The Impact of Missing Data on Heart Rate Variability Features: A Comparative Study of Interpolation Methods for Ambulatory Health Monitoring

Mouna Benchekroun, Baptiste Chevallier, Vincent Zalc, Dan Istrate, Dominique Lenne, Nicolas Vera

The Impact of Missing Data on Heart Rate Variability Features: A Comparative Study of Interpolation Methods for Ambulatory Health Monitoring

M. Benchekroun[1,2], B. Chevallier[1,3], V. Zalc[1], D. Istrate[1], D. Lenne[2] and N. Vera[3]

[1] Université de technologie de Compiègne, CNRS, Biomechanics and Bioengineering, UMR CNRS 7338, Compiègne, France

[2] Université de technologie de Compiègne, CNRS, Heudiasyc (Heuristics and Diagnosis of Complex Systems)

[3] Core for Tech

Correspondance: mouna.benchekroun02@gmail.com

baptiste.chevallier7@gmail.com

Abstract

Wearable devices have facilitated the remote measurement of heart rate variability (HRV), a promising indicator of various physiological and psychological states such as stress, sleep and other conditions. However, errors during the transmission or acquisition can lead to missing data, which can affect HRV features and cause false medical diagnosis. Interpolation is a useful technique for handling missing data, but the choice of interpolation method must be carefully considered. Therefore, the objective of this study is to investigate the impact of four interpolation methods (Nearest Neighbour, Linear, Shape-preserving piecewise cubic Hermite, and cubic spline) on HRV features when data is deliberately deleted. It is an expansion of a previously published work on HRV data imputation. The study utilizes a real-time approach to data interpolation and HRV analysis. The results indicate that the choice of interpolation method significantly affects HRV features, with varying effects depending on the percentage of missing data. Additionally, the study proposes to adapt the interpolation method based on both the percentage of missing values and the targeted HRV feature for maximum performance.

Keywords— Heart Rate Variability, HRV analysis, real time, Inter beat intervals, IBI, RR intervals, wearables, e-health.

1 Introduction

With the rise of telemedicine and healthcare wearables, scientists are eager to collect every trackable parameter from the human body throughout different physiological signals. One widely used signal is the Heart rate variability (HRV), now used as an indicator of different physiological states and pathologies [1]. Its time and frequency domain analysis can give insights into autonomic nervous function. They provide information about the sympathetic-parasympathetic balance and cardio-

vascular health [2].

HRV measures the variation in the time interval between two consecutive heartbeats, known as inter beat intervals (IBI) or RR intervals. They correspond to the time elapsed between two successive R-waves of the QRS complex, characterizing ventricular depolarization, on an ECG signal.

In an ideal situation, HRV analysis is performed with RR interval time series including only pure sinus beats, normally recorded by a 12 lead ECG. However, RR intervals are now usually measured thanks to wearable ECGs or photoplethysmographs (PPG) as a substitute of the gold standard ECG used in hospitals.

Thanks to such wearables, it is now possible to passively record heart activity continuously, opening the way to easier remote health monitoring during user's daily life.

However, for a reliable HRV analysis, these RR time-series should be carefully edited to identify gaps and abnormal heart beats beforehand.

In this paper, we investigate the impacts of data imputation using different interpolation methods on HRV features. We remove an increasing amount of data from an originally perfect HRV signal. The deleted values are then handled by four interpolation methods (Nearest Neighbour, Linear, Shape-preserving piece-wise cubic Hermite and cubic spline). Finally, an estimation error was computed to compare HRV features from reconstructed signals against those computed from the original signal. The goal is to identify the best interpolation method, that yields the lowest error in both time and frequency domains, based on the signal's quality and percentage of missing values. Ultimately, the best approach may be to choose the interpolation method according to the percentage of missing data in each HRV window analysis independently.

2 Context

The main downside to HRV assessment through wearables is the data quality that is often corrupted. Errors occur during the acquisition, the transmission or the storage, thus leading to an important data loss and unintended changes to the original HRV signal. Ectopic beats also introduce a bias into HRV features. When they are not caused by a physiological phenomenon such as premature ventricular contractions (PVC) or premature atrial contractions (PAC), they can occur due to a false QRS detection on the ECG signal or a missed beat.

Such artifacts represent a significant problem in the interpretation of HRV features making it sometimes even impossible. Therefore, they need to be addressed beforehand for a reliable HRV analysis [3].

Previous studies on the subject suggested different pre-processing methods for RR time series including filtering, deletion and interpolation. Each of these solutions however has its own disadvantages.

The main issue with the deletion approach is the signal depletion since the ectopic beats are removed without being replaced. The remaining RR-intervals are just merged together which increases the abrupt changes in the beat to beat variability and the disruptions in the natural fluctuation [4].

Deletion may be enough for HRV analysis in the time domain but is not sufficient for frequency domain analysis. Re-sampling, which is essential for analyzing HRV in the frequency domain, may produce outliers if the RRI time series contain missing values. [5]

Interpolation on the other hand roughly preserves the overall recording duration and the number of beats, but the beat manipulation does introduce changes that affect HRV analysis. The most used methods are Spline and Linear interpolation. Although they do not significantly affect the power spectral density (PSD), they may produce RRI outliers due to oscillation of the interpolation function, especially when using the spline function [5].

