA, and for the selected date range, it has 3836 records (Figure 5c).
- M_022_A_II: This MS is located in the Antofagasta region, in the Calama municipality. Its owner is the DGA, and for the selected date range, it has 4765 records (Figure 5d).

Meteorological variable mean temperature

For the meteorological variable MEDT, the range of dates between the years 1985 - 2011 was used, where two of the best MS in terms of number of records are the following:

- M_001_B_XV: This MS is located in the Arica and Parinacota region, in the Arica commune. Its owner is the DMC, and for the selected date range, it has 9855 records (Figure 6a).
- M_002_B_II: This MS is located in the Antofagasta region, in the Antofagasta commune. Its owner is the DMC, and for the selected date range, it has 9855 records (Figure 6b).

Additionally, among the MSs with a large number of missing records and for the same meteorological are the following:

- M_030_A_II: This MS is located in the Antofagasta region, in the Ollagüe commune. Its owner is the DGA, and for the selected date range, it has 3871 records (Figure 6c).
- M_028_A_XV: This MS is located in the Arica and Parinacota region, in the Putre commune. Its owner is the DGA and for the selected date range it has 4860 records (Figure 6d).

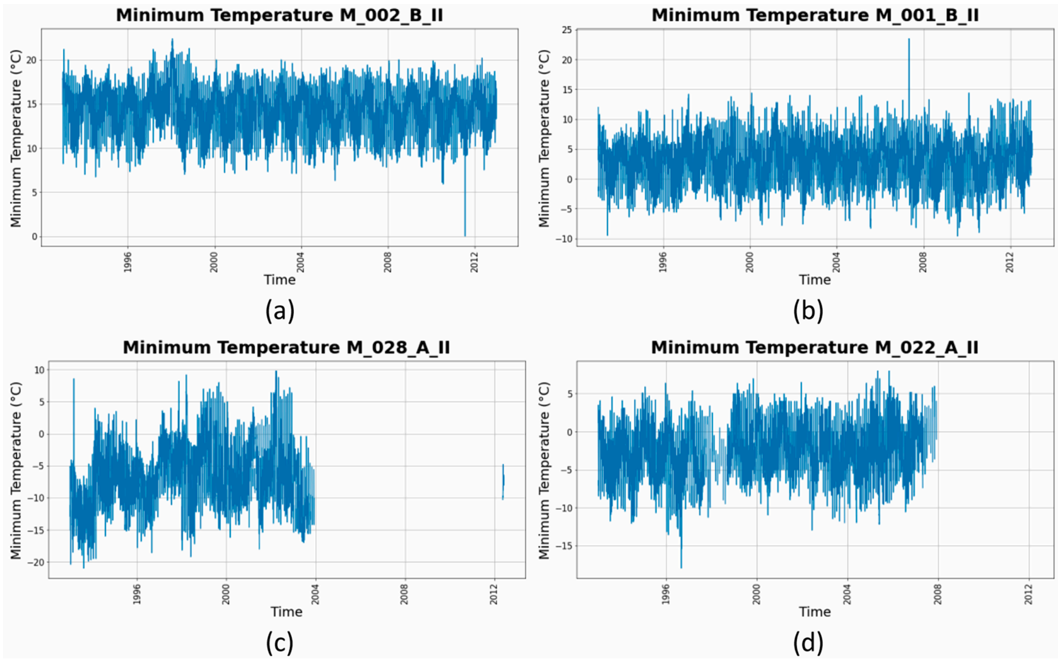


Figure 5. Quality of information on meteorological variable MINT.

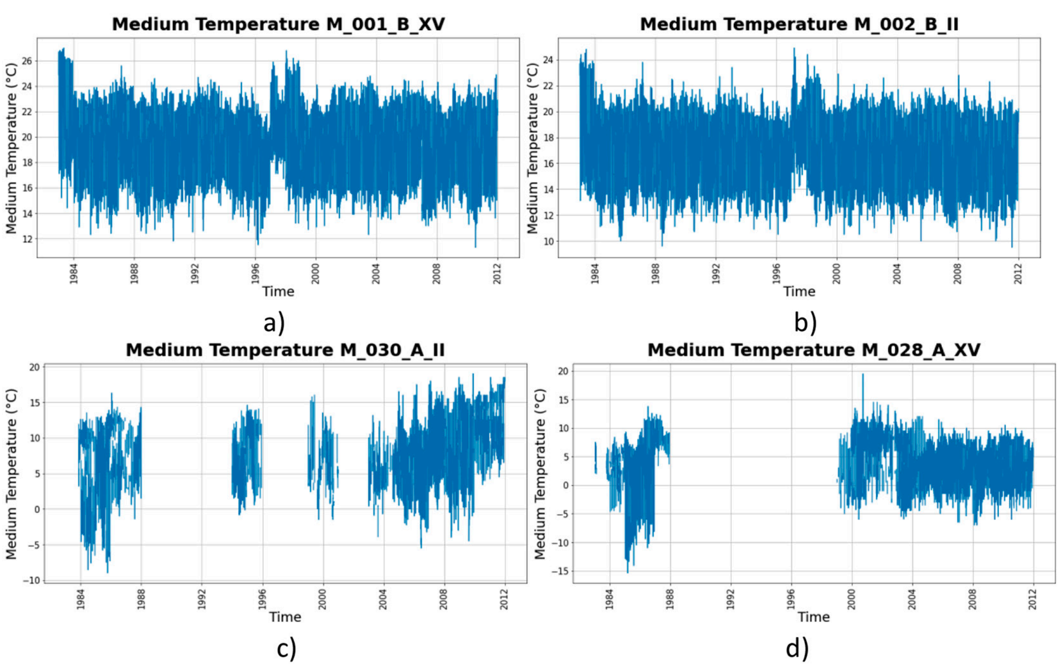


Figure 6. Quality of information in meteorological variable MEDT.

