Learning and Nonlinear Models - Journal of the Brazilian Society on Computational Intelligence (SBIC), Vol. 20, Iss. 1, pp. 31-46, 2022

Missing Data in Time Series: A Review of Imputation Methods and Case Study

Silvana Mara Ribeiro, Cristiano Leite de Castro
Graduate Program in Electrical Engineering - Universidade Federal de Minas Gerais
Av. Antonio Carlos 6627, 31270-901, Belo Horizonte, MG, Brazil
silvanaribeiro@ufmg.br, crislcastro@ufmg.br

Abstract – Dealing with missingness in time series data is a very important, but oftentimes overlooked, step in data analysis. In this paper, the nature of time series data and missingness mechanisms are described to help identify which imputation method should be used to impute missing data, along with a review of imputation methods and how they work. Recommended methods from literature are used to impute synthetic data of different nature and the results are discussed. In addition, a case study concerning the prediction (classification) of US market instability (BEAR or BULL) using a data set with mixed missingness mechanisms and mixed nature is presented to evaluate how different types of imputation methods can affect the final results of the classification task.

Keywords – Missing Data, Time Series, Imputation Methods, Missingness Mechanisms, Time Series Nature.

1. INTRODUCTION

Missing data is a common problem in data acquisition. Given that data can rarely be perfectly collected and many problems such as sensor and transmission failure, human error, and even differences in the rate of the collection might occur, learning to impute data is an important step in data analysis [1]. In cases in which it is necessary to join sequential data from different sources, the obtained data set frequently becomes full of missing data [2].

When missingness occurs in time series data, the problem can comprehend univariate or multivariate time series, different missingness mechanisms, and different time series nature. A good understanding of these factors can be helpful when choosing the appropriate imputation method. Choosing methods that take into account the temporal information intrinsic to time series data is also important [3].

A lot of research has already been done on the subject of time series imputation, treating specific data sets and comparing the effectiveness of imputation methods. Research using traffic data has been reported [4] to mitigate the challenges that Intelligent Transportation Systems encounter due to missingness that affects the ability to predict traffic. Another study used meteorological data to evaluate different (conventional and modern) imputation methods in time series [5]. A study [6] tackling data imputation in air quality data was conducted testing different imputation methods to mitigate bias and inefficiency in modeling. More recent works [7] [8] focus on comparing deep learning and/or machine learning imputation methods in time series.

The goal of this paper is to present a literature review of the proper imputation methods to deal with missing data in time series and to present the impact of different imputation methods in the results of a case study using financial indexes and instability trackers information. In a very instructive and procedural way, this paper presents the recommended steps to be taken when treating missingness in time series data, alongside notebooks to illustrate these steps and help apply them to other problems involving time series.

An experiment using synthetic data is presented to compare the effectiveness of literature recommended methods when applied to univariate time series of different nature. Regarding multivariate time series, a case study is conducted to verify how the imputation methods affect the final results. The case study aims to predict US market instability of future weeks (BEAR or BULL) using stock market indexes and instability trackers [9] [10]. The data set is full of missing data with mixed missingness mechanisms and the time series features are of different nature.

The remainder of this paper is organized as follows: Section 2 presents the missing data imputation problem formulation and theoretical background of the types of time series and missingness mechanisms, Section 3 presents the main methods of time series data imputation. Section 4 presents an experiment done with synthetic data, comparing how the literature recommended imputation methods performed when imputing univariate time series of different nature. In Section 5 the data used in the case study and the case study itself are explained. In section 6, the methodology used to compare the selected methods for the case study is presented, along with details of implementation. Section 7 shows the results found and Section 8 presents the conclusions.

2. THEORETICAL BACKGROUND

Let X be a data set of dimensions n rows (samples) $\times$ d columns (features) and M be the mask matrix, with the same dimensions $n \times d$, that only takes values in $\{0, 1\}^{d \times n}$. The mask matrix represents which values are missing from the X matrix by placing the value 1 at the position of the missing values. Each value of the imputed matrix $\tilde{X}$ for all i and j can be defined by Equation 1 [11], as follows:

$$\tilde{X}_{i,j} = \begin{cases} X_{i,j}, & \text{if } M_{i,j} = 0 \\ *, & \text{if } M_{i,j} = 1 \end{cases} \tag{1}$$

The goal is to create a new data set $\tilde{X}$ in which all positions marked as one on the mask M are imputed by some imputation method and the values from all other positions come from the corresponding position in X.

Hereafter, a row of a data set will be referred to as a sample and a column of a data set will be referred to as a feature. In addition, data that cannot be directly measured [12] will be referred to as unobserved data, whereas data that can be directly measured will be referred to as observed data. As an example, when dealing with hospital patients, observed data would be the patient's temperature, weight, and height, whereas unobserved data would be pain level, health, and well-being.

When dealing with missing data in time series it is important to identify the missingness mechanism of the data to choose the most appropriate method for imputation [2]. Further, it might be helpful to identify the characteristics of the series. The following concepts regarding the characteristics of time series data were extracted mainly from the book "Forecasting: principles and practice" [13].

A time series might be trended, meaning it presents a long decrease or increase with time, linear or not, and it can even change from increasing to decreasing. An example of trended data is presented in Figure 1.

A time series might be seasonal, meaning it is affected by seasonal factors of a fixed and known frequency. Seasonal data should not be confused with cyclic data, which is defined by data that rises and falls in intervals of unfixed length. An example of seasonal data is presented in Figure 2.

A white noise time series shows none, or close to none autocorrelation [13]. An example of white noise data is presented in Figure 3.

A time series might also be a combination of seasonal and trended data. An example of seasonal and trended data is presented in Figure 4.