Besides, authors in [6], as well as many others, found that interpolation introduces low frequency components (LF) and reduces high-frequency com-

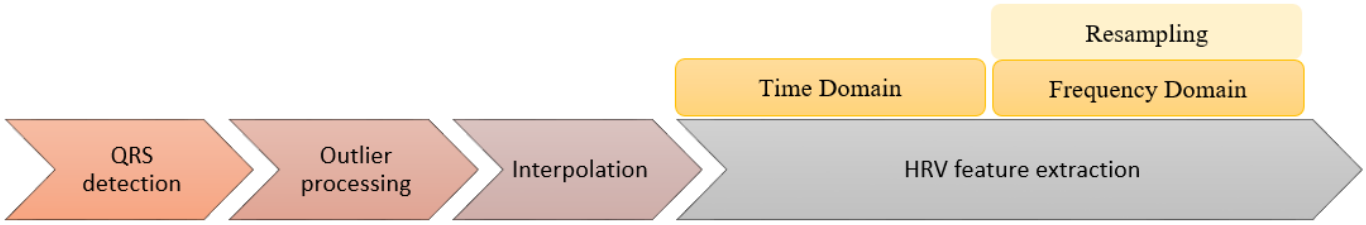

Figure 1: Processing approach for HRV analysis

ponents (HF) power, thus altering frequency domain HRV features.

Conventional HRV analysis is usually performed in four steps [1], including both deletion of ectopic beats and interpolation of missing values. If HRV is derived from an ECG sensor, the first step is R wave extraction from ECG signal and RR interval computation. Secondly, RR interval evaluation. The purpose of this step is to exclude RR intervals that do not meet some physiological criteria. The deleted RRIs are then replaced by interpolation. Time domain features can be computed from the output of the first two steps, whereas frequency domain features require data re-sampling for spectrum analysis. Another commonly used method is to compute time domain features without interpolation as done in [5].

There are other studies that propose more advanced methods for imputing missing HRV data [7, 8, 9], however, these methods tend to be quite complex and computationally expensive. They could include sophisticated mathematical algorithms, machine learning models, and other complex techniques that require significant computational resources and expertise to implement. Despite their added complexity, these methods have the potential to provide more accurate and reliable results in comparison to simpler methods. Nevertheless, the use of these methods may not be practical in certain real-world scenarios, such as in resource-limited settings or real-time analysis.

Paper contribution. The particularity of the present paper is the real-time approach to data deletion and interpolation. Both steps are performed iteratively in order to simulate a real-time HRV data acquisition and processing scenario.

This study is one of the first to investigate the effects of missing data on real-time HRV analysis. Data acquisition with missing values is simulated, and the missing values are replaced in real-time using various interpolation techniques before HRV analysis. The HRV features derived from the reconstructed signal are then compared to those from the original RR time-series.

The main purpose is to identify the best approach for processing the RR time-series in real time, based on the percentage of missing data in each HRV window. The real-time aspect is actually vital for continuous health monitoring.

Besides, to the best of our knowledge, this would be one of the first papers to investigate the effect of a very large amount of missing data (up to 70%) on HRV analysis. Recent developments in wearable devices have heightened the need for such studies since wearables produce a huge number of abnormal beats due to motion artifacts as well as missing data due to connectivity problems.

3 Methods

HRV signals used in this study are considered ideal thanks to the automatic R peak detection, on ECG

signals, which was manually corrected by a specialist. These signals did not contain missing nor ectopic peaks. The first step was to delete values from these signals in order to simulate a real-life, ambulatory low quality data acquisition. Since HRV features are usually computed from windows of the signal, data was deleted independently and iteratively from each HRV analysis window. The exact procedure is explained below and depicted in figure 3.
Next, the deleted values are replaced using different interpolation methods. In the same way as for data deletion, interpolation is also done iteratively for each window analysis.
After the signal has been degraded and reconstructed, time and frequency domain features are computed from both original and reconstructed signals; and compared using the Mean Absolute Percentage Error (MAPE). All the steps were implemented on PYTHON. Figure 2 summarizes the overall process of the study and each step is further detailed in the following sections.

3.1 Dataset

The dataset used is from the MIT-BIH Normal Sinus Rhythm RR Interval Database (nsr2db) available on PhysioNet [10].

The database includes beat annotation files for long-term ECG recordings of 54 subjects in normal sinus rhythm (30 men, aged 28.5 to 76, and 24 women, aged 58 to 73). The original ECG recordings were digitized at 128Hz, and the beat annotations were obtained by automated analysis with manual review and correction [10]. In this paper, RR segments including only normal beats between 0.3s and 1.3s were used (45-200bpm).

3.2 Missing values simulation

The objective was to delete the same percentage of data from all analysis windows. By doing so, we can directly evaluate the effect of each percentage of missing data on HRV features.

As depicted in figure 3, all windows had the same percentage of missing values but the distribution of deleted RRIs is completely randomised. Details are provided on the analysis window as well as the deletion procedure below.

HRV window :

In order to compute time domain and frequency domain HRV features, the RR timeseries were split into 5min segments, with a 1min sliding window (4min overlap). The choice of a sliding window is to address the discontinuities observed at the edges of each window. It also means a new set of HRV features is available every minute, which is closer to a real-time HRV analysis for continuous health monitoring.

Deletion procedure :

Since the goal is to evaluate the effect of interpolation on HRV features, the same percentage of missing data was removed from each window used to compute HRV features. The steps for the deletion procedure are explained in the pseudo code below.

Algorithm 1 RR deletion procedure

1: Randomly delete $P\%$ of the data in the first 5min window
2: **for** Each new window i **do**
3: Compute N, total number of values to be deleted $N = \frac{WindowLength \times P}{100}$
4: Determine $N_{overlap}$ number of deleted data in the 4min overlap.
5: Compute the number of values still to be deleted from the sliding window : $N_{sliding} = N - N_{overlap}$
6: Randomly remove $N_{sliding}$ from the last minute of the window
7: **end for**

In the first iteration, N values are randomly deleted from the first window of the signal.

