




Evaluation of the Known Sub-Sequence Algorithm (KSSA) for Optimal Use in Time Series of Tuna Fishing in the Pacific Ocean

Evaluación del algoritmo de subsecuencias conocidas (KSSA) para su uso óptimo en series de tiempo de pesca de atún en el Océano Pacífico

Julián A. Gómez ¹, Iván F. Benavides ², and John J. Selvaraj ³

ABSTRACT

An important limitation in fisheries research using time series is the presence of missing data. An inadequate handling of these data can negatively impact the results of statistical analyses, leading to erroneous decision-making. A potential solution to this problem is the estimation of missing values through imputation methods, but none of them can be universally applied to all time series. In fact, their effectiveness largely depends on the structure of the data and the distribution of the missing values. Recently, a solution to this problem was proposed which employs the known sub-sequence algorithm (KSSA), a machine learning technique designed to compare the performance of different imputation methods and validate them within the time series that contains missing data. However, due to its recent development, there is no published evidence regarding its efficiency and reliability. This research aimed to assess the efficiency of the KSSA on tuna fisheries data for the Pacific Ocean by imputing simulated missing data for seven time series with distinct structures. The results demonstrated that the algorithm is robust for accurately validating a wide combination of properties, such as length, seasonality, trend, autocorrelation structure, and percentage of missing data. Additionally, the algorithm's hyperparameters can be easily adjusted to achieve optimal results for each time series.

Keywords: missing data, machine learning, imputation, data

RESUMEN

Una limitación importante en la investigación pesquera que utiliza series temporales es la presencia de datos faltantes. Un manejo inadecuado de estos datos puede afectar negativamente los resultados de los análisis estadísticos, conllevando una toma de decisiones errónea. Una posible solución a este problema es la estimación de valores faltantes mediante métodos de imputación, pero ninguno de ellos puede aplicarse universalmente a todas las series temporales. De hecho, su eficacia depende en gran medida de la estructura de los datos y de la distribución de los valores faltantes. Recientemente, se propuso una solución a este problema que emplea el algoritmo de sub-secuencias conocidas (KSSA), una técnica de aprendizaje automático diseñada para comparar el rendimiento de diferentes métodos de imputación y validarlos dentro de la serie temporal que contiene datos faltantes. Sin embargo, debido a su desarrollo reciente, no hay evidencia publicada sobre su eficiencia y fiabilidad. Esta investigación tuvo por objetivo evaluar la eficiencia del KSSA en datos de pesca de atún para el Océano Pacífico mediante la imputación de datos faltantes simulados para siete series temporales con estructuras distintas. Los resultados demostraron que el algoritmo es robusto para validar con precisión una amplia combinación de propiedades como la longitud, la estacionalidad, la tendencia, la estructura de autocorrelación y el porcentaje de datos faltantes. Además, los hiperparámetros del algoritmo pueden ajustarse fácilmente para lograr resultados óptimos en cada serie temporal.

Palabras clave: datos faltantes, aprendizaje automático, imputación, datos

Received: July 8th, 2024

Accepted: May 14th, 2025

Introduction

The term *time series* refers to a sequence of observations of a variable recorded at regular time intervals (daily, weekly, semiannual, yearly, etc.) [1]. This has diverse applications in various disciplines such as engineering, economics, and biology, among others. Data from time series can be utilized for forecasts that aid in implementing risk prevention policies and control measures or conducting exploratory analyses [2]. However, time series often exhibit missing

¹ Environmental engineer, Universidad Nacional de Colombia, Palmira campus, Colombia. Affiliation: Student, Universidad Nacional de Colombia, Medellín campus, Colombia. Email: jugomezp@unal.edu.co

² Hydobiological Resources Research Group, Department of Engineering, Universidad Nacional de Colombia, Palmira campus, Carrera 32 No. 12-00, Palmira, Valle del Cauca, Colombia

³ Universidad Nacional de Colombia, Palmira campus, Department of Engineering, Faculty of Engineering and Administration - Carrera 32 No. 12-00 Chapinero, Vía Candelaria, Palmira, 763533 Valle del Cauca - Colombia



Attribution 4.0 International (CC BY 4.0) Share - Adapt

data (MD) due to instrument failures, record discontinuity, adverse weather conditions, and human error, among other factors [3]. Missing values can hinder statistical analysis or influence the generation of erroneous results, leading to inaccurate predictions or forecasts in different models [4]. Consequently, many analysis methods require time series to be free of MD. When there are MD, estimation or imputation is necessary, which involves the replacement of MD with logical values [5]. This issue is solved through simple calculations such as means, averages, and nearby values, or more sophisticated equations based on time series components like trends, seasonality, cyclicity, and autocorrelation structures [6].