Meteorological variable maximum temperature
For the meteorological variable MAXT, the range of dates between 1993-2009 was used, where two of the best MS in terms of number of records are the following:

- M_002_B_II: This MS is located in the Antofagasta region, in the Antofagasta commune. Its owner is the DMC, and for the selected date range, it has 6209 records (Figure 7a).

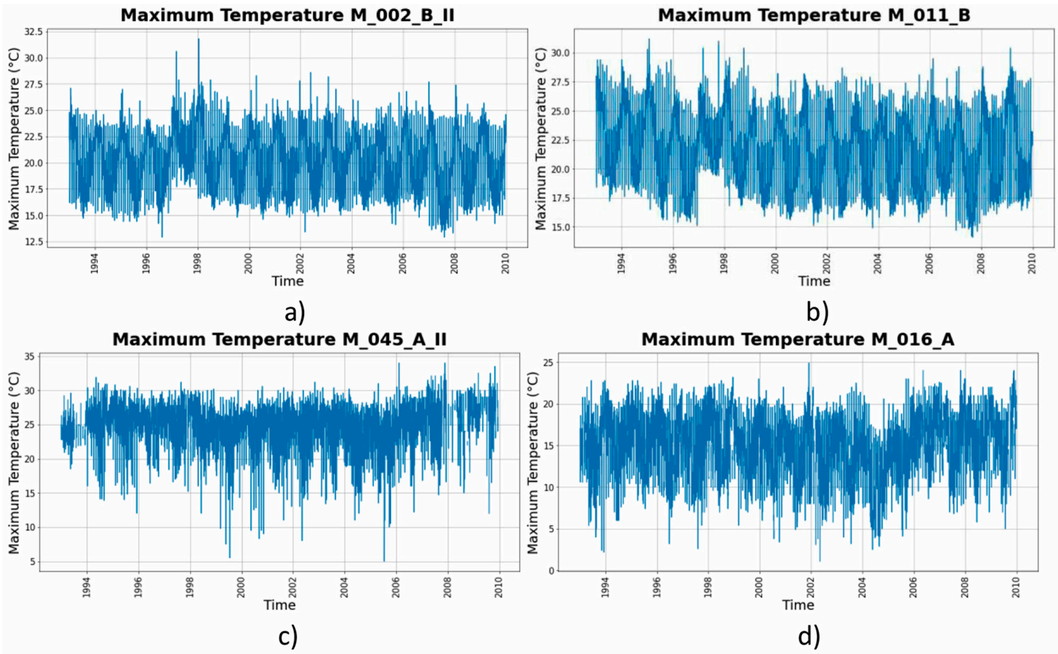


Figure 7. Quality of information on meteorological variable MAXT.

- M_011_B: This MS is located in the Tarapacá region, in the Iquique commune. Its owner is the DMC, and for the selected date range, it has 6209 records (Figure 7b).

In the case of this meteorological variable, it was observed that there were not many MS with a large number of missing records, since the MS with the fewest records were the following:

- M_045_A_II : This MS is located in the Antofagasta region, in the commune Taltal. Its owner is the DGA, and for the selected date range, it has 5587 records (Figure 7c).
- M_016_A: This MS is located in the Tarapacá region, in the Pica commune. Its owner is the DGA, and for the selected date range, it has 5625 records (Figure 7d).

As can be seen in each of the graphs presented, in the case of those MS that have better quality in terms of the number of records, it can be observed that the timeline of the records has very few gaps. In contrast, in the case of the MS that have a large amount of missing data, there are many blank spaces in the timeline of their records, except for the MS of the meteorological variable MAXT, which, for the range of dates to be established, has good data quality in terms of completeness.

Finally, the meteorological variable PP was the one that presented the most significant number of MS with information. However, many of these had significant gaps in their records. On the other hand, the meteorological variable MAXT was the one that presented the best TS in terms of completeness of information since these had much smaller gaps than the other meteorological variables.

Analysis of results

This section analyzes the results of the experiments carried out, grouped by meteorological variable (MINT, MEDT, MAXT, and PP), considering the application of the imputation methods WLC, RA, ISD, NR, and DR in each one of them. For the above, the distribution of the results of residual errors and Spearman and Kendall correlation coefficients is presented in summary form since the variables do not have a normal distribution, both obtained between the TS filled in with the methods and the TS with accurate information for the experiments.

The best results are represented by the residual errors closest to zero and positive correlations equal to or greater than 0.7 since the latter show that the data maintain the same behavior and direction, discarding negative correlations, which represent behaviors opposites.

Additionally, the EMs used for each meteorological variable are different, creating four groups through a selection process with criteria that allow them to be validated as reliable sources of information. For this, an analysis of correlation and geographic locations between EMs is carried out, accepting those that present a correlation coefficient equal to or greater than 0.7 and have coherence in the geographic characteristics where they are located. Therefore, given the above, the group of EMs selected for each meteorological variable differs since it depends on the correlation analysis, geographic location, and whether they have records of the variable of interest.

Minimum temperature result analysis

Figure 8a and Figure 8b show the distribution of both correlation coefficients; the WLC method is the one with the best results, while RA and NR have the lowest correlation.

When analyzing the results of the residual errors (Figure 9b) and the distribution of residual error means (Figure 9a) in all the methods of the different experiments performed, it is observed that DR presents the average residual error closest to zero; however, the difference with the WLC method is not significant because it does not exceed the 0.3 ºC. On the one hand, the difference with the WLC method is insignificant because it does not exceed the separation in the mean of its residual errors; this affirms that the WLC method is the best for MINT since it also presents higher average values in the correlation coefficients. On the other hand, the worst method for MINT is NR, as it presents the lowest correlations and the highest residual error.