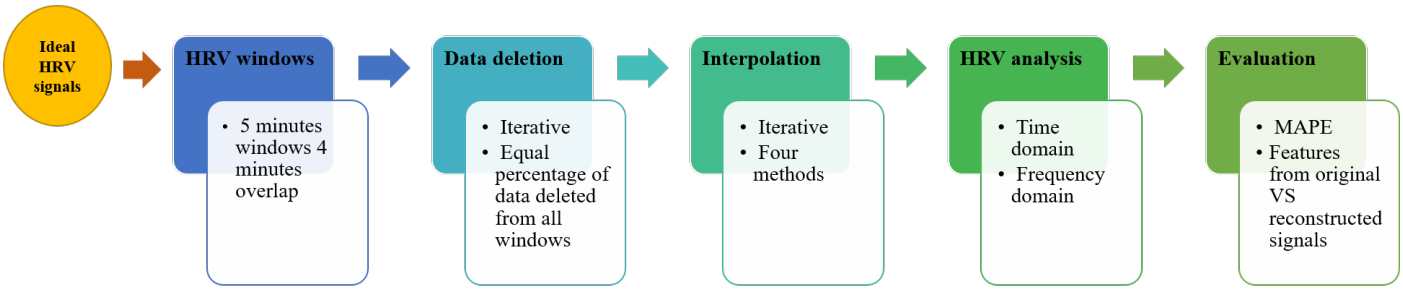

Figure 2: General flowchart of the study including the three major steps of data deletion, interpolation, followed by feature extraction and comparison using the MAPE.

$$N = \frac{\text{Window length} * P}{100} \tag{1}$$

For each window after the first one, the initial step is to compute the total number (N) of data that should be deleted in order to reach the deletion percentage (P). The number of missing values in the 4 min overlap, deleted in the previous iteration, is then computed ($N_{overlap}$), and serves to determine the number of data to randomly remove from the last minute of the window $N_{sliding} = N - N_{overlap}$.
At the end of the loop, all the analysis windows had the same percentage of randomly deleted values.
The beats were removed away from the window's edges in order to avoid extrapolation problems. Other than this, there was no condition on the number of consecutive beats to be deleted, nor on their positions. The deletion procedure is completely random. It is however obvious that the higher the percentage of deleted data, the larger (and more numerous) the gaps with successive missing beats.

3.3 Interpolation methods

The missing RR intervals deleted in the last step were then replaced using four different interpolation methods. Interpolation was also performed independently and iteratively for each HRV window. In each iteration, only the values in the last minute of the window are interpolated since the values in the first 4 minutes (overlap) were filled in the previous iteration. This is in order to simulate a real-time data acquisition and processing. Interpolation methods used in this study are listed below:

- **Nearest Neighbour (NN)**: Zero-order interpolation method that assigns the value of the nearest existing RR interval to the missing beat.
- **Linear**: First order interpolation method. Derives a straight line connecting the adjacent RR intervals and calculates the missing beats based on the line.
- **Shape-preserving piecewise cubic Hermite interpolating polynomial (PCHIP)**: A piecewise cubic polynomial determined by the given data and their specified derivatives at the interpolation points [11].

$$P(x_k) = y_k, P(x_{k+1}) = y_{k+1} \tag{2}$$
$$P'(x_k) = d_k, P'(x_{k+1}) = d_{k+1} \tag{3}$$

The main idea is to determine the slopes d_k so that the function values do not overshoot the data values [11]. One of the potential ways to determine d_k, used in this paper, is briefly explained below.

If δ_k and δ_{k-1} have opposite signs or if either of them is zero, then x_k is a discrete