It is crucial to identify the appropriate imputation method for each time series, as its efficacy depends on the structure of the existing data and the size and distribution of the MD [4], [7]. In this vein, it is important to note that there is no universally efficient MD imputation technique for all univariate time series [8]. All methods must be validated in order to determine the best approach for each particular time series, optimizing their applicability [9]. This is often achieved using external time series with no missing values, to which simulated MD are added. These MD are then imputed using different methods, and the imputed sequences are compared against the actual data, identifying the best imputation method by means of statistical performance metrics. However, considering that the efficacy of imputation methods depends on the structure of the time series and the MD, this validation approach may yield erroneous results, as the best method for imputing the complete time series used in validation is not necessarily the best option for that containing MD.

[10] offered a solution to this problem by developing a machine learning algorithm called the *known sub-sequence algorithm* (KSSA), which automatically identifies the best imputation method for each specific time series. The KSSA performs validation using the data available within the time series, eliminating the need for external time series. This algorithm leverages the sub-sequences located between the MD, which, when sufficiently informative, allow simulating new MD. These simulated MD can then be imputed using different methods and compared against the real data using performance metrics such as the root mean squared error (RMSE). This tool is used to support the selection of the best statistical imputation method for each time series, regardless of its structure, as the algorithm learns from the aforementioned sub-sequences.

The KSSA operates in six steps. The first step involves identifying the target time series and measuring its percentage of MD relative to the total data in the series. The second step corresponds to the imputation of the MD using all the available or desired imputation methods. This is referred to as *initial imputation* and is aimed at generating a complete time series for subsequent operations. In the third step, the time series is divided into equal segments, and, within each segment, a simulation window is placed whose

size and position are random but always avoid the positions of the initial imputation. Simulated MD are generated based on these positions and sizes, resulting in a new time series (step four), whose percentage of MD relative to the data in the time series is measured. In step five, the simulated MDs are imputed using the methods to be compared and validated, and, finally, performance metrics such as the RMSE are calculated (step six).

Steps three to six are repeated, randomly changing the size and position of the simulation windows at each iteration, which results in different combinations of MD sizes and positions within each segment. The hyperparameters that can be fine-tuned to optimize the results of the KSSA include the selection of initial imputation methods, the imputation methods to be compared, the number of segments, the number of iterations, the percentage of missing data, and the value of the random seed used to ensure reproducibility. This can be done using R's *kssa* package [7].

In this vein, the objective of this research was to evaluate the efficiency of the KSSA using various time series of tuna fishing in the Pacific Ocean, in comparison with a classical validation process. This validation simulated MD in complete time series, estimated them using different imputation methods, and employed performance metrics to confidently identify the best method. The classical validation approach was used as a reference to determine the reliability of the KSSA's results. Further details will be provided in the *Materials and methods* section. Optimal hyperparameter combinations were systematically sought for different time series sizes and structures, as well as for different failure thresholds, to serve as reference for future studies. The beneficiaries of these results will be resource managers, scientists, traders, and fishermen, as time series allow linking fishing information with associated environmental variables, generating risk forecasts, predicting potential fishing areas, reconstructing historical records of fishery resource abundance, and forecasting future behaviors.

Materials and methods

Data

Seven complete fisheries time series (without missing values) were collected from publicly available sources, i.e., the Inter-American Tropical Tuna Commission (IATTC) (<https://www.iattc.org/PublicDomainData/IATTC-Catch-by-species1.htm>) and the Fisheries and Aquaculture section of the Food and Agriculture Organization (FAO), through its FishStatJ application (<https://www.fao.org/fishery/en/statistics/software/fishstatj/en>) (Table I).

The IATTC time series corresponded to the monthly catch per unit effort (CPUE) of skipjack tuna (*Katsuwonus pelamis*) associated with dolphins (SKJ-DEL) and floating objects (SKJ-OBJ), as well as of free-swimming individuals (SKJ-NOA), in the Tropical Pacific Ocean. This time series also included

data on yellowfin tuna (*Thunnus albacares*) associated with dolphins (YFT-DEL) and floating objects (YFT-OBJ). The FAO time series included the annual catch tonnage of Pacific bluefin tuna (*Thunnus orientalis*) (PBF-USA) and southern bluefin tuna (*Thunnus maccoyii*) (SBF-USA) in the Pacific Ocean off the coast of the United States. Each time series was examined for outliers and analyzed to identify trends using Mann-Kendall tests with R's *trend* package [11]. Periodicity was assessed using Lomb periodograms with the *lomb* package [7]. This is shown in Table I and Fig. 1.