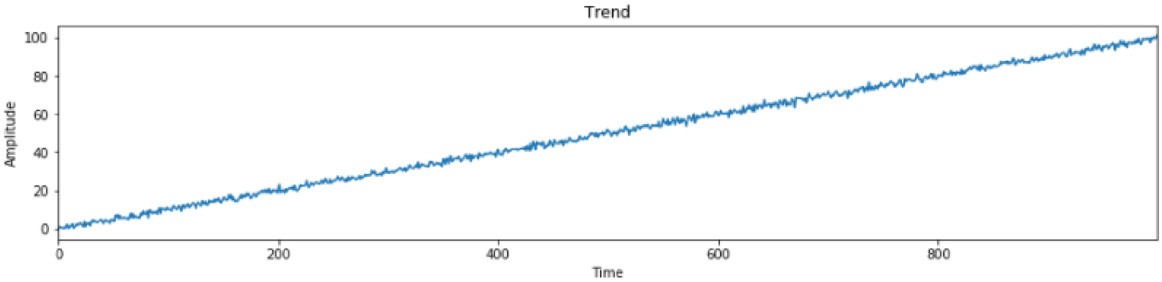

Figure 1: Time series with trend and without seasonality

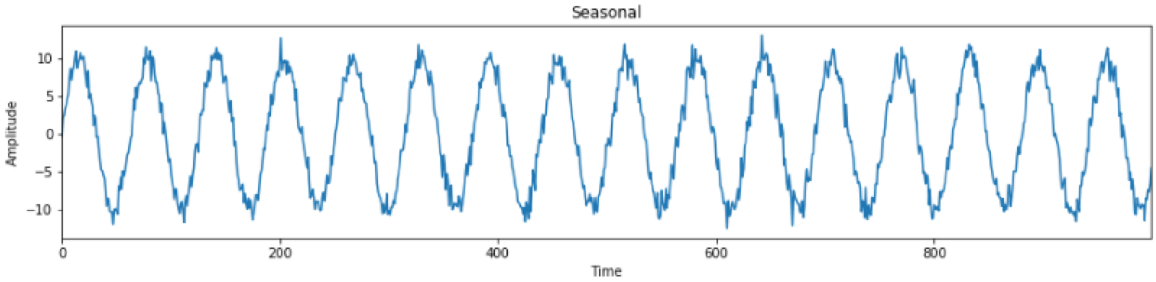

Figure 2: Time series without trend and with seasonality

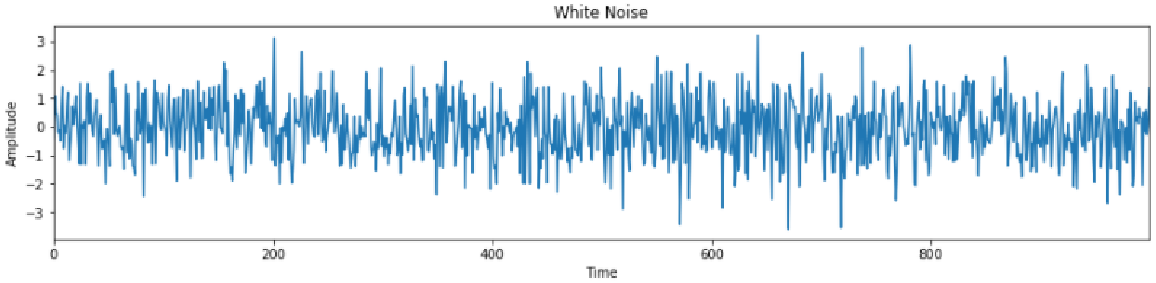

Figure 3: White Noise

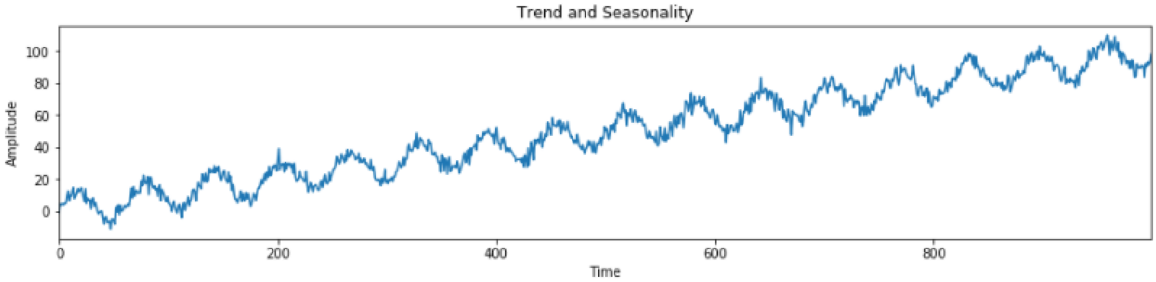

Figure 4: Time series with trend and with seasonality

An intuitive and visual way to identify trend and seasonality in time series is to calculate and plot its autocorrelation function, which measures the linear relationship between lagged values of a time series.

The autocorrelation plot for a trended time series slowly decreases as the lag increases, as can be seen in Figure 5, which is the autocorrelation function of the time series shown in Figure 1.

The autocorrelation plot for a seasonal time series presents larger values for the multiples of the seasonal frequency, as can be seen in Figure 6, which is the autocorrelation function of the time series shown in Figure 2.

The autocorrelation plot for a seasonal and trended time series presents a combination of the aforementioned effects, as can be seen in Figure 7, which is the autocorrelation function of the time series shown in Figure 4.

The autocorrelation plot for a white noise time series is expected to have 95% of the spikes to lie within an interval of $\pm 1.96/\sqrt{T}$, where T is the length of the time series, as can be seen in Figure 8, which is the autocorrelation function of the time series shown in Figure 3.

Besides visual analysis, statistical tests usually applied to verify the characteristics of the time series are:

- Cox-Stuart Test: Recommended to verify the existence of increasing or decreasing trend in time series data, this test is based on the hypothesis that there will not exist any trend if the probability of one sample being smaller than its successor is the same as the probability of this same sample being greater than its successor for the majority of the samples in the data. Otherwise, the data is trended [14].

- Autocorrelation Testing of a certain order: Recommended to verify the existence of seasonality in time series data, this test verifies if there is a correlation between the signal and itself for a certain order. For example, if there exists yearly seasonality in a series, the correlation of order twelve will be meaningful [13].

- Autocorrelation Testing: Recommended to verify if a time series is white noise, this test verifies if there is a correlation between the signal and itself only at time zero. In other words, for a signal to be white noise, the autocorrelation of the signal must be different from zero only for order zero [15].

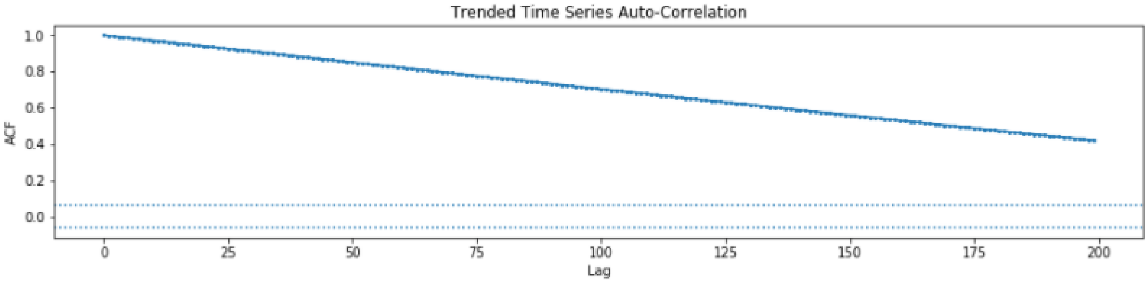

Figure 5: autocorrelation of Time series with trend and without seasonality

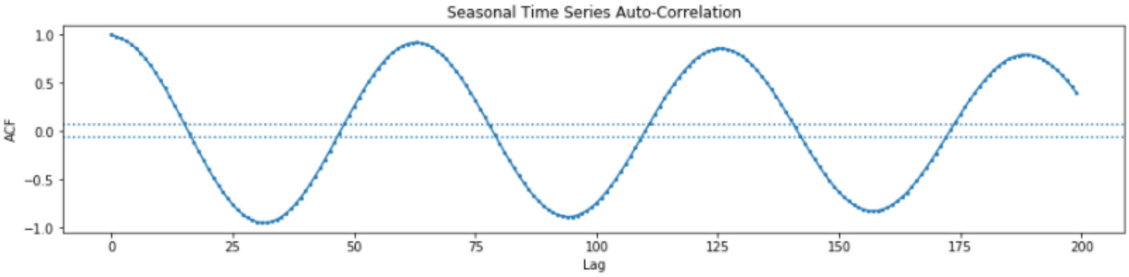

Figure 6: autocorrelation of Time series without trend and with seasonality

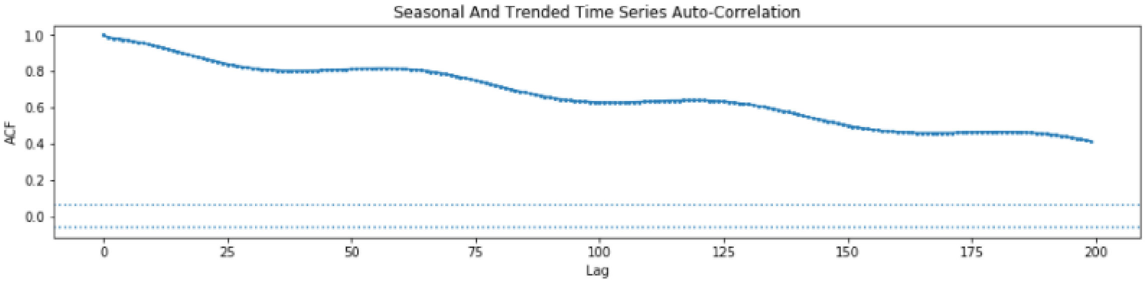

Figure 7: autocorrelation of Time series with trend and with seasonality

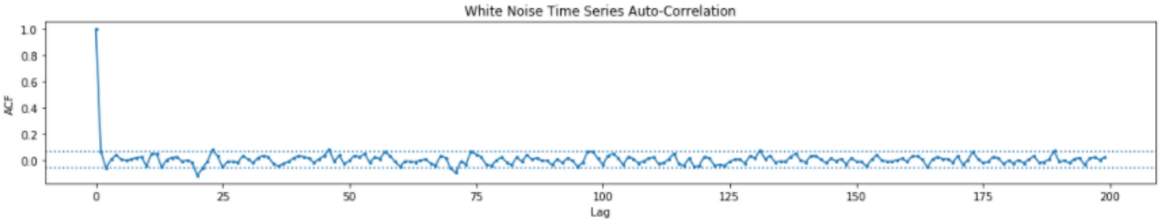

Figure 8: autocorrelation of White Noise

Concerning the missingness mechanisms, the data may suffer from a structural deficiency or the loss may occur randomly.