Table 9, Table 10, and Table 11 show further details of the results associated with the Spearman and

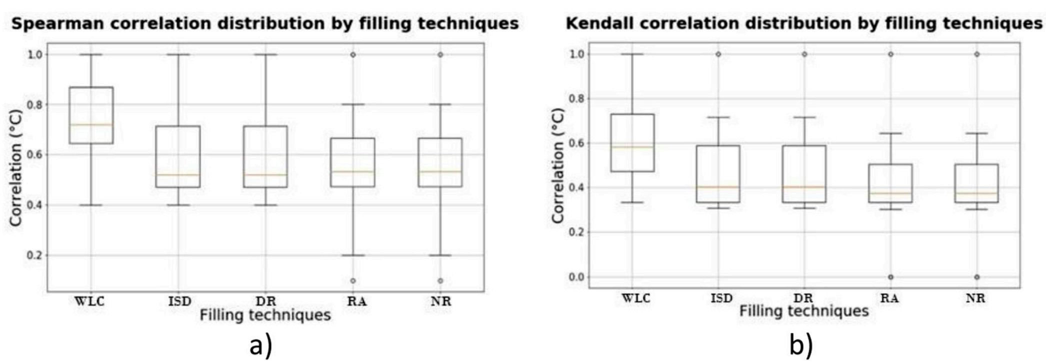


Figure 8. Distribution of Spearman and Kendall correlations according to imputation methods in MINT experiments.

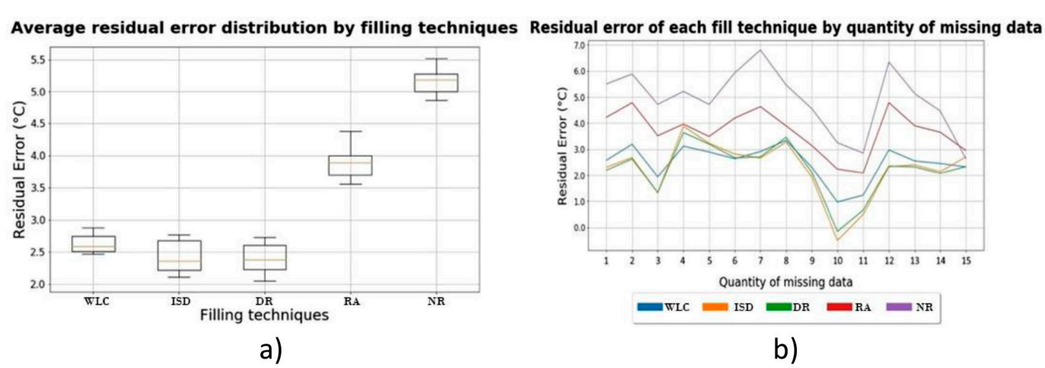


Figure 9. Distribution of residual error means according to imputation methods in MINT experiments.

Table 9. Descriptive statistics of Spearman correlations in MINT experiments.

	WLC	ISD	DR	RA	NR
Media	0.728426	0.607245	0.607245	0.554065	0.554065
Standard deviation	0.159472	0.176560	0.176560	0.223464	0.223464
Minimum	0.4	0.4	0.4	0.1	0.1
Q1	0.644128	0.471150	0.471150	0.474062	0.474062
Median	0.719741	0.520518	0.520518	0.534409	0.534409
Q3	0.869643	0.716071	0.716071	0.666984	0.666984
Maximum	1	1	1	1	1
IQR	0.225514	0.244922	0.244922	0.192922	0.192922
Kurtosis	2.480887	2.486516	2.486516	3.003914	3.003914
Skewness Coefficient	–0.272372	0.850637	0.850637	–0.142985	–0.142985

Table 10. Descriptive statistics of Kendall correlations in MINT experiments.

	WLC	ISD	DR	RA	NR
Media	0.601061	0.486204	0.486204	0.408869	0.408869
Standard deviation	0.183376	0.198478	0.198478	0.242995	0.242995
Minimum	0.333333	0.307729	0.307729	0	0
Q1	0.471216	0.333333	0.333333	0.333333	0.333333
Median	0.582039	0.40226	0.40226	0.373274	0.373274
Q3	0.730556	0.588889	0.588889	0.505224	0.505224
Maximum	1	1	1	1	1
IQR	0.259339	0.255556	0.255556	0.17189	0.17189
Kurtosis	2.535674	3.649163	3.649163	3.75086	3.75086
Skewness Coefficient	0.43081	1.241347	1.241347	0.447817	0.447817

Table 11. Descriptive statistics of residual error means of MINT experiments.

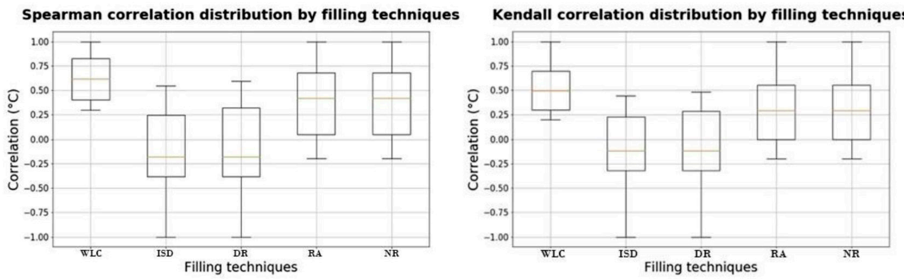
	WLC	ISD	DR	RA	NR
Media	2.640189	2.427392	2.389966	3.877176	5.152304
Standard deviation	0.131587	0.226394	0.202262	0.206808	0.183391
Minimum	2.464147	2.103245	2.050486	3.555914	4.862431
Q1	2.508217	2.214146	2.229576	3.696923	5.004064
Median	2.587091	2.359278	2.376897	3.894211	5.183658
Q3	2.746355	2.679654	2.603027	4.003007	5.276125
Maximum	2.882001	2.770307	2.726757	4.377317	5.519756
IQR	0.238137	0.465508	0.373451	0.306084	0.272061
Kurtosis	1.678614	1.413955	1.71369	3.010399	2.162846
Skewness Coefficient	0.269318	0.167399	0.18535	0.568282	0.239271

Kendall correlations, and the means of the residual errors obtained in the experiments with MINT, where the descriptive statistics are presented for each imputation method.