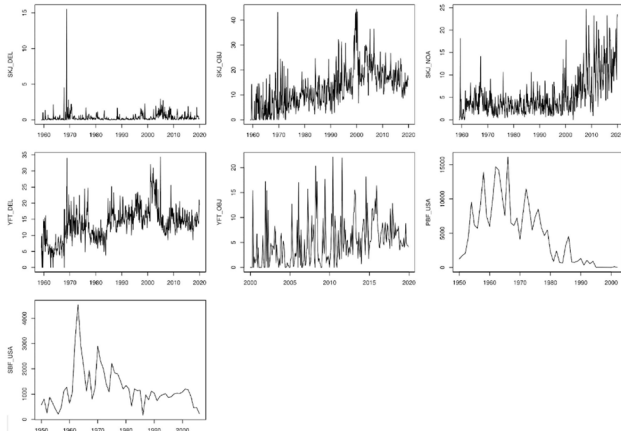


Figure 1. Time series plots used in this study. The vertical axis represents the CPUE.

Source: Authors

Experimental design

An R code was designed to compare the results of classical and KSSA validation. Classical validation involves taking a complete time series (without missing values), simulating the MD, imputing them using different methods, and selecting the best imputation method that simultaneously minimizes the error across various statistical performance metrics. This approach unequivocally identifies the best imputation method for each time series. In this vein, when the results obtained through classical validation and the KSSA converge with regard to the best or top-performing imputation method for each time series, it can be concluded that the KSSA results are reliable. In this work, we compared the performance of 11 imputation methods for MD in univariate time series available in the *kssa* package [7]. This can be seen in Table II.

Table I. Characteristics of the time series studied

Time series	Temporary extension	Length	Trend	Periodicity (dominant frequency)	Temporary resolution	Source
SKJ-DEL	1959-2019	732 months	Positive	100 months	Monthly	IACCT
SKJ-OBJ	1959-2019	732 months	Positive	100 months	Monthly	IACCT
SKJ-NOA	1959-2019	732 months	Positive	100 months	Monthly	IACCT

YFT-DEL	1959-2019	732 months	Positive	100 months	Monthly	IACCT
YFT-OBJ	2000-2019	240 months	Negative	12.5 months	Monthly	IACCT
SBF-USA	1950-2006	56 years	Null	9.09 years	Monthly	IACCT
PBF-USA	1950-2002	52 years	Negative	4.76 years	Monthly	IACCT

Source: Authors

Table II. Imputation methods used for classical and KSSA validation

Method	R function	Source in R
KALMAN smoothing over an ARIMA spatial representation	na_kalman(..., model="auto.arima", ...)	[5]
Kalman smoothing in state space structural models	na_kalman(..., model="StructTS", ...)	[5]
Exponential moving average	na_ma(..., weighting="exponential", ...)	[5]
Linear moving average	na_ma(..., weighting="linear", ...)	[5]
Simple moving average	na_ma(..., weighting="simple", ...)	[5]
Linear interpolation	na_interpolation(..., option="linear", ...)	[5]
Spline interpolation	na_interpolation(..., option="spline", ...)	[5]
Stineman interpolation	na_interpolation(..., option="stine", ...)	[5]
Last forward look	na_locf(..., option="locf", ...)	[5]
Seasonal decomposition with Kalman smoothing and state space models	na_seadec(..., algorithm="kalman", ...)	[5]
Robust decomposition by trend and seasonality with linear interpolation	na.interp	[12]

Source: Authors

Classical validation

In each time series, missing values were randomly simulated at levels of 10, 20, 30, 40, and 50%. To ensure that the positions of the missing values in each time series were different, each percentage used a different random seed by means of the *set.seed* function [13]. Subsequently, the missing values were imputed using the methods mentioned in Table II, and the RMSE was calculated.

KSSA validation

KSSA validation was conducted using different hyperparameter combinations in order to optimize the results for each time series. The imputation methods in Table II were used for both the initial imputation and the final comparison. For the series SKJ-NOA, SKJ-OBJ, SKJ-DEL, and YFT-DEL, 10, 20, and 30 segments were selected; for YFT-OBJ, 1-18 segments were used; for PBF-USA, 1-9 segments were selected; and, for SBF-USA, 1-6 segments were chosen. The percentages of MD were the same as in the classical validation. 10, 50, and 100 iterations were tested for all time series. The same random seeds used in the classical validation were employed to ensure reproducibility.

The results were recorded in a table that included the time series, the percentage of MD, the imputation method, and the performance metrics, with values arranged in ascending order. The best imputation method was the one with the lowest performance metrics. It should be highlighted that this study considered the results of the classical validation as a reference, as the simulation was carried out on a complete time series whose missing values could be accurately recovered. Fig. 2 provides a schematic representation of the classical and KSSA validation processes.

