COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 132 (2016) 29–44

journal homepage: www.intl.elsevierhealth.com/journals/cmpb

CopyMean: A new method to predict monotone missing values in longitudinal studies

Christophe Genolini [a,b,*], *Amandine Lacombe* [a], *René Écochard* [c,d], *Fabien Subtil* [c,d]

[a] *Inserm UMR U1027, Research Unit on Perinatal Epidemiology and Childhood Disabilities, Adolescent Health, Université Paul Sabatier, Toulouse III, Toulouse, France*
[b] *CeRSM (EA 2931), UFR STAPS, Université de Paris Ouest-Nanterre-La défense, 92000 Nanterre, France*
[c] *Hospices Civils de Lyon, Service de Biostatistique, F-69003 Lyon, France*
[d] *CNRS, UMR5558, Laboratoire de Biomtrie et Biologie Evolutive, Equipe Biostatistique-Sant, F-69100 Villeurbanne, France*

ARTICLE INFO

Article history:
Received 25 September 2014
Received in revised form
4 March 2016
Accepted 6 April 2016

Keywords:
Longitudinal data
Missing data
Imputation

ABSTRACT

Background: Longitudinal studies are those in which the same variable is repeatedly measured at different times. More likely than others, these studies suffer from missing values. Because the missing values may impact the statistical analyses, it is important that they be dealt with properly.
Methods: In this paper, we present "CopyMean", a new method to impute (predict) monotone missing values. We compared its efficiency to sixteen imputation methods dedicated to the treatment of missing values in longitudinal data. All these methods were tested on four datasets, real or artificial, presenting markedly different caracteristics.
Results: The analysis showed that CopyMean was more efficient in almost all situations.

1. Introduction

Longitudinal studies are those in which the same variable is repeatedly measured at different times. More likely than others, these studies suffer from missing values [1–3] since if a participant drops out at one point, his/her data are missing on subsequent measures [4]. According to Little [3], missing data may be classified into three main categories: Missing Completely At Random (MCAR) when the missingness probability is independent of the variables, Missing at Random (MAR) when the missingness probability depends only on the observed variables, and Missing Not At Random (MNAR) when the missingness probability may depend on unobserved variables.

When the main analysis involves statistical modeling of longitudinal data (such as mixed models), the model parameters are generally estimated by the maximum likelihood. It is well-known that the maximum likelihood estimation is robust to MAR data [2,5,6]. However, selection models and pattern-mixture models have been proposed when the data are MNAR or when a sensitivity analysis to this assumption is performed [2,5–8].

This paper focuses on situations where the imputed value has a particular interest (e.g., prediction of blood pressure during surgery when the device fails, subway waiting time, weather prediction) and where a decision has to be made (inject or not a given product, wait for the next metro or take a taxi, cancel or not an outdoor event). Thus, we tested methods that involve

* *Corresponding author.* Inserm UMR U1027, Research Unit on Perinatal Epidemiology and Childhood Disabilities, Adolescent Health, Université Paul Sabatier, Toulouse III, 31000 Toulouse, France. Tel.: +33 6 21 48 47 84.
E-mail address: christophe.genolini@u-paris10.fr (C. Genolini).
http://dx.doi.org/10.1016/j.cmpb.2016.04.010
0169-2607/

single imputation only and excluded model or likelihood-based methods [9,10]. These imputation methods suppose that the imputed value is correct. This is a strong assumption that might not be met. However, single imputation methods have three advantages: (i) they are mostly non-parametric, both on the temporal changes and on the distribution of the variables at each time point. They have therefore a wider scope than model-based methods or multiple imputation (especially when the trajectories are not parametric, such as BMI trajectories in Reference 11); (ii) in some cases, it may be more important to find the missing values than to estimate parameters from the imputed dataset using a model (as we said earlier, during surgery, the surgeon does not need a model but an information, as accurate as possible, to make a decision); (iii) some statistical techniques work only on complete datasets; in this case, imputing with a single imputation method allow to not exclude individuals with missing values, which is generally preferable.

Twisk [12], Engels [1] and Genolini [13] compared several imputation methods for longitudinal studies. Twisk proposed a classification of imputation methods into two categories: (i) "Cross-sectional" methods impute missing values at time *t* using information available at time *t*; and (ii) "Longitudinal" methods impute the missing values of an individual *i* using all the non-missing values of *i*.

Engels suggested four categories: (i) "No personal data" methods do not use information available on individual subjects; (ii) "Baseline data" methods use the information present at baseline but no time-dependent information; (iii) "Before data only" methods consider all the information available before the occurrence of the missing value; and (iv) "Before and after" methods impute the missing values using all available information. Regarding the evaluation of performance, Engels proposed different indices to compare the performance of imputation methods. These indices are mainly based on the difference between the imputed values and the actual values [1].

Genolini used a generalization of Twisk classifications: (i) "Cross-sectional" methods impute missing values at time *t* using information available at time *t*; (ii) "Longitudinal" methods impute the missing values of an individual *i* using all the non-missing values of *i* and (iii) "Cross-sectional and longitudinal" use both the transversal and the longitudinal information.

In Reference 13, Genolini compared different classical imputation methods with the original *CopyMean for intermittent missing values*. He showed that in almost all cases, with intermittent missing data, CopyMean is more efficient than other methods.

The aim of the present article is to generalize the results found with CopyMean for intermittent missing values to the case of monotone missing values. For that, several imputation methods for monotone missing values are compared. Section 2 provides the general framework and the methodology: a formal definition of the concept of missingness, a presentation of the imputation methods, and a presentation of the criteria used to measure performance. It reviews the classical methods used and presents the original method called *CopyMean*. Section 3 presents the design of the simulation study. Section 4 gives the results. A discussion is provided in Section 5.

2. Methods

2.1. Notations

Let us consider a set S of n subjects. For each subject, an outcome variable Y is measured at t different times. The value of Y for subject i at a specific time j is noted y_{ij}. For subject i, the sequence $y_{i.} = (y_{i1}, y_{i2}, \dots, y_{it})$ is called a trajectory. For a specific time j, vector $y_{.j} = (y_{1j}, y_{2j}, \dots, y_{nj})$ is called a cross-sectional measurement. The mean trajectory of S is noted $y_{..} = (y_{.1}, y_{.2}, \dots, y_{.t})$. When y_{ij} is missing, the value obtained by using a given imputation method IM is noted y_{ij}^{IM}. The mean trajectory of the imputed dataset is noted $y_{..}^{IM} = (y_{.1}^{IM}, y_{.2}^{IM}, \dots, y_{.t}^{IM})$. For a specific trajectory i, we note d_i (or d when there is no ambiguity) the index such that y_{id} is the first missing value of i.

2.2. Classification of missingness

In their founding documents, Rubin and Little distinguished three kinds of missingness [14,15]. They considered trajectories without missingness Y_{TRUE} (unavailable data) and trajectories with missing values Y_{OBS} (available measured longitudinal data). Then R denote the Boolean matrix of the location of a missing value and Y_{MISS} the missing part of Y_{TRUE}. Thus, $Y_{TRUE} = Y_{OBS} + Y_{MISS}$. The classification of Little and Rubin is then based on a potential link between R, Y_{TRUE}, Y_{OBS}, and Y_{MISS}:

- **MCAR**: A value is *Missing Completely At Random* when $P(y_{ij})$, the probability that y_{ij} be missing, is independent of Y_{TRUE}. In this article, we assume that $P(y_{ij}) = p_0$.
- **MAR**: A value is *Missing At Random* when the probability that y_{ij} be missing is independent of Y_{MISS}, but may depend on the observed values Y_{OBS} (and optionally on some other observed variables). For example, if patients who performed badly at time j–1 decide to miss time j, the missing data will be MAR: $P(y_{ij}) = F(Y_{OBS})$.
- **MNAR**: A value is *Missing Not At Random* when the probability that y_{ij} be missing depends on Y_{MISS}. Typically, the probability for an observation y_{ij} to be missing at time j depends on the current value of Y at time j. For example, if patients who would perform badly at time j refuse to be tested at time j, the data will be MNAR: $P(y_{ij}) = F(Y_{MISS})$.

The impact of the mechanism of missingness on the imputation of the missing values was examined by Molenberghs [16].

In the particular case of longitudinal data, the missingness mechanisms are also classified according to the position of the missing values within the trajectory:

- **Intermittent missing data** are missing within a trajectory. Formally, y_{ij} is an intermittent missing value if there exists a and b, $a < j < b$, such that y_{ia} and y_{ib} are not missing.
- **Monotone missing data** are missing either at the beginning or at the end of a trajectory. This includes the case of right (or left) censored follow-ups. When a value is missing, then all the following (respectively, preceding) values are also missing. Formally, y_{ij} is a right monotone missing value if, for all $j' > j$, $y_{ij'}$'s are missing. y_{ij} is a left monotone missing value if, for all $j' < j$, $y_{ij'}$'s are missing.

In this article, we will focus on right monotone missing data, either MCAR, MAR, or MNAR (All the results can be generalized to left monotone missing data).

Independent of the MCAR, MAR, and MNAR classification, monotonous missing values have a specificity: only the first missing value y_{id} of trajectory $y_{i.}$ is directly related to the missingness mechanism. The following missing values $y_{id'}$ with $d' > d$ are missing with probability 1. This might give the impression that only the first value is MCAR or MNAR and the others are MAR. However, the probability $Pr(R_{id'} = 1)$ depends on $R_{i,d'-1}$, the probability of $Pr(R_{I,d'-1} = 1)$ depends on $R_{i,d'-2}$ which, step by step, depends on R_{id}.

- In the case of MCAR, $Pr(R_{id} = 1)$ is p_0 and is therefore independent of Y_{OBS}. Thus, all the $Pr(R_{id'} = 1)$ are independent of Y_{OBS} and are MCAR.
- In the case of MAR, $Pr(R_{id} = 1)$ depends on Y_{OBS} but not on Y_{MISS}. So, $Pr(R_{id'} = 1)$ also depends on Y_{OBS} but not on Y_{MISS} and all the missing values are MAR.
- Similarly, in the case of MNAR, $Pr(R_{id} = 1)$ depends on Y_{OBS} and on Y_{MISS}. So, $Pr(R_{id'} = 1)$ also depends on Y_{OBS} and on Y_{MISS} and all the missing values are MNAR.

2.3. *Imputation methods*

Herein, 16 imputation methods are compared. They were grouped according to the information necessary for their implementation and are summarized in Table 1.