Mean temperature result analysis

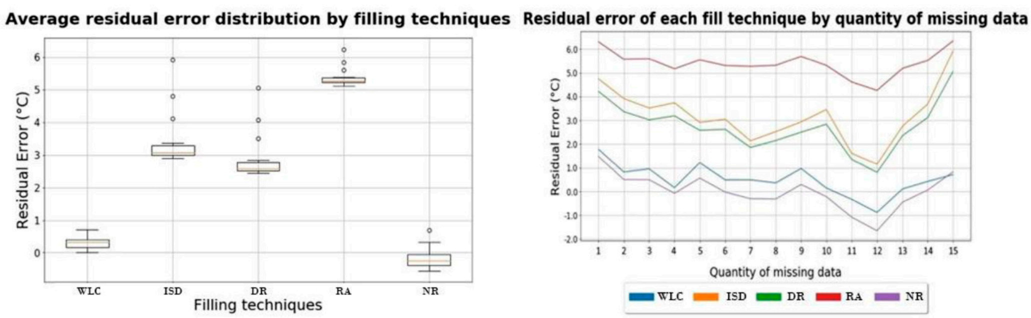
Figure 10a and Figure 10b show the distribution of both correlation coefficients; the WLC method is the one that presents the best results. In contrast, the other methods show negative correlations, meaning that filling the time series with these methods, except for WLC, leads to the meteorological variable behaving in a way that opposes its original pattern.

When analyzing the results of the residual errors (Figure 11b) and their mean distribution (Figure 11a)



a) b)

Figure 10. Distribution of Spearman and Kendall correlations according to imputation methods in MEDT experiments.



a) b)

Figure 11. Distribution of residual error means according to imputation methods in MEDT experiments.

in all the methods of the different experiments performed, it is observed that NR presents the average residual error closest to zero; however, the difference with the WLC method is not significant because it does not exceed 0.2 ºC of separation, allowing us to affirm that the WLC method is the best for MEDT since it also has the highest average value in the correlation coefficients in addition to being positive. The worst methods are ISD and DR, which have the highest negative correlations, as well as RA, which has the highest average residual error. ISD is the second method with the highest residual error, which, added to its negative correlations, is considered the worst method for MEDT, together with RA.

Table 12, Table 13, and Table 14 show more details of the results associated with the Spearman and Kendall correlations, as well as the means of the residual errors obtained in the MEDT experiments, where the descriptive statistics are presented for each imputation method.

Maximum temperature result analysis

Figure 12a and Figure 12b show that - the WLC, ISD, and DR methods have the best results- for the distribution of both correlation coefficients, while RA and NR have the lowest correlation.

When analyzing the results of the residual errors (Figure 13b) and their mean distribution (Figure 13a) in all the methods of the different experiments performed, it is observed that the ISD method is the one with the average residual error closest to zero, in addition to the fact that both correlation coefficients are the highest. Therefore, it is considered the best method for MAXT. On the other hand, the RA method has the highest average residual error, and the correlation coefficients are the lowest among all the methods, making it the worst method for MAXT.

Table 15, Table 16, and Table 17 show more details of the results associated with the Spearman and Kendall correlations, as well as the means of the residual errors obtained in the experiments with

Table 12. Descriptive statistics of Spearman correlations in MEDT experiments.

	WLC	ISD	DR	RA	NR
Media	0.628826	–0.11166	–0.08742	0.373454	0.373454
Standard deviation	0.235622	0.421058	0.446643	0.365806	0.365806
Minimum	0.3	–1	–1	–0.2	–0.2
Q1	0.40146	–0.38697	–0.38697	0.049078	0.049078
Median	0.615687	–0.18272	–0.18272	0.419251	0.419251
Q3	0.825712	0.247739	0.323422	0.68085	0.68085
Maximum	1	0.546282	0.594801	1	1
IQR	0.424252	0.634705	0.710388	0.631772	0.631772
Kurtosis	1.576002	2.388638	2.229742	1.760366	1.760366
Skewness Coefficient	0.205065	–0.19404	–0.12442	–0.00222	–0.00222

Table 13. Descriptive statistics of Kendall correlations in MEDT experiments.

	WLC	ISD	DR	RA	NR
Media	0.527506	–0.08613	–0.06582	0.287808	0.287808
Standard deviation	0.257847	0.373309	0.392439	0.342858	0.342858
Minimum	0.2	–1	–1	–0.2	–0.2
Q1	0.298123	–0.31901	–0.31901	0	0
Median	0.494949	–0.11562	–0.11562	0.294712	0.294712
Q3	0.6946	0.230075	0.286413	0.552448	0.552448
Maximum	1	0.441367	0.4806	1	1
IQR	0.396478	0.549081	0.60542	0.552448	0.552448
Kurtosis	2.091292	3.248197	2.995467	2.149295	2.149295
Skewness Coefficient	0.528397	–0.64008	–0.5498	0.352058	0.352058

Table 14. Descriptive statistics of residual error means of MEDT experiments.