The following libraries or packages were employed by the algorithm: *sjmisc* [14], *missmethods* [15], *imputeTS* [5], *trend* [11], *forecast* [12], *Metrics* [16], *kssa* [7], and *haven* [17].

Data analysis

The KSSA is considered reliable and applicable when its results converge with those of classical validation regarding the best imputation method (i.e., the one that minimizes performance metrics). For instance, if the imputation method with the lowest RMSE (the best method) according to classical validation coincides with that exhibiting the lowest average RMSE after implementing the KSSA, the efficiency of the latter is 100%. This indicates that the algorithm accurately identified the best imputation method for the time series under study. If the best method according to classical validation is ranked second by the KSSA, its efficiency is 90%, and so on. An efficiency of 0% occurs when the best method obtained via classical validation is ranked last by the KSSA. Eq. (1) expresses this efficiency.

$$Ef = \frac{m-n}{m-1} * 100 \quad (1)$$

where *Ef* represents the efficiency of KSSA imputation; *m* is the total number of imputation methods being compared (11 in this case); *n* is the position provided by the KSSA that coincides with the best method provided by classical validation; and 100 scales the value to a percentage. This formula may be regarded as a calculation of *Ef* that has been specifically adjusted for this research based on record linkage and data matching [18].

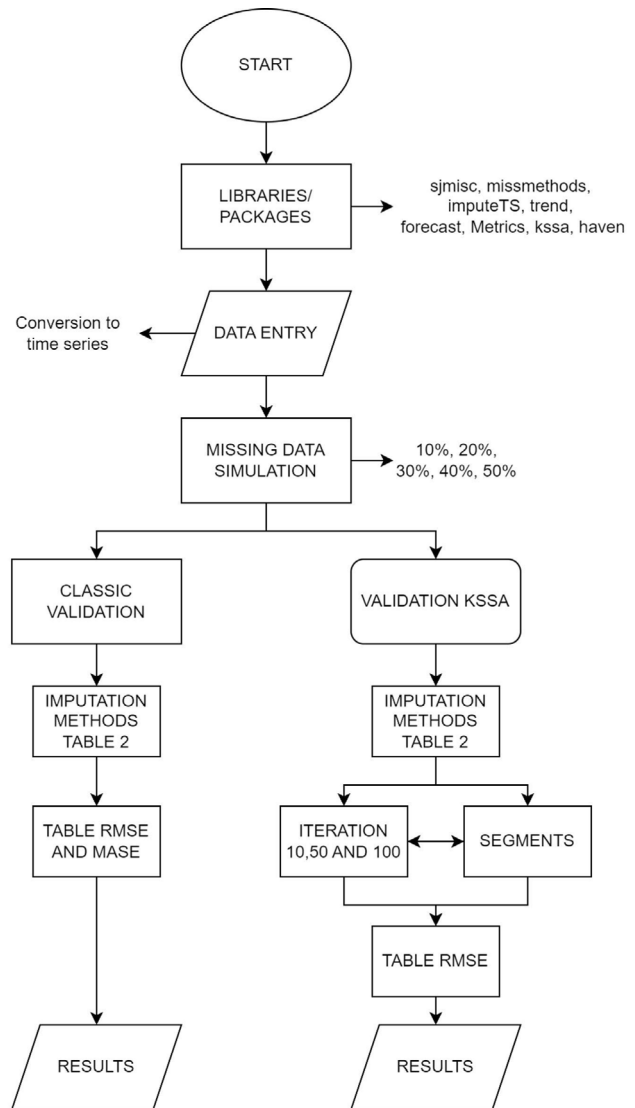


Figure 2. Flowchart summarizing the classical and KSSA validation processes

Source: Authors

The *Ef* values were analyzed using contour plots with the *filled.contour* function from R's *graphics* package [13]. Contour plots were generated for each combination of time series, number of segments, percentage of missing values, and number of iterations. The optimal efficiency area (OEA) was defined as the region within the 70% *Ef* contour, as there were no statistically significant differences regarding RMSE values between the first, second, third, and fourth positions of the KSSA (corresponding to 100, 90, 80, and 70% *Ef*, respectively). This was determined using a simple analysis of variance (ANOVA) with a significance level of 0.05. The OEA can be seen as a continuous white line in the contour plots in Fig. 3.

Finally, a multifactorial ANOVA was performed to quantify the variance explained by the number of iterations, the number of segments, and the percentage of missing values as fixed factors on *Ef*. This model considered all possible double interactions between the three factors. We selected a significance level of 0.05. These results are presented in Fig. 4.