2.3.1. *Cross-sectional imputation*

These methods use only data collected at a given time (time at which the value is missing). The imputation of a missing value at time j is made according to the values from the other individuals observed at time j, i.e., the cross-sectional measurement $y_{.j} = (y_{1j}, y_{2j}, \ldots, y_{nj})$.

1. The **Cross Mean** method replaces y_{ij} by the mean of the values observed at time j.
2. The **Cross Median** method replaces y_{ij} by the median of the values observed at time j.
3. The **Cross Hot Deck** method replaces y_{ij} by a value randomly chosen among all values observed at time j.

Table 1 – Imputation methods and their characteristics.

Imputation method	Cross-sectional	Longitudinal
1. Cross Mean	✓	
2. Cross Median	✓	
3. Cross Hot Deck	✓	
4. Traj Mean		✓
5. Traj Median		✓
6. Traj Hot Deck		✓
7. LOCF		✓
8. Interpolation *Global*		✓
9. Interpolation *Local*		✓
10. Interpolation *bisector*		✓
11. Spline Interpolation		✓
12. Linear Regression	✓	✓
13. CopyMean *LOCF*	✓	✓
14. CopyMean *Global*	✓	✓
15. CopyMean *Local*	✓	✓
16. CopyMean *bisector*	✓	✓

2.3.2. *Longitudinal imputation*

These methods use only the non-missing data of the same subject. The imputation is made independently of the data from the other individuals, only the trajectory $y_{i.} = (y_{i1}, y_{i2}, \ldots, y_{it})$ is used.

4. The **Traj Mean** replaces y_{ij} by the average of the values of trajectory $y_{i.}$.
5. The **Traj Median** replaces y_{ij} by the median of the values of trajectory $y_{i.}$.
6. The **Traj Hot Deck** replaces y_{ij} by a value chosen randomly among the values of trajectory $y_{i.}$.
7. The **Last Occurrence Carried Forward** (LOCF) replaces y_{ij} by the previous non-missing value. This method shows many strong weaknesses [17], but is still frequently used in many scientific fields.
8. The **Interpolation Global** replaces y_{ij} by drawing a line joining the first and the last non-missing values (this line is the average progression of the actual individual trajectory). The missing values are chosen on this line.
9. The **Interpolation Local** replaces y_{ij} by drawing a line joining the last and the penultimate non-missing values. The missing values are chosen on this line.
10. The **Interpolation Bisector** Interpolation Global and Interpolation Local might be sensitive to abnormal values. Interpolation Bisector offers an intermediate solution by considering the bisector of Global and Local solution. The missing values are chosen on the bisector.
11. The **Spline Interpolation** provides imputation based on spline. We used the function *splinefun* from R with the method of Forsythe, Malcolm, and Moler [18]. For details, see Fritsch and Carlson [19].

2.3.3. *Cross-sectional and longitudinal imputation*

These methods use both longitudinal information $y_{i.}$ and cross-sectional information $y_{.j}$.

12. **Linear Regression**: it imputes by contructing models iteratively. The principle is, for each j, to construct a model that predicts the values of $y_{.j}$ using the other variables $y_{ij'}$ with $j' \neq j$. Because variables $y_{.j'}$ may also contain missing values, the process is iterative by gradual approximation:
    - Initially, all the missing values are imputed (by one of the methods described above). A model regressing $y_{.1}$ as a function of $y_{.2}$, $y_{.3}$, …, $y_{.t}$ is built. The missing values in $y_{.1}$ are replaced by the values predicted by the model.
    - A model regressing $y_{.2}$ as a function of $y_{.1}$, $y_{.3}$, …, $y_{.t}$ is built. The missing values in $y_{.2}$ are replaced by the values predicted by the model.
    - In the same way, all the $y_{.j}$ are imputed using models.

Then the process is iterated: a new model is constructed for $y_{.1}$ whose values are again calculated again, then for $y_{.2}$ and so on. Each iteration allows a little more precision in estimating the missing values.

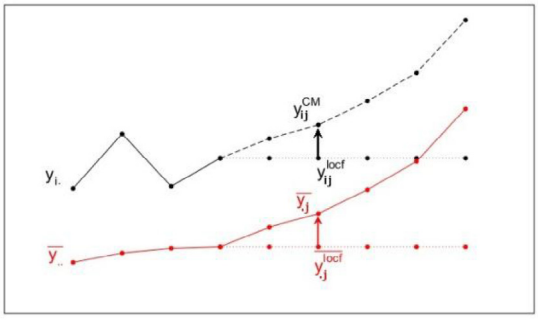


Fig. 1 – CopyMean, LOCF imputation. The individual trajectory $y_{.j}$ is in black, the mean trajectory $\overline{y_{..}}$ is in red. The dotted lines are the values imputed by LOCF. The dashed lines are the values imputed by CopyMean, LOCF. The arrows are the values of AV_j. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

After a predetermined number of iterations, the process stops. In this article, the initialization process used Cross Mean and the process was iterated 5 times.

Finally, **CopyMean** is an original method [13]. It is already included in an R package named LongitudinalData. However, its efficiency on monotone missing data has not been compared yet to other methods until today.

13. **CopyMean** (also called **CopyMean LOCF**) imputes in two steps. First, it uses the classical longitudinal imputation method LOCF to obtain an approximation of the imputed value. In a second step, it uses the information provided by the population's mean trajectory to refine the first approximation; that is, to give to the imputed trajectory the same shape as the mean trajectory.

Formally, let y_{ij} be the missing value.[1] Let $\overline{y_{..}} = (\overline{y_{.1}}, \ldots, \overline{y_{.t}})$ denote the mean trajectory of a population S. Let y_{id} be the first missing value of trajectory i. Then for all $j \geq d$, let y_{ij}^{LOCF} be the value obtained by imputing y_{ij} using *LOCF*. On the example of Fig. 1, trajectory y_{ij}^{LOCF} is in black dots.

Suppose for a while that the mean trajectory $\overline{y_{..}}$ has also some missing values for $j \geq d$. We note $\tilde{y}$ this hypothetical mean trajectory with missing values. Let $\overline{y_{.j}^{LOCF}}$ be the value obtained by applying the *LOCF* method on $\tilde{y}$ (Fig. 1, trajectory in red dots). Then the average variation AV_j at time j is the difference between $\overline{y_{.j}}$ and $\overline{y_{.j}^{LOCF}}$, i.e., $AV_j = \overline{y_{.j}} - \overline{y_{.j}^{LOCF}}$ (the red vertical arrow in Fig. 1). From there, the CopyMean imputes y_{ij} by adding the average variation AV_j to the result of the LOCF imputation: $y_{ij}^{CM} = y_{ij}^{LOCF} + AV_j$.

Note that the computation of the average variation AV_j (second step of CopyMean) depends only on the mean trajectory $y_{..}$ and on the method chosen to obtain the approximation in the first step of CopyMean. Therefore, CopyMean is not stochastic.

[1] All the notations introduced here are illustrated in Fig. 1.

In its first step, CopyMean uses LOCF to obtain a first approximation of the missing value. Many other methods are possible. We present here three possible variations of CopyMean:

14. **CopyMean Global**: in the first step, this method uses *Interpolation Global* to approximate the missing value. Then it adds the variation AV_j (the variation that makes the imputed value follow the shape of the average trajectory) to each approximated imputed value.
15. **CopyMean Local**: in the first step, this method uses *Interpolation Local*. Then it adds the variation AV_j to each approximated imputed value.
16. **CopyMean Bisector**: in the first step, this method uses *Interpolation bisector*. Then it adds the variation AV_j to each approximated imputed value.

3. Simulation

3.1. *Missing value generation*

The present simulation study was performed using two real datasets with complete data and two artificial (model-based) datasets [20].

3.1.1. *Presentation of the real datasets*

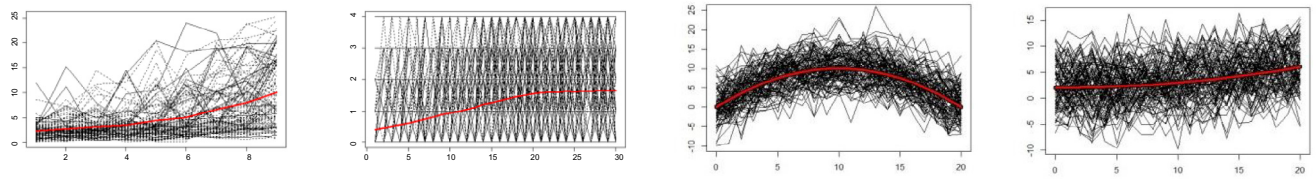
To be as general as possible, we worked on datasets with different characteristics in terms of number of individuals, lengths of trajectories, and trajectory shapes. These datasets are presented in Fig. 2.

3.1.1.1. *Pregnandiol.* The first dataset (Fig. 2a) comes from a study on human menstrual cycles [21]. The initial aim of that study was a search for biomarkers for accurate prediction of ovulation. 102 women were recruited from eight natural family planning clinics located in Aix-en-Provence, Dijon, and Lyon (France), Milano and Verona (Italy), Dusseldorf (Germany), Liège (Belgium) and Madrid (Spain). Urine pregnanediol-3a-glucuronide was measured daily. The measured were made before ovulation. This variable is continuous in the range [0.05; 26.6] mg/L (overall mean: 11.5 mg/L; overall standard deviation: 18.3). The trajectories of this variable have the characteristics of being non-stationary and increasing. Of the 102 trajectories, two (1.96% of the total) had missing values. These trajectories were removed from the present study.

3.1.1.2. *Alcohol.* The second dataset (Fig. 2b) comes from the Québec Longitudinal Study of Child Development led by the GRIP [22]. In that study, 1831 participants were interviewed retrospectively; thus, the data show a very low rate of missingness. The monthly alcohol consumption was rated on a five-point scale (0–4, overall mean: 1.18; overall standard deviation: 1.09). The main feature of this study is the stability of the values over time. Three trajectories had missing values (0.16% of the total); they were removed from the present study.