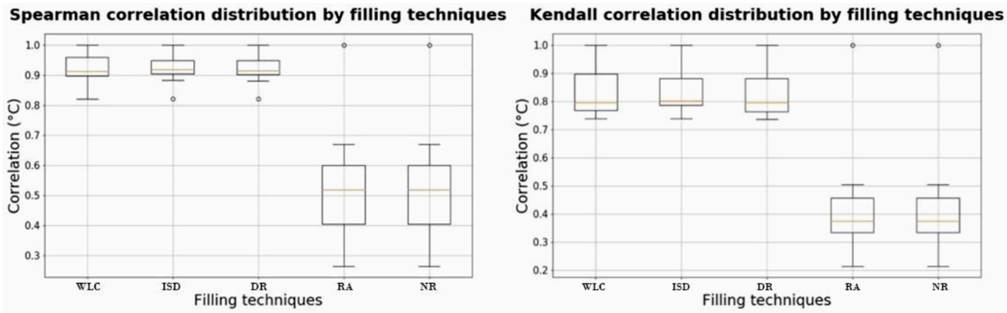
	WLC	ISD	DR	RA	NR
Media	0.302851	3.428106	2.902985	5.384131	–0.17191
Standard deviation	0.191948	0.834054	0.722335	0.28973	0.319833
Minimum	0.004303	2.904733	2.44669	5.116059	–0.55192
Q1	0.174273	2.993448	2.527293	5.241393	–0.38088
Median	0.320228	3.068654	2.573219	5.277757	–0.24009
Q3	0.406568	3.286392	2.782019	5.365601	–0.05079
Maximum	0.710221	5.914154	5.056095	6.245951	0.694601
IQR	0.232295	0.292943	0.254726	0.124208	0.330095
Kurtosis	2.470718	5.829094	5.824996	5.75784	4.253792
Skewness Coefficient	0.311041	2.000617	2.000087	1.908306	1.332717

MAXT, where the descriptive statistics are presented for each imputation method.

Precipitation result analysis

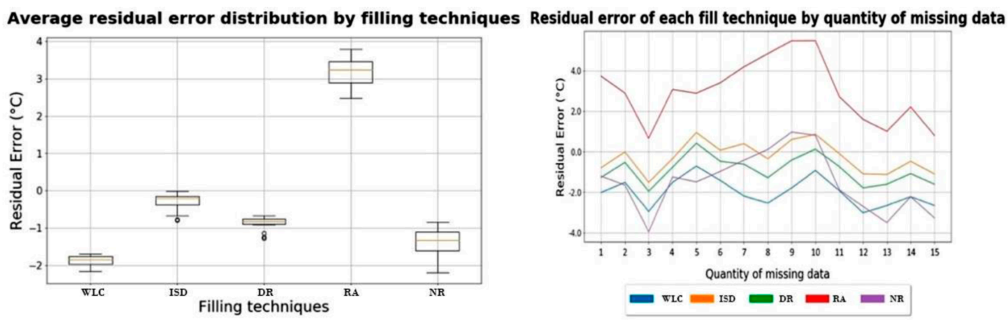
In Figure 14a and Figure 14b, it can be seen that the distribution of both correlation coefficients is similar in all methods.

When analyzing the results of the residual errors (Figure 15b) and their mean distribution (Figure 15a) for all the methods in the different experiments, it is observed that the methods have the same behavior in terms of correlation coefficients; however, the residual errors show differences where the NR method is the best for PP since it has the average



(a) (b)

Figure 12. Distribution of Spearman and Kendall correlations according to imputation methods in MAXT experiments.



a) b)

Figure 13. Distribution of residual error means according to imputation methods in MAXT experiments.

Table 15. Descriptive statistics of Spearman correlations in MAXT experiments.

	WLC	ISD	DR	RA	NR
Media	0.925414	0.927336	0.92397	0.526019	0.526019
Standard deviation	0.050274	0.047784	0.049417	0.172159	0.172159
Minimum	0.820783	0.820783	0.820783	0.26482	0.26482
Q1	0.897547	0.90278	0.902357	0.403384	0.403384
Median	0.912642	0.918047	0.913511	0.518039	0.518039
Q3	0.959546	0.948103	0.948103	0.599705	0.599705
Maximum	1	1	1	1	1
IQR	0.061999	0.045324	0.045746	0.196321	0.196321
Kurtosis	2.482743	2.98882	2.65101	4.678711	4.678711
Skewness Coefficient	–0.01511	–0.11905	0.034656	1.128224	1.128224

residual error closest to zero. On the other hand, the ISD and RP methods are the worst as they have the highest average residual error.

Table 18, Table 19, and Table 20 show further details about the results associated with the Spearman and Kendall correlations, and the means of the residual errors obtained in the experiments with PP, where the descriptive statistics are presented for each imputation method.

CONCLUSIONS

The climate of northern Chile is very variable, reflected in the different meteorological variables that characterize it, allowing us to understand its

Table 16. Descriptive statistics of Kendall correlations in MAXT experiments.

	WLC	ISD	DR	RA	NR
Media	0.842849	0.845967	0.835753	0.423531	0.423531
Standard deviation	0.094604	0.087908	0.095593	0.176488	0.176488
Minimum	0.737865	0.737865	0.73607	0.214834	0.214834
Q1	0.767596	0.78563	0.76277	0.333333	0.333333
Median	0.794317	0.801759	0.795168	0.37471	0.37471
Q3	0.896926	0.879891	0.879891	0.456287	0.456287
Maximum	1	1	1	1	1
IQR	0.12933	0.09426	0.11712	0.122953	0.122953
Kurtosis	1.943716	2.293251	2.138493	8.305527	8.305527
Skewness Coefficient	0.683668	0.879623	0.805736	2.322748	2.322748

Table 17. Descriptive statistics of residual error means of MAXT experiments.