Results and discussion

According to the results of this study, the time series exhibited a variety of structures, as detailed in Table II. This diversity included different lengths and patterns of temporal behavior, such as periodicity and trend. Specifically, periodicity and trend were observed in the SKJ-DEL, SKJ-OBJ, SKJ-NOA, YFT-DEL, and PBF-USA series, while YFT-OBJ and SBF-USA showed only periodicity. As mentioned by [19], time series can exhibit a wide variety of patterns. These characteristics can significantly affect the results, especially when there is no clear seasonality.

Starting at 40%, specific particularities were observed which differentiated each series, mainly in relation to the percentage of missing values, which affected the size, shape, and distribution pattern of the OEA. This finding is consistent with previous studies using imputation methods, wherein performance decreases when the percentage of MD is high. For example, [20] observed that, while the method generally maintains consistent performance, its accuracy deteriorates significantly when handling datasets with 50% missing values. Similarly, [21] found that the performance of imputation methods decreases when the percentage of MD varies between 5 and 30%.

The factor explaining most of the variance regarding efficiency (Ef) was the percentage of missing values – both additively (for series YFT-DEL, YFT-OBJ, SBF-USA, and PBF-USA) and multiplicatively in interacting with the number of segments (for series SKJ-DEL, SKJ-NOA, and SKJ-OBJ). Additionally, it was found that the maximum operating threshold for this factor was 50%. Above this value, the KSSA failed in all time series. Below this threshold, the efficiency increased as the percentage of MD approached 0%. The observed failure is consistent with other studies, which generally recommend working with time series exhibiting a maximum of 50% MD; it is assumed that, above this percentage, efficiency decreases dramatically. For example, [22] described a procedure to introduce random missing values into the first two datasets, varying the proportion from 2.5 to 50% in 2.5% increments. When this limit is exceeded, the precision of the algorithm is compromised.

Furthermore, increasing the number of iterations homogenized and stabilized the OEA in all time series, providing a better resolution and a more reliable decision regarding the best imputation method. This improved resolution was more noticeable in YFT-OBJ, PBF-USA, and SBF-USA. While the general structure of the OEA was similar in terms of size, position, and shape for any number of iterations, these three properties gained a better resolution with 50 and 100 iterations. This indicates that the decision on the best imputation method was more reliable with a higher number of iterations, especially when increasing from 10 to 50 – between the latter and 100, the differences were minimal or imperceptible.

Except for SKJ-DEL and SKJ-NOA, the OEA covered the majority of the contour plots, indicating the robustness of the KSSA when faced with a wide combination of factors and time series. In SKJ-DEL and SKJ-NOA, the OEA was more restricted, occupying approximately 25% of the contour plot. The issue with SKJ-DEL, whose OEA was located between 35 and 50% of MD, could be due to an atypical CPUE recorded for the 1960s, which exceeded the series average by more than 1000%. For SKJ-NOA, the OEA was located between 10 and 20% of MD. However, this series did not contain any noteworthy atypical data in comparison with SKJ-DEL. These cases highlight the importance of addressing outliers, as they may significantly impact the algorithm's efficiency.

It was observed that the OEA of shorter time series had less regularity. For example, the PBF-USA series, the shortest in length (52), showed more irregular tracings in its 70% Ef contour, even with 100 iterations. Similarly, SBF-USA and YFT-OBJ, with lengths of 56 and 240, respectively, showed patches of non-optimal Ef ($<70\%$) within the OEA for 10 and 50 iterations. Conversely, for longer time series (732), the OEA had a better resolution, even with 10 iterations in some cases (e.g., SKJ-OBJ).

It should be noted that optimal efficiency was achieved with time series exhibiting up to 50% of MD, albeit only with low or minimal segment numbers. For example, in SKJ-DEL and SKJ-OBJ, an optimal Ef at between 40 and 50% MD was reached only when the number of segments was 1 or 2. This may be related to the considerably high level of variance explained by the interaction between the number of segments and the percentage of MD in these series.

The data shown in Table I provide an overview of the various structures of the time series studied. Patterns such as periodicity and trend were identified in some of them, while others showed only periodicity. It is always important to corroborate the presence of these elements in the structure of the series, in order to better understand the functioning of the KSSA; as demonstrated in this study, its efficiency varies depending on the presence or absence of such structures. This is especially useful when seasonality, trend, cycles, and autocorrelations have been well identified.

An important observation is that the performance of the KSSA increases with longer time series but with a restricted number of segments. The algorithm exhibits total failure after surpassing the maximum threshold of this hyperparameter, which was fully identified in Table III but should be determined for each time series. The reason for this failure is that there is insufficient information in each segment for the algorithm to learn, which precludes its initialization.