The structural deficiency can be defined as a component of a feature that was omitted from the data [16]. As an example, in a data set about real state, the lack of information about the number of cars that fit in the garage might mean that, in fact, the building does not have a garage.

In case the loss is random, there are three categories of missingness mechanisms [16]:

- MCAR - Missing Completely At Random: The probability of the data being lost is the same for all data. This probability depends neither on the observed nor the non-observed data. This means that there is no logic behind the loss. The data is lost by a random process, e.g, as a result of a sensor failure.

- MAR - Missing At Random: The probability of the data being lost is not the same for all data. This probability depends on the observed data, but not on the non-observed. That means that the absence of the data can be predicted by the observed data. For example, consider a data set about income and education level, people with low education level tend to not inform their incomes. That implies that the absence of the data about income can be predicted by the information about the education level.

- NMAR - Not Missing At Random: The missing data is related to the non-observed data. In other words, the missingness is related to factors not taken into account, for instance, in a data set about income, both the lower and higher incomes are not disclosed by the respondents. It is not possible to define which of the extremities the missing data belongs to nor predict if the data will be informed or not.

3. DATA IMPUTATION METHODS FOR TIME SERIES

Applications that simply delete the samples or features that have missing values or straightforwardly ignore the missingness, can produce biased results and incorrect conclusions about the studies being conducted [6].

The deletion of features and samples might be a viable option if the missingness mechanism is MCAR and the loss rate is low [2]. If this approach cannot be used, it is recommended that the data be imputed.

Performing analyses ignoring the missingness of the data is a risky approach, especially if the loss rate is not low, and if the missingness mechanism is not MCAR [17]. In certain cases, it is not even an option to do so, depending on the application and algorithms being used on the analysis.

Some classical imputation methods work for both time series and other types of data. Some examples are mean imputation, median imputation, mode imputation, and random sample imputation [17].

- Mean imputation, median imputation, and mode imputation: These methods consist of computing the mean, median, and mode, respectively, of the features where the loss occurs and imputing those computed values whenever the missingness occurs. These methods are appropriate when dealing with stationary time series, e.g, white noise. Those methods decrease the variance of the data and therefore affect the standard deviation, which can cause bias [17].

- Random Sample Imputation: This approach uses randomly selected values from the feature in question to do the imputation. The value can be randomly selected from the whole set of possible values or it can be selected randomly from a subset of it. This approach can be appropriate in case the subset is carefully chosen and for series with seasonality and without trend. It is important to mention two methods that are part of the random sample imputation method: Hot Deck Imputation and Cold Deck Imputation [18].

  - Hot Deck Imputation: Finds samples that have similar values on the other features as the sample with a missing value and, among the selected samples, chooses one randomly to impute. This approach restricts the possible values to be imputed to values that have already occurred on the data set and increases variability, resulting in more accurate standard deviations [19].
  - Cold Deck Imputation: This method is similar to Hot Deck Imputation, however, instead of choosing randomly between similar samples within the same source, it replaces the missing values using a different source [20].

Concerning well-known imputation methods specific for time series data, the following methods should be mentioned:

- Forward and Backward Filling: Forward Filling, also known as Last observation (or value) carried forward, imputes the missing value with the last seen observation. Backward Filling, also known as Next Observation (or Value) Carried Backward, imputes the missing value with the next seen observation. These methods assume that adjacent data points are similar. However, this is especially false if there is seasonality on the data [21]. It is important to notice that for problems using streaming data the backward filling method cannot be used, since the next observation is not available immediately to impute the current missing value.

- Multiple Imputation: This method imputes a missing value n-times ($n > 1$), with values representing a probability distribution to reproduce the uncertainty of the value being imputed [22]. Since this method has a random element within itself, these values should be similar, nonetheless different. The data sets with the imputed values are then evaluated and the results are combined so that more realistic conclusions and estimates can be obtained. This method works especially well for imputing survey data, which in general have a MAR missingness mechanism [23]. If made correctly, multiple imputation generates unbiased estimates of the parameters and accurate standard deviations.

- Linear Interpolation: This method assumes that an estimated point will be on the vector that connects the nearest points on the right and left. In other words, it uses a linear function to approximate the function that represents the data well and to compute the missing value [24]. Linear Interpolation also assumes that adjacent data points are similar and, thus, have the same problems as Forward and Backward Filling.

- Spline Interpolation: This method uses polynomial functions to approximate a function that represents the data well and to calculate the missing value [24]. It solves, to a certain extent, the problem of assuming that adjacent data points are similar since it can represent more abrupt variations of the data, although it depends on the polynomials chosen to approximate the function.

- Missing Indicators: This method imputes a default value in every occurrence of missing value and creates a new feature that indicates missingness with the flag 1 on the position where it occurs and 0 where it does not. The results will be biased if the missingness mechanism is not MCAR and if the original features themselves are correlated [25].

- Expectation Maximisation: This method is an iterative procedure that uses the other features to impute a value and then checks if the imputed value is the most probable. In case it is not the most probable, re-computes a new value. This behavior is repeated until the most probable value is found. Although it preserves the relation between features, it underestimates the standard deviation [26].

A few of the most recent methods that should be mentioned when dealing with imputation in time series data are:

- ST-2SMR: This method is composed of two steps to reconstruct missingness in space-temporal data. The first step is a coarse-grained interpolation to eliminate the influence of continuous missing data on the general result. Then, based on the result obtained in the first step, a dynamic selection sliding window algorithm is used to identify the most relevant data to make a fine-grained interpolation. Finally, the results are integrated using a neural network model [27].

- TBM - Temporal Belief Memory: This method deals with missingness using recurrent neural networks [2]. It is an imputation method that, unlike conventional neural networks, does not ignore the real interval between consecutive samples, taking into account time continuity and identifying the lack of data. It computes the belief of the last observed value in time for each feature and imputes the data based on the individual belief both towards the future and the past [2].

- LSTM - Long Short-Term Memory: It is a type of Recurrent Neural Network that can remember past values, affecting the prediction of future values. It has a gating structure that allows information to be retained across time-steps [28]. The prediction is made by modeling the time series by either removing samples that contain missing time-steps, by marking missing time-steps and forcing the network to learn their meaning or by marking them and excluding them from calculations in the model. Afterwards, the values predicted can be used to impute the missing samples. Its memory allows past characteristics in data to be passed on long after it has disappeared in the data. This method has already been applied to the modeling of financial indexes data and has achieved good results [29].

- GAIN - Generative Adversarial Imputation Network: This method is an adaptation of Generative Adversarial Networks. It has a generator that observes components of the real data and generates an imputed vector. Then, a discriminator takes the imputed vector and tries to determine which values are real and which ones were generated. The discriminator is given hints of the real distribution of the data to ensure it forces the generator to improve the imputation. Once the generator succeeds in deceiving the discriminator, the process is terminated [11].

Figure 9 brings a flowchart pointing out the more appropriate imputation methods depending on the characteristics of the data and missingness mechanisms associated.