3.1.2. *Generation of artificial datasets*

To generate the artificial datasets S_A and S_B, we chose the sample size $n = 200$ or 400, the number of repeated measurements ($t = 21$



(a) Pregnandiol (b) Alcohol (c) Artificial A (d) Artificial B

Fig. 2 – Graphical representations of the datasets (the mean trajectory is in red). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

or 41), the individual variance s_m and the noise variance s_r ($s_m = s_r = 1$ or $s_m = s_r = 3$), and a generation function g:

- **Set A**: $g_A(j) = \frac{-j^2}{10} + 2j$
- **Set B**: $g_B(j) = \frac{-j^2}{100} - 2j + 30$

Each individual trajectory y_{ij} was then defined as being equal to the function $g(j)$ plus a personal variation m_i plus some noise r_{ij}: $y_{ij} = g_A(j) + m_i + r_{ij} = \frac{-j^2}{10} + 2j + 10 + m_i + r_{ij}$ with $m_i \sim \mathcal{N}(0, s_p)$ and $r_{ij} \sim \mathcal{N}(0, s_r)$. Repeating this process for each i in 1, …, n generated a dataset.

We generated 500 datasets for each condition: Set A or Set B, size 200 or 400, repeated measurement 21 or 41 and variances 1 or 3.

3.1.3. *Generation of a specific percentage of missing data*

For each dataset, we generated 9 types of incomplete datasets using 3 types of missingness mechanisms (MCAR, MAR or MNAR) and 3 different percentages of overall missing data (10%, 30% or 50%). With the real dataset, this process was repeated 500 times. With the 500 artificial dataset, this process was repeated once. Overall, in each case, we obtained 500 incomplete datasets. A summary of the missing data generation process is given in Fig. 3.

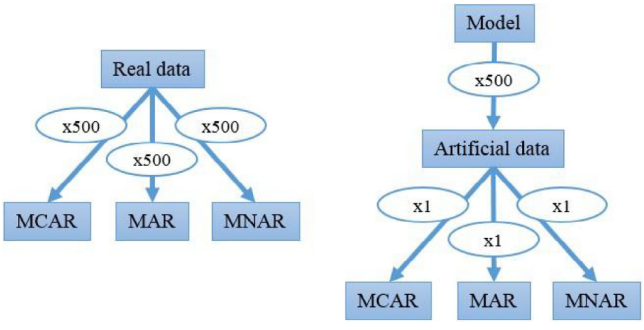


Fig. 3 – Generation of the missing values. With the real datasets, 500 datasets MCAR, 500 datasets MAR and 500 datasets MNAR were generated. With the models, 500 complete artificial datasets were built. From each of them, a dataset MCAR, a dataset MAR and a dataset MNAR were generated. The overall process was repeared for each pourcentage of missingness.

The fully detailed method used to generate the missing data is given in Appendix A.

3.2. *Imputation quality comparison criteria*

To assess the quality of each imputation method, we considered two types of criteria: (i) criteria based on the deviation between the true and the imputed values; and, in the case of artificial data, (ii) criteria based on the difference between the true and the imputed models.

The deviation is the difference between the true and the imputed value [1]: $e_{ij}^{IM} = y_{ij}^{IM} - y_{ij}$. It measures the proximity between the imputed and the true values. The deviation then leads to three sub-criteria:

1. The **Mean Absolute Deviation** (MAD) is the average of the absolute deviations: $MAD = \frac{\sum_{ij} |e_{ij}^{IM}|}{nt}$.
2. The **Root Mean Square Deviation** (RMSD) is the square root of the mean of the square of the deviation: $RMSD = \sqrt{\frac{\sum_{ij} (e_{ij}^{IM})^2}{nt}}$.
3. The **Bias** is the mean of the deviation: $Bias = \frac{\sum_{ij} e_{ij}^{IM}}{nt}$.

The bias can also be seen as a criterion for measuring the population parameters. Indeed, we can define the cross-sectional Bias at time j as $Bias_j = \frac{\sum_i e_i^{IM} j}{n}$. The Bias is the mean of all the Biases: $Bias = \frac{\sum_j Bias_j}{t}$. Then we have $y_{.j}^{IM} = \frac{\sum_i y_{ij}^{IM}}{n} = \frac{\sum_i y_{ij} + e_{ij}^{IM}}{n} = \frac{\sum_i y_{ij}}{n} + \frac{\sum_i e_{ij}^{IM}}{n} = y_{.j} + Bias_j$, that is $Bias_j = y_{.j}^{IM} - y_{.j}$. Finally, we get $Bias = \frac{\sum_j y_{.j}^{IM} - y_{.j}}{t}$. Thus, the Bias also measures the differences between the imputed trajectory $y_{..}^{IM}$ and the mean trajectory $y_{..}$.

In the case of artificial data built from a model, we also used a second criterium to measure the accuracy of the model estimation: let g^{IM} be the estimation of g obtained by imputing using imputation method IM. Then:

4. The **Mean Square Error** is the mean square difference between the model estimation g^{IM} and the model g: $MSE = \frac{\sum_j (g^{IM}(j) - g(j))^2}{t}$.

3.3. Methods and softwares

All the analyses were performed with R software [23] using S4 methods [24]. Classical and new imputation methods have been programmed in package LongitudinalData [25]. The generation of the artificial datasets was made using the function generateArtificialLongData from package kml [11,26,27]. Both packages are on CRAN [23].

4. Results

During data construction, three mechanisms of missingness (MCAR, MAR, and MNAR), three percentages of missing data (10%, 30%, and 50%) and four types of data (Pregnanediol, Alcohol, artificial dataset A and artificial dataset B) were considered. The analysis of the results showed that the missingness mechanism and the dataset had impacts on the performance of the methods, but the percentage of missing data has not. So we present the results of all the percentages merged together. On artificial data, the number of individuals, the number of repeated measurements and the variance had very little impacts on the performance. So we present here the results for $n = 200$, $t = 21$ and $s_m = s_r = 3$.

4.1. Mean absolute deviation results

Table 2 presents the mean results for each method according to the missingness mechanism and the dataset. For better readability, the performances of the best methods (the lowest MAD) are written in bold.

With Pregnanediol data, CopyMean-LOCF was the best for all mechanisms. With Alcohol dataset, classical LOCF gave the best results.

4.2. Root mean square deviation results

Table 3 presents the root mean square deviation results. The performance of the best method (the lowest RMSD) is written in bold.

CopyMean-LOCF gave the best results in all cases.

4.3. Bias results

Table 4 presents the results for bias. The results of the best methods are written in bold.

There were important differences in bias between MCAR, MAR, and MNAR mechanisms. The bias was slightly larger with the MAR than with the MCAR and even larger with the MNAR. The worst results were obtained with the MNAR mechanism. CopyMean shows once again a great efficiency, like Cross and Spline methods.

4.4. Mean square error results

Table 5 presents the results for the mean square error (artificial data only). The results of the best methods are written in bold.

4.5. Summary

Table 6 summarizes the results obtained with all the methods and criteria. Each column shows the number of times a method has been the best according to the above-defined criteria (Tables 2–4).

5. Discussion

In this article, we compare different methods for imputing trajectories. Monotone missing data were generated according to three different mechanisms (MCAR, MAR, and MNAR) in two datasets exhibiting strong structural differences. 12 conventional methods and 4 original techniques were compared according to three performance criteria: the Mean Absolute Deviation, the Root Square Mean Deviation, and Bias.

Because the evaluation criteria are numerous, it is difficult to conclude such a study with an assertion that a given method is superior to all others. Still, in many cases, this study

Table 2 – MAD (Mean Absolute Deviation) according to the imputation method.

	Pregnandiol			Alcohol			Artificial A			Artificial B		
	MCAR	MAR	MNAR	MCAR	MAR	MNAR	MCAR	MAR	MNAR	MCAR	MAR	MNAR
crossMean	4.242	4.472	4.591	0.894	0.914	0.918	1.156	1.157	1.147	0.767	0.79	0.79
crossMedian	4.064	4.192	4.558	0.864	0.885	0.893	1.163	1.165	1.153	0.772	0.795	0.797
crossHotDeck	5.515	5.718	5.872	1.161	1.177	1.182	1.605	1.617	1.605	1.077	1.085	1.084
trajMean	4.276	4.085	4.751	0.808	0.749	0.792	4.329	3.642	3.816	0.602	0.609	0.646
trajMedian	4.422	4.333	4.993	0.819	0.792	0.83	4.373	3.709	3.82	0.606	0.617	0.651
trajHotDeck	4.45	4.39	4.912	0.83	0.796	0.825	4.778	4.311	4.404	0.637	0.647	0.677
locf	4.087	4.404	4.178	0.666	**0.598**	**0.525**	3.741	3.621	3.49	0.57	0.471	0.558
linearInterpol.global	4.672	5.352	4.545	0.843	1.039	0.861	6.028	6.479	6.301	1.02	0.837	0.842
linearInterpol.local	5.199	6.447	5.161	0.727	0.918	0.689	5.425	5.557	5.495	1.773	1.629	1.629
linearInterpol.bisector	4.826	5.657	4.711	0.814	1.059	0.827	5.352	5.663	5.375	1.299	1.143	1.131
spline	7.481	8.975	7.829	0.959	1.239	1.001	7.837	8.065	8.236	2.743	2.651	2.679
regression	4.242	4.472	4.599	0.894	0.914	0.918	1.156	1.157	1.147	0.767	0.726	0.662
copyMean.locf	**3.638**	**3.979**	**3.843**	**0.626**	0.802	0.687	**1.135**	**1.149**	**1.136**	**0.348**	**0.391**	**0.368**
copyMean.global	4.549	5.685	4.461	0.865	1.002	0.85	2.923	2.175	2.322	0.967	0.876	0.789
copyMean.local	5.251	6.77	5.313	0.841	1.027	0.801	5.15	5.008	5.187	1.755	1.629	1.611
copyMean.bisector	4.708	5.963	4.676	0.851	1.057	0.837	4.181	3.838	4	1.266	1.172	1.1

Table 3 – RMSD (Root Mean Square Deviation) according to the imputation method.