These maximum thresholds varied mainly between the IATTC and FAO time series, which had monthly and annual resolutions, respectively, as well as different lengths. This suggests that both the length and the temporal resolution of the series influence the efficiency of the KSSA. This makes sense because monthly series have a greater capacity to detect

seasonal cyclicals, associated, for example, with climatic activity, which is highly relevant to fisheries. Fig. 4 shows a common pattern in all time series regarding the response of the OEA to the number of segments, percentage of MD, and iterations. This pattern is a decreasing plateau from left to right, with an inflection point starting at 40%, which is modified only by certain local particularities. The percentage of MD explained most of the of variation in the size, shape, and distribution of the OEA, highlighting its importance for the proper management of the algorithm. Thereupon, we suggest quantifying the percentage of MD before running the KSSA in order to assess the reliability of its results.

The maximum threshold of operation for this factor was 50% in all series, trials, and repetitions, which means that, when the number of MD is greater than that of existing data, the algorithm is unable to find enough information to learn and function. In this vein, we recommend that the results be cautiously interpreted and applied when the percentage of MD in the series of interest is between 40 and 50%. For this range, the decision regarding the best imputation method may have limited reliability. In any case, factors such as the structure of the time series, information on the origin and the recording process of the data, the process that generates MD (MCAR, MAR, or NMAR), and expert knowledge of the studied system should be jointly considered.

The number of iterations did not affect the size, shape, or distribution of the OEA, but it did improve its resolution in all time series. Increasing the number of iterations homogenized and stabilized the results, providing a better resolution and a more reliable decision regarding the best imputation method. We suggest using 50-100 iterations, although this will naturally depend on the processing capacity of each computer and the time available to obtain ready and reliable results. As a reference, on a standard personal-use computer with 8 GB DDR4, 2400 MHz RAM and a 2.0 GHz four-core processor, the KSSA can process a time series of 732 data points with 30% MD in five minutes using five segments.

The KSSA proved to be robust in most time series, as the OEA occupied most of the area in the contour plots. However, there were some exceptions, such as SKJ-DEL and SKJ-NOA, whose OEA was more restricted. These cases could be due to outliers in the data that affected the efficiency of the algorithm. Therefore, we suggest caution with the application of the KSSA in time series with one or a few outliers, as its results regarding the best imputation method may not be reliable. We also noted a direct relationship between the length of the time series and the regularity of the OEA: shorter series had less regular OEA, while longer ones showed a more defined OEA, even with few iterations.

This study demonstrated that the KSSA is an effective and reliable algorithm to optimally and automatically detect the best imputation method for any given time series, but caution should be exercised when the percentage of MD is between 40 and 50%. Before its publication as an R package in 2022 and as a scientific article in 2023, there was no such tool.

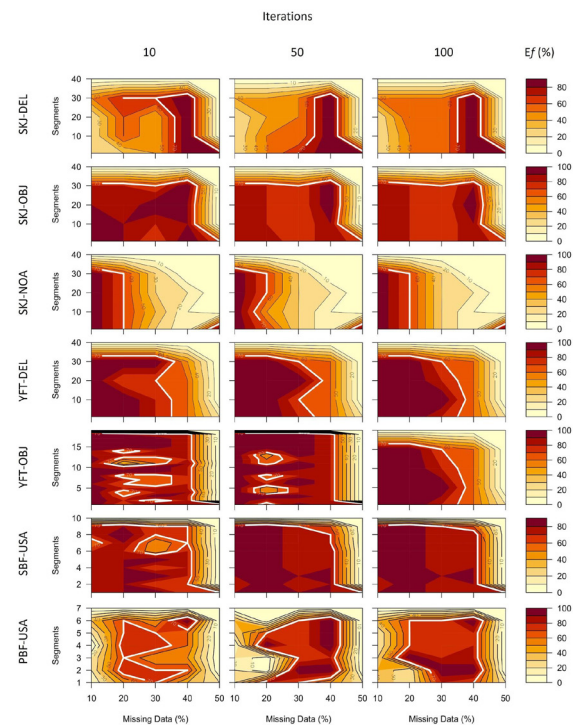


Figure 3. Contour plots representing the efficiency (E_f) of the KSSA in response to the time series, the number of segments, the percentage of missing data, and the number of iterations, using RMSE as the performance metric