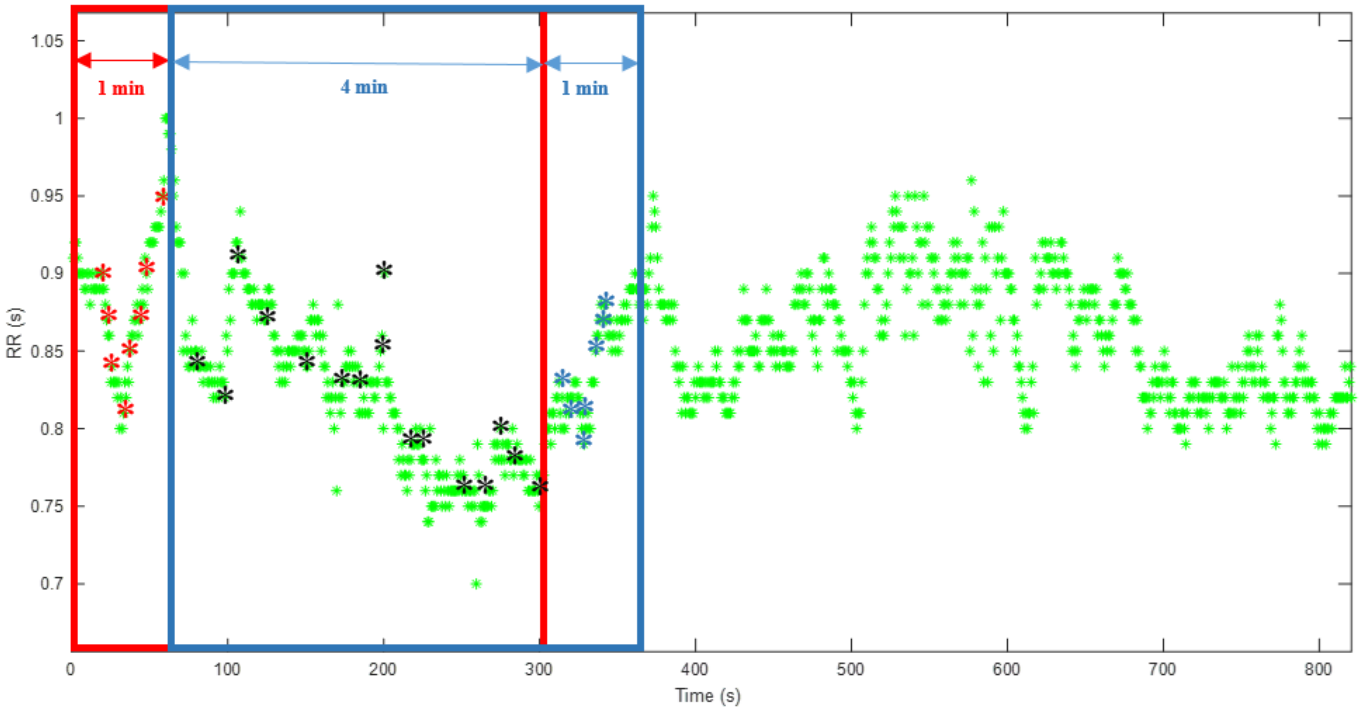

Figure 3: Example of the deletion procedure on an HRV signal (green signal). We used 5 minutes windows with 4 minutes overlap, which is the recommended and mostly used window size in literature [2]. In the first iteration, a percentage of RRIs is deleted from the first red window, depicted as red and black asterisks. In the next iteration, RRIs already deleted in the 4min overlap (black asterisks) are maintained and only RRIs located in the last minute of the blue window are deleted in order to reach the same percentage of missing data in all windows. Deleted RRI in the second iteration are depicted as blue asterisks. Since both windows have the same deleted data in the 4min overlap segment. The first minute (red arrow) of the i window and the last minute (blue arrow) of the $i+1$ window have the same percentage of missing data.

local *minimum* or *maximum*, so d_k is set to be equal to zero. In (figure 4a), the green curved line is the shape-preserving interpolant, formed from two different cubics. The two cubics interpolate the center value and their derivatives are both zero there [11]. On the other hand, if δ_k and δ_{k-1} have the same sign, then d_k is a weighted harmonic mean, with weights determined by the lengths of the two intervals around x_k.

$$\frac{w_1 + w_2}{d_k} = \frac{w_1}{\delta_{k-1}} + \frac{w_2}{\delta_k} \tag{4}$$

where $w_1 = 2h_k + h_{k-1}$, $w_2 = h_k + 2h_{k-1}$. (h_k denotes the length of the k^{th} subinterval: $h_k = x_{k+1} - x_k$) and h_{k-1} the length of the $(k-1)^{th}$ interval.

At the breakpoint, the reciprocal slope of the Hermite interpolant is the weighted average of the reciprocal slopes of the piecewise linear interpolant on either side (figure 4b). The shape-preserving interpolant is formed from the 2 cubics that interpolate the center value and that have slope equal to d_k there [11].

- **Cubic Spline**: One popular third degree in-

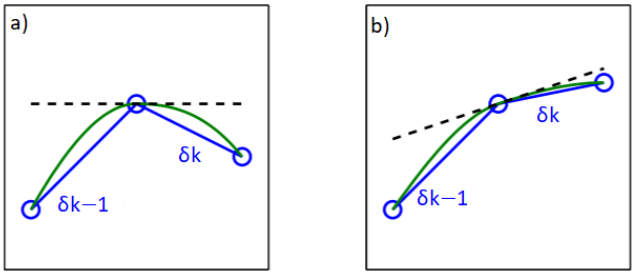

Figure 4: Slopes for PCHIP. δ_k and δ_{k-1} are the two slopes of the piecewise linear interpolant on either side of the breakpoint

terpolation method is the cubic spline interpolation, where data points are estimated by fitting a third degree polynomial. A spline is also a piecewise cubic Hermite that is exceptionally smooth, in the sense that the first and second derivatives of consecutive polynomials are equal and thus continuous, ensuring smoothness of the resulting curve. This avoids the problem of the straight polynomial interpolation that tends to induce distortions on the edges of the polynomials [11].

The Pchip and the spline methods both perform piecewise cubic Hermite interpolation. They only differ in how the slopes of the interpolant are computed, thus leading to different behaviors when the underlying data has flat areas or undulations.

After the interpolation step, HRV features were estimated on the reconstructed data and compared to the original HRV set from the original signal. The error was then estimated through the mean absolute percentage error (MAPE) in order to identify the best interpolation approach.

3.4 HRV analysis

To choose the best imputation approach for HRV signals, the impact of interpolation on multiple HRV features was evaluated through an estimation error to compare features from reconstructed signals to those from the original signal. Features mostly used in literature were selected, they can be separated into two categories, time domain and frequency domain features.

Time domain :

Two of the most known indices were chosen, which are SDNN and RMSSD, for the time domain analysis.

SDNN stands for Standard Deviation of Normal to Normal beats. Normal to normal means that ectopic and other abnormal beats have to be removed beforehand. Variations of SDNN such as Standard deviation of RR intervals (SDRR) are sometimes used. The formula is the same, the only difference is that RR time series- for SDRR- include abnormal or false beats.

(*In this study, ectopic beats created by interpolation are not filtered before HRV analysis. SDRR will be referred to as SDNN since the formula is the same.*)

SDNN is mostly computed over 24H periods, however, researchers have found significantly shorter periods of analysis to be relevant [12]. In this case 300 seconds (5min) periods were used. Considered as gold standard in quantification of the cardiac risk [2], reflection of both sympathetic nervous system (SNS) and parasympathetic nervous system (PNS) activity can be measured on SDNN which makes it one of the most useful features of HRV analysis.

$$SDNN = \sqrt{\frac{\sum_{i=1}^{N}(RR_i - \overline{RR})^2}{N-1}} \tag{5}$$

Where :

$$\overline{RR} = \frac{1}{N}\sum_{i=1}^{N}(RR_i) \tag{6}$$

RMSSD is the root mean square of successive differences between normal heartbeats. Like SDNN, it takes only normal IBI as an input. This feature reflects more PNS activation than SDNN does.

$$RMSSD = \sqrt{\frac{\sum_{i=1}^{N-1}(RR_i - RR_{i+1})^2}{N-1}} \quad (7)$$

where N is the number of RR intervals in the signal.