	Pregnandiol			Alcohol			Artificial A			Artificial B		
	MCAR	MAR	MNAR	MCAR	MAR	MNAR	MCAR	MAR	MNAR	MCAR	MAR	MNAR
crossMean	29.318	30.958	34.523	1.138	**1.177**	1.2	2.109	2.109	2.072	3.113	3.277	3.294
crossMedian	32.857	32.815	39.906	1.357	1.381	1.426	2.132	2.137	2.095	3.156	3.32	3.36
crossHotDeck	56.485	59.419	61.726	2.268	2.307	2.323	4.057	4.095	4.044	6.108	6.184	6.164
trajMean	40.947	37.187	46.157	1.476	1.2	1.343	25.016	18.031	19.563	1.75	1.746	1.906
trajMedian	43.877	42.154	50.984	1.658	1.561	1.66	25.807	19.245	20.098	1.773	1.79	1.935
trajHotDeck	43.973	42.707	49.748	1.713	1.627	1.687	31.922	27.265	28.123	2.009	2.049	2.198
locf	37.869	33.521	37.528	1.253	1.236	1.076	21.218	21.429	19.665	1.67	1.171	1.58
linearInterpol.global	47.095	57.993	42.635	1.707	2.381	1.798	52.162	58.361	56.453	6.803	4.774	4.851
linearInterpol.local	55.611	79.664	52.539	1.56	2.437	1.544	43.995	46.071	45.24	15.83	13.357	13.823
linearInterpol.bisector	49.508	64.006	45.157	1.671	2.63	1.767	43.656	47.9	44.602	9.589	7.386	7.516
spline	100.09	133.447	105.646	2.193	3.298	2.355	77.917	80.748	84.197	30.736	28.935	29.57
regression	29.318	33.633	34.307	1.138	1.177	1.2	2.109	2.109	2.072	3.113	2.995	2.757
copyMean.locf	**24.853**	**30.932**	**27.027**	**0.953**	1.428	**1.039**	**2.023**	**2.086**	**2.028**	**0.637**	**0.807**	**0.72**
copyMean.global	43.445	61.868	39.724	1.699	2.238	1.721	16.649	8.956	10.312	6.2	4.977	4.35
copyMean.local	55.246	84.018	53.694	1.655	2.471	1.587	40.418	38.182	40.288	15.481	13.259	13.479
copyMean.bisector	46.195	67.674	43.016	1.694	2.562	1.726	28.849	24.241	25.788	9.034	7.585	7.006

showed the particular efficiency of the CopyMean method. These results comfort those already found in the case of intermittent missing values [13].

More precisely, method CopyMean LOCF outperforms the other methods. Methods CopyMean global, CopyMean local, and CopyMean bisector showed lower performance because global, local, and bisector methods are highly dependent on some specific values of the trajectories. For example, a slight change in the last known value $y_{i,d-1}$ changes the slope of the line used for the first step of CopyMean. In the case of a missing value y_{ij} distant from $y_{i,d-1}$, the change is multiplied by $j - d$. Thus a "small" change in $y_{i,d-1}$ has a significant impact on the imputed value at time j. CopyMean LOCF depends also on a specific value (the last known value), but there is no multiplication of the effect. This makes CopyMean LOCF more reliable than CopyMean local, global, or bisector.

LOCF exhibited few good results on "stationary" dataset, i.e., those in which the trajectories are almost constant (like alcohol consumption, a phenomenon that changes very little over time) but showed some strong weaknesses on data that vary.

On the contrary, the Cross-sectional imputation method showed better performances on dataset in which the variance is small regarding the variation of the mean trajectory. On such datasets, all the individuals can be considered as "close" to the mean trajectory; thus, these methods perform correctly.

Imputation using spline was one of the worst. It is linked to the fact that this method imputes using polynomials. This amplifies the tendency observed on the very last non missing values of the trajectory: if the slope increases, the imputed value will be very high, if the slope decreases, the imputed value will be very low. Finally, all the method based on some specific individual value (all the interpolation, copyMean.global, copyMean.local and copyMean.bissector) showed poor performance. This is due to a leverage effect: a small variation in the last non-missing value can have a significant impact on the imputation of the missing values at the end of the trajectories.

Table 4 – Bias according to the imputation method.

	Pregnandiol			Alcohol			Artificial A			Artificial B		
	MCAR	MAR	MNAR	MCAR	MAR	MNAR	MCAR	MAR	MNAR	MCAR	MAR	MNAR
crossMean	**–0.025**	**0.533**	–0.62	**–0.001**	0.037	–0.023	–0.009	0.016	0.02	0.02	0.023	**0.002**
crossMedian	–1.62	–0.922	–2.171	–0.359	–0.241	–0.337	**–0.002**	0.024	0.028	0.025	0.033	0.005
crossHotDeck	–0.048	0.535	–0.641	–0.002	**0.036**	–0.022	–0.013	**0.007**	**0.014**	0.017	**0.022**	0.004
trajMean	–3.76	–3.3	–4.366	–0.685	–0.532	–0.631	–2.458	–1.239	–1.807	–0.548	–0.557	–0.611
trajMedian	–4.013	–3.776	–4.68	–0.701	–0.608	–0.684	–2.066	–0.718	–1.32	–0.55	–0.563	–0.614
trajHotDeck	–3.761	–3.308	–4.367	–0.684	–0.531	–0.632	–2.453	–1.245	–1.794	–0.548	–0.556	–0.612
locf	–3.083	–1.254	–3.362	–0.43	0.174	–0.143	0.012	2.05	1.239	–0.46	–0.271	–0.463
linearInterpol.global	–1.518	1.301	–1.914	–0.113	0.688	0.274	5.559	6.393	6.08	–0.356	0.095	–0.345
linearInterpol.local	–1.106	2.449	–1.282	–0.325	0.501	0.021	2.927	3.665	2.666	–0.286	0.489	–0.309
linearInterpol.bisector	–1.464	1.64	–1.781	–0.18	0.682	0.206	4.496	5.329	4.714	–0.333	0.299	–0.331
spline	1.437	4.485	1.229	–0.302	0.411	–0.036	–0.271	–0.341	–1.221	–0.34	0.069	–0.418
regression	–0.025	0.537	–0.596	–0.001	0.038	**–0.019**	–0.009	0.016	0.02	0.02	0.023	–0.019
copyMean.locf	–0.041	1.93	–0.557	–0.007	0.523	0.21	–0.004	0.256	–0.035	**0.006**	0.172	–0.024
copyMean.global	–0.132	2.804	–0.466	–0.236	0.562	0.156	–0.145	0.488	–0.079	–0.057	0.361	–0.059
copyMean.local	–0.061	3.505	**–0.255**	–0.414	0.404	–0.067	–0.192	1.077	–0.296	–0.157	0.583	–0.203
copyMean.bisector	–0.169	2.999	–0.459	–0.288	0.578	0.104	–0.201	0.98	–0.17	–0.104	0.484	–0.121

Table 5 – MSE (Mean Square Error) according to the imputation method.

	Artificial A			Artificial B		
	MCAR	MAR	MNAR	MCAR	MAR	MNAR
crossMean	0.022	**0.022**	0.017	0.156	0.218	0.202
crossMedian	0.022	0.023	0.018	0.166	0.223	0.212
crossHotDeck	0.023	0.023	0.018	0.164	0.224	0.206
trajMean	2.728	2.067	2.251	1.053	1.192	1.33
trajMedian	2.636	2.071	2.154	1.058	1.213	1.341
trajHotDeck	2.726	2.078	2.257	1.055	1.192	1.335
locf	1.925	2.803	2.225	0.796	0.49	0.842
linearInterpol.global	7.715	10.377	9.43	0.603	**0.165**	0.547
linearInterpol.local	2.754	4.202	2.79	0.63	0.56	0.563
linearInterpol.bisector	5.576	7.848	6.487	0.598	0.26	0.529
spline	1.322	1.398	1.478	1.03	0.725	1.102
regression	0.022	0.022	0.017	0.156	0.212	0.194
copyMean.locf	**0.018**	0.023	**0.015**	**0.118**	0.184	**0.162**
copyMean.global	0.094	0.077	0.044	0.214	0.337	0.183
copyMean.local	0.378	0.614	0.328	0.442	0.754	0.385
copyMean.bisector	0.232	0.367	0.161	0.286	0.544	0.226

5.1. *Limitations*

In the present study, we used four datasets with marked differences in terms of shape, number of individuals, number of repeated measurements, and type of the outcome variable. Nevertheless, because these datasets were only examples, a generalization of our results to other datasets should be examined with caution.

Another important point is that all the methods presented here made the assumption that the population is uniform. It might be interesting to study the performance of imputation methods considering that the population consists of subgroups. CopyMean would calculate *AV*, the average variation, relative to the mean trajectory of its group. Furthermore, imputation could be combined with the partitioning process, which is usually iterative: At each iteration, the missing data would be imputed relatively to their cluster. Then the clusters would be modified (according to the clustering algorithm) and the values of imputed data would be recalculated.

Table 6 – Number of times a method has been particularly performant.

Imputation method	MAD	RMSD	Bias	MSE	Total
1. crossMean	0	1	4	1	**6**
2. crossMedian	0	0	1	0	**1**
3. crossHotDeck	0	0	4	0	**4**
4. trajMean	0	0	0	0	**0**
5. trajMedian	0	0	0	0	**0**
6. trajHotDeck	0	0	0	0	**0**
7. locf	2	0	0	0	**2**
8. linearInterpol.global	0	0	0	1	**1**
9. linearInterpol.local	0	0	0	0	**0**
10. linearInterpol.bisector	0	0	0	0	**0**
11. spline	0	0	0	0	**0**
12. regression	0	0	1	0	**1**
13. copyMean.locf	10	11	1	4	**26**
14. copyMean.global	0	0	0	0	**0**
15. copyMean.local	0	0	1	0	**1**
16. copyMean.bisector	0	0	0	0	**0**

We compared the performance of the methods using different indices either based on the differences between the true values and the imputed values or indices that assess the quality of the modeling (Bias, MAD, RMSE, and MSE). It might be also interesting to study the imputed methods with regard to some applications. Within the context of artificial data, when the data characteristics are known (such as the partition of the population), one may compare the imputation methods regarding their abilities to find these characteristics.

Finally, the performance of the methods we used varied from one dataset to another. In the literature, most studies used single datasets without detailed descriptions of their characteristics. Here, the datasets were described with a few parameters but it is probably possible to be much more accurate. The list of the necessary and sufficient parameters that fully characterize longitudinal data is still to be defined.

Acknowledgements

This research was funded by the ANR grant "IDOL: ANR-12-BSV1-0036". We would like to thank Jean Iwaz (Hospices Civils de Lyon) for his conscientious proofreading. We also thank the reviewers whose careful readings and insightful comments allowed us to greatly improve the quality of this article.

Appendix

Appendix ADetails of the method used to generate the missing values

To generate missing values in a complete dataset, we defined R_{ij} the missing value indicator [28] and a probability function $P(R_{ij}=1)$ that y_{ij} (and all the subsequent values, since the missing values are monotones) be missing for j in [2, t].

$$R_{ij}=\begin{cases}1 & \text{if } y_{ij} \text{ is missing}\\ 0 & \text{otherwise}\end{cases}$$

Note that the first value is always observed. The probability function depends on the missingness mechanism, see section 2.2. In the MCAR case, this probability is independent of Y: $P^{MCAR}(R_{ij}=1)=p_0$. In the MAR case, the probability depends on $y_{i,d-1}$ where $y_{i,d-1}$ is the last observed value preceding y_{ij}: $P^{MAR}(R_{ij}=1)=f(y_{i,d-1})$. Finally, in the MNAR case, the probability depends on y_{id} where y_{id} is the first missing value preceding y_{ij}: $P^{MNAR}(R_{ij}=1)=f(y_{id})$.