Source: Authors

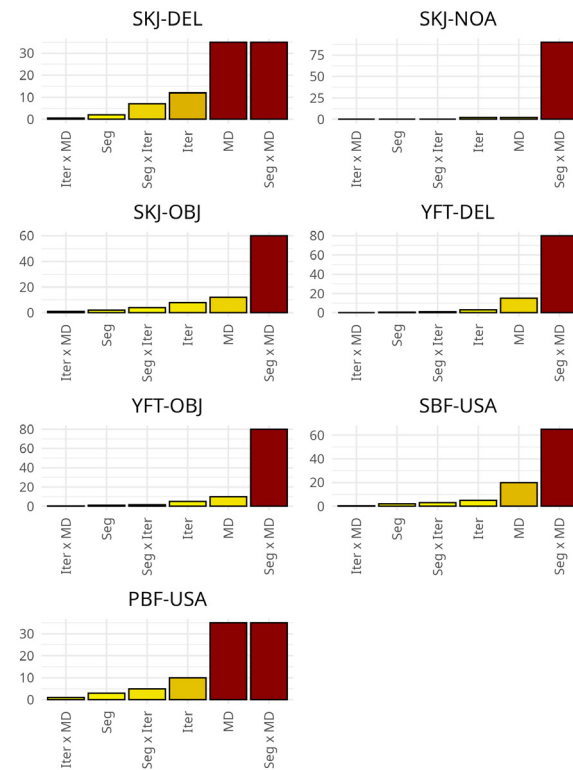


Figure 4. Portion of the variance explained by each source of variation and their double interactions regarding efficiency (E_f) for each time series. The values on the vertical axis represent the percentage of variance explained, according to a multifactorial ANOVA.

Source: Authors

Conclusions

The efficiency of the KSSA varies depending on the structure of the time series under study, including elements such as periodicity, trend, cycles, and autocorrelation. It performs better with longer time series but requires a restricted number of segments. The performance of the algorithm is significantly affected when the number of segments exceeds a certain threshold, as identified in this study. This failure occurs due to insufficient information in each segment for the algorithm to learn, which hinders its initialization.

Both the length and the temporal resolution of the time series influence the efficiency of the KSSA. Monthly series, for example, are more effective in detecting seasonal cycles that are relevant to fisheries. The results indicated that both length and temporal resolution are crucial for the algorithm's performance, as demonstrated by the differences between the IACCT and FAO time series, with monthly and annual resolutions, respectively.

The performance of the KSSA decreases significantly when the percentage of MD exceeds 50%. For MD percentages between 40 and 50%, the imputation results may be less reliable, underscoring the need to quantify this aspect before running the algorithm. Additionally, increasing the number of iterations improves the resolution and stability of the KSSA's results. We suggest using between 50 and 100 iterations depending on the processing capacity and time available.

The KSSA proved to be robust for most time series, with the optimal efficiency area (OEA) occupying most of the contour plots. However, caution is advised when applying the KSSA to time series with outliers, as these can affect the algorithm's efficiency. We also noted a direct relationship between the length of the time series and the regularity of the OEA, with shorter series showing a less regular OEA. This study demonstrated that the KSSA is an effective and reliable algorithm for optimally detecting the best imputation method for each time series of interest. However, caution should be exercised when the percentage of MD is between 40 and 50%. Prior to its publication as an R package in 2022 and as a scientific article in 2023, there was no tool that efficiently performed this task.

Given its early success, with 2340 downloads and three citations within a year of publication, it is likely that this algorithm will be widely used in time series research across various fields. Selecting the appropriate values for its parameters is crucial for obtaining optimal results, as each time series has unique characteristics. Future studies should include comprehensive analyses aimed at identifying additional sources of variation that may affect the algorithm's performance. For now, the use of KSSA is recommended for imputing MD in univariate tuna fishing time series.

Acknowledgements

The authors acknowledge the funding provided by Centro de Estudios Interdisciplinarios Básicos y Aplicados (CEIBA), as well as the logistical support from the Pacific Studies Institute of Universidad Nacional de Colombia, Tumaco campus, and from the authors of R's kssa package.

CRedit author statement

- Author 1 conceived the idea and was responsible for the statistical analysis, visualization, and original draft of the article.
- Author 2 supervised the work, was in charge of the methodology and data curation, and was involved in writing the manuscript.
- Author 3 supervised the work and was involved in writing and revising the manuscript.