Frequency domain :

Several methods can be used for frequency domain analysis such as Fast Fourier Transform (FFT), auto regressive modeling (AR) or wavelet transform. In this study, we tested both FFT and AR and compared the results to KUBIOS, the reference software for HRV analysis for validation purposes. Beside the simplicity of implementation, we opted for FFT since the results from Python were the closest to KUBIOS.

The goal of frequency domain analysis using any of the methods cited above is always to separate HRV signal spectrum into four components which are Ultra Low Frequency ($\leq 0.003Hz$), Very Low Frequency ($0.003 - 0.04Hz$), Low Frequency ($0.04 - 0.15Hz$), and High Frequency ($0.15 - 0.4Hz$) [2], (respectively ULF, VLF, LF and HF).

Since ULF and VLF generally require long periods of recording not suitable for real-time analysis, they will not be included in this study. Also, their physiological correlates are still unknown which makes them less relevant for e-health applications.

HF and LF on the other hand can be assessed on 1 to 2 min windows respectively [2]. Their ability to reflect the overall cardiac health and the state of the autonomic nervous system (ANS) has been proven by many studies [13, 1], in different contexts including stress [14, 15] and sleep [16].

3.5 Evaluation metrics

The difference between HRV features from the reconstructed data and those from the original signal was assessed by the Mean Absolute Percentage Error (MAPE) (8). Other metrics can also be used but the idea behind choosing the MAPE is to avoid mutual cancellation of the positive and negative errors. Moreover, since each HRV parameter has a wide range [17], normalization by the actual value allows the comparison of differently scaled time-series data.

$$Mape = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - E_t}{A_t}\right| \quad (8)$$

where :
n = number of times the summation iteration happens, which corresponds to the number of HRV windows.
A_t = Actual value, from the original RR time-series.
E_t = Estimated value, from reconstructed signal.

Another interesting parameter to look at is the number of ectopic beats created by the interpolation. As explained before, non physiological beats should be filtered and, eventually, replaced before HRV analysis. The replacement method (ie : interpolation) should not be creating more ectopic beats. We assessed the percentage of abnormal RR intervals ($P_{ectopic}$) in the reconstructed signals as follows :

$$P_{ectopic} = \frac{\text{Number of ectopic beats}}{\text{Signal Length}} \quad (9)$$

4 Results and discussion

In this paper, 24 RR time-series of 50min duration were analysed for a total of 1104 HRV windows of 5min duration. To investigate the effect of missing data on HRV features, the same percentage of RR-intervals was removed from each window starting from 10% up to 70% of missing values with a 10% step.

The deleted beats were then replaced by four different interpolation methods explained in section

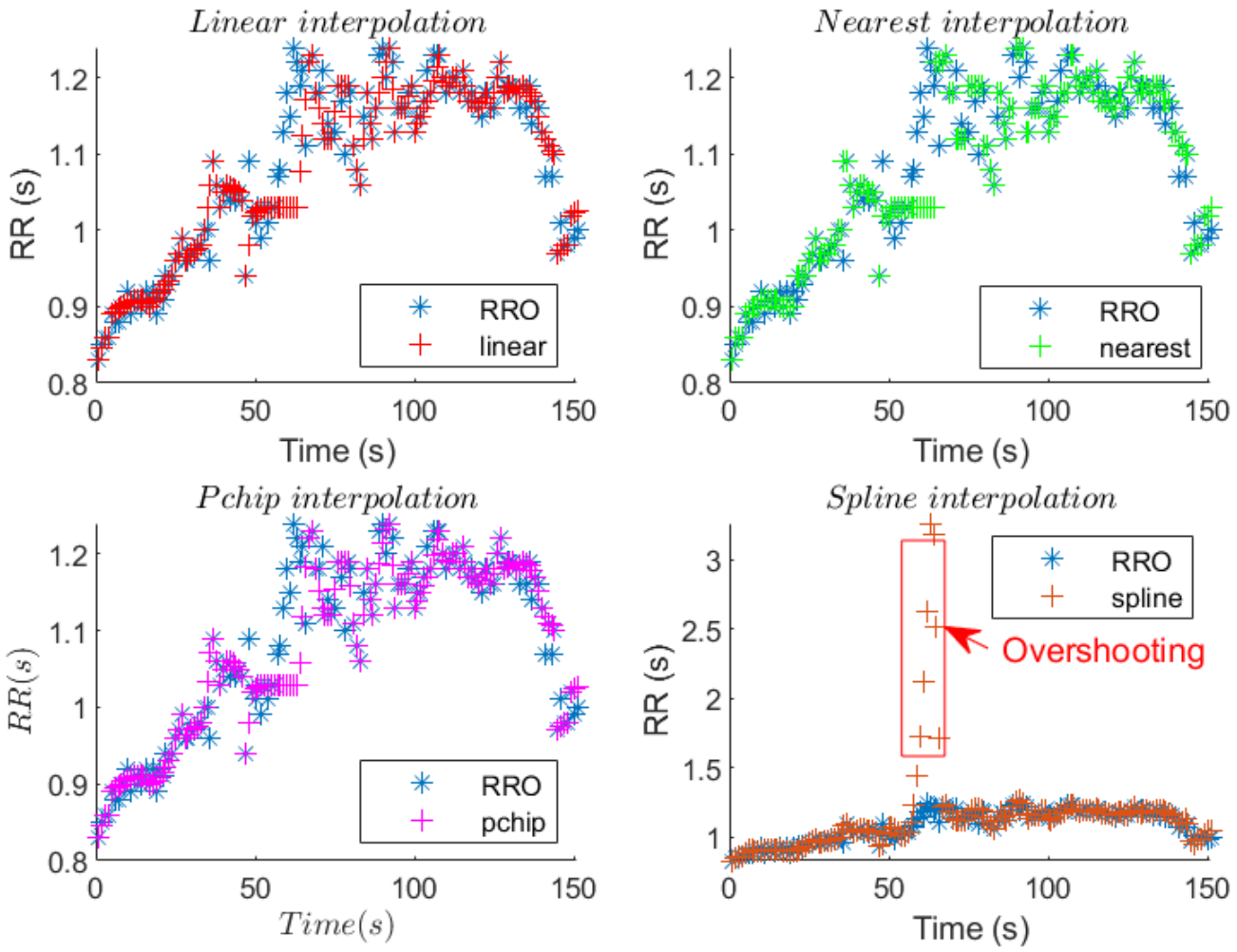

Figure 5: Example of interpolation for 50% missing data. The red arrow indicates the ectopic beats created by the Cubic Spline interpolation

3.3. An example of data interpolation is shown in Figure 5.

The cubic spline interpolation overshoots the data at some points as can be seen in figure 5. This is due to the requirement for equal second order derivatives at every point. By eliminating this condition, it is possible to prevent, or at least reduce, the overshooting as done by the Pchip method.