	WLC	ISD	DR	RA	NR
Media	–1.87606	–0.30664	–0.88083	3.1795	–1.40125
Standard deviation	0.128805	0.242945	0.182622	0.388889	0.371079
Minimum	–2.16284	–0.79243	–1.27189	2.484457	–2.18739
Q1	–1.97822	–0.37554	–0.90394	2.901211	–1.60394
Median	–1.85658	–0.2126	–0.82229	3.239945	–1.32916
Q3	–1.7623	–0.1482	–0.76642	3.465411	–1.09891
Maximum	–1.68831	–0.01783	–0.67924	3.802356	–0.8508
IQR	0.215929	0.227341	0.137526	0.564199	0.505027
Kurtosis	2.40459	2.64178	3.017049	1.951535	2.395325
Skewness Coefficient	–0.50771	–0.93186	–1.13544	–0.17419	–0.51206

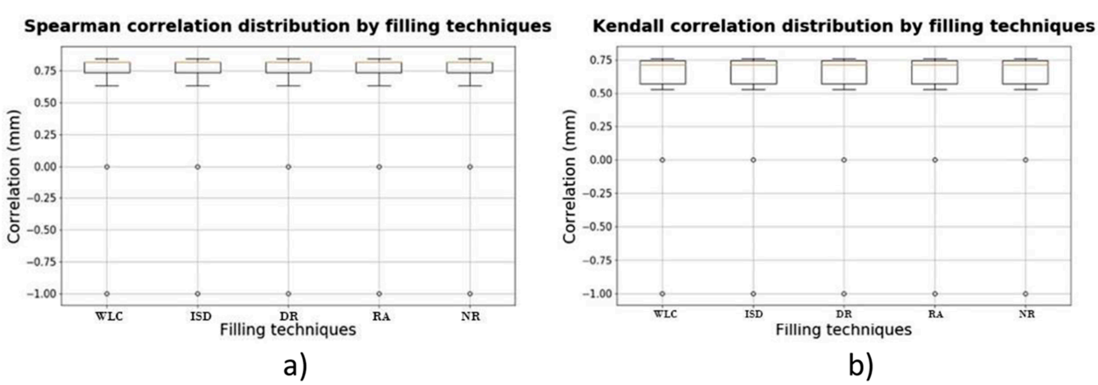


Figure 14. Distribution of Spearman and Kendall correlations according to imputation methods in PP experiments.

natural behavior. In this context, the experiments carried out in the present work have been familiar with this variability. The exploration of the data corroborated the diagnosis made about the MS located in northern Chile, which has data quality problems, mainly manifested as missing data in the registered TS, preventing them from complying with the WMO recommendations to calculate the ECI in a reliable and representative way.

Based on the exploratory analysis performed, the analysis of data imputation methods, and the results

Table 18. Descriptive statistics of Spearman correlations in PP experiments.

	WLC	ISD	DR	RA	NR
Media	0.612451	0.612451	0.612451	0.612451	0.612451
Standard deviation	0.494905	0.494905	0.494905	0.494905	0.494905
Minimum	–1	–1	–1	–1	–1
Q1	0.737168	0.737168	0.737168	0.737168	0.737168
Median	0.820137	0.820137	0.820137	0.820137	0.820137
Q3	0.823743	0.823743	0.823743	0.823743	0.823743
Maximum	0.850127	0.850127	0.850127	0.850127	0.850127
IQR	0.086575	0.086575	0.086575	0.086575	0.086575
Kurtosis	8.240134	8.240134	8.240134	8.240134	8.240134
Skewness Coefficient	–2.54993	–2.54993	–2.54993	–2.54993	–2.54993

Table 19. Descriptive statistics of Kendall correlations in PP experiments.

	WLC	ISD	DR	RA	NR
Media	0.523105	0.523105	0.523105	0.523105	0.523105
Standard deviation	0.463846	0.463846	0.463846	0.463846	0.463846
Minimum	–1	–1	–1	–1	–1
Q1	0.571506	0.571506	0.571506	0.571506	0.571506
Median	0.712415	0.712415	0.712415	0.712415	0.712415
Q3	0.741145	0.741145	0.741145	0.741145	0.741145
Maximum	0.760204	0.760204	0.760204	0.760204	0.760204
IQR	0.169639	0.169639	0.169639	0.169639	0.169639
Kurtosis	8.448857	8.448857	8.448857	8.448857	8.448857
Skewness Coefficient	0.43081	1.241347	1.241347	0.447817	0.447817

Table 20. Descriptive statistics of residual error means of PP experiments.

	WLC	ISD	DR	RA	NR
Media	0.472858	0.87992	0.623762	0.880603	0.154545
Standard deviation	0.534568	0.877728	0.875569	0.526793	0.810051
Minimum	–1.07359	–1.94185	–2.2573	–0.48132	–2.30843
Q1	0.447432	0.718422	0.586912	0.634699	0.301162
Median	0.584281	0.968539	0.791244	0.86784	0.411356
Q3	0.694546	1.344168	1.064917	1.197599	0.515556
Maximum	1.188207	1.972769	1.606758	1.820014	0.843352
IQR	0.247114	0.625746	0.478006	0.5629	0.214394
Kurtosis	5.488716	7.414555	8.05197	3.976049	6.454641
Skewness Coefficient	–1.66461	–1.98996	–2.27205	–0.61888	–2.15561

obtained, it is concluded that there is more than one method with minimum residual error and positive correlation that can be used for data imputation in the meteorological variables of northern Chile. For MINT and MEDT, the method that shows to be most suitable for these purposes is WLC, while for MAXT and PP, it is ISD and NR, respectively. The above is attributed to the fact that a different group of MS was used for each meteorological variable as a result of the selection process carried out based on the correlation analysis, geographic location, and whether these had records of the variable under study, creating four groups of MS where each one is used for a specific variable; this