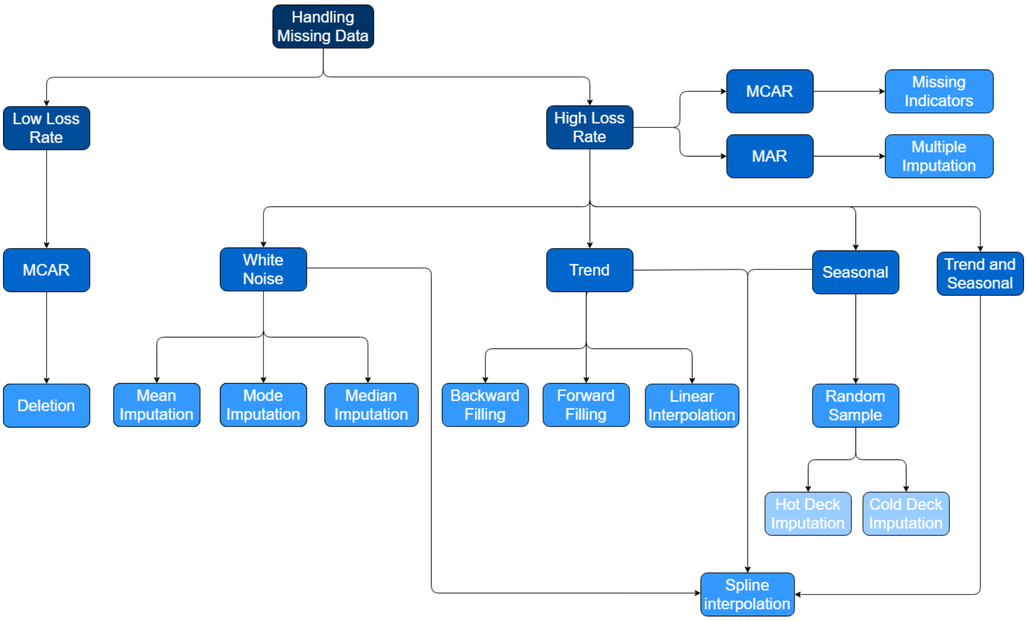

Figure 9: Flowchart of methods to deal with missing data

4 SYNTHETIC DATA EXPERIMENT

To illustrate the effectiveness of some of the methods and verify the recommendations given by the literature, an experiment was conducted using univariate times series such as the ones of Figures 1, 2, 3, and 4. A flowchart representing the steps of the experiment is presented in Figure 10.

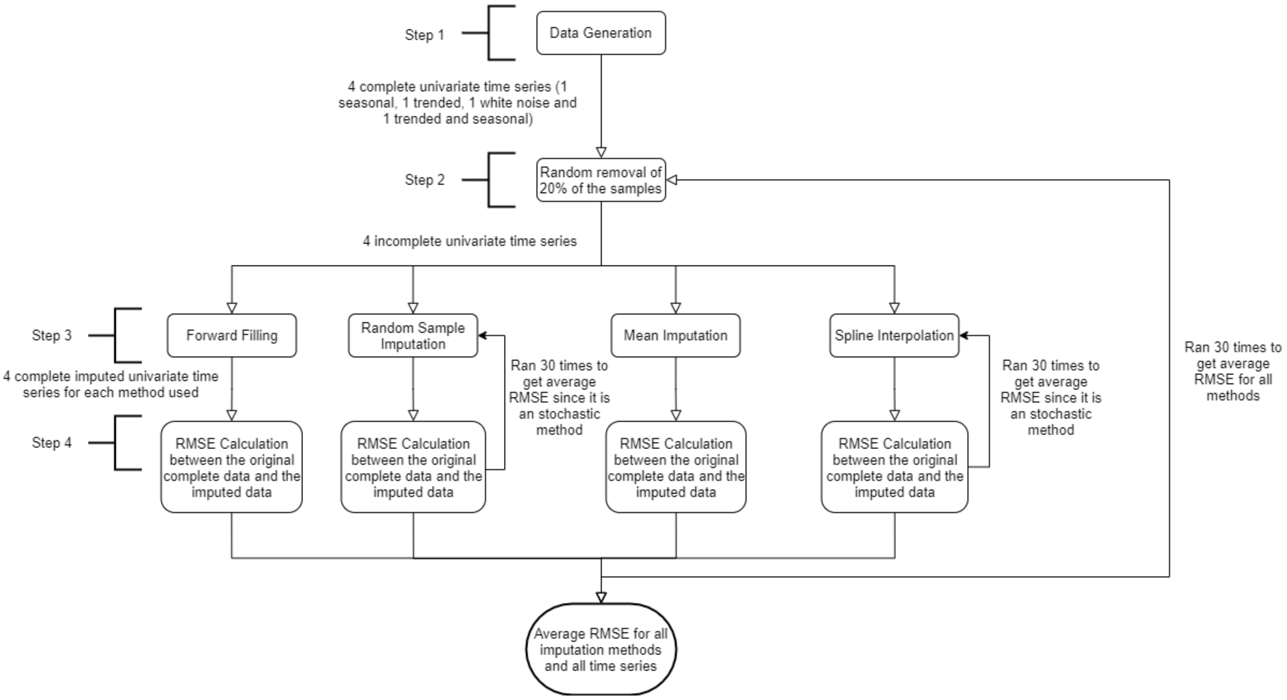
Step 1
Data Generation
4 complete univariate time series (1 seasonal, 1 trended, 1 white noise and 1 trended and seasonal)
Step 2
Random removal of 20% of the samples
4 incomplete univariate time series
Step 3
Forward Filling
Random Sample Imputation
Mean Imputation
Spline Interpolation
4 complete imputed univariate time series for each method used
Ran 30 times to get average RMSE since it is an stochastic method
Ran 30 times to get average RMSE since it is an stochastic method
Ran 30 times to get average RMSE for all methods
Step 4
RMSE Calculation between the original complete data and the imputed data
RMSE Calculation between the original complete data and the imputed data
RMSE Calculation between the original complete data and the imputed data
RMSE Calculation between the original complete data and the imputed data
Average RMSE for all imputation methods and all time series

Figure 10: Flowchart of the Synthetic Data Experiment

The first step of the experiment was to generate four synthetic univariate time series of each nature: one trended, one seasonal, one white-noise, and one trended and seasonal time series.

The second step was the random removal of 20% of the samples of each one of the four time series to emulated missingness, making the missingness mechanism MCAR. Then, using the flowchart shown in Figure 9, one appropriate method was chosen for each type of data. For the seasonal data, the random sample method was chosen. For the trended data, the forward filling method was chosen. For the trended and seasonal data, the spline interpolation method was chosen. For the white-noise data, the mean imputation method was chosen.

Each one of the time series was imputed using all four methods previously chosen, accounting for the third step of the experiment. For reproducibility purposes, the library created with implementations of the imputation methods aforementioned is available[1] and it is called ImputationLibrary. The fourth step consisted of calculating the Root Mean Square Error between the imputed data and the real data.

Considering that Spline Interpolation and Random Sample are stochastic methods, they were run 30 times to compute the average RMSE.

The whole process of removing random samples and imputing them using the chosen imputation methods (steps 2 through 4) was repeated 30 times to calculate the average RMSE. That means that the stochastic methods were actually run 900 times.

For reproducibility purposes, the notebooks in which the experiment was conducted are available[2] and correspond to the notebooks starting with the prefix 00.

The results obtained are shown in Table 1.

Data \ Method	Forward Filling	Random Sample	Mean Imputation	Spline Interpolation
Trend & Seasonal	$\mathbf{3.675 \pm 8.882e^{-16}}$	23.616 ± 2.077	$171.020 \pm 8.527e^{-14}$	$8.421 \pm 1.776e^{-15}$
Seasonal	$\mathbf{0.578 \pm 0.0}$	20.943 ± 1.636	$11.289 \pm 3.553e^{-15}$	$0.624 \pm 2.220e^{-16}$
Trend	$\mathbf{0.414 \pm 1.110e^{-16}}$	19.771 ± 1.497	$9.717 \pm 3.553e^{-15}$	0.432 ± 0.0
White Noise	$0.481 \pm 5.551e^{-17}$	0.403 ± 0.035	$\mathbf{0.213 \pm 0.0}$	0.542 ± 0.0

Table 1: Average RMSE of the imputed data sets for each method

[1] https://github.com/silvanaribeiro/imputationLibrary
[2] https://github.com/silvanaribeiro/MissingDataInTimeSeries

Observing Table 1 it can be seen that the best RMSE obtained for the time series with trend and seasonality was achieved by the Forward filling method, followed by the Spline Interpolation method, which is the one recommended by literature. For the seasonal data, the best result was not achieved by the literature recommended random sample method, it was also achieved by the forward filling method. However, this result was followed closely by the result achieved by the Spline Interpolation method, which is recommended for all cases. For the data with trend, the best result achieved was by the forward filling method, which is the one recommended by the literature. Finally, for white noise data, the best result achieved was by the mean imputation method, which is also the one recommended by literature.