Time domain features According to the results in table 1, SDNN seems to be less sensitive to interpolation. It was the least affected with an estimation error not greater than 5% even with a huge number (70%) of missing data. The same conclusion was found by authors in [18].

RMSSD on the other hand is much more sensitive to interpolation. The estimation error increases almost linearly with the percentage of missing data in the original signal.
As can be seen in table 1, the Nearest Neighbour interpolation yields the least error for both SDNN and RMSSD compared to other interpolation methods.

		Mape (%)			
Missing %	HRV Features	NN	Linear	Pchip	Spline
10%	RMSSD SDNN % ectopic	3.84 ± 1.15 0.87 ± 0.31 0	6.69 ± 0.96 0.91 ± 0.40 0	6.56 ±0.99 0.86 ± 0.38 0	5.13 ±1.03 1.03 ±0.44 0
20%	RMSSD SDNN % ectopic	6.98 ± 2.88 1.36 ± 0.59 0	12.76 ± 2.59 1.43 ± 0.55 0	12.49 ±2.65 1.3 ± 0.48 0	9.51 ±2.38 1.89 ±1.01 0
30%	RMSSD SDNN % ectopic	10.17 ± 3.94 1.70 ± 0.52 0	19.89 ± 3.23 2.28 ± 1.03 0	19.39 ± 3.29 1.96 ± 0.87 0	14.84 ± 3.17 2.93 ± 1.42 0.5
40%	RMSSD SDNN % ectopic	13.99 ± 4.45 2.08 ± 0.52 0	27.63 ± 3.91 3.18 ± 1.04 0	26.92 ±4.05 2.56 ± 0.83 0	26.11 ± 26.4 7.42 ± 15.25 0.7
50%	RMSSD SDNN % ectopic	17.3 ± 6.34 2.63 ± 0.91 0	33.66 ± 7.66 4.21 ± 1.57 0	32.54 ± 7.48 3.45 ± 1.32 0	25.83 ± 6.94 6.60 ± 3.01 1.3
60%	RMSSD SDNN % ectopic	20.7 ± 7.87 3.47 ± 1.39 0	41.63 ± 8.98 5.31 ± 2.09 0	40.87 ± 9.07 4.35 ± 1.83 0	31.9 ± 7.37 9.59 ± 6.58 1.5
70%	RMSSD SDNN % ectopic	25.57 ± 8.25 2.63 ± 0.91 0	50.58 ± 10.6 4.21 ± 1.57 0	49.4 ± 10.4 3.45 ± 1.32 0	40 ± 8.4 6.60 ± 3.01 1.5

Table 1: Mean absolute percentage error of estimated time domain HRV features from 10 to 70% missing data

		Mape (%)	
%	HRV_{feat}	NN	No interp
50%	RMSSD	17.3±6.34	17.3±7.52
60%	RMSSD	20.7±7.87	15.44±6.81
70%	RMSSD	25.57±8.25	12.38±6.60

Table 2: Mean absolute percentage error of estimated RMSSD for 50%, 60% and 70% missing data

Since SDNN is the standard deviation of each RR interval from the mean RR duration, it reflects the LF component in some way whereas the RMSSD correlates with the HF since it uses the difference between successive beats. This may explain why SDNN is much less sensitive to interpolation than RMSSD. In fact, NN interpolation acts as a low-pass filter since it produces flat-like shapes [18]. In situations where the heart rate is relatively stable and does not vary abruptly, the NN interpolation is most likely to preserve the heart rate variability.

When the percentage of missing data exceeds 50% however, it has been found that the best results for RMSSD estimation are achieved without editing the RR tachograms, i.e without replacing the missing data by any of the interpolation methods used in the study.

[19] also concluded that RMSSD does not require any interpolation to obtain reliable estimations, but they found the threshold to be at 30% instead.

Table 2 summarises RMSSD estimation errors by nearest neighbour approach against no interpolation. Not editing RR time-series yields better RMSSD estimation than editing more than half the data. This however should be verified when the acquisition includes different contexts that may cause the heart rate to vary a lot.

The decrease of the Mape when the percentage of missing data increases may be due to the lower number of compared windows. When the missing values are not replaced by any interpolation, remaining RR intervals are just merged. This makes the RR signal much shorter and thus reduces the number of HRV windows.

Frequency domain features are clearly much more sensitive to interpolation as can be seen from table 3. Linear and Pchip interpolation perform almost equally and yield the least estimation error for LF, HF and LF/HF. They are thus considered to be the best interpolation methods for frequency domain features.

Generally speaking, physiological variables such as the Autonomic cardiovascular regulation operates at sufficiently low frequencies [20] that nothing would be lost using a linear or a Pchip approach. Unless there is a physiological reason to suppose a non-linear trend, linear seems to assume less than the other methods.

Contrary to the time domain analysis, the cubic spline interpolation gives the worst results with an error almost two times greater than all the other interpolation methods for frequency domain features. This can be explained by the fact that cubic splines are prone to severe oscillation and they overshoot at intermediate points. The overshooting introduces many ectopic beats thus increasing the HF components. It has been found in [21] that the presence of only one ectopic beat in a 2 min ECG recording introduces an increase in the HF power of around 10%.