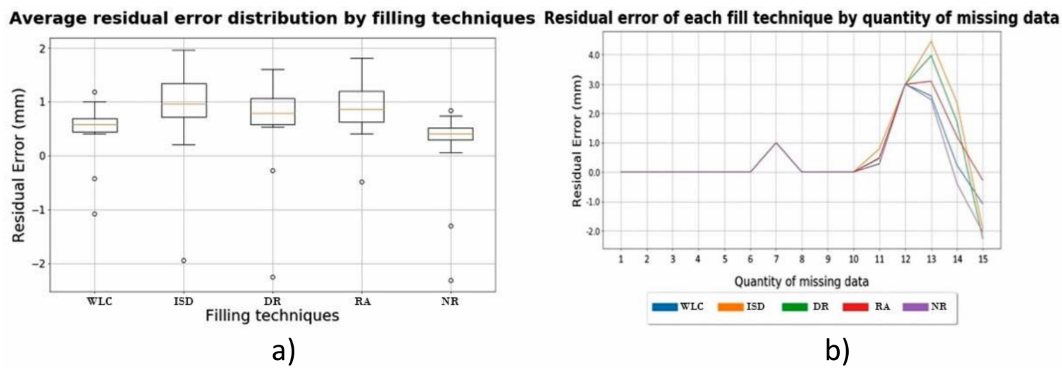


Figure 15. Distribution of residual error means according to imputation methods in PP experiments.

implies that each group of MS is concentrated in different geographic sectors with their climatic characteristics, contributing to the fact that different methods can present better results according to the meteorological variable.

In general, it is also concluded that the imputation methods that use correlation coefficients as weights for neighboring MS data work very well in all variables except for PP, where NR is more suitable since it is based on the averages of neighboring MS. Since PP is scarce in northern Chile, the average value is similar in most sectors, helping this method to be positioned above the others.

The case of MAXT, where ISD is presented as a better method than WLC for MINT and MEDT, is attributed to the fact that in the MAXT experiment, the neighboring MS were geographically closer than the neighboring MS used in MINT and MAXT. Since ISD uses the distance between MS as a factor, it is natural that it can deliver better results since it weights them better.

Regarding the negative correlations observed in the MEDT experiments, and in comparison to MINT and MAXT that presented only positive correlations, it is attributed to the fact that MINT and MAXT correspond to meteorological variables of extreme values due to the climatic characteristics of northern Chile, so that when completed with the imputation methods, they manage to capture very well the behavior of the MINT and MAXT data of the neighboring MS used since they are also extreme values of similar behavior. However, MEDT has more variability in its data since it represents the average behavior of temperature, which is affected by extreme MINT and MAXT that occur in the study area, explaining the difficulty of some imputation methods to reproduce the same behavior.

Future work will include the incorporation of more imputation methods to broaden the comparison, and the use of reanalysis data in the experiments to observe the behavior of the methods with meteorological data simulated by these models.

ACKNOWLEDGMENTS

This work was partially funded by the postdoctoral project (grant No. 74200094) from ANID, Chile, Fondecyt de Iniciación 11230961 from ANID, Chile, and by INICI-UV 20993 of the University of Valparaiso, Chile.

REFERENCES

[1] T. Aljuaid and S. Sasi, “Proper imputation techniques for missing values in data sets,” in *International Conference on Data Science and Engineering (ICDSE)*, Cochin, India, 2016, doi: 10.1109/ICDSE.2016.7823957.

[2] S.I. Khan and A.S.M. Hoque, “SICE: an improved missing data imputation technique,” *Journal of big Data*, vol. 7, no. 1, Art. no. 37, pp. 1-21, 2020, doi: 10.1186/s40537-020-00313-w.

[3] A. Jadhav, D. Pramod, and K. Ramanathan, “Comparison of Performance of Data Imputation Methods for Numeric Dataset,” *Applied Artificial Intelligence*, vol. 33, no. 10, pp. 913-933, 2019, doi: 10.1080/08839514.2019.1637138.

[4] D. Papale, “Data gap filling,” in *Eddy Covariance. A Practical Guide to Measurement and Data Analysis*, M. Aubinet, T. Vesala, and D. Papale, Eds. Dordrecht, Netherlands: Springer, Nov. 2011, pp. 159-172, doi: 10.1007/978-94-007-2351-1.

[5] M.S. Osman, A.M. Abu-Mahfouz, and P.R. Page, “A Survey on Data Imputation Techniques: Water Distribution System as a Use Case,” in *IEEE Access*, vol. 6, pp. 63279-63291, 2018, doi: 10.1109/ACCESS.2018.2877269.

[6] B.L. Smith, W.T. Scherer, and J.H. Conklin, “Exploring Imputation Techniques for Missing Data in Transportation Management Systems,” *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1836, no. 1, pp. 132-142, 2003, doi: 10.3141/1836-17.

[7] A.W. Liew, N.F. Law, and H. Yan, “Missing value imputation for gene expression data: computational techniques to recover missing data from available information,” *Briefings in Bioinformatics*, vol. 12, no. 5, pp. 498-513, 2011, doi: 10.1093/bib/bbq080.

[8] J.M. Jerez, I. Molina, P.J. García-Laencina, E. Alba, N. Ribelles, M. Martín, and L. Franco, “Missing data imputation using statistical and machine learning methods in a real breast cancer problem,” *Artificial intelligence in medicine*, vol. 50, no. 2, pp. 105-115, Octubre, 2010, doi: 10.1016/j.artmed.2010.05.002.