It is important to notice that, as expected, Spline interpolation did a fairly good job for all types of time series. Even for white noise data, where it performed the worst, it is still close to the other methods. This result is expected since the spline interpolation method is recommended by the literature to be used with time series data of all nature.

The results for the random sample imputation method show how difficult it is to choose an appropriate subset from which to randomly select from.

Generally speaking, using the literature recommended methods is the best choice to be made when deciding how to impute missing data.

5 THE CASE STUDY AND THE DATA

The case study consists of generating a model to predict (classify) US Market instability (BEAR or BULL regimens shown in Figure 11) through supervised learning and given a group of stock market indexes and a group of trackers that correlate market volatility to other subjects and keywords mentioned in newspaper articles [9] [10]. However, the data is full of missingness, to begin with, and, as most data sets have different granularities, once merged the problem becomes even greater. Being so, treating the missingness of the data set becomes crucial to achieve the final objective of predicting market instability.

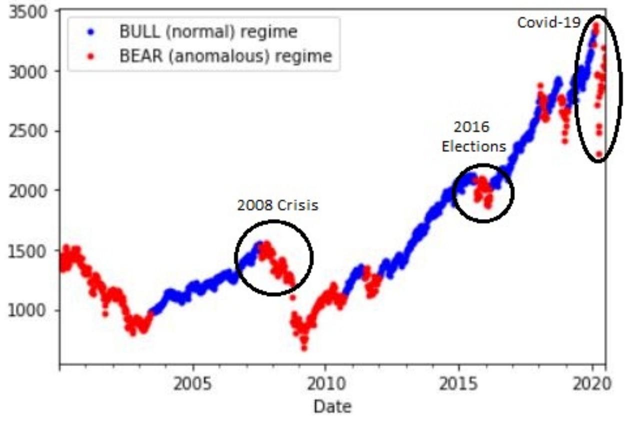

Figure 11: S&P 500 close index classified as BEAR or BULL

The data from financial indexes were taken from Yahoo Finance, and are described as follows [30]:

- Standard and Poor's 500 (S&P 500): Stock Market index that tracks the stocks of the 500 biggest large-capital US Companies.
- Volatility Index (VIX): Index based on S&P 500 that measures the market's expectation of future volatility.
- Dow Jones Industrial Average (Dow 30): Stock Market index that combines the stock price of 30 large, publicly-traded companies to determine the industrial average.
- NASDAQ 100: Stock Market Index that includes the shares of the 100 largest American and international non-financial companies that are traded on the Nasdaq electronic stock exchange.
- NIKKEI 225: Japanese Stock Market Index that includes the shares of Japan's top 225 companies traded on the Tokyo Stock Exchange.
- FTSE 100: Stock Market Index that consists of the 100 most highly capitalized companies in the UK.
- Hang Seng (HSI): Stock Market Index that consists of the largest companies that trade on the Hong Kong Exchange. It covers approximately 65% of the total market capitalization of the Hong Kong Exchange.
- Euronext 100: European Stock Market Index that consists of the 100 largest and most liquid blue-chip stocks traded on Euronext exchanges.

The trackers were taken from Economic Policy Uncertainty and are, as follows:

- US Equity Market Volatility Index: Newspaper-based Equity market Volatility tracker that is constructed based on the counts of the monthly occurrences of the terms ”economic”, ”economy”, ”financial”, ”stock market”, ”equity”, ”equities”, ”Standard and Poors” (and variants), ”volatility”, ”volatile”, ”uncertain”, ”risk” and ”risky” on eleven major US newspapers [31]. This tracker has monthly granularity.

- Daily Infectious Disease Equity Market Volatility Tracker: Newspaper-based Infectious Disease Equity Market Volatility Tracker that is constructed based on the counts of the daily occurrences of the terms ”economic”, ”economy”, ”financial”, ”stock market”, ”equity”, ”equities”, ”Standard and Poors” (and variants), ”volatility”, ”volatile”, ”uncertain”, ”risk”, ”risky”, ”epidemic”, ”pandemic”, ”virus”, ”flu”, ”disease”, ”coronavirus”, ”mers”, ”sars”, ”ebola”, ”H5N1” and ”H1N1” on approximately 3,000 US Newspapers [31]. This tracker has daily granularity.

- Geopolitical Risk Index: Newspaper-based Geopolitical Risk Index that is constructed by counting the occurrence of words related to geopolitical tensions in 11 leading international newspapers [32]. This index has monthly granularity.

- Trade Policy Uncertainty and Market Volatility: Newspaper-based Trade Policy Uncertainty tracker that is constructed based on the counts of the frequency of occurrences of trade policy and uncertainty terms across major newspapers over the world [33]. This tracker has monthly granularity.

6 METHODOLOGY OF THE CASE STUDY

As the granularity of the data being analyzed is different in each data set used, once they are merged the resulting data set becomes full of missing values. Some of the individual data sets already present missingness even before the merge. As the missingness of the data can, in some cases, be predicted by the temporal feature, but there is also missingness that cannot be predicted by the observed data, the missingness mechanism is considered a combination of MAR and MCAR.

The methodology can be described by the following steps and illustrated by Figure 12:

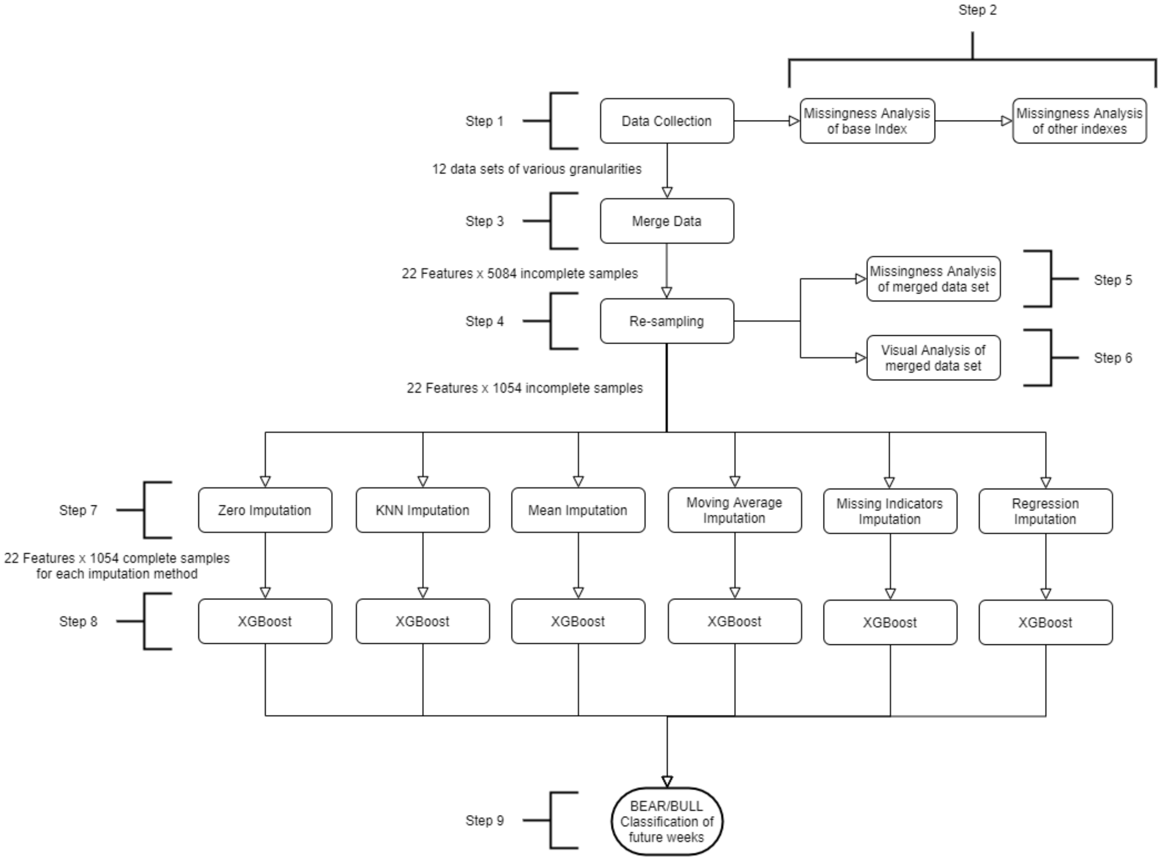
Step 2
Step 1
Data Collection
Missingness Analysis of base Index
Missingness Analysis of other indexes
12 data sets of various granularities
Step 3
Merge Data
22 Features x 5084 incomplete samples
Step 4
Re-sampling
Missingness Analysis of merged data set
Step 5
Visual Analysis of merged data set
Step 6
22 Features x 1054 incomplete samples
Step 7
Zero Imputation
KNN Imputation
Mean Imputation
Moving Average Imputation
Missing Indicators Imputation
Regression Imputation
22 Features x 1054 complete samples for each imputation method
Step 8
XGBoost
XGBoost
XGBoost
XGBoost
XGBoost
XGBoost
Step 9
BEAR/BULL Classification of future weeks