Choosing f and its parameters in order to obtain a specific percentage of missing data is a complex task. Most authors give no details on the method they use and there is almost no literature on the subject. In this article, we use the following method:

Let P_g be the global percentage of missing data that we would like to obtain. For the i^{th} trajectory, we call M_i the number of missing data for the sequence $y_{i.}$ (Note that since d_i is the index of the first missing value for trajectory i, we have $M_i=t-d_i+1$). For n subjects and t times, we want to generate missing values such that:

$$\frac{\sum_{i=1}^{n} M_i}{n\cdot t}=P_g \tag{A1}$$

That is

$$\mathbb{E}[M_i]=t\cdot P_g \tag{A2}$$

Appendix A.1 Case MCAR

For the safeness of the notation, we note $p_{ij}=P^{MCAR}(R_{ij}=1)$.

In the case of monotone data and MCAR missingness mechanism, we know that the probability of being missing P_{ij} is constant over time and is independent of the trajectory. Let $p_0=p_{ij}$ be the probability of missingness at $j=2,\ldots,t$.

- At time $j=2$, y_{ij} can be missing with probability p_0, then the number of missing values is $M_i=(t-1)p_0$.
- At time $j=3$, y_{ij} can be missing with probability $p_0(1-p_0)$ (not missing at time $t=2$ but missing at time $t=3$) then the number of missing values is $M_i=(t-2)p_0(1-p_0)$.
- At time $j=4$, y_{ij} can be missing with probability $p_0(1-p_0)^2$ (not missing at times $t=2,3$ but missing at time $t=4$) then the number of missing values is $M_i=(t-3)p_0(1-p_0)^2$.
- At time $j=k$, y_{ij} can be missing with probability $p_0(1-p_0)^{k-2}$ and the number of missing values is $M_i=(t+1-k)p_0(1-p_0)^{k-2}$.

Overall, the expected number of missing data for the sequence $y_{i.}$ is

$$\mathbb{E}[M_i]=(t-1)p_0+\sum_{k=3}^{t}(t+1-k)p_0(1-p_0)^{k-2} \tag{A3}$$

Combining (2) and (3), we get:

$$t\cdot P_g=(t-1)p_0+\sum_{k=3}^{t}(t+1-k)p_0(1-p_0)^{k-2} \tag{A4}$$

This is a polynomial of degree $k-1$. It is not possible to solve it in the general case but, in our specific situation, an approximative solution is sufficient. We have:

Table A1 – Parameters for creating missing data in the MCAR case.

Pg	Pregnanediol	Alcohol
	p_0	p_0
10% missing	0.02658423	0.007379685
30% missing	0.09224992	0.0258696
50% missing	0.1849447	0.05055199

Appendix A.2 Case MAR

In case MAR, the missingness probability depends on the last observed variables y_{id}. Without loss of generality, we took two assumptions: (i) $P(R_{ij}=1$ depends on the value of $y_{i,j-1}$ and (ii) the higher is $y_{i,j-1}$, the higher is $p_{ij}=P^{MAR}(R_{ij}=1)$ too. This models the fact that, for example, a patient who had a bad exam result is more prone to not going to the next exam.

More precisely, let y_{min} and y_{max} be the lowest and the highest value of the whole dataset. We note $p_0=P(y_{ij}$ following y_{min} is missing) and we arbitrarily choose $P(y_{ij}$ following y_{max} is missing$)=2p_0$. We also assume linearity:

$$p_{ij}=\left(\frac{y_{i,j-1}-y_{min}}{y_{max}-y_{min}}+1\right)p_0 \tag{A5}$$

As previously said, our problem is to find p_0 such that $\mathbb{E}[M_i]=t\cdot P_g$.

In the case MAR, we have:

- At time $j=2$, y_{ij} can be missing with probability p_{i2} then the number of missing values is $M_i=(t-1)p_{i2}$.
- At time $j=3$, y_{ij} can be missing with probability $p_{i3}(1-p_{i2})$ (not missing at time $t=2$ but missing at time $t=3$) then the number of missing values is $M_i=(t-2)p_{i3}(1-p_{i2})$.
- At time $j=4$, y_{ij} can be missing with probability $p_{i4}(1-p_{i3})(1-p_{i2})$ (not missing at times $t=2,3$ but missing at time $t=4$) then the number of missing values is $M_i=(t-3)p_{i4}(1-p_{i3})(1-p_{i2})$.
- At time $j=k$ (for $k\geq 3$), y_{ij} can be missing with probability $p_{ik}(1-p_{i,k-1})\ldots(1-p_{i2})$ then the number of missing values is $M_i=(t+1-k)p_{i,k}(1-p_{i,k-1})\ldots(1-p_{i2})$

Equation (A2) becomes:

$$\begin{aligned}\mathbb{E}[M_i]&=(t-1)p_{i2}+\sum_{k=3}^{t}\left[(t+1-k)p_{ik}\times\prod_{m=k-2}^{1}(1-p_{i,m+1})\right]\\&=(t-1)p_{i2}+\sum_{k=2}^{t-1}\left[(t-k)p_{i,k+1}\times\prod_{m=k-1}^{2}(1-p_{im})\right]\end{aligned} \tag{A6}$$

This equation is not solvable in the general case; so, to go further, some simplifications are necessary. In the definition of p_{ij}, we replaced $y_{i,j-1}$ by $\overline{y_{i,j-1}}$, the mean of the y_{ij} at time j.

Because of our data pattern (see the red line on Fig. 2), we also took the arbitrary assumption that the mean trajectory $\overline{y_{ij}}$ can be modeled by a linear regression. Let a and b be the coefficients of the regression $y=a.j+b$. With these assumptions, Equation (A5) becomes:

$$p_{ij} = \left(\frac{a(j-1) + b - y_{min}}{y_{max} - y_{min}} + 1 \right) p_0 \tag{A7}$$

For the safeness of the notations, let $A = \frac{a}{y_{max} - y_{min}}$ and $B = \frac{b - y_{min}}{y_{max} - y_{min}} + 1$. So (A7) can be rewritten

$$p_{ij} = (A(j-1) + B) p_0 \tag{A8}$$

Combining (A2), (A6) and (A8), we get:

$$t \cdot P_g = (t-1)(A+B) p_0 + \sum_{k=2}^{t-1} \left[(t-k)(A \cdot k + B) p_0 \times \prod_{m=k-1}^{1} 1 - (A \cdot m + B) p_0 \right] \tag{A9}$$

This is a polynomial of degree $t+1$. It still cannot be solved but we can approximate it easily using a root-finding algorithm. We get:

Table A2 – Parameters p_0 for creating missing data on MAR case.

	Pregnanediol	Alcohol
	p_0	p_0
10% missing	0.023739	0.005978
30% missing	0.082959	0.021199
50% missing	0.170553	0.044403

Appendix A.3 Case MNAR

In the MNAR case, the missingness probability depends on the unobserved variables. The process is very similar to the MAR process. The only difference is that p_{ij} depends on j (in the MAR case, it was depending of j–1). With the same hypothesis, we have:

$$p_{ij} = \left(\frac{y_{ij} - y_{min}}{y_{max} - y_{min}} + 1 \right) p_0 \tag{A10}$$

The problem is once again to find p_0 such that $\mathbb{E}[M_i] = t \cdot P_g$. In the case of MNAR, we have:

- At time $j = 2$, y_{ij} can be missing with probability p_{i2} so the number of missing values is $M_i = (t-1) p_{i2}$.
- At time $j = 3$, y_{ij} can be missing with probability $p_{i3}(1 - p_{i2})$ (not missing at time $t = 2$ but missing at time $t = 3$) so the number of missing values is $M_i = (t-2) p_{i3}(1 - p_{i2})$.
- At time $j = 4$, y_{ij} can be missing with probability $p_{i4}(1 - p_{i3})(1 - p_{i2})$ (not missing at times $t = 2,3$ but missing at time $t = 4$) so the number of missing values is $M_i = (t-3) p_{i4}(1 - p_{i3})(1 - p_{i2})$.
- At time $j = k$ (for $k \geq 3$), y_{ij} can be missing with probability $p_{ik}(1 - p_{i,k-1}) \ldots (1 - p_{i2})$ so the number of missing values is $M_i = (t+1-k) p_{ik}(1 - p_{i,k-1}) \ldots (1 - p_{i2})$

Equation (A2) becomes:

$$\begin{aligned} \mathbb{E}[M_i] &= (t-1) p_{i2} + \sum_{k=3}^{t} \left[(t+1-k) p_{ik} \times \prod_{m=k-1}^{2} (1 - p_{im}) \right] \\ &= (t-1) p_{i2} + \sum_{k=2}^{t-1} \left[(t-k) p_{i,k+1} \times \prod_{m=k}^{2} (1 - p_{im}) \right] \end{aligned} \tag{A11}$$

We replaced y_{ij} by the mean trajectories $\overline{y_{ij}}$ and we assumed that it can be modeled by a linear regression. The equation is identical to that of the MAR case, (A5). We get:

$$p_{ij} = (A \cdot j + B) p_0 \tag{A12}$$

Combining (A2), (A11) and (A12), we get:

$$t \cdot P_g = (t-1)(2 \cdot A + B) p_0 + \sum_{k=2}^{t-1} \left[(t-k)(A(k+1) + B) p_0 \times \prod_{m=k-1}^{2} 1 - (A \cdot m + B) p_0 \right] \tag{A13}$$

This is a polynomial of degree $t+1$. We can approximate it using a root-finding algorithm. We get:

Table A3 – Parameters p_0 for creating missing data on MAR case.

	Pregnanediol	Alcohol
	p_0	p_0
10% missing	0.023739	0.005978
30% missing	0.082959	0.021199
50% missing	0.170553	0.044403

Appendix B Detailed results

Appendix B.1 MAD

Table B1 – MAD on dataset Pregnandiol according to the imputation method.