[22] however compared linear, spline backward and forward interpolation. They opted for spline as the best interpolation method. One potential explanation for this is the small percentage of missing data (10%) simulated in the signal and the deletion approach that was not completely random. A maximum of four successive missing values was set.

Table 4 summarises the best interpolation approach for some HRV feature at a specific range of missing data. At exactly 50% of missing beats, NN and no interpolation approach perform equally

		Mape (%)			
Missing %	HRV Features	NN	Linear	Pchip	Spline
10%	LF	5.86 ± 2.59	4.69 ± 2.00	4.82 ± 2.26	7.77 ± 5.01
	HF	5.9 ± 2.49	5.07 ± 2.04	5.09 ± 2.10	6.1 ± 2.43
	LF/HF	9.58 ± 3.56	7.45 ± 2.49	7.57 ± 2.7	11.22 ± 5.2
20%	LF	8.46 ± 4.39	7.07 ± 4.39	7.15 ± 3.78	13.45 ± 9.40
	HF	7.53 ± 2.69	6.8 ± 2.82	6.89 ± 2.72	8.94 ± 3.61
	LF/HF	12.64 ± 4.93	10.67 ± 3.87	10.89 ± 3.99	18.70 ± 9.85
30%	LF	11.19 ± 5.74	9.47 ± 4.09	9.61 ± 4.7	20.21 ± 13.44
	HF	11.30 ± 5.48	11.22 ± 5.34	11.35 ± 5.34	14.38 ± 7.51
	LF/HF	16.63 ± 5.93	14.96 ± 4.67	15.12 ± 4.70	27.02 ± 11.67
40%	LF	14.14 ± 6.16	12.50 ± 4.14	12.09 ± 4.65	26.18 ± 19.33
	HF	13.39 ± 5.24	14.36 ± 7.17	13.72 ± 6.63	21.45 ± 21.8
	LF/HF	20.70 ± 6.73	19.32 ± 5.34	18.51 ± 5.20	30.84 ± 15.56
50%	LF	16.55 ± 8.00	16.31 ± 5.02	15.24 ± 5.6	36.43 ± 26.65
	HF	17.1 ± 7.73	18.56 ± 9.17	18.67 ± 9.4	26.99 ± 14.15
	LF/HF	23.95 ± 7.4	24.15 ± 6.08	23.44 ± 6.63	40.51 ± 17.08
60%	LF	19.42 ± 7.16	22.24 ± 5.6	21.0 ± 8.19	35.29 ± 27.10
	HF	21.7 ± 10.79	23.57 ± 11.7	25.4 ± 13.7	39.0 ± 21.45
	LF/HF	27.28 ± 7.4	31.26 ± 7.9	29.8 ± 7.79	40.76 ± 12.1
70%	LF	25.29 ± 7.6	31.24 ± 7.61	27.6 ± 6.88	39.15 ± 29.9
	HF	31.0 ± 11.52	32.57 ± 12.8	33.0 ± 12.1	52.4 ± 21.63
	LF/HF	33.7 ± 9.42	40.47 ± 10.0	37.7 ± 9.56	40.58 ± 10.7

Table 3: Mean absolute percentage error of estimated frequency domain HRV features from 10% to 70% missing data

Missing %	HRV feat	Best interp
1^{st}category: $10\% - 50\%$	RMSSD SDNN LF HF LF/HF	NN NN / Pchip Lin / Pchip Lin / Pchip Lin / Pchip
2^{nd}category: $50\% - 70\%$	RMSSD SDNN LF HF LF/HF	No interpolation NN NN /Pchip NN / Lin NN / Pchip

Table 4: Best interpolation approach for HRV features based on the percentage of missing data.

with regards to RMSSD estimation (Table 2). The latter method outperforms the first one when the percentage crosses the 50% threshold.

Generally speaking, the Pchip interpolation seems to do well in most cases. It preserves the linear trend of the data while adding very light waves. As explained in [1], the structure generating the RR signal is not only simply linear, but also involves nonlinear contributions. The Pchip interpolation thus seems to better mimic the RR timeseries trend.

5 Conclusion

In the time domain, nearest neighbour interpolation gives the best results for up to 50% of edited data. Beyond 50%, the best estimation was achieved when the deleted data was not replaced. It seems better not to use any interpolation for RMSSD beyond this threshold. In the frequency domain however, the lowest errors of HRV feature estimation are obtained using linear or Pchip interpolation.

If only one approach had to be chosen for a good overall estimation, the Pchip would be privileged because it preserves the linear trend and the slightly non linear contributions in the RR timeseries.

Since HRV features are used for preventive health and users' well-being, it is fundamental to know the effect of missing data on these parameters. The findings of this study, namely the best interpolation methods based on the percentage of missing beats could be used for a data-driven decision-making strategy to decide whether reliable conclusions can be drawn from the signal.

This preprocessing step, including filtering and interpolation, is fundamental before any HRV analysis can be performed. It enables continuous passive monitoring of users' cardiovascular activity in a non-obtrusive way despite a relatively poor data quality.

6 Limits and Perspectives

It is worth bearing in mind that interpolation remains at a mathematical level. Physiological implications and interpretations could further be explored but are outside the scope of this paper. The need and efficacy of interpolation in general should be assessed against the end-goal of HRV analysis. Moreover, in real-life acquisitions, the exact number of missing data in a time gap is unknown.

On the other hand, many additional aspects could be investigated in a future work. Based on the findings described above and table 4, a potential good approach may be using a combination of different interpolation methods chosen based on the HRV feature and the percentage of missing data in each HRV segment. It would be interesting to measure the estimation error of such an approach including different interpolation methods based on the percentage of missing data in each window. The effect of interpolation on other HRV features such as the total spectral power, and Non linear features could also be investigated.