Figure 12: Methodology illustrated as the steps taken to conduct the case study

1. Collection of the data: The data was collected manually from all sources.

2. Missingness Analysis of each data set: Using the S&P500 index as a base, the dates for which a value was expected for all indexes were set and the analysis was made. This step is available at the repository[3] and corresponds to the notebooks starting with prefix 02. Some of the missingness occurs because different markets open on different days and some occur because the data is not available. The results obtained for each data set were as shown in Table 2.

Data set	Missing dates	Missing values
S&P 500	0	0
VIX	0	0
Dow 30	0	0
NASDAQ 100	0	0
NIKKEI 225	173	124
FTSE 100	283	61
HSI	165	84
Euronext 100	37	32
Us Market Volatility Index	NA	0
Daily Infectious Disease	NA	0
Geopolitical Risk Index	NA	0
Trade Policy uncertainty	NA	24

Table 2: Missingness analysis of each feature in relation to S&P 500 before merging the data sets

3. Merging all data: The data were merged as illustrated by Figure 13. From the financial Indexes data sets, the close and volume features were used from each index. From the trackers, the features used were the Overal EMV tracker, Daily Infectious EMV index, GPR, US Trade Policy Uncertainty, Japanese Trade Policy Uncertainty, and Trade Policy Uncertainty EMV Fraction. The obtained data set has 22 features and 7383 samples. The period of analysis ranges from January 1, 2000, to March 13, 2020. This step is available at the repository and corresponds to the notebook starting with the prefix 03.

4. Re-sampling: The daily data set was re-sampled to weekly granularity by calculating each week's average. The final data set is composed of 1054 dates (rows or samples) and 22 features (columns). The period of analysis ranges from January 1, 2000, to March 13, 2020. This step is available at the repository and also corresponds to the notebook starting with prefix 03.

5. Missingness Analysis of the merged and re-sampled data set: After merging the 12 data sets of different granularity and re-sampling it, another analysis was made to verify missingness. This step is available at the repository and corresponds to the notebook starting with the prefix 04. The results are presented by Table 3

Feature	Missing values
S&P 500 Close and Volume	0
VIX Close and Volume	0
Dow 30 Close and Volume	0
NASDAQ 100 Close and Volume	0
NIKKEI 225 Close and Volume	2
FTSE 100 Close and Volume	45
HSI Close and Volume	0
Euronext 100 Close and Volume	0
Us Market Volatility Index	813
Daily Infectious Disease	0
Geopolitical Risk Index	813
Trade Policy uncertainty	820
US Trade Policy uncertainty	820
Japanese Trade Policy uncertainty	820

Table 3: Missingness analysis of each feature after merging and re-sampling the data sets

6. Visual Analysis of the features: The autocorrelation plot of every feature was analyzed as shown in Table 4. It was

[3] `https://github.com/silvanaribeiro/MissingDataInTimeSeries`

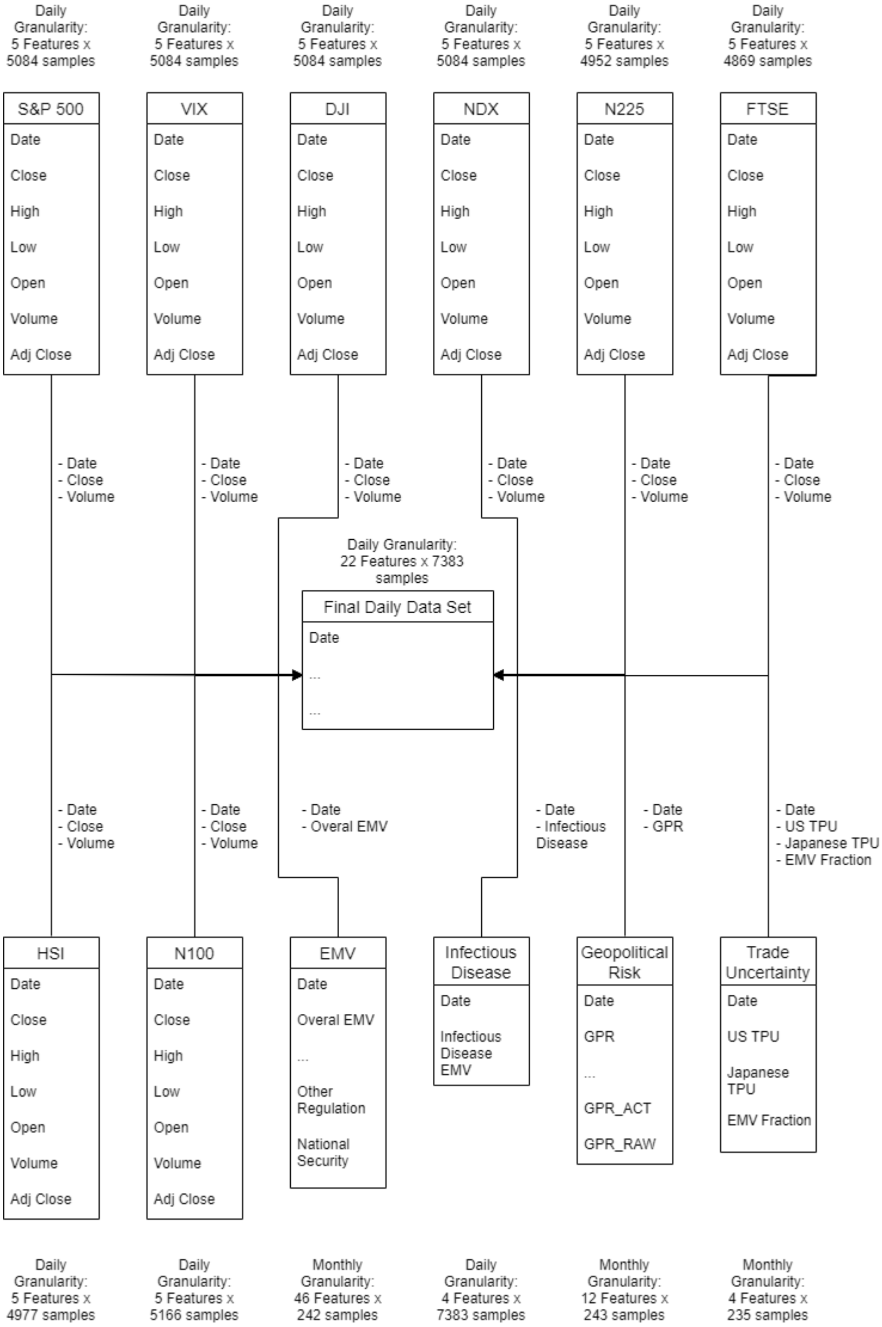

Figure 13: Diagram showing information about the original data sets and the final daily data set after the merger

concluded that the data set is a mixture of all types of time series, which makes it difficult to choose one method that would fit all cases. The autocorrelation plots of the features can be found at the repository and correspond to the notebook 05.