[9] Y. Julien and J.A. Sobrino, “Optimizing and comparing gap-filling techniques using simulated NDVI time series from remotely sensed global data,” *International Journal of Applied Earth Observation and Geoinformation*, vol. 76, pp. 93-111, 2019. doi: 10.1016/j.jag.2018.11.008

[10] S. Dengel, D. Zona, T. Sachs, M. Aurela, M. Jammet, F.J.W. Parmentier, W. Oechel, and T. Vesala, “Testing the applicability of neural networks as a gap-filling method using CH 4 flux data from high latitude wetlands,” *Biogeosciences*, vol. 10, no. 12, pp. 8185-8200, 2013, doi: https://doi.org/10.5194/bg-10-8185-2013.

[11] G. Wang, D. Garcia, Y. Liu, R. De Jeu, and A.J. Dolman, “A three-dimensional gap filling method for large geophysical datasets: Application to global satellite soil moisture observations,” *Environmental Modelling & Software*, vol. 30, pp. 139-142, 2012, doi: 10.1016/j.envsoft.2011.10.015.

[12] M.N. Norazian, Y.A. Shukri, R.N. Azam, and A.M.M. Al Bakri, “Estimation of missing values in air pollution data using single imputation techniques,” *ScienceAsia*, vol. 34, no. 3, pp. 341-345, 2008, doi: 10.2306/scienceasia1513-1874.2008.34.341.

[13] D.V. Carrera Villacrés, P.V. Guevara García, L.C. Tamayo Bacacela, A.L. Balarezo Aguilar, C.A. Narváez Rivera y D.R. Morocho López, “Relleno de series anuales de datos meteorológicos mediante métodos estadísticos en la zona costera e interandina del Ecuador, y cálculo de la precipitación media”, *Idesia*, vol. 34, no. 3, pp. 81-90, 2016, doi: 10.4067/S0718-34292016000300010.

[14] R. Pizarro, P. Ausensi, D. Aravena, C. Sangüesa, L. León y F. Balocchi, “Evaluación de métodos hidrológicos para la completación de datos faltantes de precipitación en estaciones de la región del Maule, Chile”, *Aqua-LAC*, vol. 1, no. 2, pp. 172-184, 2009, doi:10.29104/phi-aqualac/2009-v1-2-07.

[15] B. Cho, T. Dayrit, Y. Gao, Z. Wang, T. Hong, A. Sim, and K. Wu, “Effective Missing Value Imputation Methods for Building Monitoring Data,” in *2020 IEEE International Conference on Big Data, Atlanta,*GA, USA, 2020, doi: 10.1109/BigData50022.2020.9378230.

[16] E. Afrifa-Yamoah, U.A. Mueller, S.M. Taylor, and A.J. Fisher, “Missing data imputation of high-resolution temporal climate time series data,” *Meteorologital Applications*, vol. 27, no. 1, 2020, doi: 10.1002/met.1873.

[17] J. Parra-Plazas, P. Gaona-Garcia, and L. Plazas-Nossa, “Time series outlier removal and imputing methods based on Colombian weather stations data,” *Environmental Science and Pollution Research*, vol. 30, no. 28, 2023, doi: 10.1007/s11356-023-27176-x.

[18] N. Kannegowda, S. Udayar Pillai, C. Kommireddi, and Fousiya, “Comparative assessment of univariate and multivariate imputation models for varying lengths of missing rainfall data in a humid tropical region: a case study of Kozhikode, Kerala, India,” *Acta Geophysica*, vol. 72, no. 4, 2024, doi: 10.1007/s11600-023-01152-y.

[19] D. Machiwal and M. Kumar Jha, *Hydrologic Time Series Analysis: Theory and Practice*, Dordrecht, Netherlands: Springer, 2012, doi: 10.1007/978-94-007-1861-6.

[20] J.C. Gonzalez-Hidalgo, M. De Luis Arrillaga, P. Stepanek, J. Raventós y J.M. Cuadrat, "Reconstrucción, estabilidad y proceso de homogeneización de series de precipitación en ambientes de elevada variabilidad pluvial", en *VII Reunión Nacional de Climatología. La Información Climática Como Herramienta de Gestión Ambiental*, Albarracín, España, 2002.

[21] A. Barrera, "Técnicas de completado de series mensuales y aplicación al estudio de la influencia de la NAO en la distribución de la precipitación en España", Trabajo de Doctorado, Departamento de Astronomía y Meteorología, Universidad de Barcelona, España, 2004.

[22] T.C. Peterson and D.R. Easterling, "Creation of homogeneous composite climatological reference series," *International Journal of Climatology*, vol. 14, no. 6, pp. 671-679, 1994, doi: 10.1002/joc.3370140606.

[23] A.E. Luna Romero y W.S. Lavado Casimiro, "Evaluación de métodos hidrológicos para la completación de datos faltantes de precipitación en estaciones de la cuenta Jetepeque, Perú", *Revista Tecnológica ESPOL*, vol. 28, no. 3, pp. 42-52, 2015. [En línea]. Disponible en: https://rte.espol.edu.ec/index.php/tecnologica/article/view/400/261

[24] C. Guevara Ochoa, N. Briceño, E. Zimmermann, L. Vives, M. Blanco y G. Ares, "Relleno de series de precipitación diaria para largos periodos de tiempo en zonas de llanura: Caso de estudio cuenca superior del arroyo del Azul", *Geoacta*, vol. 42, no. 1, pp. 38-62, 2017, [En línea]. Disponible en: https://revistas.unlp.edu.ar/geoacta/article/view/13448

[25] R. Aguilera *et al. Guía metodológica para la elaboración del mapa de zonas áridas, semiáridas, subhúmedas secas de América Latina y el Caribe*, La Serena, Chile: Editorial del Norte, 2006.