	MCAR			MAR			MNAR		
	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
crossMean	4.356	4.228	4.14	4.576	4.499	4.341	4.802	4.611	4.361
crossMedian	4.142	4.037	4.013	4.192	4.222	4.162	4.702	4.578	4.393
crossHotDeck	5.622	5.553	5.37	5.877	5.777	5.501	6.225	5.941	5.449
trajMean	4.362	4.294	4.172	**4.025**	4.143	4.087	5.001	4.799	4.454
trajMedian	4.534	4.448	4.285	4.32	4.408	4.271	5.323	5.053	4.604
trajHotDeck	4.545	4.476	4.33	4.405	4.459	4.305	5.193	4.959	4.583
locf	4.124	4.087	4.05	4.768	4.372	4.072	4.201	4.188	4.144
linearInterpol.global	4.676	4.665	4.674	5.735	5.333	4.988	4.529	4.536	4.569
linearInterpol.local	5.306	5.209	5.082	7.247	6.464	5.629	5.363	5.194	4.927
linearInterpol.bisector	4.863	4.819	4.797	6.156	5.649	5.166	4.742	4.714	4.675
spline	7.878	7.608	6.955	10.045	9.138	7.741	8.581	8.006	6.899
regression	4.356	4.228	4.14	4.576	4.499	4.337	4.802	4.611	4.328
copyMean.locf	**3.667**	**3.621**	**3.626**	4.087	**3.964**	**3.887**	**3.956**	**3.804**	**3.77**
copyMean.global	4.51	4.527	4.61	6.173	5.653	5.229	4.448	4.415	4.519
copyMean.local	5.332	5.256	5.165	7.569	6.786	5.956	5.536	5.332	5.07
copyMean.bisector	4.699	4.685	4.739	6.521	5.949	5.42	4.722	4.65	4.656

Best values are in bold.

Table B2 – MAD on dataset Alcohol according to the imputation method.

	MCAR			MAR			MNAR		
	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
crossMean	0.897	0.895	0.891	0.918	0.913	0.909	0.925	0.917	0.911
crossMedian	0.862	0.866	0.865	0.865	0.894	0.896	0.879	0.9	0.899
crossHotDeck	1.165	1.162	1.158	1.18	1.178	1.174	1.187	1.182	1.176
trajMean	0.799	0.809	0.817	0.722	0.754	0.77	0.762	0.797	0.816
trajMedian	0.812	0.819	0.826	0.761	0.798	0.816	0.801	0.836	0.853
trajHotDeck	0.824	0.831	0.836	0.78	0.803	0.807	0.802	0.831	0.841
locf	0.641	0.664	0.692	**0.663**	**0.586**	**0.545**	**0.508**	**0.525**	**0.543**
linearInterpol.global	0.819	0.842	0.867	1.107	1.029	0.981	0.844	0.868	0.872
linearInterpol.local	0.699	0.728	0.756	1.031	0.897	0.827	0.669	0.688	0.709
linearInterpol.bisector	0.789	0.815	0.84	1.161	1.042	0.975	0.809	0.831	0.841
spline	0.944	0.961	0.971	1.338	1.22	1.16	0.982	1.003	1.018
regression	0.897	0.895	0.891	0.918	0.913	0.909	0.925	0.917	0.911
copyMean.locf	**0.598**	**0.623**	**0.655**	0.846	0.788	0.772	0.666	0.689	0.707
copyMean.global	0.834	0.863	0.898	1.046	0.989	0.97	0.823	0.852	0.876
copyMean.local	0.809	0.84	0.874	1.114	1.004	0.963	0.766	0.798	0.838
copyMean.bisector	0.819	0.849	0.884	1.141	1.036	0.995	0.806	0.837	0.868

Best values are in bold.

Table B3 – MAD on dataset Artificial A according to the imputation method.

	MCAR			MAR			MNAR		
	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
crossMean	1.157	1.158	1.154	1.159	1.154	1.159	**1.137**	1.152	1.153
crossMedian	1.162	1.162	1.165	1.162	1.159	1.173	1.138	1.158	1.164
crossHotDeck	1.598	1.61	1.607	1.632	1.608	1.611	1.607	1.607	1.6
trajMean	4.193	4.317	4.478	3.518	3.615	3.794	3.685	3.768	3.996
trajMedian	4.257	4.368	4.495	3.655	3.677	3.797	3.73	3.764	3.966
trajHotDeck	4.696	4.774	4.865	4.252	4.303	4.378	4.343	4.364	4.506
locf	3.728	3.714	3.781	3.779	3.617	3.468	3.567	3.465	3.438
linearInterpol.global	6.18	6.027	5.876	6.692	6.484	6.26	6.461	6.356	6.086
linearInterpol.local	5.447	5.375	5.454	5.586	5.523	5.563	5.502	5.516	5.468
linearInterpol.bisector	5.42	5.358	5.278	5.753	5.634	5.603	5.401	5.435	5.29
spline	8.03	7.804	7.679	8.127	8.099	7.969	8.409	8.244	8.057
regression	1.157	1.158	1.154	1.159	1.154	1.159	1.137	1.152	1.153
copyMean.locf	**1.139**	**1.122**	**1.143**	**1.148**	**1.148**	**1.151**	1.148	**1.127**	**1.134**
copyMean.global	2.731	2.9	3.138	2.016	2.167	2.341	2.18	2.273	2.514
copyMean.local	5.069	5.099	5.281	4.985	4.973	5.065	5.221	5.111	5.229
copyMean.bisector	4.019	4.134	4.389	3.699	3.812	4.004	3.949	3.918	4.132

Best values are in bold.

Table B4 – MAD on dataset Artificial B according to the imputation method.

	MCAR			MAR			MNAR		
	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
crossMean	0.257	0.762	1.283	0.263	0.783	1.324	0.264	0.776	1.33
crossMedian	0.257	0.764	1.295	0.263	0.788	1.333	0.264	0.779	1.348
crossHotDeck	0.358	1.083	1.79	0.36	1.085	1.809	0.362	1.076	1.813
trajMean	0.202	0.61	0.995	0.211	0.618	0.998	0.227	0.659	1.053
trajMedian	0.203	0.614	1	0.215	0.627	1.008	0.229	0.665	1.058
trajHotDeck	0.214	0.643	1.052	0.224	0.657	1.06	0.237	0.688	1.105
locf	0.19	0.567	0.954	0.152	0.461	0.799	0.177	0.544	0.952
linearInterpol.global	0.327	0.991	1.741	0.247	0.784	1.48	0.237	0.8	1.49
linearInterpol.local	0.589	1.768	2.964	0.551	1.607	2.73	0.522	1.613	2.753
linearInterpol.bisector	0.424	1.264	2.208	0.365	1.098	1.966	0.338	1.094	1.961
spline	0.956	2.813	4.461	0.926	2.721	4.306	0.94	2.741	4.357
regression	0.257	0.762	1.283	0.263	0.783	1.322	0.264	0.776	1.343
copyMean.locf	**0.111**	**0.342**	**0.592**	**0.123**	**0.369**	**0.68**	**0.112**	**0.345**	**0.646**
copyMean.global	0.304	0.934	1.661	0.258	0.817	1.552	0.216	0.742	1.41
copyMean.local	0.585	1.743	2.936	0.549	1.603	2.733	0.516	1.596	2.721
copyMean.bisector	0.414	1.233	2.152	0.37	1.13	2.016	0.325	1.065	1.91

Best values are in bold.

Appendix B.2 RMSD

Table B5 – RMSD on dataset Pregnandiol according to the imputation method.

	MCAR			MAR			MNAR		
	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
crossMean	30.195	28.931	28.828	**31.237**	**30.99**	30.648	34.986	34.269	34.316
crossMedian	33.476	32.338	32.757	32.2	32.835	33.41	39.866	40.082	39.771
crossHotDeck	58.263	56.967	54.226	62.363	60.209	55.686	66.803	62.816	55.558
trajMean	42.165	40.934	39.743	35.747	37.906	37.909	48.799	46.742	42.93
trajMedian	45.653	43.975	42.002	41.85	43.152	41.461	55.227	51.808	45.916
trajHotDeck	45.482	44.069	42.37	42.654	43.704	41.764	53.374	50.398	45.474
locf	38.352	37.636	37.619	33.653	33.244	33.665	36.899	37.586	38.098
linearInterpol.global	47.144	46.983	47.158	64.85	57.528	51.601	41.72	42.502	43.683
linearInterpol.local	57.725	55.789	53.319	96.943	79.599	62.451	55.541	53.114	48.962
linearInterpol.bisector	50.196	49.386	48.942	73.833	63.638	54.547	45.031	45.254	45.186
spline	109.238	102.853	88.197	160.647	136.703	102.992	123.1	109.396	84.442
regression	30.195	28.931	28.828	38.433	32.926	**29.54**	34.986	34.269	33.495
copyMean.locf	**25.109**	**24.527**	**24.924**	**31.237**	**30.99**	30.559	**27.119**	**26.219**	**27.743**
copyMean.global	43.008	43.169	44.157	71.276	61.16	53.167	38.973	39.118	41.08
copyMean.local	57.148	55.384	53.206	102.026	83.897	66.13	57.161	54.144	49.778
copyMean.bisector	46.432	45.948	46.204	79.431	67.136	56.457	43.22	42.752	43.077

Best values are in bold.

Table B6 – RMSD on dataset Alcohol according to the imputation method.

	MCAR			MAR			MNAR		
	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
crossMean	1.138	1.138	1.138	1.174	**1.175**	1.184	1.202	1.194	1.205
crossMedian	1.363	1.36	1.347	1.354	1.394	1.394	1.409	1.431	1.437
crossHotDeck	2.275	2.27	2.26	2.314	2.308	2.3	2.343	2.321	2.306
trajMean	1.435	1.474	1.519	**1.11**	1.203	1.287	1.253	1.342	1.434
trajMedian	1.638	1.657	1.68	1.487	1.574	1.623	1.59	1.668	1.722
trajHotDeck	1.695	1.715	1.73	1.587	1.638	1.655	1.632	1.698	1.732
locf	1.191	1.249	1.317	1.403	1.204	**1.101**	1.084	1.088	**1.055**
linearInterpol.global	1.637	1.704	1.779	2.601	2.351	2.19	1.791	1.809	1.795
linearInterpol.local	1.494	1.564	1.622	2.864	2.363	2.085	1.528	1.538	1.565
linearInterpol.bisector	1.604	1.673	1.738	3.017	2.57	2.302	1.763	1.774	1.765
spline	2.163	2.202	2.214	3.712	3.226	2.955	2.348	2.355	2.361
regression	1.138	1.138	1.138	1.174	1.175	1.184	1.202	1.194	1.204
copyMean.locf	**0.926**	**0.952**	**0.981**	1.625	1.4	1.258	**1.008**	**1.037**	1.071
copyMean.global	1.62	1.695	1.783	2.407	2.204	2.103	1.694	1.727	1.741
copyMean.local	1.587	1.655	1.722	2.864	2.394	2.157	1.552	1.579	1.631
copyMean.bisector	1.622	1.692	1.766	2.923	2.498	2.266	1.706	1.729	1.744

Best values are in bold.