Feature	Trended	Seasonal	White-noise
S&P 500 Close and Volume	X		
VIX Close and Volume	X		
Dow 30 Close and Volume	X		
NASDAQ 100 Close and Volume	X		
NIKKEI 225 Close and Volume	X		
FTSE 100 Close and Volume	X		
HSI Close and Volume	X		
Euronext 100 Close and Volume	X		
Us Market Volatility Index	X	X	
Daily Infectious Disease			X
Geopolitical Risk Index	X	X	
Trade Policy uncertainty	X	X	
US Trade Policy uncertainty	X	X	
Japanese Trade Policy uncertainty	X		

Table 4: Visual analysis of the autocorrelation plot of each feature

7. Imputation of missing data: Using six different imputation methods to impute data sets of weekly granularity. It is important to notice that at this point the data set had already been separated into training (75% of all data - 782 samples) and test (25% of all data - 272 samples) data sets. Six different complete data sets were generated after this step. This step is available at the repository and corresponds to the notebooks starting with the prefix 06.

   - Zero Imputation: The missing values were all imputed using the value zero. This is the baseline method for comparison.
   - K-Nearest Neighbours: Although this particular method is usually designed for matrix data instead of time series [34], it was decided to test it to evaluate how such method would perform compared to more appropriate ones according to literature. The KNN Imputer from the Scikit Learn library[4] was used. Each missing value is imputed using the mean of the K nearest neighbors to the sample being imputed. For the test set, only data from the training set was used for the imputation. Before using this method it was necessary to scale the data set.
   - Mean Imputation: The Simple Imputer from the Scikit Learn Library[5] was used. For the imputation of the test set only the mean from the training set was used.
   - Moving Average Imputation: A simple bilateral moving average algorithm was implemented. Three non-null values from before and after the missing value are averaged and the result is imputed. If the empty values occurred at the beginning of the series, it was used as many values prior to the missing one as possible (perhaps smaller than 3) and the same is true for the ending of the series (it was used as many values posterior to the missing one as possible). To impute the test data, only values prior to the one being imputed were used.
   - Missing Indicators Imputation: All missing values were imputed by the value zero. A new mask feature was added to the data set with the value one indicating that there is missingness in that position on the original feature. Otherwise, the position on the mask feature has value zero.
   - Regression Imputation: The Iterative Imputer from the Scikit Learn Library[6] was used. At each step, a feature is designated as the output y, and the other features are treated as the input X. Then, a regressor is fit on (X,y) for known y and this regressor is used to predict the missing values of y. For the test set, only data from the training set was used for the imputation. Before using this method it was necessary to scale the data set.

8. Train Models using XGBoost Algorithm for each imputed data set: XGBoost[7] is an optimized gradient boosting library. It uses the Gradient Boosting framework [35], providing a parallel tree boosting. XGBoost is such a successful framework primarily because of its scalability in all scenarios. Some of the advantages of using XGBoost are its speed, customization, and performance. As XGBoost can automatically do parallel computing on most operational systems, its speed is generally

[4] KNN Imputer. Scikit Learn Version 23. https://scikit-learn.org/stable/modules/generated/sklearn.impute.KNNImputer.html?highlight=knn#sklearn.impute.KNNImputer

[5] Simple Imputer. Scikit Learn Version 23. https://scikit-learn.org/stable/modules/generated/sklearn.impute.SimpleImputer.html

[6] Iterative Imputer. Scikit Learn Version 23. https://scikit-learn.org/stable/modules/generated/sklearn.impute.IterativeImputer.html

[7] ”XGBoost Version 1.1. ”https://xgboost.readthedocs.io/en/latest/build.html

10 times faster than gbm. It is also possible to customize the objective and evaluation functions. As it has been used in several competitions, such as the ones in Kaggle, and it has won several times while being used with different types of data sets, its performance has been proven to be better in several instances [36]. This step is available at the repository and corresponds to the notebooks starting with the prefix 07.

The tunning of the hyperparameters was done manually and separately for each imputed data set. The usage of each hyperparameter can be found in the library documentation[8]. The hyperparameters that were tunned, as well as the ranges tested, were as follows:

- colsample_bytree: 0.3-1, step 0.1.
- gamma: 0-10, step 1.
- learning_rate: 0.1, 0.01, 0.001, 0.0001, 0.00001.
- max_delta_step: 0-10, step 1.
- max_depth: 0-10, step 1.
- min_child_weight: 1-10, step 1.
- n_estimators: 30-100, step 10, 100-400, step 100.
- reg_alpha: 0.1, 0.01, 0.001, 0.0001, 0.00001.
- scale_pos_weight: 1, 10, 20, 50-1000, step 50.
- subsample: 0.3-1, step 0.1.

9. Make predictions: After finding the best parameters for each one of the six data sets, the test set for each one was used to make predictions and evaluate the models. This step is available at the repository and corresponds to the notebooks starting with the prefix 07.

7. RESULTS

For reproducibility purposes, the notebooks in which the experiments were conducted are available at the repository[9] and comprehend the notebooks starting with prefix 01 through 07.

Zero Imputed		True	False
	True	49	19
	False	16	188
KNN Imputed		True	False
	True	49	24
	False	16	183
Mean Imputed		True	False
	True	49	17
	False	16	190
Moving Average Imputed		True	False
	True	0	0
	False	65	207
Missing Indicators Imputed		True	False
	True	50	20
	False	15	187
Regression Imputed		True	False
	True	50	21
	False	15	186

Table 5: Confusion Matrices of predictions made using all six imputation methods

Zero Imputed	KNN	Mean	Moving Average	Regression	Missing Indicators
0.7368	0.7101	0.7480	0	0.7407	0.7352

Table 6: Comparison of the F1 Scores found by the models generated using each imputed data set.

[8] ”XGBoost Parameters” https://xgboost.readthedocs.io/en/latest/parameter.html

[9] https://github.com/silvanaribeiro/MissingDataInTimeSeries

As it can be observed in Table 6, the best result was achieved by the model generated using the Mean Imputed data set, followed closely by the model generated using the Regression and Missing Indicators Imputed data sets. As expected, the model generated using the KNN Imputed data set did not achieve the best result, but it did not achieve the worst either. The model generated using the Moving Average Imputed data set could not classify any positive labeled sample correctly. The model generated using the Missing Indicators data set, KNN imputed data set and the one generated using the Moving Average Imputed data set achieved worse results than the model generated using the baseline data set, zero imputed. Therefore, those methods did not work well for imputing this problem's data. The prediction versus true values of the two best models can be seen in Figures 14 and 15.