Additionally, it would be very interesting to identify an upper limit for missing beats, in each

HRV window, beyond which any interpolation would be pointless. This upper limit would depend once again on the context and the purpose of HRV analysis in the first place. It would help decide whether an HRV segment can be used for a reliable diagnosis or should be discarded.

Conflict of interest :

None

Acknowledgment :

The study was carried out in the frame of a collaboration between two PhD researches. The first one is funded by a grant from the French Ministry of Education The second PhD is funded by the industrial sponsor, Core for Tech and the National Association of Research and Technology in France (ANRT).

References

[1] U. R. Acharya, K. P. Joseph, N. Kannathal, C. M. Lim, and J. S. Suri, "Heart rate variability: a review," *Medical and biological engineering and computing*, vol. 44, no. 12, pp. 1031–1051, 2006.

[2] Task Force of the European Society of Cardiology *et al.*, "Heart rate variability: standards of measurement, physiological interpretation and clinical use," *circulation*, vol. 93, pp. 1043–1065, 1996.

[3] K. K. Kim, Y. G. Lim, J. S. Kim, and K. S. Park, "Effect of missing RR-interval data on heart rate variability analysis in the time domain," *Physiological Measurement*, vol. 28, pp. 1485–1494, oct 2007.

[4] M. Peltola, "Role of editing of rr intervals in the analysis of heart rate variability," *Frontiers in physiology*, vol. 3, p. 148, 2012.

[5] K. Eguchi, R. Aoki, S. Shimauchi, K. Yoshida, and T. Yamada, "RR interval outlier processing for heart rate variability analysis using wearable ECG devices," *Advanced Biomedical Engineering*, vol. 7, pp. 28–38, 2018.

[6] G. D. Clifford and L. Tarassenko, "Quantifying errors in spectral estimates of hrv due to beat replacement and resampling," *IEEE transactions on biomedical engineering*, vol. 52, no. 4, pp. 630–638, 2005.

[7] M. Benchekroun, B. Chevallier, D. Istrate, V. Zalc, and D. Lenne, "Preprocessing methods for ambulatory hrv analysis based on hrv distribution, variability and characteristics (dvc)," *Sensors*, vol. 22, no. 5, p. 1984, 2022.

[8] V. Stankus, P. Navickas, A. Slušnienė, I. Laucevičienė, A. Stankus, and A. Laucevičius, "A novel adaptive noise elimination algorithm in long rr interval sequences for heart rate variability analysis," *Sensors*, vol. 22, no. 23, p. 9213, 2022.

[9] Q. Zhang, D. Zhou, and X. Zeng, "A novel machine learning-enabled framework for instantaneous heart rate monitoring from motion-artifact-corrupted electrocardiogram signals," *Physiological measurement*, vol. 37, no. 11, p. 1945, 2016.

[10] A. L. Goldberger, L. A. Amaral, and L. Glass, "Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[11] C. Moler, "Splines and pchips," *Cleve's Corner: Cleve Moler on Mathematics and Computing, 16-Jul-2012. [Online];* Available: https://blogs.mathworks.com/cleve/2012/07/16/splines-and-pchips/#59fe8852-238e-4351-9285-8f9a17018a89 [Accessed: Mar 2019].

[12] H. J. Baek, C.-H. Cho, J. Cho, and J.-M. Woo, "Reliability of ultra-short-term analysis as a surrogate of standard 5-min analysis of heart rate variability," *Telemedicine and e-Health*, vol. 21, no. 5, pp. 404–414, 2015.

[13] F. Shaffer and J. Ginsberg, "An overview of heart rate variability metrics and norms," *Frontiers in public health*, vol. 5, p. 258, 2017.

[14] L. Salahuddin, M. G. Jeong, D. Kim, S.-K. Lim, K. Won, and J.-M. Woo, "Dependence of heart rate variability on stress factors of stress response inventory," in *2007 9th international conference on e-health networking, application and services*, pp. 236–239, IEEE, 2007.

[15] M. Benchekroun, B. Chevallier, H. Beaouiss, D. Istrate, V. Zalc, M. Khalil, and D. Lenne, "Comparison of stress detection through ECG and PPG signals using a random forest-based algorithm," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 3150–3153, IEEE, 2022.

[16] E. Michail, A. Kokonozi, I. Chouvarda, and N. Maglaveras, "Eeg and hrv markers of sleepiness and loss of control during car driving," in *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 2566–2569, IEEE, 2008.

[17] K. K. Kim, J. S. Kim, Y. G. Lim, and K. S. Park, "The effect of missing rr-interval data on heart rate variability analysis in the frequency domain," *Physiological measurement*, vol. 30, no. 10, p. 1039, 2009.

[18] M. A. Salo, H. V. Huikuri, and T. Seppanen, "Ectopic beats in heart rate variability analysis: effects of editing on time and frequency domain measures," *Annals of noninvasive electrocardiology*, vol. 6, no. 1, pp. 5–17, 2001.

[19] D. Morelli, A. Rossi, M. Cairo, and D. A. Clifton, "Analysis of the impact of interpolation methods of missing rr-intervals caused by motion artifacts on hrv features estimations," *Sensors*, vol. 19, no. 14, p. 3163, 2019.

[20] J. P. Saul, "Beat-to-beat variations of heart rate reflect modulation of cardiac autonomic outflow," *Physiology*, vol. 5, no. 1, pp. 32–37, 1990.

[21] G. G. Berntson and J. R. Stowell, "Ecg artifacts and heart period variability: don't miss a beat!," *Psychophysiology*, vol. 35, no. 1, pp. 127–132, 1998.

[22] A. Tlija, K. Wegrzyn-Wolska, and D. Istrate, "Missing-data imputation using wearable sensors in heart rate variability," *Bulletin of the Polish Academy of Sciences. Technical Sciences*, vol. 68, no. 2, 2020.