Table B7 – RMSD on dataset Artificial A according to the imputation method.

	MCAR			MAR			MNAR		
	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
crossMean	2.121	2.108	2.097	2.107	2.107	2.114	**2.035**	2.09	2.093
crossMedian	2.14	2.123	2.133	2.117	2.128	2.166	2.042	2.109	2.134
crossHotDeck	4.043	4.076	4.051	4.143	4.083	4.06	4.029	4.068	4.036
trajMean	23.768	24.812	26.467	17.085	17.791	19.216	18.431	19.077	21.183
trajMedian	24.781	25.681	26.958	19.029	18.942	19.764	19.382	19.577	21.336
trajHotDeck	31.01	31.891	32.864	26.648	27.181	27.966	27.46	27.791	29.116
locf	21.325	20.958	21.371	23.062	21.423	19.801	20.571	19.543	18.882
linearInterpol.global	54.552	52.106	49.828	61.677	58.372	55.034	58.811	57.297	53.252
linearInterpol.local	44.614	43.341	44.03	46.748	45.696	45.768	45.638	45.672	44.411
linearInterpol.bisector	44.879	43.786	42.302	49.246	47.548	46.905	45.159	45.509	43.137
spline	81.01	77.367	75.373	81.663	81.213	79.367	87.136	84.097	81.358
regression	2.121	2.108	2.097	2.107	2.107	2.114	2.035	2.09	2.093
copyMean.locf	**2.031**	**1.99**	**2.049**	**2.105**	**2.07**	**2.083**	2.077	**1.989**	**2.017**
copyMean.global	14.795	16.445	18.708	7.64	8.825	10.404	9.031	9.949	11.956
copyMean.local	39.76	39.602	41.892	38.033	37.707	38.805	40.554	39.554	40.756
copyMean.bisector	27.204	28.192	31.152	22.838	23.96	25.924	25.139	25.046	27.179

Best values are in bold.

Table B8 – RMSD on dataset Artificial B according to the imputation method.

	MCAR			MAR			MNAR		
	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
crossMean	1.049	3.074	5.216	1.086	3.204	5.541	1.076	3.192	5.614
crossMedian	1.055	3.098	5.316	1.089	3.25	5.623	1.083	3.221	5.777
crossHotDeck	2.024	6.169	10.132	2.046	6.175	10.332	2.062	6.107	10.322
trajMean	0.584	1.78	2.887	0.608	1.773	2.855	0.676	1.953	3.089
trajMedian	0.592	1.806	2.922	0.632	1.825	2.914	0.69	1.991	3.122
trajHotDeck	0.676	2.042	3.308	0.728	2.097	3.323	0.788	2.26	3.545
locf	0.558	1.654	2.799	0.369	1.126	2.019	0.479	1.511	2.751
linearInterpol.global	2.177	6.512	11.721	1.317	4.269	8.736	1.201	4.558	8.794
linearInterpol.local	5.283	15.71	26.496	4.682	13.186	22.201	4.373	13.588	23.507
linearInterpol.bisector	3.15	9.196	16.42	2.299	6.942	12.916	2.048	7.176	13.325
spline	10.983	31.967	49.259	10.454	30.244	46.108	10.801	30.567	47.341
regression	1.049	3.074	5.216	1.086	3.204	5.496	1.076	3.192	5.722
copyMean.locf	**0.195**	**0.61**	**1.105**	**0.237**	**0.71**	**1.474**	**0.198**	**0.623**	**1.338**
copyMean.global	1.931	5.923	10.746	1.386	4.447	9.098	1.04	4.058	7.954
copyMean.local	5.192	15.321	25.931	4.643	13.052	22.083	4.278	13.284	22.875
copyMean.bisector	2.949	8.672	15.481	2.349	7.144	13.264	1.88	6.709	12.428

Best values are in bold.

Appendix B.3 Bias

Table B9 – Bias on dataset Pregnandiol according to the imputation method.

	MCAR			MAR			MNAR		
	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
crossMean	**–0.014**	–0.008	–0.053	0.647	**0.511**	0.441	–0.393	–0.645	–0.822
crossMedian	–1.71	–1.667	–1.485	–1.029	–0.975	–0.763	–2.198	–2.293	–2.023
crossHotDeck	–0.082	–0.01	**–0.052**	0.648	0.512	0.446	–0.446	–0.651	–0.826
trajMean	–3.863	–3.775	–3.641	–3.114	–3.372	–3.415	–4.635	–4.419	–4.044
trajMedian	–4.154	–4.036	–3.848	–3.737	–3.854	–3.737	–5.038	–4.743	–4.258
trajHotDeck	–3.868	–3.78	–3.635	–3.127	–3.38	–3.416	–4.636	–4.423	–4.041
locf	–3.091	–3.07	–3.087	**–0.424**	–1.283	–2.055	–3.282	–3.374	–3.429
linearInterpol.global	–1.61	–1.482	–1.463	2.316	1.288	**0.3**	–1.907	–1.887	–1.949
linearInterpol.local	–1.11	–1.065	–1.144	3.786	2.469	1.091	–1.131	–1.223	–1.493
linearInterpol.bisector	–1.537	–1.428	–1.426	2.745	1.638	0.536	–1.742	–1.747	–1.856
spline	1.671	1.596	1.045	5.744	4.618	3.094	1.542	1.412	0.733
regression	–0.014	–0.008	–0.053	0.647	0.511	0.449	–0.393	–0.645	–0.792
copyMean.locf	–0.046	–0.015	–0.061	2.616	1.925	1.249	–0.399	–0.53	–0.742
copyMean.global	–0.155	–0.071	–0.171	3.791	2.834	1.787	–0.352	–0.397	–0.651
copyMean.local	–0.082	**–0.004**	–0.097	4.711	3.538	2.266	**–0.147**	**–0.188**	**–0.429**
copyMean.bisector	–0.208	–0.111	–0.187	4.034	3.03	1.932	–0.378	–0.398	–0.601

Best values are in bold.

Table B10 – Bias on dataset Alcohol according to the imputation method.

	MCAR			MAR			MNAR		
	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
crossMean	0.003	**0**	**–0.005**	0.077	**0.032**	0.001	0.028	**–0.017**	–0.079
crossMedian	–0.426	–0.359	–0.292	–0.323	–0.217	–0.182	–0.388	–0.299	–0.322
crossHotDeck	**0.001**	–0.001	–0.006	**0.075**	0.032	**0**	0.031	–0.017	–0.079
trajMean	–0.665	–0.684	–0.705	–0.46	–0.534	–0.601	–0.578	–0.63	–0.687
trajMedian	–0.685	–0.7	–0.718	–0.546	–0.61	–0.669	–0.634	–0.682	–0.734
trajHotDeck	–0.664	–0.684	–0.705	–0.458	–0.534	–0.601	–0.579	–0.63	–0.687
locf	–0.388	–0.427	–0.476	0.339	0.174	0.01	–0.058	–0.126	–0.245
linearInterpol.global	–0.068	–0.109	–0.164	0.841	0.688	0.535	0.348	0.297	0.179
linearInterpol.local	–0.287	–0.321	–0.368	0.706	0.491	0.305	0.097	0.037	–0.071
linearInterpol.bisector	–0.138	–0.175	–0.226	0.869	0.676	0.502	0.277	0.225	0.114
spline	–0.26	–0.298	–0.348	0.611	0.402	0.219	0.043	–0.025	–0.125
regression	0.003	0	–0.005	0.077	0.032	0.002	0.028	–0.017	–0.079
copyMean.locf	–0.003	–0.007	–0.011	0.619	0.514	0.435	0.238	0.224	0.167
copyMean.global	–0.199	–0.232	–0.276	0.686	0.556	0.444	0.2	0.174	0.095
copyMean.local	–0.385	–0.411	–0.447	0.593	0.385	0.233	**–0.013**	–0.059	–0.13
copyMean.bisector	–0.254	–0.284	–0.325	0.747	0.564	0.422	0.152	0.118	**0.041**

Best values are in bold.

Table B11 – Bias dataset Artificial A according to the imputation method.

	MCAR			MAR			MNAR		
	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
crossMean	0.037	–0.029	–0.034	0.043	–0.007	0.013	**0.008**	0.032	0.021
crossMedian	0.041	**–0.019**	**–0.029**	0.053	**0.004**	0.014	0.009	0.036	0.041
crossHotDeck	0.027	–0.029	–0.039	**0.035**	–0.025	**0.01**	0.009	**0.026**	**0.008**
trajMean	–2.003	–2.426	–2.944	–0.647	–1.189	–1.881	–1.363	–1.724	–2.335
trajMedian	–1.534	–2.047	–2.617	–0.046	–0.655	–1.454	–0.814	–1.219	–1.928
trajHotDeck	–1.991	–2.433	–2.935	–0.629	–1.209	–1.897	–1.363	–1.693	–2.325
locf	0.448	0.072	–0.485	2.584	2.099	1.468	1.663	1.361	0.693
linearInterpol.global	5.825	5.556	5.295	6.64	6.405	6.134	6.253	6.159	5.829
linearInterpol.local	3.038	3.023	2.719	3.726	3.569	3.699	2.512	2.835	2.651
linearInterpol.bisector	4.699	4.537	4.252	5.485	5.3	5.2	4.758	4.828	4.556
spline	0.059	–0.271	–0.601	–0.212	–0.268	–0.544	–1.343	–1.092	–1.228
regression	0.037	–0.029	–0.034	0.043	–0.007	0.013	0.008	0.032	0.021
copyMean.locf	**0**	0.022	–0.034	0.287	0.241	0.241	–0.053	–0.028	–0.025
copyMean.global	–0.1	–0.111	–0.225	0.52	0.473	0.471	–0.028	–0.127	–0.083
copyMean.local	0.037	–0.103	–0.509	1.382	0.9	0.948	–0.379	–0.172	–0.336
copyMean.bisector	–0.007	–0.138	–0.457	1.228	0.868	0.844	–0.153	–0.112	–0.247

Best values are in bold.

Table B12 – Bias dataset Artificial B according to the imputation method.