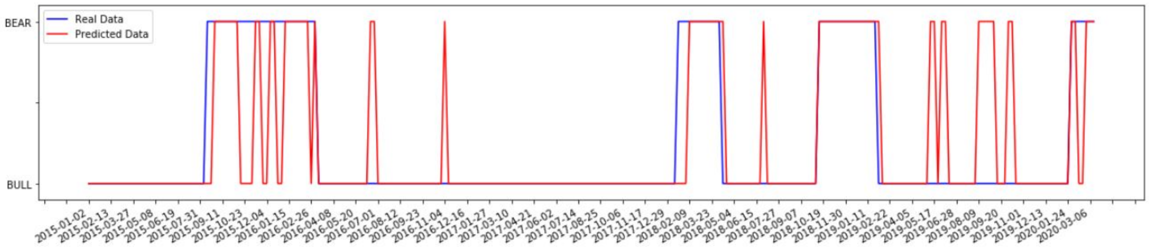

Figure 14: BULL and BEAR real data and prediction made by model generated using the Mean Imputed data set

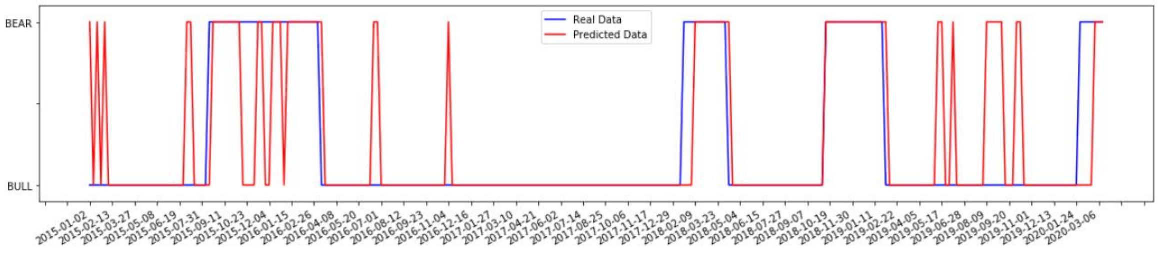

Figure 15: BULL and BEAR real data and prediction made by model generated using the Regression Imputed data set

8. CONCLUSION

Literature recommended methods work very well for well-behaved data. However, when trying to impute a real data set with mixed types of time series and/or different types of missingness mechanisms it is very important to try several different methods and choose the one that fits the problem better. In the specific case of the case study of this paper, an Imputation method as simple as Mean Imputation did a better job than a relatively complex imputation method such as Moving Average. Not always using more complex and sophisticated methods means that better results will be achieved for all problems.

The process of Imputing missing values is a commonly overlooked step when dealing with data and the case study showed that the wrong imputation method can heavily affect the results of predictions. Therefore it is recommended that some time be spent to figure out the characteristics of the data features and their missingness to choose an appropriate imputation method, and in case of mixed data sets, testing different methods before settling for one.

REFERENCES

[1] Y. Hang, S. Fong and W. Chen. "Aerial Root Classifiers for Predicting Missing Values in Data Stream Decision Tree Classification". 04 2011.

[2] Y. J. Kim and M. Chi. "Temporal Belief Memory: Imputing Missing Data during RNN Training". *IJICAI, Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, 2018.

[3] Y. Luo et al. "Multivariate Time Series Imputation with Generative Adversarial Networks". *32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*, 2019.

[4] H. Z. Wu P., Xu L. "Imputation Methods Used in Missing Traffic Data: A Literature Review". *ISICA 2019. Communications in Computer and Information Science*, vol. 1205, 2020.

[5] A. S. I. C. e. a. Yozgatligil, C. "Comparison of missing value imputation methods in time series: the case of Turkish meteorological data." *Theor Appl Climatol 112, 143â167*, vol. 1205, 2013.

[6] L. L. Wijesekara W.M.L.K.N. "Comparison of Imputation Methods for Missing Values in Air Pollution Data: Case Study on Sydney Air Quality Index." *Advances in Intelligent Systems and Computing,*, vol. 1130, 2020.

[7] M. Saad, L. Nassar, F. Karray and V. Gaudet. "Tackling Imputation Across Time Series Models Using Deep Learning and Ensemble Learning". In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 3084–3090, 2020.

[8] M. Saad, M. Chaudhary, F. Karray and V. Gaudet. "Machine Learning Based Approaches for Imputation in Time Series Data and their Impact on Forecasting". In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 2621–2627, 2020.

[9] J. Chen and E. P. K. Tsang. "Classification of Normal and Abnormal Regimes in Financial Markets." *MDPI, Multidisciplinary Digital Publishing Institute*, 2018.

[10] J. Chen. "Studying Regime Change Using Directional Change". Ph.D. thesis, University of Essex, 2019.

[11] J. Yoon, J. Jordon and M. van der Schaar. "GAIN: Missing Data Imputation using Generative Adversarial Nets". In *Proceedings of the 35th International Conference on Machine Learning*, edited by J. Dy and A. Krause, volume 80 of *Proceedings of Machine Learning Research*, pp. 5689–5698, StockholmsmÃ¤ssan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[12] M. A. Santos Curado et al. "Analysis of Variables That Are Not Directly Observable: Influence on Decision-Making during the Research Process." *Rev. Esc. Enferm. USP*, 2014.

[13] R. Hyndman and G. Athanasopoulos. *"Forecasting: principles and practice", 2nd edition*. OTexts, Melbourne, Australia, 2018.

[14] M. V. Silva Gurgel do Amaral. "Ajuste De Modelos e Comparacao De Series Temporais Para Dados De Vazao Especefica Em Microbacias Pareadas."

[15] L. A. Aguirre. *"Introducao a Identificacao de Sistemas: Tecnicas Lineares e nao-Lineares Aplicadas a Sistemas Reais."*. Ed. UFMG, 2004.

[16] M. Kuhn and K. Johnson. *"Feature Engineering and Selection: a Practical Approach for Predictive Models"*. CRC Press, 2020.

[17] I. Pratama et al. "A Review of Missing Values Handling Methods on Time-Series Data." *International Conference on Information Technology and Innovation (ICITSI)*, 2016.

[18] G. Kalton and L. Kish. "Some efficient random imputation methods." *Communications in Statistics - Theory and Methods*, 1984.

[19] R. R. Andridge and R. J. A. Little. "A Review of Hot Deck Imputation for Survey Non-Response." *Wiley Online Library, John Wileya and Sons*, 2010.

[20] M. Duma et al. "Partial Imputation of Unseen Records to Improve Classification Using a Hybrid Multi-Layered Artificial Immune System and Genetic Algorithm." *Applied Soft Computing, Elsevier*, 2013.

[21] M. F. J. et al. "Does analysis using last observation carried forward introduce bias in dementia research?" *CMAJ*, 2008.

[22] M. J. Azur et al. "Multiple Imputation by Chained Equations: What Is It and How Does It Work?" *International Journal of Methods in Psychiatric Research*, 2011.

[23] R. D. Little and Rublin. *"Statistical Analysis with Missing Data."*. Wiley, New York, 1987.

[24] R. K. Elissavet. "Missing Data in Time Series and Imputation Methods". Master's thesis, University of the Aegean, Samos, February 2017.

[25] R. Groenwold et al. "Missing Covariate Data in Clinical Research: When and When Not to Use the Missing-Indicator Method for Analysis." *Canadian Medical Association*, 2012.

[26] A. P. Dempster et al. "Maximum Likelihood from Incomplete Data Via the EM Algorithm". *Journal of the Royal Statistical Society: Series B (Methodological)*, 1977.

[27] S. Cheng and F. Lu. "A Two-Step Method for Missing Spatio-Temporal Data Reconstruction." *International Journal of Geo-Information*, 2017.

[28] J. Patterson and A. Gibson. *Deep Learning: a Practitioner's Approach - LSTM Networks*. PHI Learning Private Limited, 2017.

[29] S. Siami-Namini and A. S. Namin. “Forecasting Economics and Financial Time Series: ARIMA vs. LSTM.” *ArXiv.org*, 2018.

[30] R. A. Haugen. *”Modern Investment Theory”*. PHI Learning Private Limited, 2013.

[31] S. R. Baker et al. “Policy News and Stock Market Volatility”. *NBER*, 2019.

[32] D. Caldara and M. Iacoviello. “Measuring Geopolitical Risk”. *International Finance Discussion Papers 1222*, 2019.

[33] D. Caldara et al. “The Economic Effects of Trade Policy Uncertainty”. *Journal of Monetary Economics*, 2019.

[34] P. G. G. B. Bin Sun, Wei Cheng. “Short-term traffic forecasting using self-adjusting k-nearest neighbours”. *Economic Policy Uncertainty Index*, vol. 12, 2018.

[35] J. H. Friedman. “Greedy Function Approximation: A Gradient Boosting Machine.” *The Annals of Statistics*, vol. 29, no. 5, pp. 1189â1232, 2001.

[36] T. Chen and C. Guestrin. “Greedy Function Approximation: A Gradient Boosting Machine.” 2016.