Data availability

Link: <https://github.com/gjuliannp/KSSA.git>

References

- [1] F. Parra, "Estadística y machine learning con R," 2019. [Online]. Available: <https://bookdown.org/content/2274/series-temporales.html>
- [2] N. Bokde, M. W. Beck, F. M. Álvarez, and K. Kulat, "A novel imputation methodology for time series based on pattern sequence forecasting," *Pattern Recognit. Lett.*, vol. 116, pp. 88–96, 2018. <https://doi.org/10.1016/j.patrec.2018.09.020>
- [3] E. A. Yamoah, U. A. Mueller, S. M. Taylor, and A. J. Fisher, "Missing data imputation of high-resolution temporal climate time series data," *Meteorol. Appl.*, vol. 27, no. 1, Jan. 2020. <https://doi.org/10.1002/met.1873>
- [4] M. W. Beck, N. Bokde, G. Asencio-Cortés, and K. Kulat, "R package imputetestbench to compare imputation methods for univariate time series," *The R Journal*, vol. 10, no. 1, pp. 218–233, 2018. <https://doi.org/10.32614/rj-2018-024>
- [5] S. Moritz, A. Sardá, T. Bartz-Beielstein, M. Zaefferer, and J. Stork, "Comparison of different Methods for Univariate Time Series Imputation in R," 2015. [Online]. Available: <https://doi.org/10.48550/arXiv.1510.03924>
- [6] H. Demirhan and Z. Renwick, "Missing value imputation for short to mid-term horizontal solar irradiance data," *Appl. Energy*, vol. 225, pp. 998–1012, Sep. 2018. <https://doi.org/10.1016/j.apenergy.2018.05.054>
- [7] I. F. Benavides, M. Santacruz, J. P. Romero-Leiton, C. Barreto, and J. J. Selvaraj, "kssa: Known sub-sequence algorithm," *Aquac. Fish.*, vol. 8, no. 5, pp. 587–599, Jun. 2022. <https://doi.org/10.1016/j.AAF.2021.12.013>
- [8] J. Honaker et al., "What to do about missing values in time-series cross-section data," *Am. J. Polit. Sci.*, vol. 54, no. 2, pp. 561–581, 2010.

- [9] N. Golyandina and A. Korobeynikov, "Basic singular spectrum analysis and forecasting with R," *Comput. Stat. Data Anal.*, vol. 71, pp. 934–954, Mar. 2014. <https://doi.org/10.1016/j.csda.2013.04.009>
- [10] I. F. Benavides, M. Santacruz, J. P. Romero-Leiton, C. Barreto, and J. J. Selvaraj, "Assessing methods for multiple imputation of systematic missing data in marine fisheries time series with a new validation algorithm," *Aquac. Fish.*, vol. 8, no. 5, pp. 587–599, Sep. 2023. <https://doi.org/10.1016/j.AAF.2021.12.013>
- [11] T. Pohlert, "Non-parametric trend tests and change-point detection [R package trend version 1.1.5]," 2023, [Online]. Available: <https://CRAN.R-project.org/package=trend>
- [12] R. J. Hyndman and Y. Khandakar, "Automatic time series forecasting: The forecast package for R," *J. Stat. Softw.*, vol. 27, no. 3, pp. 1–22, July 2008. <https://doi.org/10.18637/jss.v027.i03>
- [13] R. C. Team, "R: A language and environment for statistical computing," 2022. [Online]. Available: <https://www.r-project.org/>
- [14] D. Lüdtke, "sjmisc: Data and variable transformation functions," *J. Open Source Softw.*, vol. 3, no. 26, p. 754, Jun. 2018. <https://doi.org/10.21105/JOSS.00754>
- [15] R. Tobias, "missMethods: Methods for Missing Data. R package version 0.3.0," 2022. [Online]. Available: <https://cran.r-project.org/web/packages/missMethods/index.html>
- [16] H. Ben and F. Michael, "Metrics: Evaluation metrics for machine learning. R package metrics version 0.1.4," 2018, [Online]. Available: <https://CRAN.R-project.org/package=Metrics>
- [17] W. Hadley, M. Evan, and S. Danny, "haven: Import and export 'SPSS', 'Stata' and 'SAS' files," 2022, [Online]. Available: <https://CRAN.R-project.org/package=haven>
- [18] P. Christen, "Data linkage: The big picture," *Harv. Data Sci. Rev.*, vol. 1, no. 2, p. 2019, Nov. 2019. <https://doi.org/10.1162/99608F92.84DEB5C4>
- [19] S. Moritz, A. Sardá, T. Bartz-Beielstein, M. Zaefferer, and J. Stork, "Comparison of different methods for univariate time series imputation in R," 2015. [Online]. Available: <https://doi.org/10.48550/arXiv.1510.03924>
- [20] C. Yozgatligil, S. Aslan, C. Iyigun, and I. Batmaz, "Comparison of missing value imputation methods in time series: The case of Turkish meteorological data," *Theor. Appl. Climatol.*, vol. 112, Apr. 2012. <https://doi.org/10.1007/s00704-012-0723-x>
- [21] N. Savarimuthu and S. Karesiddaiah, "An unsupervised neural network approach for imputation of missing values in univariate time series data," *Concurr. Comput. Pract. Exp.*, vol. 33, no. 9, art. e6156, 2021. <https://doi.org/10.1002/cpe.6156>
- [22] R. Wei et al., "Missing value imputation approach for mass spectrometry-based metabolomics data," *Sci. Rep.*, vol. 8, no. 1, art. 663, Jan. 2018. <https://doi.org/10.1038/s41598-017-19120-0>