	MCAR			MAR			MNAR		
	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
crossMean	–0.004	0.008	0.057	0.008	0.007	0.053	**0**	–0.011	0.016
crossMedian	**0**	0.023	0.052	0.011	**0.005**	0.083	0.002	**0.002**	**0.012**
crossHotDeck	–0.008	**0.004**	0.057	**0.005**	0.012	**0.049**	0.004	–0.014	0.022
trajMean	–0.185	–0.559	–0.9	–0.197	–0.57	–0.903	–0.219	–0.628	–0.987
trajMedian	–0.186	–0.561	–0.901	–0.201	–0.578	–0.909	–0.22	–0.633	–0.988
trajHotDeck	–0.186	–0.558	–0.901	–0.198	–0.572	–0.899	–0.219	–0.627	–0.989
locf	–0.157	–0.456	–0.766	–0.079	–0.26	–0.474	–0.145	–0.447	–0.796
linearInterpol.global	–0.139	–0.362	–0.568	0.036	0.087	0.161	–0.104	–0.315	–0.617
linearInterpol.local	–0.122	–0.235	–0.503	0.163	0.494	0.81	–0.107	–0.237	–0.583
linearInterpol.bisector	–0.137	–0.312	–0.548	0.094	0.289	0.513	–0.111	–0.285	–0.597
spline	–0.14	–0.261	–0.618	0.024	0.059	0.124	–0.123	–0.334	–0.798
regression	–0.004	0.008	0.057	0.008	0.007	0.068	0	–0.011	–0.083
copyMean.locf	–0.008	–0.008	**0.034**	0.057	0.171	0.287	–0.013	–0.027	–0.031
copyMean.global	–0.042	–0.084	–0.045	0.12	0.357	0.608	–0.021	–0.044	–0.113
copyMean.local	–0.081	–0.114	–0.275	0.189	0.582	0.977	–0.079	–0.136	–0.394
copyMean.bisector	–0.062	–0.1	–0.149	0.152	0.475	0.825	–0.052	–0.086	–0.226

Best values are in bold.

Appendix B.4 MSE

Table B13 – MSE dataset Artificial A according to the imputation method.

	MCAR			MAR			MNAR		
	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
crossMean	0.014	0.018	0.033	**0.011**	**0.018**	**0.037**	**0.011**	0.015	0.026
crossMedian	0.014	0.019	0.034	0.011	0.019	0.039	0.011	0.014	0.027
crossHotDeck	0.014	0.02	0.036	0.011	0.02	0.04	0.011	0.015	0.028
trajMean	0.236	2.004	5.942	0.188	1.533	4.482	0.191	1.611	4.95
trajMedian	0.237	1.977	5.693	0.214	1.627	4.374	0.197	1.59	4.677
trajHotDeck	0.236	2.012	5.932	0.188	1.549	4.499	0.191	1.62	4.96
locf	0.213	1.592	3.97	0.358	2.579	5.471	0.26	1.982	4.433
linearInterpol.global	0.794	6.424	15.928	1.109	8.852	21.169	0.943	8.058	19.288
linearInterpol.local	0.265	2.343	5.653	0.405	3.178	9.024	0.249	2.343	5.778
linearInterpol.bisector	0.558	4.673	11.497	0.794	6.438	16.313	0.614	5.5	13.348
spline	0.169	1.091	2.707	0.166	1.194	2.833	0.184	1.285	2.967
regression	0.014	0.018	0.033	0.011	0.018	0.037	0.011	0.015	0.026
copyMean.locf	**0.013**	**0.015**	**0.026**	0.013	0.018	0.039	0.013	**0.012**	**0.021**
copyMean.global	0.021	0.077	0.184	0.02	0.066	0.146	0.017	0.036	0.08
copyMean.local	0.05	0.307	0.778	0.105	0.476	1.26	0.072	0.241	0.671
copyMean.bisector	0.034	0.19	0.473	0.067	0.295	0.738	0.041	0.119	0.322

Best values are in bold.

Table B14 – MSE dataset Artificial B according to the imputation method.

	MCAR			MAR			MNAR		
	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
crossMean	0.11	0.143	0.214	0.112	0.144	0.399	0.103	0.141	0.361
crossMedian	0.108	0.146	0.244	0.112	0.154	0.404	0.103	0.141	0.392
crossHotDeck	0.115	0.153	0.223	0.113	0.143	0.416	0.102	0.143	0.372
trajMean	0.212	0.912	2.033	0.25	1.085	2.241	0.278	1.226	2.485
trajMedian	0.214	0.918	2.043	0.256	1.11	2.272	0.282	1.244	2.496
trajHotDeck	0.213	0.911	2.041	0.251	1.086	2.238	0.279	1.228	2.498
locf	0.183	0.65	1.555	0.134	0.392	0.943	0.183	0.673	1.669
linearInterpol.global	0.179	0.519	1.111	0.098	0.143	**0.253**	0.153	0.414	1.074
linearInterpol.local	0.195	0.446	1.248	0.134	0.45	1.095	0.17	0.37	1.149
linearInterpol.bisector	0.185	0.466	1.143	0.103	0.208	0.469	0.159	0.372	1.056
spline	0.292	0.787	2.013	0.235	0.538	1.402	0.272	0.718	2.318
regression	0.11	0.143	0.214	0.112	0.144	0.381	0.103	0.141	0.338
copyMean.locf	**0.097**	**0.106**	**0.15**	**0.09**	**0.121**	0.34	**0.101**	**0.115**	**0.271**
copyMean.global	0.12	0.198	0.324	0.102	0.242	0.669	0.111	0.132	0.306
copyMean.local	0.165	0.315	0.846	0.149	0.59	1.524	0.153	0.256	0.745
copyMean.bisector	0.136	0.228	0.494	0.119	0.405	1.11	0.125	0.159	0.396

Best values are in bold.

REFERENCES

[1] J.M. Engels, P. Diehr, Imputation of missing longitudinal data: a comparison of methods, J. Clin. Epidemiol. 56 (10) (2003) 968–976.

[2] N.M. Laird, Missing data in longitudinal studies, Stat. Med. 7 (1–2) (1988) 305–315.

[3] R.J.A. Little, Pattern-mixture models for multivariate incomplete data, J. Am. Stat. Assoc. 88 (421) (1993) 125–134.

[4] Y. Dong, C.Y. Joanne Peng, Principled missing data methods for researchers, Springerplus 2 (1) (2013) 1–17.

[5] R.J.A. Little, Modeling the drop-out mechanism in repeated-measures studies, J. Am. Stat. Assoc. 90 (431) (1995) 1112–1121.

[6] S.L. Zeger, K.-Y. Liang, An overview of methods for the analysis of longitudinal data, Stat. Med. 11 (14–15) (1992) 1825–1839.

[7] E. Dantan, C. Proust-Lima, L. Letenneur, H. Jacqmin-Gadda, Pattern mixture models and latent class models for the analysis of multivariate longitudinal data with informative dropouts, Int. J. Biostat. 4 (1) (2008) 1–26.

[8] W.J. Shih, H. Quan, Testing for treatment differences with dropouts present in clinical trials – a composite approach, Stat. Med. 16 (11) (1997) 1225–1239.

[9] M.S. Gold, P.M. Bentler, Treatments of missing data: a Monte Carlo comparison of RBHDI, iterative stochastic regression imputation, and expectation-maximization, Struct. Equ. Modeling 7 (3) (2000) 319–355.

[10] R.J.A. Little, D.B. Rubin, The analysis of social science data with missing values, Sociol. Methods Res. 18 (2–3) (1989) 292–326.

[11] C. Genolini, B. Falissard, Kml: k-means for longitudinal data, Comput. Stat. 25 (2) (2010) 317–328.

[12] J. Twisk, W. de Vente, Attrition in longitudinal studies: how to deal with missing data, J. Clin. Epidemiol. 55 (4) (2002) 329–337.

[13] C. Genolini, R. Écochard, H. Jacqmin-Gadda, Copy mean: a new method to impute intermittent missing values in longitudinal studies, Open J. Stat. 3 (2013) 26.
[14] R.J.A. Little, D.B. Rubin, Statistical Analysis with Missing Data, John Wiley, New York, 1987.
[15] D.B. Rubin, Inference and missing data, Biometrika 63 (3) (1976) 581–592.
[16] G. Molenberghs, H. Thijs, I. Jansen, C. Beunckens, M.G. Kenward, C. Mallinckrodt, et al., Analyzing incomplete longitudinal clinical trial data, Biostatistics 5 (3) (2004) 445–464.
[17] M.K. Olsen, K.M. Stechuchak, J.D. Edinger, C.S. Ulmer, R.F. Woolson, Move over LOCF: principled methods for handling missing data in sleep disorder trials, Sleep Med. 13 (2) (2012) 123–132.
[18] G.E. Forsythe, C.B. Moler, M.A. Malcolm, Computer Methods for Mathematical Computations, Prentice-Hall, Upper Saddle River, 1977.
[19] F.N. Fritsch, R.E. Carlson, Monotone piecewise cubic interpolation, SIAMJ Numer. Anal. 17 (2) (1980) 238–246.
[20] A. Burton, D.G. Altman, P. Royston, R.L. Holder, The design of simulation studies in medical statistics, Stat. Med. 25 (24) (2006) 4279–4292.
[21] R. Ecochard, H. Boehringer, M. Rabilloud, H. Marret, Chronological aspects of ultrasonic, hormonal, and other indirect indices of ovulation, BJOG 108 (8) (2001) 822–829.
[22] R.E. Tremblay, R.O. Pihl, F. Vitaro, P.L. Dobkin, Predicting early onset of male antisocial behavior from preschool behavior, Arch. Gen. Psychiatry 51 (9) (1994) 732.
[23] R Core Team, R: A Language and Environment for Statistical Computing, r foundation for statistical computing, Vienna, Austria, 2012. ISBN 3-900051-07-0, 2012. Technical report.
[24] C. Genolini. A (not so) short introduction to S4, 2008.
[25] C. Genolini. LongitudinalData, R package version 2.3, 2014.
[26] C. Genolini, X. Alacoque, M. Sentenac, C. Arnaud, kml and kml3d: R packages to cluster longitudinal data, Journal of Statistical Software 65 (4) (2015) 1–34.
[27] C. Genolini, B. Falissard, Kml: a package to cluster longitudinal data, Comput. Methods Programs Biomed. 104 (3) (2011) e112–e121.
[28] J.L. Schafer, J.W. Graham, Missing data: our view of the state of the art, Psychol. Methods 7 (2) (2002) 147.